Lenovo Validated Design: On-Prem High Performance AI Data Platform

Last update: **06 March 2026**

Version 1.0

Scale unstructured data with validated, sovereign-ready AI

High-performance object storage engineered for low-latency AI

Seamless scale-out storage built for expanding AI workloads.

Accelerate AI deployment with Lenovo Hybrid AI Factory

Vanita Meyer

Sinan Atan

LENOVO PRESS

Table of Contents

Introduction

Executive Summary

The Lenovo Validated Design (LVD) for High-Performance AI Data Platform with Cloudian HyperStore provides a secure, scalable architecture for deploying enterprise AI workloads on-premises. The solution enables organizations to operationalize large volumes of unstructured data, including documents, images, logs, audio, and video, while maintaining full control over data sovereignty, security, and performance.

Built on Lenovo infrastructure, the design integrates Cloudian HyperStore object storage with Cloudian AI Data Platform (AIDP) services to support high-throughput data ingestion, embedding generation, vector indexing, and low-latency retrieval for generative AI applications. Validated with NVIDIA Enterprise RAG blueprints, the architecture delivers predictable performance and repeatable deployments for organizations moving AI from pilot to production.

Intended Audience

The LVD provides customers, partners, and system integrators with a tested blueprint that reduces deployment risk, accelerates time-to-value, and enables confident adoption of on-premises Generative AI for sensitive, performance-critical workloads.

Challenges and Opportunity

Enterprise AI Challenges to Solve

Organizations seeking to deploy production AI applications, from document intelligence to video analytics, face several common challenges as they scale beyond pilots and experiments:

Storage Throughput and GPU Utilization

AI workloads require sustained, high-volume data access as embeddings, vectors, and source documents are retrieved during inference. Traditional TCP-based data paths introduce CPU overhead and latency, limiting throughput and preventing GPUs from operating at full efficiency. As a result, storage bottlenecks often lead to GPU underutilization and reduced ROI on AI infrastructure.

Latency and consistency under concurrent access

Enterprise AI systems must support multiple users, agents, and applications simultaneously across diverse workload types. As concurrency increases, organizations encounter unpredictable latency spikes, degraded time-to-first-token (TTFT), and inconsistent user experience. Without a validated architecture, scaling AI workloads often leads to over-provisioning, inefficiency, or missed SLAs.

Durability, Resilience, and Data Integrity at Scale

Enterprise AI environments cannot tolerate data corruption, loss, or downtime. RAG systems rely on:

- Durable storage of source documents
- Consistent object integrity
- Protection against hardware failures
- Multi-site replication and disaster recovery

Legacy storage systems may require costly replication strategies or introduce downtime during expansion. Enterprises need storage platforms that provide cloud-scale durability and automatic healing without performance trade-offs.

Security, governance, and data sovereignty requirements

Many enterprises, particularly in regulated industries, cannot send sensitive data to public cloud LLM services. They require on-premises or sovereign AI deployments with:

- Encryption at rest and in transit
- Role-based access control (RBAC)
- Immutable storage and audit trails
- Policy-driven data lifecycle management
- Geographic data residency enforcement

Balancing performance with governance requirements significantly increases architectural complexity when storage platforms are not AI-ready.

Explosive growth of unstructured enterprise data

Enterprises generate rapidly growing volumes of unstructured data—documents, images, audio, and video—often fragmented across file systems, NAS, and cloud repositories. Traditional storage architectures struggle to deliver the throughput and concurrency required for modern AI pipelines, leaving valuable data inaccessible to generative AI applications.

Operational complexity and deployment risk

Building production-ready AI applications requires integrating storage, networking, compute, vector databases, orchestration frameworks, and inference engines across multiple workload types. Each new AI use case can introduce additional components and configuration dependencies. Without a validated design that supports multiple NVIDIA AI blueprints on a common infrastructure foundation, organizations face long integration cycles, inconsistent configurations, and compounding operational risk as they expand from a single pilot to a portfolio of AI applications.

Opportunity

This LVD addresses these challenges by delivering a proven, production-ready architecture optimized for enterprise RAG workloads.

Unlock enterprise data for GenAI Securely and at Scale

By combining Cloudian HyperStore's S3-compatible object storage with AI-native data services in AIDP, organizations can:

- Consolidate fragmented unstructured data into a unified object storage platform
- Enable massively parallel read access for AI pipelines
- Scale from terabytes to petabytes without disruption
- Preserve data sovereignty and compliance

Align Storage Performance with GPU Compute

The validated architecture ensures a high-performance data path from storage to inference engines, delivering:

- Sustained, predictable read bandwidth
- Low-latency object retrieval
- Stable P95/P99 performance under concurrency
- Efficient GPU utilization

This eliminates GPU starvation and maximizes ROI on AI infrastructure investments.

Deliver consistent, low-latency user experiences

Through validated concurrency testing and end-to-end performance measurement, the LVD demonstrates:

- Stable TTFT under load
- Consistent token throughput
- Predictable scaling across users and agents
- Reliable performance during peak demand

This supports enterprise chatbots, copilots, and agent-based workflows with production-grade reliability.

Maintain full control, governance, and compliance

The on-premises design enables organizations to meet strict regulatory, privacy, and security requirements. Integrated access controls, encryption, and auditability allow enterprises to deploy GenAI with confidence, knowing sensitive data remains protected throughout the RAG pipeline.

Reduce deployment risk and accelerate time-to-value

By providing a pre-validated blueprint across Lenovo infrastructure and Cloudian software, the LVD eliminates guesswork, shortens deployment cycles, and enables repeatable, scalable rollouts across teams, business units, and regions.

Technical Overview

As GPU-accelerated computing continues to advance, AI workloads require faster and more efficient access to large datasets. Modern AI pipelines depend on scalable storage platforms capable of delivering high throughput, low latency, and efficient data management across distributed environments.

Technologies such as NVIDIA GPUDirect Storage (GDS) and RDMA networking enable direct data movement between storage and GPU memory, bypassing CPU bottlenecks and significantly improving data transfer performance. This architecture ensures that GPU resources remain fully utilized while processing large-scale AI workloads.

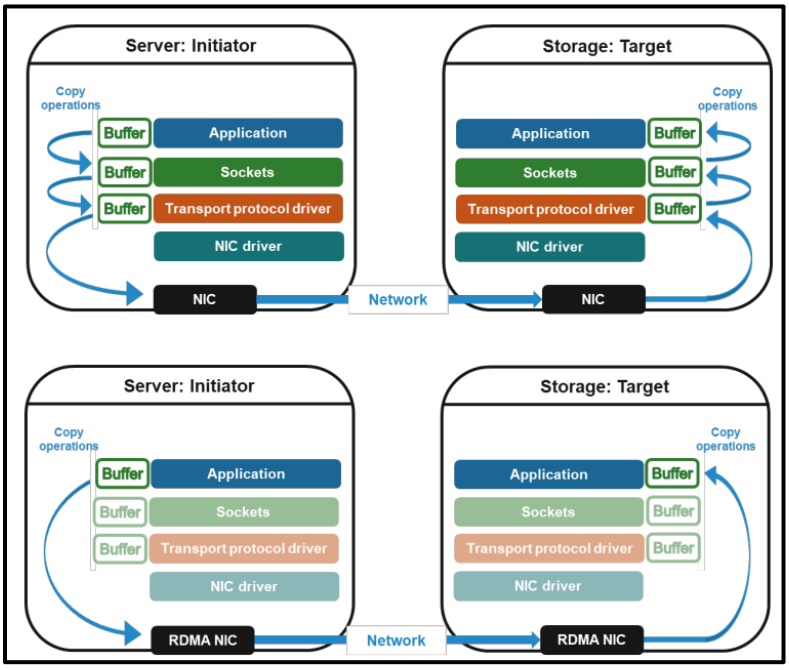

Figure 1 - GPU Direct IO Flow

NVIDIA GPUDirect Storage (GDS) enables a direct data path between storage systems and GPU memory, allowing data to move using RDMA without passing through the CPU or system memory. By bypassing the traditional operating system data path, GDS reduces CPU overhead and significantly lowers latency while increasing overall system bandwidth.

In this architecture, Cloudian HyperStore delivers data directly to GPU-enabled compute nodes over RDMA-enabled networking. Parallel data transfers from multiple storage nodes allow the system to sustain high throughput and ensure GPUs remain fully utilized during AI training and inference workloads.

GDS maintains the standard S3 control path for authentication and object management while accelerating the underlying data path. This separation enables applications to continue using familiar S3 APIs while benefiting from significantly improved data movement performance.

Requirements

The requirements for high-performance AI inferencing with Cloudian are divided into Functional Requirements and Non-Functional Requirements.

Functional requirements describe the core capabilities the platform must deliver, including scalable storage, high-throughput data access, and integration with AI data pipelines supporting tasks such as document processing, semantic search, and retrieval-augmented generation (RAG).

Non-functional requirements define the operational characteristics of the platform, including performance, scalability, security, reliability, and manageability. These requirements ensure that Cloudian HyperStore operates efficiently, securely, and at scale in enterprise AI environments.

Functional Requirements

Cloudian HyperStore is an S3-compatible object storage platform shall provide scalable, durable, secure, and high-performance storage for AI and enterprise workloads. It should serves as the foundational data layer for storing and delivering large volumes of unstructured data required but for training, inferencing, and retrieval-augmented generation (RAG).

Performance and Throughput

The platform must:

- Deliver high throughput required for AI training, data ingestion, and inference workloads
- Provide low-latency data access for real-time inferencing and RAG queries
- Maintain consistent performance under concurrent ingestion and retrieval workloads
- Support simultaneous access from multiple AI workloads, applications, and users
- Enable efficient parallel read and write operations across storage nodes

Scalability

The platform must:

- Support horizontal scale-out by adding storage nodes without downtime
- Scale from terabytes to petabytes of data while maintaining performance
- Allow non-disruptive capacity expansion without impacting running workloads

Data Storage and Access

The platform must:

- Provide S3-compatible object storage APIs for integration with AI platforms and data pipelines
- Store large volumes of unstructured data including documents, images, video, logs, and training datasets
- Support efficient storage and retrieval of large objects ranging from gigabytes to terabytes

Data Ingestion and Preparation

The platform must:

- Support automated ingestion of enterprise data from multiple sources
- Integrate with file systems, databases, object storage, and enterprise applications
- Support both batch and real-time ingestion workflows

AI Data Pipeline Integration

The platform must:

- Support integration with vector databases used for semantic search and RAG
- Enable data access for embedding generation and vector indexing
- Integrate with LLM inference systems and AI processing pipelines

Metadata and Indexing

The platform must:

- Support metadata extraction and indexing of enterprise data
- Maintain searchable indexes for efficient data discovery
- Allow continuous updates to metadata and indexing as data changes

Data Management

The platform must:

- Support lifecycle management policies such as retention, deletion, and tiering
- Provide bucket-based logical organization of data
- Enable metadata tagging for classification and discovery

Data Protection and Durability

The platform must:

- Provide erasure coding for efficient and resilient data protection
- Ensure high durability through distributed storage architecture
- Support replication across nodes, racks, or geographic locations
- Automatically detect and repair failures to maintain data integrity

Integration with Modern AI Platforms

The platform must:

- Integrate with Kubernetes-based AI infrastructure
- Support AI pipelines for training and inference workloads
- Integrate with data protection platforms such as Veeam Kasten
- Support integration with enterprise data processing and orchestration tools

Non-Functional Requirements

Cloudian HyperStore must meet enterprise-grade requirements for performance, scalability, reliability, security, and operational manageability to support production AI workloads.

Integration with Existing Infrastructure

The platform must:

- Provide standardized APIs for integration with enterprise systems and data pipelines
- Support flexible deployment models including on-premises and hybrid cloud environments

Security and Compliance

The platform must:

- Ensure secure data processing with encryption in transit and at rest
- Provide role-based access control (RBAC) for user and application permissions
- Support secure multi-tenant environments with logical data separation
- Enable data immutability through object locking to protect against ransomware or unauthorized modification
- Maintain audit logs for compliance and security monitoring

Reliability and Fault Tolerance

The platform must:

- Provide a fault-tolerant architecture with no single point of failure
- Support automatic failover during node or hardware failures
- Maintain continuous data availability during infrastructure disruptions

Deployment and Operations

The platform must:

- Support simplified installation and deployment processes
- Provide centralized monitoring and operational visibility
- Enable automated updates, diagnostics, and lifecycle management tools

Cloudian HyperStore

Cloudian HyperStore is a clustered, scale-out object storage platform that provides a single storage namespace across multiple storage locations. The platform enables organizations to store, manage, and access large volumes of unstructured data while supporting site redundancy and flexible deployment models ranging from edge environments to core data centers.

As AI workloads scale, storage systems must deliver both high performance and large capacity. GPU-accelerated workloads such as model training and inference require sustained high-throughput data access to ensure that GPUs remain fully utilized. HyperStore addresses this requirement by providing distributed, parallel access to enterprise data at scale.

The architecture described in this document supports deployments ranging from small clusters with three Lenovo ThinkSystem servers to large-scale environments supporting thousands of GPU-enabled compute nodes. The Cloudian object storage platform is designed to scale to exabyte-level capacity while maintaining consistent performance.

Cloudian HyperStore uses object-based storage architecture with the Amazon S3 API as the primary data access interface. The platform also supports file protocols such as NFS and SMB, enabling both object and file-based access to shared datasets.

Cloudian clusters are built on cloud-scale design principles and can be deployed across multiple architectures, including private cloud environments, globally distributed storage clusters, edge-to-core data infrastructures, and AI data platforms.

Within the storage cluster, physical storage resources are virtualized into logical containers called **S3** buckets, which provide a namespace for storing objects. The system automatically manages data placement across storage nodes based on configurable storage policies. These policies define how data is distributed, protected, and managed across the cluster.

Figure 2 illustrates the high-level architecture of the Cloudian HyperStore scale-out cluster.

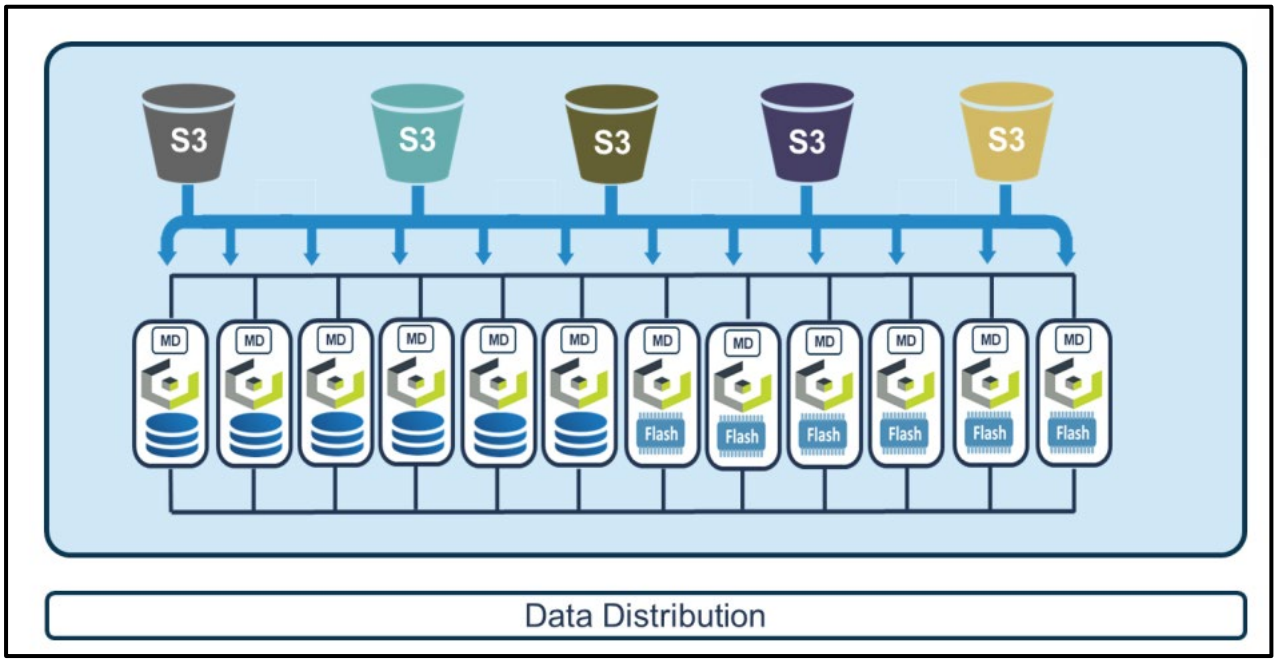

Figure 2 - Cloudian Object Storage Scale-out Cluster

Key Characteristics of Cloudian HyperStore

Performance

Traditional object storage systems were often used for archival storage rather than high-performance workloads. However, modern distributed object storage architectures enable highly parallel data access across multiple storage nodes. This distributed design allows HyperStore to deliver high throughput for AI workloads, analytics pipelines, and large-scale data processing.

Parallel Data Access

HyperStore uses a scale-out architecture that enables parallel data transfers across multiple nodes and storage devices. Large objects are divided into smaller segments and distributed across the cluster, allowing data requests to be processed concurrently. This parallel processing model eliminates storage bottlenecks and allows performance to scale as additional nodes are added.

Metadata-driven Data Management

Each stored object includes metadata that describes its attributes, location, and access policies. Metadata enables efficient indexing and search capabilities across diverse datasets. By combining metadata with stored objects, HyperStore allows organizations to manage and analyze large volumes of unstructured data more effectively.

Multi-tenancy

Cloudian HyperStore supports cloud-style multi-tenancy, enabling multiple users or applications to securely share the same storage infrastructure. Logical separation is achieved through independent namespaces, access controls, and storage policies.

User authentication and access control can integrate with enterprise identity systems such as LDAP, Active Directory, and IAM services. Storage policies can be applied at the bucket level to control data placement, replication, and protection.

HyperStore also supports strong security capabilities including:

- Server-side encryption with AES-256
- KMIP-compatible key management
- Secure communication via HTTPS
- S3-compatible access controls and identity federation

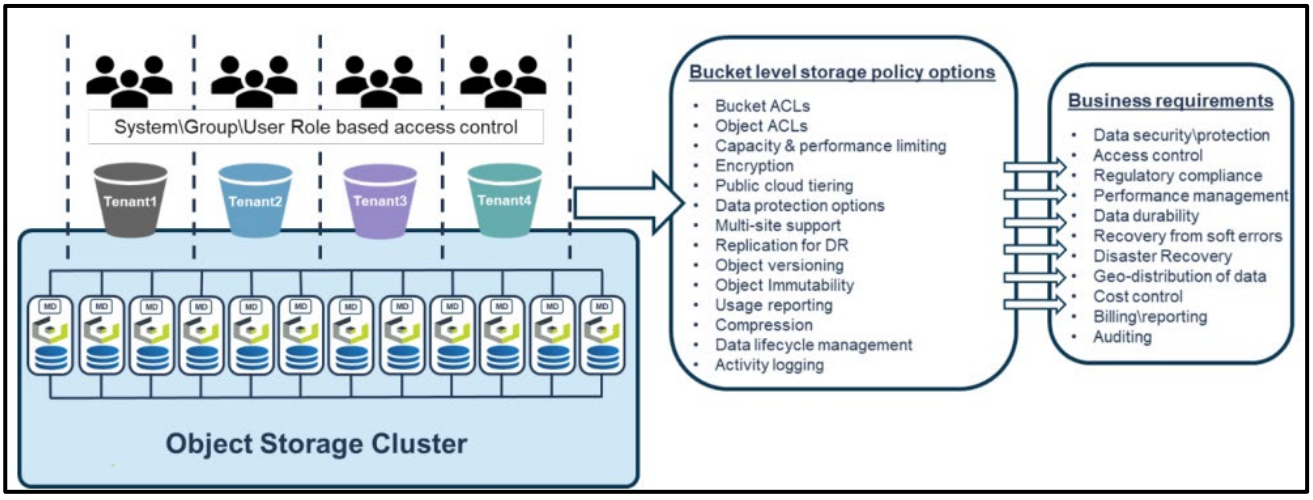

Figure 3 - Logical Multi-Tenancy Architecture

Physical Multi-Tenancy Options

In some environments, organizations may require dedicated hardware resources to guarantee performance for specific workloads. HyperStore supports deployments where storage resources are physically separated for certain tenants or workloads.

This model can be used for AI training workloads that require sustained high-bandwidth data access while allowing other tenants to share general-purpose storage resources.

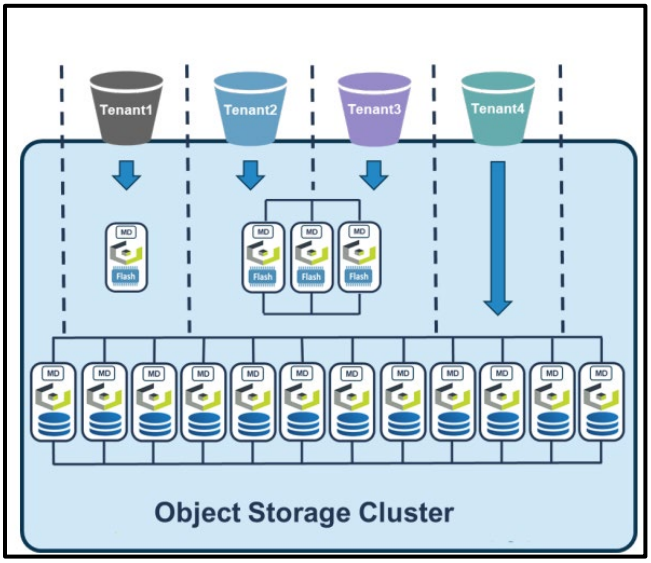

Figure 4 - Physical Multi-tenancy with dedicated HW Resources

HyperStore and the Cloudian AI Data Platform

HyperStore provides the foundational storage layer for the AI data platform. It delivers secure, scalable storage and high-performance access to enterprise data.

On top of HyperStore, the Cloudian AI Data Platform (AIDP) provides AI data services that enable data ingestion, preparation, indexing, and retrieval. These services allow organizations to operationalize enterprise data for generative AI, retrieval-augmented generation (RAG), and AI inference workloads.

Together, HyperStore and AIDP form a complete enterprise AI data platform that enables organizations to store, manage, and activate large datasets for AI-driven applications.

Architecture Overview

Cloudian HyperStore high-performance data platform resides under multiple layers of AI stack functional architecture. Physical storage is achieved in infrastructure layer in a distributed manner integrated to workloads and other nodes within the storage cluster over high-speed networking where logical layer of data management is achieved via AI Software stack. Regardless of the workload Cloudian HyperStore can expose data management capabilities for different AI functions utilizing inferencing, training or insight functions for the workloads.

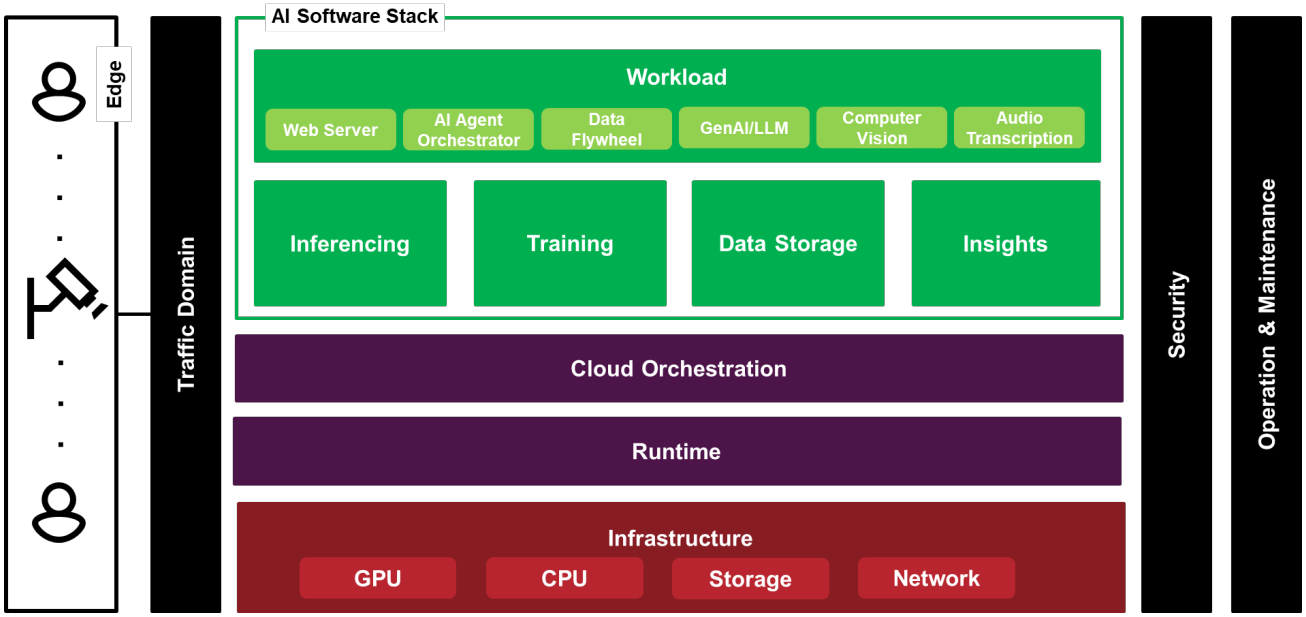

Figure 5 - High-Level Functional Architecture for AI Stacks

On the deployment side, Workload and Data Storage clusters are placed on top of Lenovo ThinkSystem where high speed RDMA data traffic networking is achieved via Nvidia Spectrum-X. Cloudian's HPS software runs both in workload and data storage domains providing data access to application layer over S3 APIs.

Storage area can be enabled both via high speed SSD/NVMe disks or traditional HDDs based on throughput requirements of client application. Spectrum-X network cards are placed both in Workload and Data Storage clusters to minimize processing latency in client and server side of storage access functions. High-level view of Deployment Architecture is given in Figure 6.

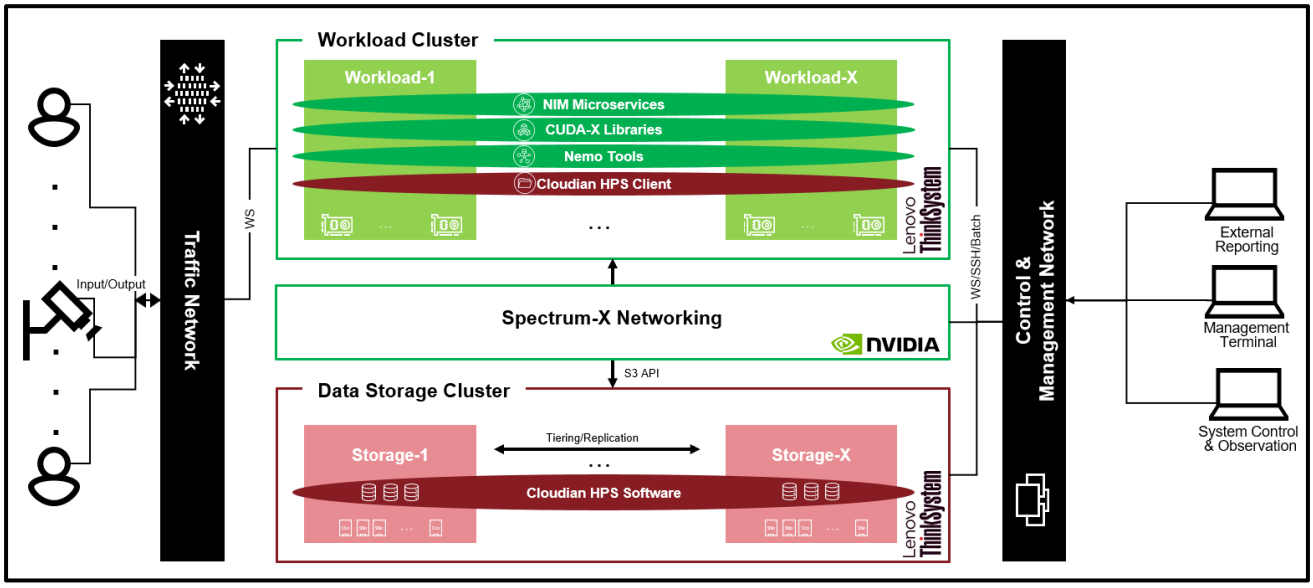

Figure 6 - Deployment Architecture Overview

Storage technology proposed in this solution is agnostic to any use-case that can span from entry level computer vision use cases to very large-scale language models and GenAI. It is possible to easily integrate those on top of Nvidia software suite such as NIM, Nemo Tools and CUDA-X libraries.

Table 1 given below summarizes the technologies utilized in each of the layers as part of this validation study.

Table 1 - LVD Technology Mapping

Layer	Technology	Version	Role
Hardware Infra	ThinkSystem	SR630v3	Hardware platform for Storage Nodes
		SR675v3	Hardware platform for AI Compute (workload)
Operating System	Helm	3.17.3	OS for AI Compute
Operating System	Kubernetes	1.32.3	OS for AI Compute OS for Storage Nodes
Operating System	Ubuntu	22.04 LTS	
Operating System Drivers	Rocky Linux	8.10	
Operating System	Nvidia	nvidia-driver-590	GPU Drivers
Drivers Object Storage	AIDP	8.2.6.1	Hyperstore AIDP software version
Elbencho	Benchmarking Software	v.3.0	Any other open-source software
AIDP	1.0.0	1.0.0	Any other plugin, SDK, etc.
AI Workload	Milvus	2.6	Information about RAG application, Vector DB, etc.

Operational Model

Cloudian HyperIQ

HyperIQ provides centralized management and monitoring for the object storage data lake infrastructure. It collects operational data from logs, system metrics, and events to deliver actionable insights into system performance and health. This unified view enables administrators to monitor user activity, track resource utilization, and efficiently manage the storage environment.

Figure 7 - Cloudian HyperIQ Infrastructure insights and monitoring

HyperIQ provides a unified view of operational health with end-to-end performance visibility from client access to storage media, enabling optimized resource utilization across the cluster. This level of monitoring is particularly important for AI and ML workloads, where GPUs must be continuously supplied with data to minimize idle time and maintain high utilization. High-speed read performance is essential to sustain inference and data retrieval workloads. Similarly, training checkpoints—used to periodically save model states—must complete quickly, as training processes pause during checkpoint creation to ensure consistency across GPUs participating in distributed training. HyperIQ provides full observability across the storage cluster, allowing administrators to detect, diagnose, and address potential issues before they impact workload performance.

Lenovo Systems Operational Model

Deployment orchestration and management of the stack can be done in container level integrating any validated solution in the environment. Lenovo Xclarity & LOC-A provides united dashboards to manage IT infrastructure and workloads with advanced automation capabilities both in hardware and Containers-as-a-Service (CaaS) layers. These tools also support open APIs, enabling integration with northbound systems to federate operations, streamline maintenance, and enhance end-to-end visibility.

Lenovo Xclarity

The Lenovo XClarity software family provides foundational management services for data center and edge infrastructure. It enables automated infrastructure management, faster service delivery, and simplified operational workflows.

Key capabilities include:

• **XClarity Controller** – Embedded management engine available on ThinkSystem and ThinkEdge servers providing a web-based interface and Redfish-compliant REST APIs for remote system management and integration.

• **XClarity Essentials** – A suite of management utilities such as OneCLI, UpdateXpress, and Bootable Media Creator used for system configuration, firmware updates, and diagnostics.

• **XClarity Administrator** – Centralized infrastructure management platform that supports deployments of up to 1,000 nodes, providing hardware monitoring, alerting, firmware management, configuration, and operating system deployment.

• **XClarity Orchestrator** – Multi-domain management platform that scales to environments with up to 10,000 servers, providing centralized control, automation, and predictive analytics for improved operational efficiency.

• **XClarity Energy Manager** – Provides monitoring and optimization of energy consumption across infrastructure, enabling capacity planning, thermal management, and energy-efficient workload scheduling.

• **XClarity Integrators** – Integration modules that connect XClarity with virtualization and IT management platforms such as VMware, Microsoft, Red Hat, Ansible, Puppet, Splunk, and Azure analytics tools.

Lenovo Open Cloud Automation (LOC-A)

Lenovo Open Cloud Automation (LOC-A) automates deployment, provisioning, and lifecycle management of infrastructure platforms, reducing operational overhead and enabling scalable infrastructure rollouts.

Key capabilities include:

• **Near-Zero Touch Provisioning (nZTP)** – Enables automated discovery and deployment of new servers, minimizing the need for field technician involvement and allowing administrators to manage deployments through the LOC-A portal.

• **Automated OS Deployment** – LOC-A remotely provisions operating systems on bare-metal servers, ensuring consistent configuration across nodes without requiring pre-built golden images.

• **Partner Integration Framework** – A plugin architecture enables partners to integrate automated deployment workflows for their platforms, supporting functions such as bare-metal provisioning, edge cluster deployment, node onboarding, and automated hostname/FQDN assignment.

• **Northbound APIs** – Secure APIs enable integration with external orchestration platforms and OSS/BSS systems, allowing deeper automation and accelerated feature integration.

Deployment Considerations

Cloudian HyperStore is a highly scalable object storage platform, starting from a minimum three-node storage cluster suitable for enterprise deployments. The cluster and network can be expanded without limitation based on capacity and throughput requirements. HyperStore provides S3-compatible object storage using the local disks of each storage node, scaling horizontally as nodes are added.

Storage nodes should be equipped with dual-socket CPUs and ConnectX Ethernet adapters to support RDMA. The network switches must also support RDMA to offload CPU overhead and reduce latency during storage operations. Additional details on storage node hardware requirements are provided in the next section, Server/Compute Nodes.

AI compute nodes are deployed based on workload requirements. HyperStore deployments require NVIDIA GPUs and ConnectX Ethernet adapters to ensure compatibility and performance. The next section will also detail the AI compute configuration used in this validation.

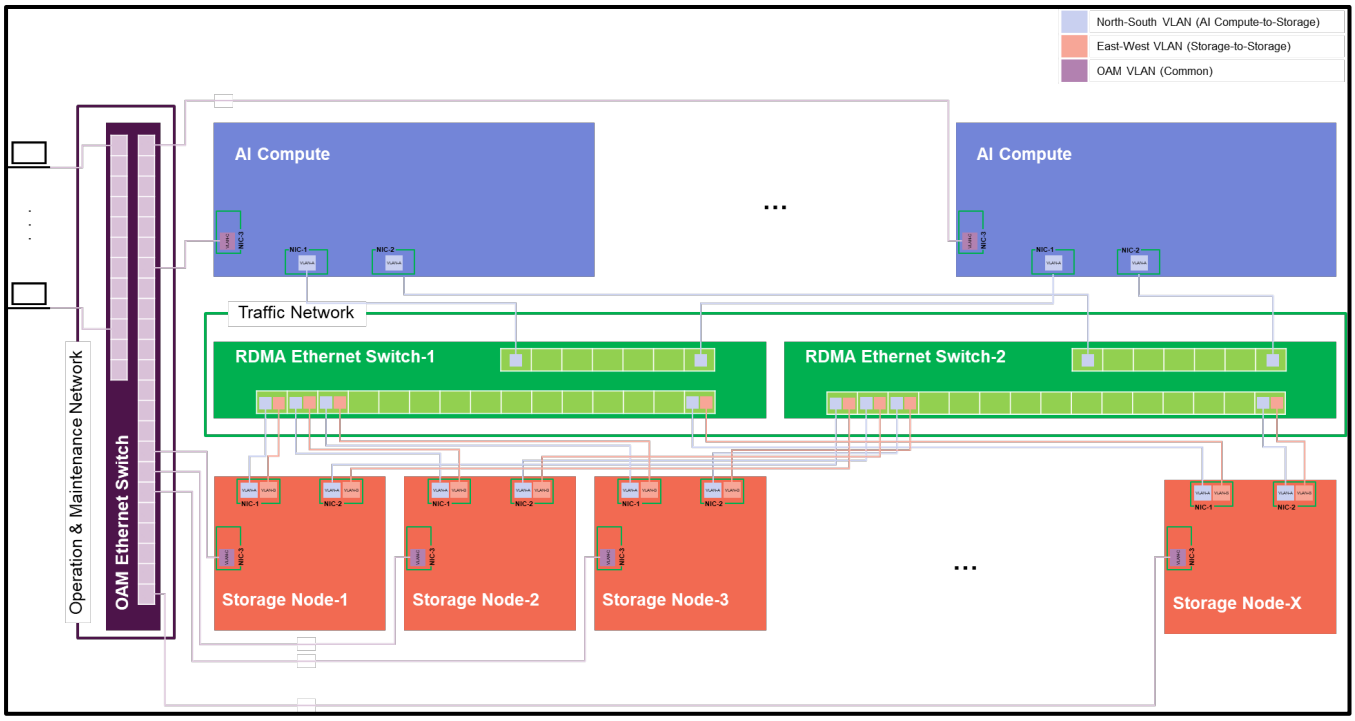

Figure 8 - Deployment Network Overview

For simplicity, the east-west communication paths between AI compute nodes are not shown in Figure 8, as the AI-to-AI network is outside the scope of this study and typically requires additional design considerations based on the specific workloads being deployed.

VLAN tagging is used to separate different traffic types over the same network interfaces, ensuring clean segmentation without additional physical ports. In addition, interface bonding is implemented to increase

available bandwidth and provide higher availability, eliminating single points of failure in both the network switches and Ethernet adapters.

High availability for the Operation and Maintenance (O&M) network is not implemented at the host level in this design, as redundant console access is already provided through the XCC interface. However, host-level O&M network redundancy can be added if required.

Table 2 - Validation Scope Hardware Setup

Qty	Model	Purpose of Use
1	Thinksystem SR675 V3	AI Compute / Compute Plane
1	Nvidia Spectrum SN5600	RDMA Switch / Network Plane
6	Thinksystem SR630 V3	Storage Node / Storage Plane

Further information on the hardware configuration is provided in the next section, and the detailed server Bill of Materials (BoM) can be found in the Appendix.

Server / Compute Nodes

Storage Node

The storage node used in this LVD is based on the Lenovo ThinkSystem SR630 V3, providing an excellent balance of footprint, storage density, and performance for enterprise environments. The SR630 V3 is a versatile 1U, dual-socket rack server designed for organizations ranging from small businesses to large enterprises that require high reliability, strong management capabilities, and robust security. It delivers the performance and flexibility needed to support future growth and is built to handle a broad set of workloads, including databases, virtualization, cloud computing, infrastructure security, systems management, enterprise applications, collaboration and email services, streaming media, web hosting, and HPC workloads.

.

Figure 9 - ThinkSystem SR630 V3 Front-View

The SR630 V3 delivers an ideal balance of performance and flexibility, making it well suited for enterprises of any size. It supports a wide range of drive and expansion slot configurations and incorporates numerous high-performance features. Its strong reliability, availability, and serviceability (RAS) capabilities, combined with an energy-efficient design, help optimize operational environments and reduce ongoing operational costs.

Table 3 - ThinkSystem SR630 V3 Features and Benefits

Feature	SR630 V3	Benefits
Processor	• 2x 4th/5th Gen Intel Xeon Scalable CPUs • Up to 64 cores • Up to 385W TDP • 80 PCIe 5.0 lanes per CPU	Higher core density • Improved compute performance • Faster networking and NVMe connectivity
Memory	• DDR5 memory operating up to 5600 MHz • 8 channels per CPU • 32 DIMMs • Up to 8TB of system memory	• New DDR5 memory offers significant performance improvements over DDR4 • Support for lower-cost 9x4 DIMMs
Internal storage	• Up to 10x 2.5" NVMe drives • Up to 16x E1.S NVMe bays • Flexible SAS/SATA/NVMe • Internal M.2 boot support	• Flexible storage configurations • High-performance NVMe support
RAID	• 8/16-port RAID adapters • Lenovo/Broadcom support • Tri-mode NVMe compatibility •	• Flexible storage protection options • High storage performance
Networking	• OCP 3.0 slot (PCIe Gen5 x16) • Additional PCIe NIC support • 1GbE management port	• High-speed networking support • Flexible adapter options
PCIe	• Up to 2x PCIe Gen5 + 1x Gen4 slots • Optional front-accessible slots • OCP 3.0 support	• High I/O bandwidth • Flexible expansion options
Management and security	• Integrated XClarity Controller 2 • Support for full XClarity toolset including XClarity Administrator • Platform Firmware Resiliency (PFR) hardware Root of Trust (RoT) • Tamper Switch security solution (intrusion switch)	• New XCC2 offers improved management capabilities • Same system management tool with previous generation • Silicon-level security solution
Power	• 750W, 1100W, 1800W AC Platinum/Titanium Hot Plug PSU • 1100W -48VDC Platinum general support • Active-Standby mode	• Flexible power configurations • Energy-efficient operation

AI Compute

Compute Node in this LVD utilizes a standalone ThinkSystem SR675 V3 that is part of Lenovo's Hybrid AI 285 platform that enables enterprises of all sizes to quickly deploy hybrid AI factory infrastructure, supporting Enterprise AI use cases as either a new, greenfield environment or an extension of their existing IT infrastructure.

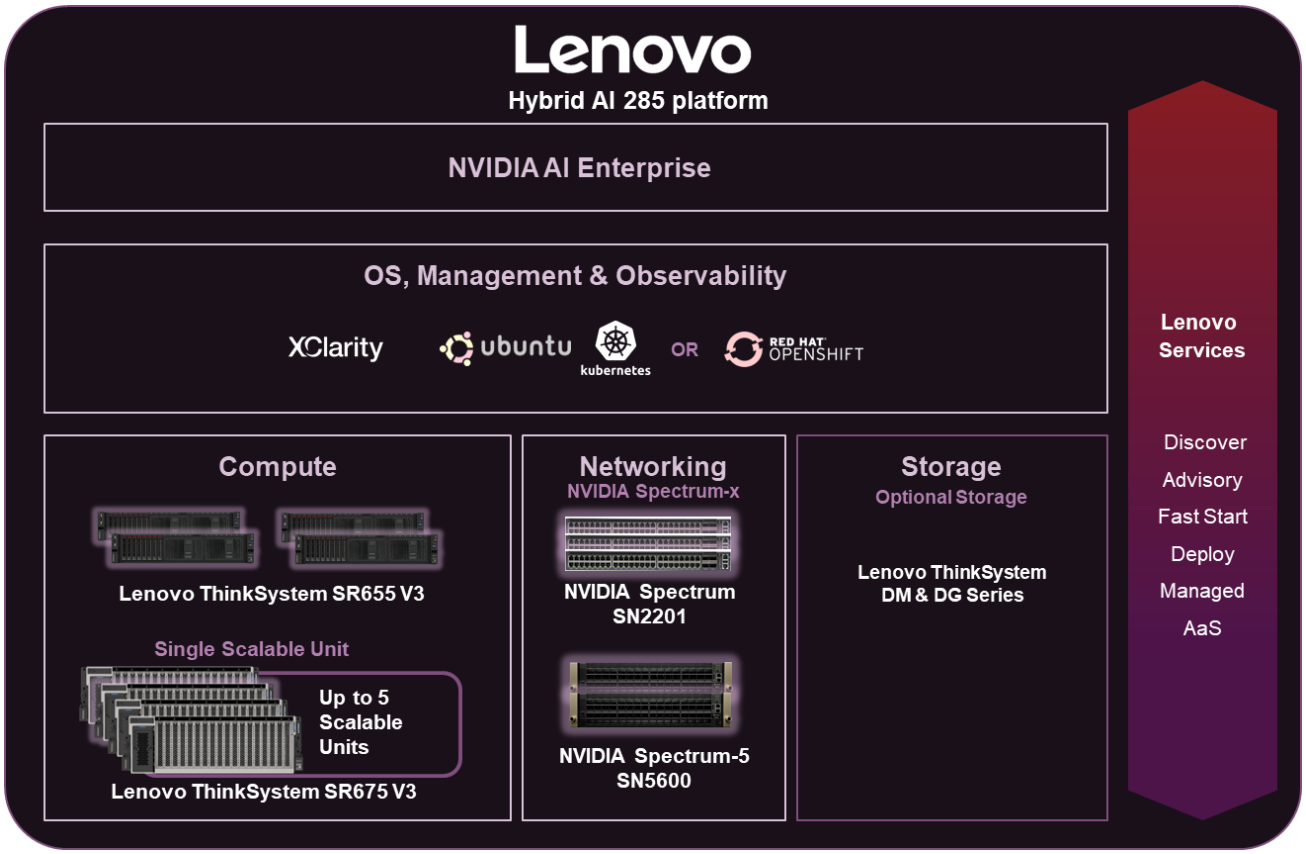

Figure 10 - Hybrid AI 285 Platform Overview

The Lenovo ThinkSystem SR675 V3 is a flexible 3U rack server designed for GPU-accelerated workloads. It supports up to eight double-wide or single-wide GPUs, including NVIDIA H200 and L40S Tensor Core GPUs, as well as the NVIDIA HGX H200 4-GPU configuration with NVLink connectivity and Lenovo Neptune hybrid liquid-to-air cooling technology. The platform is powered by AMD EPYC™ 9004 Series processors ("Genoa," "Genoa-X," and "Bergamo") and the latest 5th Gen AMD EPYC™ 9005 Series processors ("Turin").

Engineered for demanding Artificial Intelligence (AI), High Performance Computing (HPC), and data-intensive workloads, the SR675 V3 delivers the compute density and GPU acceleration required for modern machine learning and deep learning pipelines. Its high GPU capacity and flexible configuration make it well suited for enterprise AI deployments across industries such as retail, manufacturing, financial services, and healthcare. The ThinkSystem SR675i V3 is a specialized variant optimized specifically for AI inference workloads.

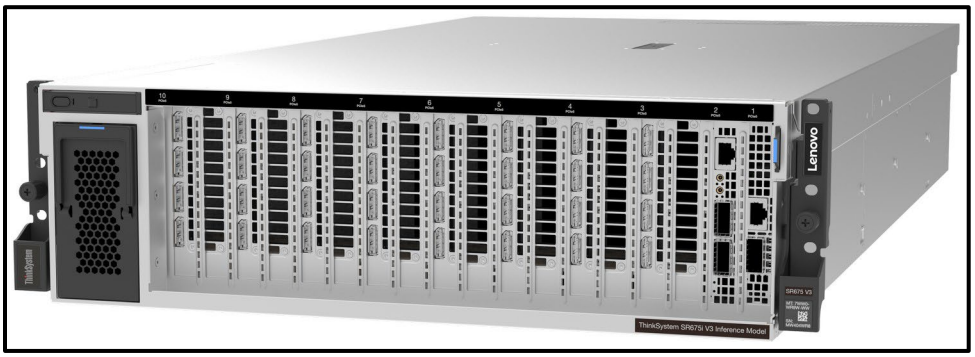

Figure 11 - Lenovo ThinkSystem SR675 V3 configured to support eight double-wide GPUs

The SR675 V3 features a modular design for ultimate flexibility. Multiple configurations are supported, including:

- One or two 4th or 5th Generation AMD EPYC™ Processors
- Up to eight double-wide or single-wide GPUs with NVLink bridges
- NVIDIA HGX H200 4-GPU with NVLink and Lenovo Neptune hybrid liquid cooling
- AMD Instinct™ MI Series Accelerators
- Choice of front or rear high-speed networking
- Choice of local high speed NVMe storage

Table 4 - Standard Specifications of ThinkSystem SR675 V3

Components	Specification
Form factor	3U rack
Processor	• Up to 2× AMD EPYC CPUs (Turin 9005 or Genoa 9004) • Up to 160 cores, 4.1 GHz, 400W TDP
Memory	24 DIMM slots (12 per CPU), up to 3TB, 6400 MHz.
Storage	• Up to 245.76 TB via 8× 30.72TB SAS SSDs • Up to 122.88 TB via 8× 15.36TB NVMe SSDs • Supports 12Gb SAS/SATA RAID & HBA
Storage controller	• 12 Gb SAS/SATA RAID adapters • 12 Gb SAS/SATA non-RAID HBAs (JBOD support only)
Network interfaces	OCP 3.0 slot (PCIe 4.0 x8/x16)
GPU support	8× double-wide GPUs (600W each) 8× single-wide GPUs
Ports	Front: USB 3.1, USB 2.0 (XCC), VGA, diagnostics Rear: 3× USB 3.1, VGA, 1GbE management, optional serial
Power & Cooling	Up to 4× hot-swap PSUs (1800W/2400W/2600W, Titanium/Platinum) 5× dual-rotor 80mm fans

Video	AST2600 BMC graphics, up to 1920×1200 @ 60H
Hot-swap parts	Drives and power supplies.
Management	XClarity Controller 2, XClarity ecosystem, XCC Platinum option
Security	TPM 2.0, chassis intrusion, passwords
OS Support	Windows Server, RHEL, SLES, VMware ESXi, Ubuntu
Dimensions	448 × 131 × 892 mm (W × H × D)

RDMA Switch

The NVIDIA Spectrum-X SN5600 spine switch provides the traffic plane for storage nodes, supporting both east-west and north-south data flows. It delivers 64 ports of 800 GbE in a 2U form factor with a total switching capacity of 51.2 Tb/s. The switch supports a range of connectivity options including 100, 200, 400, and 800 GbE, and integrates seamlessly into middle-of-row (MoR) or end-of-row (EoR) architectures optimized for Spectrum-X deployments.

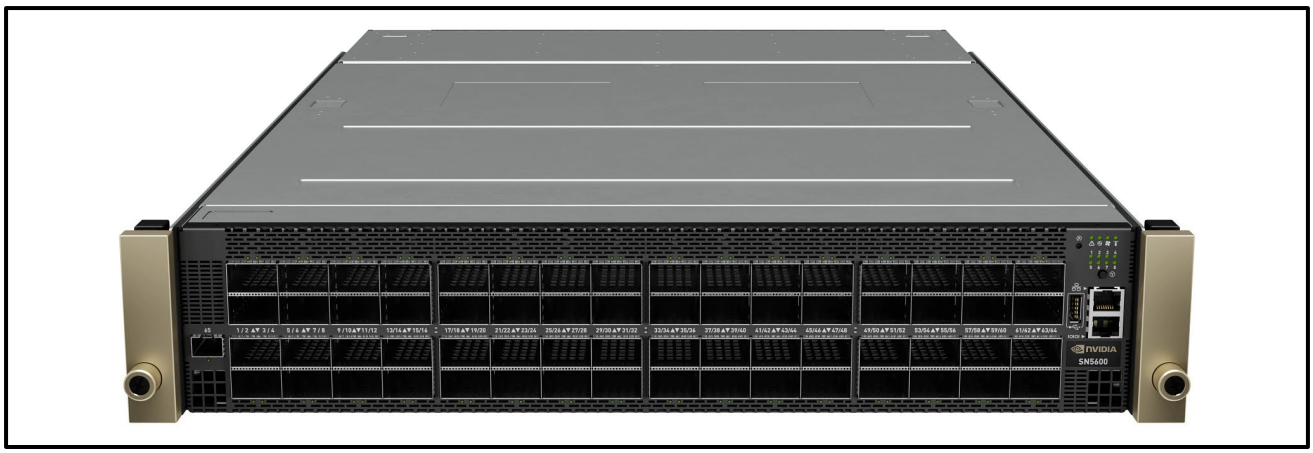

Figure 12 - Nvidia Spectrum-X SN5600 Switch Front-View

Networking

Cloudian HyperStore requires a high-performance, low-latency network to support AI inference and training workloads. The following guidelines support maximum performance, scalability, and reliability.

Bandwidth Planning

Designing a forward-looking network architecture is essential. Interface planning for storage nodes and the traffic switch layer should be done independently from the AI compute layer, while still reflecting current and future workload demands. For data-intensive inference or training workloads, allocate at least 200 Gbps connectivity per storage node.

Network Interfaces and High Availability

Each storage node should have a minimum of two Ethernet ports to ensure redundancy and load balance. For mission-critical applications, two separate Ethernet cards per server are recommended. RDMA switches should be deployed in redundant pairs to avoid single points of failure.

RDMA Support

AI workloads benefit from high-bandwidth, low-latency networking. Spectrum-X uses RDMA over Converged Ethernet (RoCE) to deliver efficient throughput, multi-tenant isolation, and minimal CPU overhead during data movement.

Security Best Practices

Firewalls, encryption, and VLAN segmentation should be used to secure network traffic, especially between DMZ or OAM networks and the storage traffic plane. Active inspection or monitoring of traffic within the data-plane (east-west or north-south) is not recommended, as it can significantly degrade throughput.

Network Segmentation

Separating management, data-plane, and training traffic domains ensures predictable performance and prevents congestion across functional boundaries. This segmentation can be implemented through either physical network separation or VLAN-based logical segmentation.

VPN Connectivity for Hybrid Workloads

High-bandwidth and reliable VPN connectivity is required for hybrid deployments that integrate private and public cloud environments, particularly for scenarios involving application-to-inference and retraining workflows.

Additional Recommendations

Management networks should be physically separated from data-plane traffic to ensure that administrative access remains responsive under heavy system load. East-west and north-south data-plane traffic should be split either physically or via VLAN tagging to prevent congestion in one direction from affecting the other.
On server NICs, ensure that the appropriate RDMA drivers are installed and enable jumbo frames (MTU 4200) on NIC bonds and VLAN-tagged bonded interfaces.

Storage Integration

AI workloads generate and consume high amounts of data from different sources that both throughput and input/output operations capacity are critical to ensure a future proof, scalable data plane. Cloudian's Hyperstore technology achieves a high performing, scalable (scale-out & scale-up) object-oriented storage as well reducing the operational complexity of persistence layer. For this purpose, data plane further takes following design considerations:

- **Fast & Efficient Storage –** Regardless of state-of-the-art object storage technology used, physical disks as part of data plane should be high-performing units to avoid any bottlenecks. Hence NVMe SSDs are recommended for high-speed data access and processing.

- **Scalable & Flexible Storage Options –** Hyperstore can be deployed both on-premise and cloud based for efficient long-term data management and easy scalability as data volumes grow. However high-speed networking is key for AI applications to get trained and interfere at par.
- **Organized Data Management –** Object managed storage access makes it easier to classify and store data in different format ensuring quick access to critical information.
- **Back Up, Retention & Restoration –** Apart from resiliency and data consistency, creation of highly customizable back-up policies and restoration processes are quite simple, fast and straightforward with standard tools.

Systems Management

A structured and efficient systems management strategy is crucial for maintaining Cloudian Hyperstorage AIDP stack and ensuring continuous operations. To optimize system reliability and security, organizations should focus on:

- **Monitoring & Performance Tracking** – Connect AIDP KPIs into the existing monitoring solutions, or use Cloudian HyperIQ, to track system health, detect potential issues early and capture detailed operational logs.
- **Automated Updates & Security Patching** – Deploy automated management tools to keep the software environment up to date with the latest security patches and feature updates, minimizing vulnerabilities and downtime.
- **Remote Administration** – Lenovo XClarity offers remote management capabilities, enhancing operational oversight for ThinkEdge-based deployments.

Solution Validation Results

Validation Scope

The validation scope for this Lenovo Validated Design focused on confirming the successful deployment, interoperability, and baseline performance of the Cloudian HyperScale AI Data Platform on Lenovo infrastructure. Testing validated the end-to-end data path between Lenovo compute systems, the Cloudian HyperStore object storage platform, and the supporting network infrastructure. Performance testing was conducted using representative high-throughput data workloads to verify sustained storage throughput and system stability under parallel access conditions. The validation demonstrates that the architecture can reliably support data-intensive AI and analytics workflows, while additional configuration options and performance optimizations supported by the architecture may be evaluated separately depending on customer deployment requirements.

Testing Methodology

Performance validation was conducted using Elbencho, a distributed storage benchmarking tool designed to evaluate high-throughput, parallel I/O environments. The benchmark generates concurrent read and write workloads across multiple client threads and nodes to simulate the access patterns commonly seen in large-scale data analytics and AI pipelines. Sequential read and write tests were executed using large block sizes to measure sustained throughput across the storage, network, and compute infrastructure. The test dataset size was configured to exceed available system memory to minimize caching effects and ensure that results reflected true storage and network performance. Aggregate throughput, IOPS, and latency metrics were collected across all participating nodes to provide a representative view of system behavior under sustained, production-like workloads.

Validation Results

Benchmark tests were performed for both read and write operations using three data-chunk sizes: 1 MB, 5 MB, and 32 MB. Each test case was repeated multiple times to smooth out irregularities and ensure consistent performance results. Measurements collected included elapsed time, objects per second, IOPS, throughput (MiBps), total data transferred (MiB), total object count, and minimum, maximum, and average latency for both I/O and object operations.

This validation confirms seamless integration and efficient data-handling capabilities when using accelerated compute clusters with object storage for AI workloads. Key observations include:

- GPU-accelerated applications can fully leverage the performance capabilities of HyperStore
- Performance is evenly distributed across all storage nodes in the cluster
- The platform scales linearly as additional storage nodes are added

This validation focused on Object Storage over TCP for the current test cycle. The architecture also supports RDMA-based data paths and NVIDIA GPUDirect Storage for low-latency GPU-to-storage communication. Validation of RDMA and GPUDirect Storage capabilities will be included in the next release of this LVD.

Solution Summary

This Lenovo Validated Design (LVD) delivers a production-ready architecture for deploying a high-performance enterprise AI data platform on Lenovo infrastructure using Cloudian HyperStore and the Cloudian AI Data Platform (AIDP). The solution enables organizations to consolidate large volumes of unstructured data into a scalable object storage platform capable of supporting modern AI workloads such as retrieval-augmented generation (RAG), semantic search, document intelligence, and advanced analytics.

The architecture integrates Lenovo ThinkSystem compute, NVIDIA GPU-accelerated nodes, high-performance networking, and Cloudian HyperStore scale-out object storage to form a robust data pipeline for AI workloads. While the architecture supports advanced capabilities such as RDMA-based data paths and NVIDIA GPUDirect Storage, the current validation focused on Object Storage over TCP to establish baseline performance and interoperability.

Validation testing confirmed reliable integration between Lenovo compute platforms, Cloudian object storage, and the network infrastructure. Results demonstrated consistent throughput, stable parallel access, and linear performance scaling as additional storage nodes were added, ensuring readiness for data-intensive AI workflows.

In addition to performance, the design addresses enterprise requirements for durability, security, and operational efficiency. Cloudian HyperStore provides S3-compatible storage with integrated erasure coding, encryption, and policy-driven lifecycle management, enabling organizations to maintain full control of sensitive data. Centralized management via Cloudian HyperIQ and Lenovo XClarity simplifies monitoring and operational automation across compute and storage resources.

By providing a validated blueprint that unifies compute, storage, networking, and AI data services, this LVD streamlines deployment and accelerates time-to-value for enterprise AI initiatives. Organizations can confidently scale from initial pilots to large-scale production environments while maintaining predictable performance, efficiency, and data sovereignty.

Appendix A: Lenovo Bill of materials (BOM)

SR630 V3 BoM for Storage Node

Part Number	Product Description	Qty
7D73CTO1WW	Server : ThinkSystem SR630 V3-3yr Base Warranty	1
BLK4	ThinkSystem V3 1U 10x2.5" Chassis	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
BYVY	Intel Xeon Gold 6542Y 24C 250W 2.9GHz Processor	2
BWHU	ThinkSystem 128GB TruDDR5 5600MHz (4Rx4) 3DS RDIMM	4
5977	Select Storage devices - no configured RAID required	1
C0ZS	ThinkSystem 2.5" U.2 VA 15.36TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	10
BRQX	ThinkSystem 1U 2.5" 10 NVMe Gen5 Backplane	1
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Adapter	1
BTTY	M.2 NVMe	1
CBT0	ThinkSystem M.2 VA 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 InfiniBand Adapter	1
BMXB	ThinkSystem Intel X710-T4L 10GBase-T 4-Port PCIe Ethernet Adapter	1
BLKF	ThinkSystem V3 1U x16/x16 BF PCIe Gen4 Riser1	1
BLK9	ThinkSystem V3 1U MS LP+LP BF Riser Cage	1
BLKH	ThinkSystem 1100W 230V Titanium Hot-Swap Gen2 Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
BLKD	ThinkSystem 1U V3 10x2.5" Media Bay w/ Ext. Diagnostics Port	1
BH9M	ThinkSystem V3 1U Performance Fan Option Kit v2	8
B8LA	ThinkSystem Toolless Slide Rail Kit v2	1
BPKR	TPM 2.0	1
B7XZ	Disable IPMI-over-LAN	1
BLK5	ThinkSystem SR630 V3 MB	1
BVMC	Trigger MFG to scan the SN from the CPU Board via this MI	1
C8WP	SR630 V3 Laser service indicator	1
BRPJ	XCC Platinum	1
B8KJ	ThinkSystem 1U 10x2.5" NVMe HDD Type Label	1
BCH1	ERP LOT9 Solution Country	1
BPCH	ThinkSystem SR630 V3 PCIe5.0 Cable from MB to BP, left-exit, 580mm	1
BRLS	ThinkSystem MCIO8x CBL, PCIe5.0, NY, MB PCIe 2-CFF C3 IN, 470mm	1

BRLT	ThinkSystem MCIO8x flat CBL, PCIe5, NY PCIe1 to 10AnyBay NVMe 8-9, 360mm	1
BRLW	ThinkSystem MCIO8x CBL, PCIe5, NY PCIe2/3 to 10AnyBay NVMe 6-7/4-5,340mm	1
BRQF	ThinkSystem SR630 V3,Gen5 CBL-4x2.5" AB,MCIO8x,PCIe 5/6 (MB) to NVMe0-1/2-3(BP),570mm	1
BRF3	ThinkSystem Cable132	1
BS3A	ThinkSystem SR645 V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 sideband, 730/300/250mm	1
BE0E	N+N Redundancy With Over-Subscription	1
BQBP	ThinkSystem MCC CPU Clip	2
BP50	ThinkSystem SR630 V3 Performance Heatsink	2
B984	ThinkSystem 1U PLV Top Cover Sponge	1
BYQN	ThinkSystem SR630 V3 Firmware and Root of Trust Security Module v2	1
B989	ThinkSystem V2 1U Package	1
AURS	Lenovo ThinkSystem Memory Dummy	28
B8NK	ThinkSystem 1U Super Cap Holder Dummy	1
B8NB	ThinkSystem 1U MS LP Riser Filler	1
AUWG	Lenovo ThinkSystem 1U VGA Filler	1
B5WJ	ThinkSystem OCP3 Filler	1
BK15	High voltage (200V+)	1
BPK3	ThinkSystem WW Lenovo LPK	1
B97B	XCC Label	1
BPDF	ThinkSystem 1100W Ti Power rating Label WW	1
BPD7	ThinkSystem SR630 V3 Service Label for WW	1
BPD6	ThinkSystem SR630 V3 Model Name Label	1
AWF9	ThinkSystem Response time Service Label LI	1
BPD5	ThinkSystem SR630 V3 Agency Label	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
BQPS	ThinkSystem logo Label	1
7S0XCTO5WW	XClarity Controller Platin-FOD	1
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
3444	Registration only	1
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1

QAA5	SR630 V3	1
QA18	Premier	1
QA0Y	Months	36
QA12	24x7 4hr Resp	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1
QAA5	SR630 V3	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTO2WW	SERVER CO2 OFFSET	1
QABN	CO2 Offset 20 Metric Tonnes	1
QABD	CO2 Offset	1

SR675 V3 BoM for AI Compute

Part Number	Product Description	Qty
7D9RCTO1WW	Server : ThinkSystem SR675 V3-3yr Base Warranty for AI	1
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	1
BFYB	Operating mode selection for: "Maximum Performance Mode"	1
C2AL	AMD EPYC 9535 64C 300W 2.4GHz Processor	2
CA1M	ThinkSystem 128GB TruDDR5 6400MHz (2Rx4) RDIMM-A v2	24
5977	Select Storage devices - no configured RAID required	1
BPKW	ThinkSystem E1.S 5.9mm 7450 PRO 7.68TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	6
BFTQ	ThinkSystem 1x6 E1.S EDSFF Backplane Option Kit	1
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Adapter	1
CBT1	ThinkSystem M.2 VA 1.92TB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BVBG	ThinkSystem NVIDIA BlueField-3 B3220 VPI QSFP112 2P 200G PCIe Gen5 x16 Adapter	1
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 InfiniBand Adapter	4
BE4T	ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port OCP Ethernet Adapter	1
CBK8	ThinkSystem NVIDIA RTX PRO 6000 Blackwell Server Edition 96GB PCIe Gen5 Passive GPU	8
C2RK	ThinkSystem SR675 V3 2 x16 Switch Cabled PCIe Rear IO Riser	2
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	1

BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	1
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	2
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	4
6252	2.5m, 16A/100-250V, C19 to C20 Jumper Cord	4
C3KA	ThinkSystem SR670 V2/SR675 V3 Heavy Systems Toolless Slide Rail Kit	1
B7XZ	Disable IPMI-over-LAN	1
C3EF	ThinkSystem SR675 V3 System Board v2	1
C8WW	SR675 V3 Laser service indicator	1
BK15	High voltage (200V+)	1
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	1
BE0D	N+1 Redundancy With Over-Subscription	1
BYT3	PCIe DUMMY BKT w/o Air hole	18
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	1
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	1
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	1
BRUC	ThinkSystem SR675 V3 CPU Heatsink	2
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	1
BABV	ThinkSystem Screw for fix M.2 Adapter	1
BXB6	ThinkSystem SR675 V3 BlueField-3 Power Cable	1
C5WW	ThinkSystem SR675 V3 Dual Rotor System High Performance Fan	5
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	8
C2RN	ThinkSystem SR675 V3 PCIe Rear Riser Support Bracket	2
C2RP	ThinkSystem SR675 V3 Mylar Perforation Cover	1
BR8W	ThinkSystem SR675 V3 Front PCIe Riser Cable 3	1
BR8R	ThinkSystem SR675 V3 Front PCIe Riser Cable 4	1
C2RL	ThinkSystem SR675 V3 Rear PCIe Riser Cable 7	1
C2RM	ThinkSystem SR675 V3 Rear PCIe Riser Cable 8	1
BR8H	ThinkSystem SR675 V3 Front OCP Cable	1
BRUL	ThinkSystem SR675 V3 EDSFF Drive Sequence Label	1
BFGZ	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 4	1
BRUQ	ThinkSystem SR675 V3 EDSFF to Riser Cables	1
BFTM	ThinkSystem SR670 V2/ SR675 V3 EDSFF Cage	1
BRNM	ThinkSystem SR670 V2/SR675 V3 2600W Power Supply Caution Label	1
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	1

BR7U	ThinkSystem SR675 V3 Root of Trust Module	1
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	1
BR80	ThinkSystem SR675 V3 Agency Labels	1
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
3444	Registration only	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1
QAAK	SR675 V3	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
QAAK	SR675 V3	1
QA18	Premier	1
QA0Y	Months	36
QA12	24x7 4hr Resp	1

Appendix B: Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
AIDP	AI Data Platform
AMD	Advanced Micro Devices
API	Application Programming Interface
BoM	Bill of Materials
CPU	Central Processing Unit
DC	Data Center
EoR	End-of-Row
GDS	GPUDirect Storage
GigE	Gigabit Ethernet
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HPC	High Performance Computing
HTTP	Hyper-Text Transfer Protocol
HTTPS	Hyper-Text Transfer Protocol Secure
HW	Hardware
I/O	Input/Output
IO	Input Output
KPI	Key Performance Indicator
LLM	Large Language Modelling
LOC-A	Lenovo Open Cloud Automation
LTS	Long-Term Support
LTSC	Long-Term Servicing Channel
LVD	Lenovo Validated Design
ML	Machine Learning
MoR	Middle-of-Row
MQTT	MQ Telemetry Transport
NLP	Natural Language Processing

NVMe	Non-Volatile Memory
nZTP	Near-Zero Touch Provisioning
OCP	Open Compute Project
OS	Operating System
RAG	Retrieval Augmented Generation
RAID	Redundant Array of Independent Disks
RAM	Random Access Memory
RAS	Reliability, Availability, and Serviceability
RDMA	Remote Direct Memory Access
REST	Representational State Transfer
RoCE	RDMA over Converged Ethernet
SFTP	Secure File Transfer Protocol
SSD	Solid State Drive
SSH	Secure Shell
SUT	System Under Test
TDP	Thermal Design Power
TLS	Transport Layer Security
TTFT	Time-to-First-Token
UDP	User Datagram Protocol
UEFI	Unified Extensible Firmware Interface
UI	User Interface
VLAN	Virtual Local Area Network
VM	Virtual Machine
VPN	Virtual Private Network
VROC	Virtual RAID on CPU
WLAN	Wireless Local Area Network
WS	Web Service
XCC	XClarity Controller

Resources

Resources	Links
Product Guide: Lenovo ThinkSystem SR630 V3 Server	Lenovo ThinkSystem SR630 V3 Server Product Guide > Lenovo Press
Product Guide: Lenovo ThinkSystem SR675 V3 and SR675i V3 Servers	Lenovo ThinkSystem SR675 V3 and SR675i V3 Servers Product Guide > Lenovo Press
Cloudian	Cloudian Website
Cloudian HyperStore	Cloudian HyperStore
Cloudian AI Data Platform	Cloudian AI Data Platform
Lenovo XClarity	Systems Management
LOC-A	Lenovo Open Cloud Automation (LOC-A)

Document history

Version 1.0	March 2026	Lenovo SR630 V3, Cloudian AIDP 8.2.6.1

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