Reference Architecture: Lenovo ThinkEdge for AI

Last update: **18 June 2025**

Version 1.0

Reference Architecture for Edge AI inference workloads

Describe solution use cases for edge AI inference

Recommended configuration and sizing models

Contains Bill of Materials for compute servers and storage

Vanita Meyer
Jorge Conejero Alberzoni
Connor Blumsack

LENOVO PRESS

Table of Contents

Introduction

As organizations generate massive volumes of data at the network edge from smart sensors, cameras, and IoT devices, extracting real-time business value has become essential. To meet this need, companies are increasingly adopting edge computing and real-time event-streaming technologies. This shift away from centralized data centers has made AI inference at the edge a critical enabler of intelligent decision making.

Unlike traditional AI architectures that depend on cloud-based processing, edge AI allows data to be processed locally, at the point of generation. This localized approach reduces latency, strengthens data privacy and sovereignty, and addresses growing concerns around compliance with regulations such as GDPR, particularly in scenarios where transferring sensitive data to the cloud is restricted. Additionally, by minimizing the need to transmit large volumes of raw data over the backhaul to central servers, edge AI significantly reduces bandwidth consumption and associated operational costs. The result is faster, context-aware insights that support real-time actions and improved efficiency.

To support this evolution, Lenovo's ThinkEdge servers provide a robust, secure, efficient, and scalable infrastructure for edge AI deployments. Beyond performance and compliance, Lenovo ThinkEdge servers offer environmental and operational benefits with their compact size, low acoustic output, and energy-efficient design, ideal for space-constrained and noise-sensitive environments.

The reference architecture includes integrated data services, built-in protection mechanisms, seamless scalability, and hybrid connectivity for cloud integration. With this solution, organizations can confidently deploy and manage AI workloads at the edge, unlocking transformative value across industries such as retail, manufacturing, healthcare, and smart cities.

Target Audience

This document is intended for the following audiences:

- Business leaders and enterprise architects who want to productize AI at the edge.
- Platform architects who design solutions for the development of AI models and software
- Data scientists, data engineers, architects, and developers of AI systems
- Edge device managers & administrators responsible for deployment & management of edge inferencing models.

Solution Overview

Lenovo's Edge AI architecture is built on a comprehensive foundation that integrates infrastructure, compute, software, and services, shown in the diagram below.

Figure 1 - End-to-End Lenovo Architecture for AI at the Edge

Flexible Architecture Framework for All Verticals

To streamline solution design and validation, this architecture document outlines nine standard edge server sizing models that address a broad range of use cases. Customers can use these models to identify the configuration that best aligns with their specific requirements. Each sizing model includes recommended workloads optimized for the available hardware; however, a clear understanding of the intended use case remains the most critical factor in selecting the appropriate configuration.

Solution Areas

The key advantage of AI inferencing and edge computing lies in their ability to process and analyze data locally with minimal latency. This enables faster decision-making and real-time responsiveness across industries. Below are leading use cases across five major sectors:

Retail: Self-Checkout and Shopper Analytics

Edge computing enables AI-driven self-checkout systems that reduce wait times and enhance the customer experience. These systems manage real-time authentication (linking a shopper to an account), product selection tracking via computer vision, and inventory monitoring using RFID and sensors. Edge servers are deployed at each checkout station to process visual and sensor data locally, while shared storage centrally

manages AI models and transaction logs across the store. The visual data processing can also help manage checkout queues, abandoned shopping carts, and alerts for required cleanups.

Manufacturing: Quality Inspection and Predictive Maintenance (Industry 4.0)

In smart factories, AI and edge computing support predictive maintenance and real-time quality inspection. Edge servers collect data from a wide array of sensors and vision systems to detect defects, monitor equipment health, and manage production lines with minimal delay. Updated inference models can be centrally stored and pushed from a shared storage system. This setup helps reduce downtime, extends equipment life, and improves operational efficiency, supporting ISO-aligned quality management and automated decision-making without constant cloud reliance.

Smart Cities: Traffic Management and Public Safety

Urban infrastructure now benefits from real-time AI insights at the edge. Use cases include smart traffic lights, pedestrian safety monitoring, license plate recognition, and video analytics for crowd or threat detection. Edge servers placed at intersections or public buildings can locally process high-resolution video and sensor data. Shared storage provides a central repository for AI models, event history, and analytics dashboards, enabling city planners to respond proactively to evolving conditions.

Healthcare: Remote and In-Hospital Patient Monitoring

AI at the edge supports continuous health monitoring in both clinical and home settings. Devices tracking vitals, such as heart rate, respiratory patterns, glucose levels, and neurological signals, require immediate inferencing for timely interventions. Edge servers analyze incoming patient data in real-time, triggering alerts or adjusting treatment automatically. Shared storage systems maintain model libraries and anonymized health data, ensuring rapid access while meeting regulatory requirements for data security.

Financial Services: Smart ATMs and Fraud Detection

Banks are transforming customer experience and safety using AI-powered kiosks and surveillance at the edge. Interactive ATMs use facial recognition, behavioral analysis, and object detection to assess potential threats, identify fraud attempts, and provide enhanced services. Edge compute nodes integrated with kiosk cameras process data locally, ensuring real-time inferencing without depending on cloud latency. Shared storage supports centralized model updates and synchronization across multiple branches or kiosk networks.

Edge Location in a Hybrid AI Platform

Enabling Real-Time Intelligence at the Edge

Edge AI empowers real-time decision-making by reducing the latency, bandwidth usage, and privacy concerns often associated with data center– or cloud-based inference. While centralized AI infrastructure remains essential for training large models and handling complex batch workloads, many real-time applications require immediate responses at the point of data generation. Lenovo ThinkEdge servers are purpose-built to meet these demands, enabling scalable AI deployments at the edge, from compact form factors for constrained spaces to high-performance systems supporting multimodal inference workloads. This section outlines the technologies and frameworks that make this possible, including vLLM, TensorRT-LLM, and OpenVINO, each optimized to deliver high-throughput, low-latency inference tailored to specific application needs.

vLLM: Unlocking Efficiency LLM Serving at the Edge

Virtual Large Language Model (vLLM) technology is designed for efficient and high throughput serving of Large Language Models (LLMs). While traditionally used in cloud or data center and cloud environments with powerful GPUs, vLLM is now emerging as a practical option for edge deployments.

Why vLLM at the Edge?
vLLM significantly improves the usability and efficiency of large language models (LLMs) on advanced edge devices equipped with GPUs. It is especially valuable in scenarios that demand:

- Real-time, low-latency inference
- High throughput for multiple simultaneous requests
- On-device processing to maintain privacy and data sovereignty
- Resilience in environments with intermittent or unreliable cloud connectivity

By enabling efficient LLM deployment on edge hardware—such as industrial PCs, intelligent robotics, or high-performance mobile devices, vLLM makes real-time AI more accessible and practical outside the data center.

These capabilities enable smarter, real-time decision-making in environments such as smart cities, retail, industrial automation, and security operations.

TensorRT-LLM: High-Throughput LLM Inference on Edge GPUs

TensorRT-LLM, built on NVIDIA's TensorRT inference engine, is optimized for running large language models (LLMs) with extremely low latency and high efficiency. It is particularly suited for powerful edge servers like the ThinkEdge SE360 V2 and SE455 V3, equipped with NVIDIA L4 or L40S GPUs. TensorRT-LLM is ideal for demanding use cases such as:

- Vision-Language Model (VLM) inference
- Real-time video analytics with natural language reasoning
- Text-to-text inference

Edge servers powered by TensorRT-LLM can support secure, high-performance LLM inference entirely on-premises, ensuring minimal latency and greater control over data privacy.

OpenVINO™: Accelerated Vision and AI Inference with Intel CPUs

OpenVINO™, Intel's open-source AI toolkit, delivers optimized inferencing performance on Intel CPU platforms—including those in the ThinkEdge SE100 and SE350 V2 servers. OpenVINO enables efficient execution of models such as YOLO, SSD, and ResNet for vision applications, even without GPUs. Its lightweight footprint and high compatibility make it ideal for:

- Video and image analytics in constrained edge environments
- Object detection and classification
- Industrial automation and predictive maintenance

OpenVINO offers quantization and graph optimization capabilities that reduce compute overhead while maintaining accuracy, which is especially valuable in power- and cost-sensitive deployments.

Computer Vision: Foundational for Edge AI Insights

Computer vision is one of the foundational workloads for edge AI platforms. It enables machines to interpret and act on visual data captured from cameras, drones, or industrial sensors. When deployed at the edge, these capabilities can deliver real-time insights critical for safety, efficiency, and customer engagement.

Edge-based computer vision supports a wide range of applications:

- **Retail**: Queue detection, shelf analytics, and theft prevention
- **Manufacturing**: Defect detection and assembly verification
- **Smart Cities**: License plate recognition, crowd analytics, and traffic flow monitoring
- **Healthcare**: Patient monitoring and gesture recognition

Lenovo's ThinkEdge portfolio provides the ideal infrastructure for deploying vision AI workloads, ranging from compact, fanless SE100 models to GPU-powered SE360 V2 and SE455 V3 servers, supporting models accelerated by vLLM, TensorRT, and OpenVINO.

Audio Environmental Intelligence (Sound AI)

While the human ear is remarkably adept at recognizing a wide range of sounds and their sources, it is still prone to error, distraction, or limitations in noisy environments. AI-powered auditory systems, however, can continuously listen, classify, and interpret audio signals—often with greater precision and consistency than humans.

This capability is known as Auditory AI or Sound AI, which involves training models to detect and respond to various sound categories, such as:

- Human-generated sounds (e.g., speech, shouting, coughing)
- Animal sounds (e.g., barking, birdsong)
- Mechanical or object sounds (e.g., glass breaking, engines)
- Music and structured audio
- Ambient or natural sounds (e.g., rain, wind, fire)
- Unclassified or ambiguous audio sources

Training AI systems on a growing library of environmental sounds continuously enhances their accuracy and responsiveness. This expanding intelligence is the foundation of Sound AI's value, enabling real-time monitoring, anomaly detection, and proactive responses across industries such as security, healthcare, manufacturing, retail, and smart cities.

Asset Location – Real-Time Location System (RTLS)

Real-time tracking of critical assets with low-latency performance enhances both operational efficiency and security. **RTLS** enables organizations to locate, monitor, and manage assets across facilities, campuses, or field environments using a variety of access technologies, including:

- Ultra-Wideband (UWB) over 5G
- Bluetooth Low Energy (BLE)
- Wi-Fi Positioning
- Passive RFID
- LoRaWAN

Each of these technologies offers unique advantages depending on the use case, environment, and performance requirements.

RTLS is not a one-size-fits-all solution—it requires a thoughtful strategy that considers business objectives, physical layout, infrastructure constraints, and the trade-offs of each technology.

Lenovo can integrate and process data from these access technologies directly at the edge, enabling fast, secure, and scalable RTLS deployments tailored to industry-specific needs across manufacturing, healthcare, logistics, retail, and smart cities.

Data Analytics at the Edge

Combining computer vision, real-time asset location (RTLS), and environmental audio data enables a powerful and context-rich IoT experience at the edge. This convergence of modalities drives intelligent, situational awareness and enables actionable insights across industries.

Lenovo Validated Designs (LVD) demonstrate how these integrated data streams can be analyzed and utilized to power smart solutions that deliver:

- Safer and more secure environments
- Enhanced customer experiences

- Continuous quality assurance and compliance

Lenovo Hybrid AI Vision Strategy

Lenovo's Hybrid AI strategy unifies edge, data center, and cloud capabilities to deliver secure, scalable, and context-aware intelligence. This approach prioritizes data privacy, GenAI integration, real-time profiling, and multimodal analysis, allowing customers to deploy AI workloads across the most appropriate environments.

By enabling LLM-driven contextual understanding and semantic video analysis, while keeping sensitive data local or selectively processed in the cloud or data center, Lenovo's Hybrid AI empowers organizations to maintain control over their information without sacrificing performance or innovation.

Scalable Multi-Node, Multi-GPU Architecture

Our architecture supports multi-node, multi-GPU edge deployments, enabling distributed AI processing closer to the data source. This configuration can either:

- Host GPUs directly in the edge servers, ideal for on-prem, high-performance inferencing
- Support Virtual Desktop Infrastructure (VDI) to dynamically allocate GPU resources from a centralized private cloud

By integrating technologies like vLLM and LLM-d, Lenovo solutions provide scalable, high-throughput inferencing with low latency—ideal for mission-critical applications requiring rapid decision-making and edge autonomy.

Architecture Overview

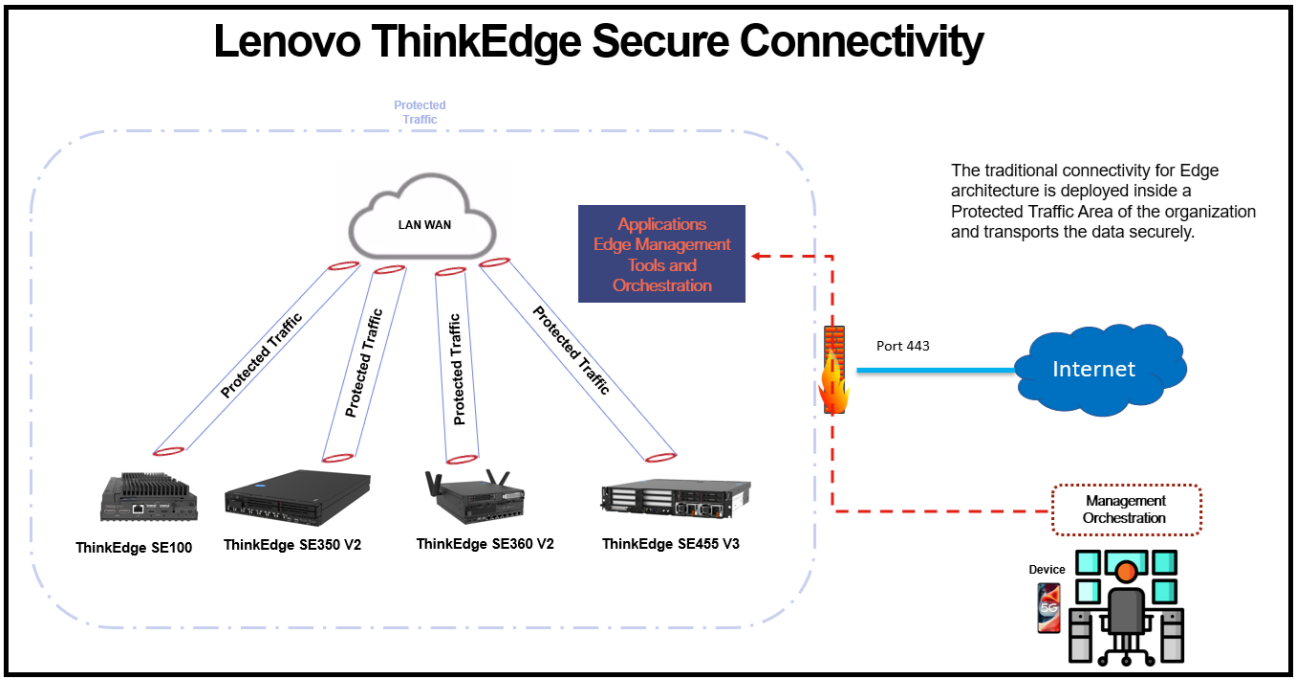

Figure 1- Connectivity for ThinkEdge servers in an Edge Location

The traditional connectivity for Edge architecture is deployed inside a Protected Traffic Area of the organization and transports the data securely. Utilizing Tools such as Lenovo Open Cloud Automation (LOC-A) and XClarity for Management and REST APIs integration, the Edge nodes can be deployed and managed from a remote/central location. The Edge nodes make data actionable performing tasks even if the back-haul connectivity is interrupted.

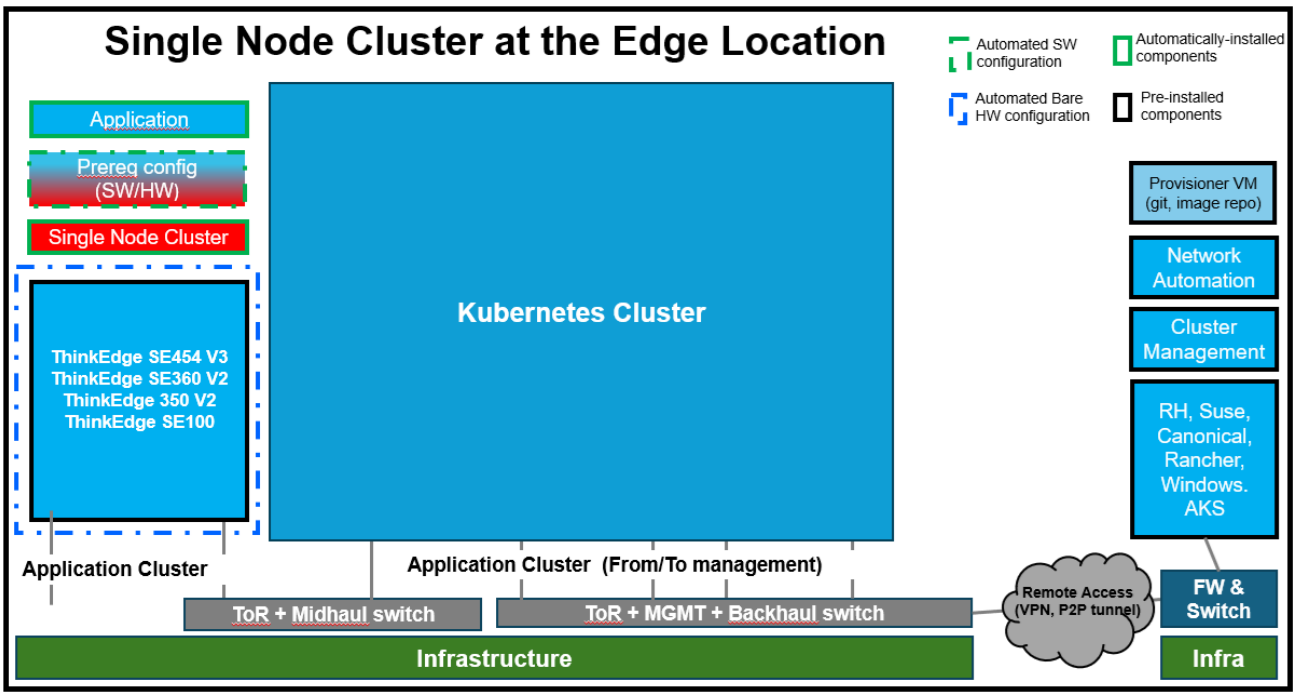

Figure 2 - Single Node Kubernetes Cluster in an Edge Location

The Edge stack is designed as a fully optimized, end-to-end solution capable of supporting a wide range of operating systems—including Red Hat, SUSE, Canonical, Windows, and AKS—within Kubernetes containers. This flexible architecture ensures compatibility with existing IT investments while enabling rapid adoption of emerging technologies.

To support evolving workloads and future scalability, the infrastructure can be deployed as a single-node cluster or expanded to thousands of nodes across distributed edge locations. This flexibility allows applications to be deployed and orchestrated in a containerized environment, ensuring services are delivered precisely when and where they are needed.

With support for automated provisioning, bare metal deployment, and multi-cloud orchestration, the platform is built for future-proofing and operational efficiency. Applications can seamlessly run on ThinkEdge systems, including SE454 V3, SE360 V2, SE350 V2, and SE100, integrated into a Kubernetes-managed cluster with secure remote access and network automation.

This solution empowers organizations to manage edge computing workloads with centralized visibility, simplified operations, and consistent performance from edge to cloud.

.

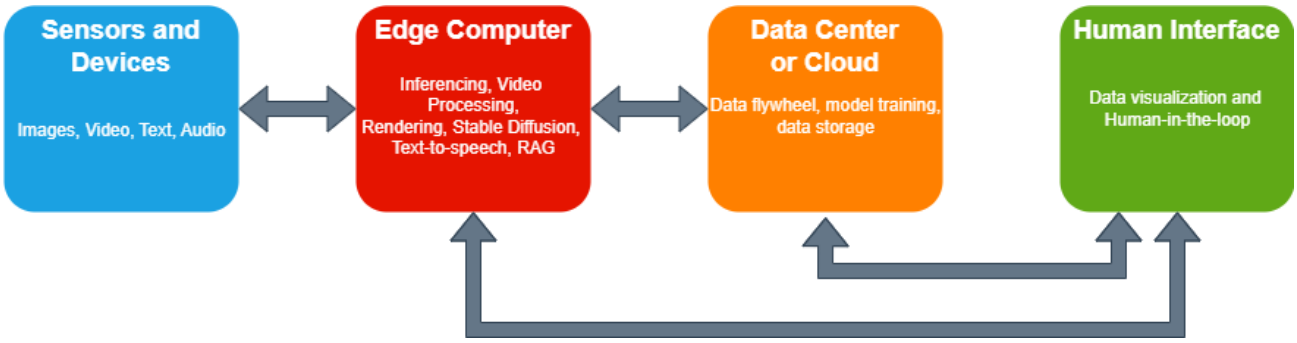

Figure 3 – High level overview of edge platform

Edge computing plays a critical role in a unified AI Hybrid Strategy, seamlessly integrating with data center and cloud environments to deliver a consistent and complementary platform experience.

By deploying sensors, cameras, microphones, and other far-edge devices, data can be captured and processed locally at the Edge—close to where it's generated. This reduces latency, minimizes bandwidth usage by avoiding unnecessary backhaul, and enables real-time inferencing, rendering, stable diffusion, text-to-speech, and retrieval-augmented generation (RAG) directly at the edge nodes.

Modern edge computing has evolved to match the growing demands of AI workloads, enabling intelligent decision-making at the point of action while remaining tightly coupled with centralized infrastructure for model training, data storage, and visualization.

This distributed architecture not only enhances performance and responsiveness but also ensures the infrastructure is scalable, adaptable, and future-ready, supporting a wide range of AI use cases from the edge to the cloud.

Lenovo ThinkEdge Server Portfolio

Lenovo ThinkEdge servers combine cutting-edge hardware, software, and services to address today's most pressing customer challenges while offering a modular, future-ready design to meet evolving business needs. Built on industry-standard x86 technologies and enhanced with Lenovo's unique innovations, these servers deliver exceptional flexibility, scalability, and performance at the edge. Designed for durability and adaptability, ThinkEdge servers can be deployed in a wide range of constrained environments, including space-limited locations, thanks to their compact form factor, low acoustic output, and support for wall-mount installation. This makes them ideal for retail, industrial, and remote edge use cases where space, power, and environmental constraints are critical.

Figure 4 - Lenovo ThinkEdge server wall-mounted in an industrial manufacturing environment

Key advantages of deploying Lenovo ThinkEdge servers include:

- Highly scalable, modular designs to grow with your business
- Industry-leading resilience to save hours of costly unscheduled downtime
- Fast flash technologies for lower latencies, quicker response times, and smarter data management in real time

In the AI area, Lenovo is taking a practical approach to helping enterprises understand and adopt the benefits of ML and AI for their workloads. Lenovo customers can explore and evaluate Lenovo AI offerings in Lenovo AI Innovation Centers to fully understand the value for their particular use case. To improve time to value, this customer-centric approach gives customers proofs of concept for solution development platforms that are ready to use and optimized for AI.

Lenovo ThinkEdge SE100 Edge Server

Edge computing allows data from internet of things devices to be analyzed at the edge of network before being sent to data center or cloud. The Lenovo ThinkEdge SE100 is a purpose-built server that is 1/3 width and significantly shorter than a traditional server, making it ideal for deployment in tight spaces. It can be mounted on a wall, desktop, or mounted in a rack. The ThinkEdge SE100 server is Artificial Intelligence optimized with increased processing power, storage and network closer to where data is generated

Key characteristics:

- **Compact Form Factor** – The SE100 is the smallest Edge server in the Lenovo portfolio, built to be extremely versatile for various environments
- **Low power usage** – Able to operate with less energy requirements compared to most servers
- **Encrypted Data Storage** – Since Edge servers are deployed outside the data center, extra protection is needed to prevent tampering
- **Dust and Vibration Protection** – Added environmental durability ensures the SE100 will not be damaged by more rugged environments
- **Remote Management** – Managing edge servers without being onsite makes IT operations easier

Table 1 – Lenovo ThinkEdge SE100

Form Factor	Base node: Height: 53mm, Width: 142mm, Depth: 278mm, 2.1L With Expansion Kit: Height: 53mm, Width: 214mm, Depth: 278mm, 3.1L
RAID Support	N/A
Mounting Options	Single node mounting options for desktop, VESA, DIN, wall, or ceiling 1U 2N or 1U 3N rackmount
Power	Dual-redundant external power supplies 140W for stand alone 1U2N, 1U3N in 1U enclosure with 300W power adapter
Network Interfaces (Wired)	2x 1GbE 1GbE RJ45 management port
Front I/O	2x USB 3.2 Gen2 (Type-A) 1x USB 3.2 Gen2 (Type-C, with XCC display), 1x USB 3.2 Gen2 (type-C, with CPU display) 2x HDMI 2.0 1x RJ45 for RS232 serial COM

Rear I/O	1x Type-C (Power), with locking screw 1x Type-C (Power, BMC USB 2.0), support power redundancy with locking screw 2x USB 3.2 Gen2 Type-A 2x 1GbE+ 1x GbE management
Systems Management	Lenovo XClarity Administrator
Environmental	Extended operating temperature of 5 to 45°C Satisfy 15G shock & 0.15 Grms vibration, IEC 60068 IP50 dust protection; noise level: 35 dBA (base node) MERV5 with expansion kit
Security	Security 2.0 with ThinkShield Key Vault or XCC management Optional Key Vault SED encrypted storage for boot and data drives Lenovo Trusted Supplier Program, Secure boot, and Smart USB Protections NIST SP800-193 compliance using hardware Root of Trust and Platform Firmware Resilience TPM 2.0 Intrusion tamper protection Optional Kensington keyed lock compatible chassis Type-C Locking screw
Operating Systems	Microsoft Windows 11 IoT Enterprise LTSC, Microsoft Windows 11 Enterprise, Ubuntu 24.04, RHEL 10.0*. Visit lenovopress.lenovo.com/osig for details.
Limited Warranty	3-year customer replaceable unit and onsite service, next business day 9x5; optional service upgrades

Lenovo ThinkEdge SE350 V2 Edge Server

The ThinkEdge SE350 V2 server puts increased processing power, storage and network closer to where data is generated, allowing actions resulting from the analysis of that data to take place more quickly. The SE350 V2 is targeted toward hybrid clouds at the edge, virtualization, NFV, web host, and management server workloads. Note that the SE350 V2 cannot be configured with a GPU.

Key Characteristics:

- **Specialized for CPU workloads** – Additional memory and CPU speed makes this server well suited for use cases revolving around the CPU
- **Hyperconverged infrastructure support** – Either as a ThinkAgile solution or a user-configured system
- **Deployment versatility** – The SE350 V2 can be either deployed in a rack or mounted on a wall

Table 2 – Lenovo ThinkEdge SE350 V2

Form Factor	1U height, half width edge server; Height: 41.7mm, Width: 209mm, Depth: 384mm
RAID Support	RAID 0, 1 for boot drives RAID 0, 1, 5, 10 for data drives
Mounting Options	Single node mounting options for desktop, VESA, DIN, wall, or ceiling 1U2N Enclosure for two nodes side-by-side and internal power supply. Depth: 476.1mm, Height: 1U 1U2N Enclosure for two nodes side-by-side and 4x external power supplies. Depth: 771.1mm, Height: 1U 2U2N Short Depth Enclosure for two nodes sides by side + 4x power supplies: Depth: 476.1mm, Height: 2U Locking bezels and dust filter options
Power	Dual-redundant external power supplies 300W 115V/230V AC Dual DC supply: 12V-48VDC Single internal AC power supply: 500W
Network Interfaces (Wired)	4x 10GbE/25GbE SFP+/SFP28 2x 2.5GbE TSN 1GbE RJ45 management port; or 4x 1GbE 2x 2.5GbE TSN 1GbE RJ45 management port
I/O	Front: 2x USB 3.2 Gen 1 (Type-A) + 1x BMC USB 2.0 (Type-C), 1x Display USB-C (USB 2.0 + Display Port (video) / USB 3.2 Gen 1 auto switch), 1x BMC serial RJ45 management port Rear: 1x RJ45 Console Serial Port (can be disabled)
Systems Management	Lenovo XClarity Administrator with mobile option
Environmental	Extended operating temperature of 0-55°C, up to 40G shock & 1.91Grms vibration, IEC 60068, optional dust filter
Security	ThinkShield Key Vault secure management with motion and intrusion tamper protection Optional Key Vault SED encrypted storage for boot and data drives Lenovo Trusted Supplier Program, Secure boot, and Smart USB Protections Optional Kensington keyed lock compatible chassis Cable locking bezel

Operating Systems	Microsoft Windows Server, SLES, Ubuntu, RHEL, VMware ESXi, vSAN. Visit lenovopress.lenovo.com/osig for details.
Limited Warranty	3-year customer replaceable unit and onsite service, next business day 9x5; optional service upgrades

Lenovo ThinkEdge SE360 V2 Edge Server

The ThinkEdge SE360 V2 is a versatile solution supporting a wide range of workloads, including Augmented Reality, Edge AI & MRP, CDN, NFV, Gaming, and Video Streaming. All the compute power comes inside a 2U height and half width server, making the SE360 V2 a great option for GPU workloads that must be processed close to the source, saving network bandwidth and reducing latency.

Key characteristics:

- **Intelligent System Cooling** – Optimized for quiet operations in occupied spaces & dust filtering for ultimate reliability being twice as quiet compared to all other competitive Edge servers on the market
- **Broad Networking, I/O and GPU** – Expansion for integrating control systems and deploying AI being the smallest edge AI server on the market, rick networking options with WLAN support and up to 100gbE ethernet connection
- **Energy Efficient** – Configurable Energy Optimized Mode for workloads performance to reduce energy consumption
- **Enhanced Security** – Lockable bezel with smart filtering technology supported with disk encryption, ThinkShield Activation, system lockdown, movement detection and tamper protection
- **Compact Form Factor** – 78% smaller than competitive products for easy deployment
- **Data Protection** – If the system is moved or tampered, beyond the configuration threshold, the data is automatically lock-up
- **Unique Mounting Options** – Short-depth for ultimate space savings and secure cable routing

- **Shock & Vibration Resistant** – Designed to perform with high level of shock & vibration up to 40G

Table 3 – Lenovo ThinkEdge SE360 V2

Form Factor	2U height, half width edge server; Height: 84.5mm, Width: 212mm, Depth: 317.5mm
RAID Support	RAID 0, 1 for boot drives RAID 0, 1, 5, 10 for data drives
Mounting Options	Single node mounting options for desktop, VESA, DIN, wall, or ceiling 2U2N Short Depth Enclosure for two nodes sides by side: Depth: 466mm, Height: 2U Locking bezels and dust filter options
Power	Dual-redundant external power supplies 300W 115V/230V AC Dual DC supply: 12V-48VDC Single internal AC power supply: 500W
Network Interfaces (Wired)	4x 10GbE/25GbE SFP+/SFP28 2x 2.5GbE TSN 1GbE RJ45 management port; or 4x 1GbE 2x 2.5GbE TSN 1GbE RJ45 management port
Network Interfaces (Wireless)	Four wireless SMA connectors for WLAN 1 wireless SMA connector for Bluetooth WLAN 128/192-bits encrypted WPA2, WPA3 802.11 a/b/g/n/ac/ax Geotracking
I/O	Front: 2x USB 3.2 Gen 1 (Type-A) + 1x BMC USB 2.0 (Type-C), 1x Display USB-C (USB 2.0 + Display Port (video) / USB 3.2 Gen 1 auto switch), 1x BMC serial RJ45 management port Rear: 1x RJ45 Console Serial USB and Console ports can be disabled
Systems Management	Lenovo XClarity Administrator with mobile option
Environmental	Extended operating temperature of 0-55°C, support -20~65°C with certain configuration, up to 40G shock & 1.91Grms vibration, IEC 60068, optional dust filter, IP3X, Marine certification

Security	ThinkShield Key Vault secure management with motion and intrusion tamper protection Optional Key Vault SED encrypted storage for boot and data drives Lenovo Trusted Supplier Program, Secure boot, and Smart USB Protections, Lenovo WLAN Security, Lenovo Bluetooth Security Optional Kensington keyed lock compatible chassis Nationz TPM 2.0 for customers in China Cable locking bezel Geotracking
Operating Systems	Microsoft Windows Server, SLES, Ubuntu, RHEL, VMware ESXi, vSAN.
Limited Warranty	3-year customer replaceable unit and onsite service, next business day 9x5; optional service upgrades

Lenovo ThinkEdge SE455 V3 Edge Server

The ThinkEdge SE455 V3 is Lenovo's most powerful edge server, containing an AMD EPYC CPU and up to 2 NVIDIA L40s GPUs. As a 2U server with a short depth case, it can be mounted in a 2-post or 4-post rack. This server is well suited for transformative AI systems, especially those that need video processing and AI inference at the edge.

Key characteristics:

- **Short depth chassis for use outside the data center** – 2U short-depth form factor with front access I/O ensuring dust filtering, extended temperature operation (5 to 55 C), shock & vibration resistance
- **Very flexible I/O and storage configuration to support most demanding edge applications** – Optional internal drive bays, optional second PCIe riser, high speed Gen5 x16 OCP achieving 2x more strage bays delivering over 490 TB of total storage, 20% more PCIe slots than competitive edge servers
- **Quiet Operation** – Acoustic modes, enhanced heat dissipation via noise controlling with UEFI settings, larger CPU heatsink
- **Energy Efficient** – AMD Edge optimized processers fine-tuned with UEFI operating modes resulting in up to 64 cores in socket achieving 32% better energy efficiency.
- **GPU & Memory Rich** – Up to 6x single-width GPUs or 2x double-width GPUs supported with up to 768GB of system memory

- **Increased security to protect data** – Encrypted disk and ThinkShield activation achieving tamper protection and system lockdown in case of security incidents

Table 4 – Lenovo ThinkEdge SE455 V3

Form Factor	2U rack server 440mm depth
Network Interface	1/10/25/100 Gb LOM adapter in OCP 3.0 slot Up to 6x 1/10/25/100/200 Gb PCIe network adapters
Ports, Buttons	Front: 1x Power Button (with green LED), 1x System Locator (with blue LED), 1x NMI button, 1x USB-C, 2x USB 3.0, 1x USB 2.0 for XCC2, 1x RJ45 for XCC2, 1x Diagnostic handset, COM port via PCI slot
LED	Security (green), Attention (yellow), XCC2 Ethernet (Link and Activity),
HBA/RAID Support	HW RAID with/without cache or SAS HBAs 4350-8i, 5350-8i, 440-8i, 540-8i, 940-8i with Supercap 2 and 4 port Qlogic QLE277x 32Gb Fibre Channel HBA
Power	Dual redundant power supplies AC (1100W Platinum/Titanium, 1800W Platinum) or Dual redundant power supplies -48V DC 1100W
Security	Security 2.0 with ThinkShield or XCC Management, Security Bezel, Tamper detection, Rack Security Bracket, Encrypted SSD, System Lockdown, Silicon Root of Trust, TPM 2.0, System Guard, AMD Infinity Guard, NIST SP800-193 compliance using hardware Root of Trust and Platform Firmware Resilience
Easy to Deploy	ThinkShield or XCC Managed lockdown mode and SED Lenovo Open Cloud Automation (LOC-A) and XClarity Administrator/Pro
Environmental	5°C to 55°C standard support ; NEBS 3 -5°C to 55°C (< 96 hours) 40/45dBA Acoustic Modes MERV2 Air Filter with clog detecting airflow sensor
Systems Management	Lenovo XClarity Controller (AST2600) DC-SCM 2.0
OS Support	Tier 1/1.5: Microsoft 2019/2022, Red Hat 8.x, 9.x, SLES 15, Ubuntu 20.x/22.x,VMware 7.x/8.x Tier 2: XenServer 8.2 Tier 3: Alma Linux 8.x/9.x, Rocky Linux 8.x/9.x Future Support: Windows 11 IOT Enterprise LTSC
Limited Warranty	3-year customer replaceable unit and onsite service, next business day 9x5 Optional service upgrades

Workload-Optimized AI Hardware Guide

This Edge AI Server Sizing Guide provides recommended configurations for Lenovo ThinkEdge platforms, optimized for a range of AI workloads**.** Each model size from X-Small to X-Large is tailored to specific performance needs, balancing compute power, memory, storage, and GPU acceleration. These reference configurations are designed to support key use cases such as computer vision, natural language processing (NLP), self-checkout systems, and LLM-based inference. By matching workloads with the appropriate hardware profile, customers can achieve efficient, scalable AI deployments at the edge.

Table 5 – Edge AI Server Sizing Guide

Server	Size	GPU	CPU	RAM	Storage	AI Task	Framework	Model type examples
SE100	X-Small	None	14 core Intel Ultra 5-225H	32GB (2 x 16)	2 x 480GB	Image identification, Natural Language Processing, Data Analytics	OpenVINO	BERT, YOLO
		1x RTX A1000				Machine Learning, Small Scale Inference	PyTorch, TensorFlow, vLLM, TensorRT	Convolutional Neural Network, other Computer Vision Models
	Small	1x RTX 2000E Ada						
SE350 V2	Small	None	8 core Intel Xeon D-2733NT	128 GB (2 x 64)		Video and Image Analysis	OpenVINO	Specialized Machine Learning Models
SE360 V2	Small	1x NVIDIA A2	8 core Intel Xeon D-2733NT	32GB (2 x 16)		Text-to-speech, Image and Video AI	vLLM, TensorRT, Pytorch, TensorFlow	EfficientDet, BERT-Large, Llama 3.2 3B
	Medium	1x NVIDIA L4	16 core Intel Xeon D-2775TE		2 x 960 GB	LLM/VLM inference, Stable Diffusion at 512x512 on Per-GPU basis (See Test Results)	vLLM, TensorRT-LLM	LlaVA-NeXT, LlaVa-OneVision, MiniCPM-V
	Large	2x NVIDIA L4			2 x 1.92 TB			
SE455 V2	Large	3x NVIDIA L4	32 core AMD EPYC 8324P	64 GB (2 x 32)				
	X-Large	2x NVIDIA L40S	32 core AMD EPYC 8324PN	512 GB (6 x 96)	2 x 3.84 TB	Multi-tenancy, Fine-tuning, RAG, Stable diffusion at 1024x1024		

Note: All SE360 V2 models support optional Qualcomm A100 or Intel Flex 140. SE455 V3 supports only the Qualcomm A100.

Test Overview

Vision Language Models (VLMs) are a class of generative AI models designed to interpret and respond to visual inputs such as images and videos through natural language. In smart city and public safety applications, VLMs can be used for tasks like situational awareness, anomaly detection, and identifying persons of interest—all in real time.

To assess the performance of these models in edge environments, the Lenovo team has developed a purpose-built benchmark that evaluates the throughput of VLMs running on Lenovo ThinkEdge servers. This benchmark provides valuable insights into how efficiently these models can process and analyze video frames at the edge, enabling scalable, low-latency AI deployments for critical use cases.

The benchmark evaluates the time taken for large language models (LLMs) to generate a textual response from a 6-second video processed at 4 frames per second (24 frames total). Each model is prompted with 5 different sample videos and 5 sample prompts randomly a total of 10 times. The 5 prompts are: "Describe the characters/objects in this video", "Is anything dangerous occurring in this video?", "Answer yes or no: is there an animal in this video?", "Why is this video funny?", "What is shown in this video? Be concise.” These prompts were chosen to cover a variety of scenarios that VLMs could be used for, ensuring that contextual reasoning is a factor in the benchmark.

Tests were conducted on the Lenovo ThinkEdge SE360 V2 equipped with 2 NVIDIA L4 GPUs, as found in the out-of-the-box configuration with no optimizations or tuning. Two inference platforms were used, vLLM and TensorRT-LLM. Since model support varies, three models were tested on vLLM: LLaVA-NeXT, MiniCPM-V, and LLaVA-OneVision, while the latter was used as the common model for comparison across both platforms.

The LLaVA models are chosen to fit within a single L4 GPU, enabling dual-model concurrency, while MiniCPM-V requires both GPUs. Frame rate and video length were set to achieve a consistent input size (24 frames), but results are model-dependent due to limitations like context length, token output (capped at 10 for this test), and model architecture.

Both platforms were installed using default configurations via pip, with the sampling temperature set to 0 to ensure deterministic, repeatable outputs. The results primarily reflect each platform's inference efficiency under consistent conditions. However, selecting the right platform and model should also consider factors such as model compatibility, accuracy requirements, and the specific context of the deployment.

This benchmark helps identify which platform and model combinations offer the best balance of performance, efficiency, and model fidelity for edge inferencing use cases in bandwidth-constrained, latency-sensitive environments.

Test Results

The Lenovo SE360 V2 with 2 L4 GPUs was able to achieve impressive video throughput in frames per second, and the Video Language Models demonstrate contextual understanding of actions that occur in the videos. Consider the following example.

The image below shows a single frame from a longer video of a vehicle incident, with the model's response to the question "Is anything dangerous occurring in this video?". This is just one of the scenarios tested here and is used for demonstrative purposes. The model's textual response is superimposed onto the image.

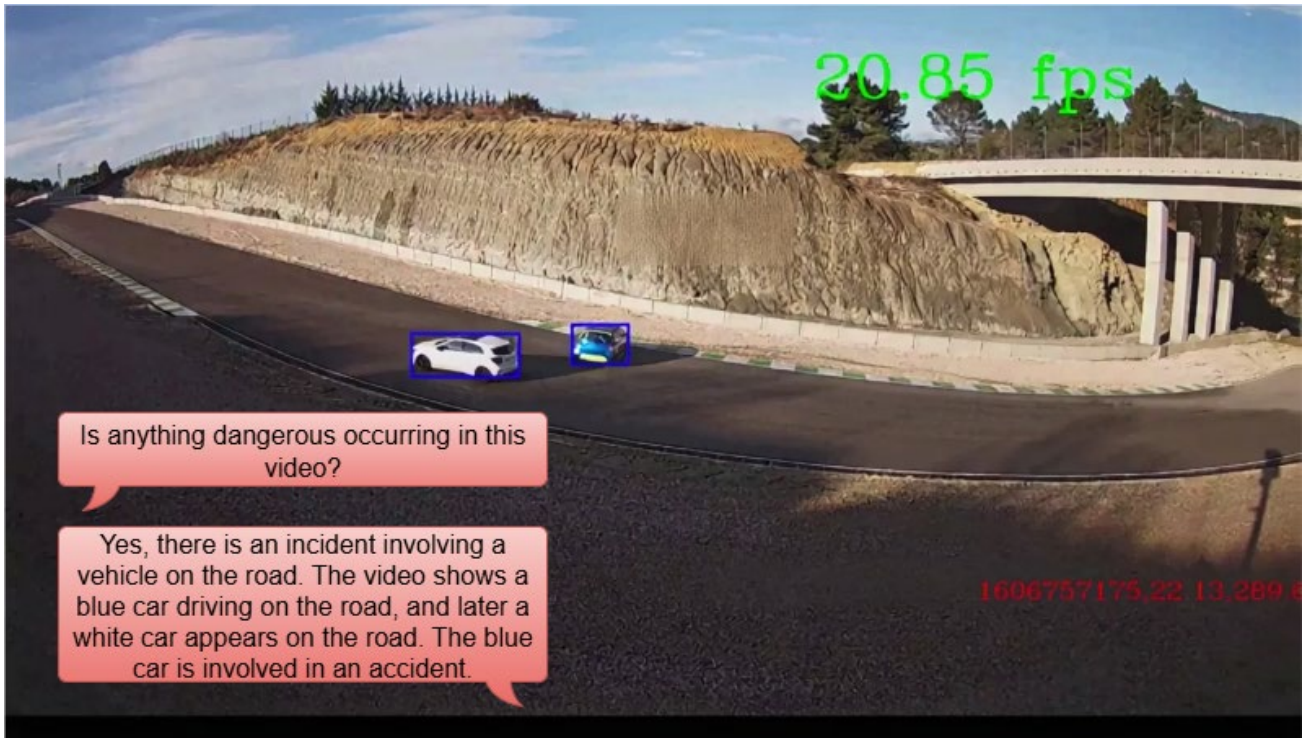

Figure 5 - AI video analytics detecting and explaining a vehicle incident in real time.

The model accurately identifies a vehicle incident in the video and assesses its potential danger. It recognizes the sequence of events and tracks objects in relation to one another. This showcases the model's capability to contextualize visual information and respond to natural language queries with precision. The green and red text on the screen originates from the source video feed and is not relevant to the analysis.

For benchmark results on CPU-only AI workloads with OpenVino, please see Lenovo Press Article 2249: Implementing Generative AI Using Intel Xeon 6 CPUs

To evaluate TensorRT-LLM's performance relative to vLLM, we compared results using the LLaVA-OneVision model, which is supported on both platforms. TensorRT-LLM consistently demonstrated slightly higher throughput for the same model, indicating greater efficiency in frame processing. While this advantage may not extend to all models, similar trends have been reported in external benchmarking efforts. For instance, the

Llava-NeXT model shows exceptional performance on vLLM, highlighting that both platforms can deliver impressive results depending on the use case.

It is important to note that model compatibility varies between TensorRT-LLM and vLLM. While TensorRT-LLM may deliver higher throughput, vLLM may support a broader range of models or offer features better suited for certain applications. For inference scenarios where model fidelity and accuracy are more critical than speed, models such as MiniCPM-V may be more appropriate. These larger models often produce higher-quality outputs but require increased computational resources, as observed in our test results.

Ultimately, selecting the optimal inference platform and model depends on the characteristics of the video input and the desired outcomes. Benchmarking across a representative dataset is recommended to identify the best fit for the specific use case and deployment constraints.

For full test data, please contact your sales representative.

Life Cycle Management Software

Deployment orchestration and management of the ThinkEdge servers can be done in container level integrating any validated solution in the environment. Lenovo XClarity One & 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 One

Lenovo XClarity One is a management-as-a-service offering for hybrid-cloud management of on-premises data-center assets. XClarity One makes use of local management hubs across multiple sites to collect inventory, incidents, and service data, and to provision resources. XClarity One provides a modern, intuitive interface that centralizes IT orchestration, deployment, automation, and support from edge to cloud, with enhanced visibility into infrastructure performance, usage metering, and analytics.

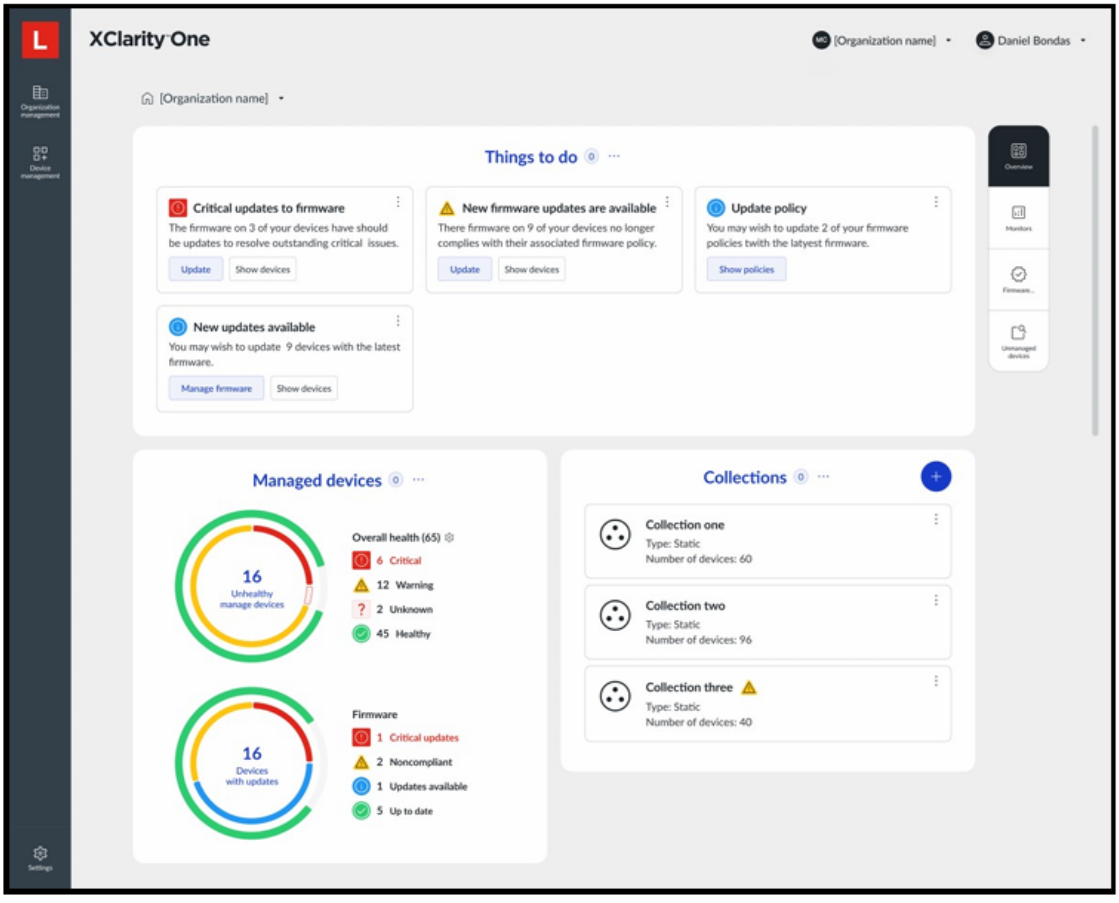

Figure 6 – Lenovo XClarity One Dashboard

Features

Lenovo XClarity One offers the following key features:

- **Unified Dashboard**: Centralized view for monitoring and managing infrastructure.
- **Firmware Management**: Streamlined updates across supported devices.
- **Security**: Built-in security for device communication and data handling.

Lenovo XClarity One manages devices through lightweight software components known as Management Hubs. These hubs operate as virtual appliances deployed on-premises, typically within customer datacenters across one or more locations. This architecture enables low-latency communication, rapid response times, and enhanced data privacy by keeping sensitive operations local.

The supported hub, Lenovo XClarity Management Hub 2.0, facilitates secure provisioning and management of Lenovo devices across distributed environments.

The following figure illustrates the XClarity One infrastructure architecture, highlighting the logical placement of Management Hubs within the environment.

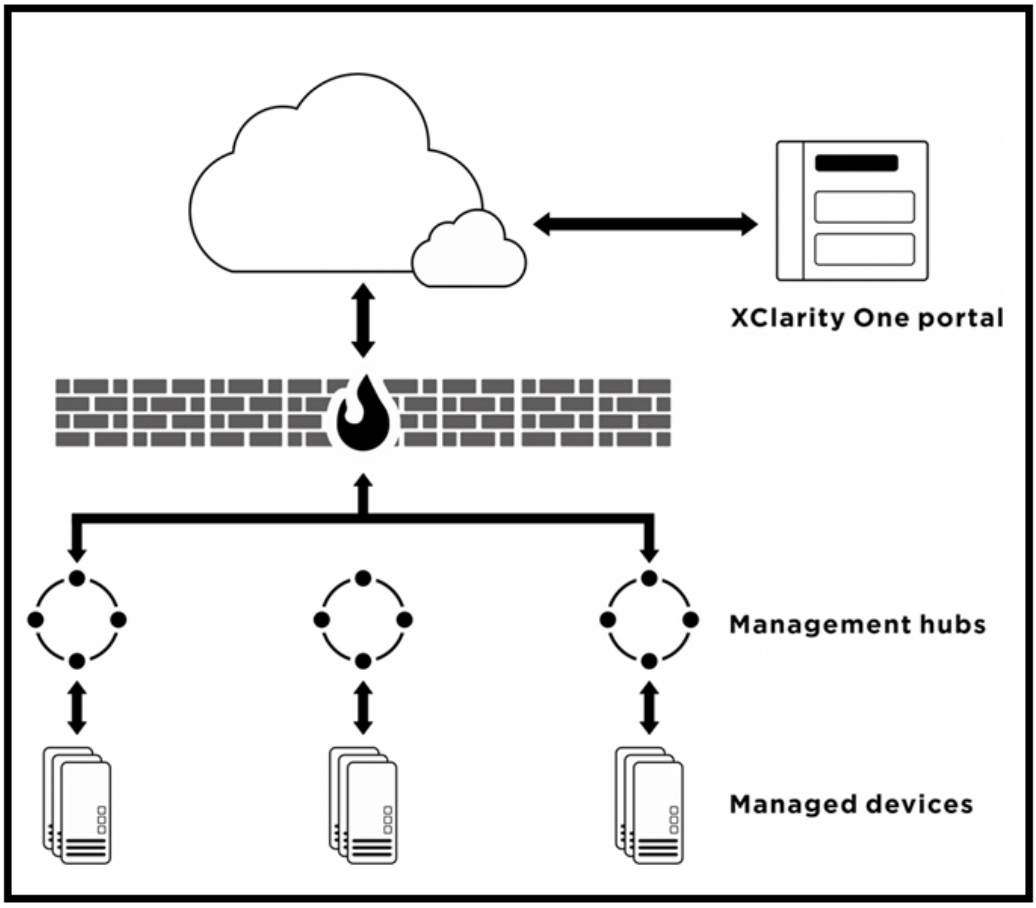

Figure 7 - XClarity One infrastructure

XClarity One Overview

The XClarity One Dashboard is a cloud-based interface designed for fast, efficient resource discovery and task execution. It provides a centralized, intuitive view of your infrastructure, enabling administrators to quickly locate devices and act.

XClarity One enhances security by requiring only a single secure connection between the cloud portal and your on-premises or private cloud-managed devices. This reduces exposure and simplifies security management while maintaining robust connectivity.

Key Capabilities of Lenovo XClarity One

With its simplified, unified dashboard, Lenovo XClarity One enables a broad set of administrative functions, including:

- Discovery – Automatically detect Lenovo rack servers and other infrastructure.
- Inventory Management – View and track hardware inventory and device status at a glance.
- System Health & Monitoring – Real-time monitoring of managed endpoints.
- Alerts & Events – Centralized view of system events and alerts for faster issue identification.
- Firmware Updates – Simplified management and deployment of firmware across devices.
- Remote Control – Access and control systems securely from a single console.
- Call Home & Support Integration – Automatically upload service data and monitor ticket status.
- Predictive Failure Analytics – Identify potential issues before they impact operations.
- Upward Integration – Easily integrate with higher-level management systems.
- Risk Mitigation & Resiliency – Improve uptime with proactive detection and failover readiness.
- Controlled Infrastructure Accessibility – Secure, role-based access to infrastructure resources.

Operational Benefits

- Fast Time to Value: Automatic discovery of new and existing Lenovo rack servers accelerates onboarding and management.
- Comprehensive Visibility: The dashboard offers a centralized view of infrastructure health, alerts, and events. When an issue is detected on a managed device, it is reported to XClarity One and displayed across the dashboard, status bar, and system-specific detail views.
- Security and Simplicity: The use of a single, secure connection ensures minimal attack surface while maintaining full visibility and control.

Firmware management

Firmware management is streamlined using firmware compliance policies, which can be assigned to support managed endpoints to ensure that firmware remains up to date and aligned with compliance standards. When validated firmware levels differ from pre-defined policies, custom firmware compliance policies can be created or edited to reflect approved configurations. In addition to policy-based updates, firmware versions that are newer than the currently installed versions can be applied and activated directly, either on individual managed

endpoints or across groups, without requiring compliance policies. This flexibility supports both structured policy-based management and on-demand updates as needed.

Security

Lenovo XClarity One provides robust security features to protect infrastructure and support compliance:

- Standards Compliance: Supports NIST SP 800-131A and FIPS 140-3 with self-signed or external SSL certificates.
- Two-Factor Authentication: Required for all users to ensure secure access.
- Access Control: Individual management functions can be disabled to reduce exposure.
- Audit Logging: Tracks user actions such as logins, account creation, and password changes for full traceability.

Lenovo Open Cloud Automation (LOC-A)

LOC-A is integrated into Lenovo XClarity One, a unified, cloud-based platform for managing and automating infrastructure from edge to cloud. LOC-A enables streamlined Edge IT automation, authentication, and provisioning within XClarity One, making it easier to deploy and manage large, distributed compute environments. Key benefits include:

- **Near-Zero Touch Provisioning** – Near-Zero Touch Provisioning (nZTP) enables scalable, late-binding deployment of edge servers, minimizing the need for field technician involvement. Instead of confirming early binding, LOC-A focuses on discovering what has arrived, with most tasks handled by the deployment admin via the LOC-A portal.
- **OS Deployment** – LOC-A performs remote Operating System (OS) deployment on bare-metal nodes, automatically triggering and monitoring the process to ensure consistency and security across the fleet. This approach eliminates the need for Golden OS images in manufacturing or staging by provisioning the OS directly in the field. Additionally, LOC-A uses OS side-loading, transferring the OS image during provisioning, saving bandwidth by avoiding low-throughput WAN streaming.
- **Plugin Mechanism for Partner Integration** – LOC-A features a plugin mechanism that allows Lenovo partners to easily create and integrate their own automated deployment configurations. This system leverages LOC-A's advanced features, enabling key functions for partner platforms, including:
    - Bare-metal server provisioning and onboarding
    - Orchestration of edge-node deployments
    - Edge cluster and node instance creation (including OS deployment)
    - Access to a smart naming convention engine for hostname, Fully Qualified Domain Name (FQDN), etc. The Edge Platform uses this plugin mechanism to integrate seamlessly with LOC-A.
- **Northbound API –** LOC-A also provides a secure, public northbound API for deeper integration with application orchestrators like Edge Platform or Operations Support System/Business Support System (OSS/BSS) platforms. This API enhances integration capabilities beyond the plugin mechanism, accelerating new feature enablement across platforms

Summary

This reference architecture presents Lenovo's approach to deploying and scaling AI inference workloads at the edge using ThinkEdge servers. It details validated server configurations, across small, medium, and large sizing models tailored for specific edge AI workloads across industries such as retail, manufacturing, healthcare, smart cities, and financial services.

Key highlights include:

- **Edge AI Use Cases**: Supports diverse applications such as self-checkout, quality inspection, patient monitoring, public safety, asset tracking, and fraud detection.
- **Workload Optimization**: Offers nine pre-sized configurations (X-Small to X-Large), aligning compute, memory, storage, and GPU acceleration with AI workloads like computer vision, NLP, audio intelligence, and LLM-based inference.
- **AI Frameworks and Tools**: Supports TensorRT-LLM, vLLM, and OpenVINO for optimized deployment across NVIDIA- and Intel-based ThinkEdge systems.
- **Hybrid AI Strategy**: Enables distributed inferencing at the edge while maintaining centralized training and governance, ensuring data privacy, low latency, and efficient bandwidth usage.
- **Platform Integration**: Features robust lifecycle management with Lenovo XClarity One and LOC-A for zero-touch provisioning, secure firmware updates, and unified orchestration across multi-node, multi-GPU environments.

Through this flexible and scalable architecture, Lenovo empowers organizations to deliver real-time AI insights where data is generated, driving faster decisions, enhanced customer experiences, and operational efficiency at the edge.

.

Appendix: Lenovo Bill of Materials

This appendix contains the bill of materials (BOMs) for computational servers and a storage server.

X-Small SE 100 No GPU

7DGRCTO1WW	Node : ThinkEdge SE100 - 3 Year Warranty	1
C31D	ThinkEdge SE100 Chassis	1
C30L	ThinkEdge SE100 Planar with Intel Core Ultra 5 225H ,14C, 28W, 1.7GHz	1
C39J	ThinkEdge 16GB TruDDR5 6400MHz CSODIMM	2
C8V3	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	1
BYF8	ThinkSystem M.2 ER3 480GB Read Intensive SATA 6Gb NHS SSD	1
C30D	ThinkEdge SE100 Expansion Connection Cover	1
C39R	ThinkEdge 140W 230V/115V External Power Supply	2
A4VP	1.0m, 10A/100-250V, C13 to C14 Jumper Cord	2
C31J	ThinkEdge SE100 Bottom Rubber Feet	1
C31A	ThinkEdge SE100 Fan Module	1
C31C	ThinkEdge SE100 Port Dust Cover Kit	1
BRPJ	XCC Platinum	1
C8U9	Top-Cover Thermal Gap Pad Kit	1
C319	ThinkEdge SE100 Node Cosmetic Cover	1
C308	ThinkEdge SE100 M.2 Holder	1
C8UC	Front I/O Panel	1
C8UA	Bottom-Cover Thermal Gap Pad Kit	1
7S0XCTO5WW	XClarity Controller Platin-FOD	1
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	1

X-Small SE100

7DGRCTO1WW	SE100 X-Small : ThinkEdge SE100 - 3 Year Warranty	1
C31D	ThinkEdge SE100 Chassis	1
C30L	ThinkEdge SE100 Planar with Intel Ultra 5-225H ,14C, 28W, 1.7GHz	1
C39J	ThinkEdge 16GB TruDDR5 6400MHz CSODIMM	2
C8V3	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	1
BS2P	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	1
C39N	ThinkSystem NVIDIA RTX A1000 8GB PCIe Gen4 Active GPU	1
C39R	ThinkEdge 140W 230V/115V External Power Supply	1
A4VP	1.0m, 10A/100-250V, C13 to C14 Jumper Cord	1
C31J	ThinkEdge SE100 Bottom Rubber Feet	1
C31A	ThinkEdge SE100 Fan Module	1
C31C	ThinkEdge SE100 Port Dust Cover Kit	1
BRPJ	XCC Platinum	1
C8U9	Top-Cover Thermal Gap Pad Kit	1
C319	ThinkEdge SE100 Node Cosmetic Cover	1
C308	ThinkEdge SE100 M.2 Holder	1
C8UC	Front I/O Panel	1
C8UB	Expansion Kit Rubber Feet	1
C8UA	Bottom-Cover Thermal Gap Pad Kit	1
C30F	SE100 Expansion Kit for Active Cooling GPU	1
7S0XCTO5WW	XClarity Controller Platin-FOD	1
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	1

Small SE100

7DGRCTO1WW	Node : ThinkEdge SE100 - 3 Year Warranty	1
C31D	ThinkEdge SE100 Chassis	1
C30L	ThinkEdge SE100 Planar with Intel Core Ultra 5 225H ,14C, 28W, 1.7GHz	1
C39J	ThinkEdge 16GB TruDDR5 6400MHz CSODIMM	2
C8V3	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	1
BYF8	ThinkSystem M.2 ER3 480GB Read Intensive SATA 6Gb NHS SSD	1
C39P	ThinkSystem NVIDIA RTX 2000E Ada 16GB PCIe Active GPU	1
C39R	ThinkEdge 140W 230V/115V External Power Supply	2
A4VP	1.0m, 10A/100-250V, C13 to C14 Jumper Cord	2
C31J	ThinkEdge SE100 Bottom Rubber Feet	1
C31A	ThinkEdge SE100 Fan Module	1
C31C	ThinkEdge SE100 Port Dust Cover Kit	1
BRPJ	XCC Platinum	1
C8U9	Top-Cover Thermal Gap Pad Kit	1
C319	ThinkEdge SE100 Node Cosmetic Cover	1
C308	ThinkEdge SE100 M.2 Holder	1
C8UC	Front I/O Panel	1
C8UB	Expansion Kit Rubber Feet	1
C8UA	Bottom-Cover Thermal Gap Pad Kit	1
C30F	SE100 Expansion Kit for Active Cooling GPU	1
7S0XCTO5WW	XClarity Controller Platin-FOD	1
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	1

Small SE350 V2 No GPU

7DA9CTO1WW	Node : ThinkEdge SE350 V2 - 3 Year Warranty	1
BS3S	ThinkEdge SE350 V2 Chassis	1
BS3T	ThinkEdge SE350 V2 4x 10/25Gb, 2x 2.5Gb(TSN) I/O Module	1

BS41	ThinkEdge SE350 V2/SE360 V2 Planar with Intel Xeon D-2733NT 8C 80W 2.1 GHz	1
B966	ThinkSystem 64GB TruDDR4 3200 MHz (2Rx4 1.2V) RDIMM	2
BQ1V	ThinkSystem 7mm 5400 PRO 480GB Read Intensive SATA 6Gb HS SSD	1
BS48	ThinkEdge SE350 V2 7mm SSD Module	1
BS46	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD (with Heatsink)	1
BUGP	ThinkEdge SE350 V2 AC Power Input Board	1
BWK7	ThinkEdge SE350 V2 500W 230V/115V Non-Hot Swap Power Supply	1
6201	1.5m, 10A/100-250V, C13 to C14 Jumper Cord	1
BS4E	ThinkEdge 130mm USB-C to VGA Display Cable	1
B6Q3	ThinkEdge Rubber Feet	1
BRPJ	XCC Platinum	1
B8L4	ThinkSystem 7mm Tray Filler	3
BS4V	ThinkEdge SE350 V2 Front IO Bezel (25G/10G) Assembly	1
BS4M	ThinkEdge SE350 V2 Operational Panel Module	1
BS4L	ThinkEdge SE350 V2 Bridge Board	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
7S0YCTO1WW	Lenovo Open Cloud Automation w/Support	1
SD2S	Lenovo Open Cloud Automation - nZTP with Device Management platform onboarding for 1-socket ThinkEdge server with 1 year support. Price per node	1
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1

Small SE360 V2

7DAMCTO1WW	Node : ThinkEdge SE360 V2 - 3 Year Warranty	1
BS56	ThinkEdge SE360 V2 Chassis	1
BS58	ThinkEdge SE360 V2 4x 1Gb, 2x 2.5Gb(TSN) I/O Module	1

BS41	ThinkEdge SE350 V2/SE360 V2 Planar with Intel Xeon D-2733NT 8C 80W 2.1 GHz	1
B963	ThinkSystem 16GB TruDDR4 3200 MHz (2Rx8 1.2V) RDIMM	2
BSW6	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD (with Heatsink)	1
BS5M	ThinkEdge SE360 V2 M.2 Cabled Adapter Module	1
BS46	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD (with Heatsink)	1
BQZT	ThinkSystem NVIDIA A2 16GB PCIe Gen4 Passive GPU w/o CEC	1
BS5J	ThinkEdge SE360 V2 Riser Assembly (PCIe Riser + 7mm Backplane)	1
BUGU	ThinkEdge SE360 V2 AC Power Input Board	1
BW8U	ThinkEdge SE360 V2 500W 230V/115V Non-Hot Swap Power Supply	1
6313	2.8m, 10A/120V, C13 to NEMA 5-15P (US) Line Cord	1
BS5W	ThinkEdge SE360 V2 Fan Assembly (Front to Rear)	1
BS4E	ThinkEdge 130mm USB-C to VGA Display Cable	1
B6Q3	ThinkEdge Rubber Feet	1
BRPJ	XCC Platinum	1
BUGS	ThinkEdge SE350 V2/ SE360 V2 7mm Tray Filler	2
BS69	ThinkEdge SE360 V2 Top Cover	1
BTJK	ThinkEdge SE360 V2 Air Baffle for Processor	1
BS66	ThinkEdge SE360 V2 IO Cover Assembly for 1GbE I/O Module	1
BS64	ThinkEdge SE360 V2 Rear Operational Panel Module	1
BS63	ThinkEdge SE360 V2 Operational Panel Module	1
BUGV	ThinkEdge SE360 V2 AC Power Module Board Air Baffle	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
7S0YCTO1WW	Lenovo Open Cloud Automation w/Support	1
SD2S	Lenovo Open Cloud Automation - nZTP with Device Management platform onboarding for 1-socket ThinkEdge server with 1 year support. Price per node	1
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1

Medium SE360 V2

7DAMCTOAWW	Node : ThinkEdge SE360 V2 - 3 Year Warranty with Controlled GPU	1
BS56	ThinkEdge SE360 V2 Chassis	1

BS58	ThinkEdge SE360 V2 4x 1Gb, 2x 2.5Gb(TSN) I/O Module	1
BS42	ThinkEdge SE350 V2/SE360 V2 Planar with Intel Xeon D-2775TE 16C 100W 2.0 GHz	1
B963	ThinkSystem 16GB TruDDR4 3200 MHz (2Rx8 1.2V) RDIMM	2
BZEF	ThinkSystem M.2 N-30m2 960GB Read Intensive NVMe PCIe 3.0 x4 NHS SSD (Industrial)	1
BS5M	ThinkEdge SE360 V2 M.2 Cabled Adapter Module	1
BQUJ	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD (with Heatsink)	1
BS2C	ThinkSystem NVIDIA L4 24GB PCIe Gen4 Passive GPU	1
BS5J	ThinkEdge SE360 V2 Riser Assembly (PCIe Riser + 7mm Backplane)	1
BUGU	ThinkEdge SE360 V2 AC Power Input Board	1
BW8U	ThinkEdge SE360 V2 500W 230V/115V Non-Hot Swap Power Supply	1
6313	2.8m, 10A/120V, C13 to NEMA 5-15P (US) Line Cord	1
BS5W	ThinkEdge SE360 V2 Fan Assembly (Front to Rear)	1
BS4E	ThinkEdge 130mm USB-C to VGA Display Cable	1
B6Q3	ThinkEdge Rubber Feet	1
BRPJ	XCC Platinum	1
BUGS	ThinkEdge SE350 V2/ SE360 V2 7mm Tray Filler	2
BS69	ThinkEdge SE360 V2 Top Cover	1
BTJK	ThinkEdge SE360 V2 Air Baffle for Processor	1
BS66	ThinkEdge SE360 V2 IO Cover Assembly for 1GbE I/O Module	1
BS64	ThinkEdge SE360 V2 Rear Operational Panel Module	1
BS63	ThinkEdge SE360 V2 Operational Panel Module	1
BUGV	ThinkEdge SE360 V2 AC Power Module Board Air Baffle	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
7S0YCTO1WW	Lenovo Open Cloud Automation w/Support	1
SD2S	Lenovo Open Cloud Automation - nZTP with Device Management platform onboarding for 1-socket ThinkEdge server with 1 year support. Price per node	1

Large SE360 V2

7DAMCTOAWW	SE360 V2 Large : ThinkEdge SE360 V2 - 3 Year Warranty with Controlled GPU	1
BS56	ThinkEdge SE360 V2 Chassis	1

BS58	ThinkEdge SE360 V2 4x 1Gb, 2x 2.5Gb(TSN) I/O Module	1
BS42	ThinkEdge SE350 V2/SE360 V2 Planar with Intel Xeon D-2775TE 16C 100W 2.0 GHz	1
B963	ThinkSystem 16GB TruDDR4 3200 MHz (2Rx8 1.2V) RDIMM	2
BZEG	ThinkSystem M.2 N-30m2 1.92TB Read Intensive NVMe PCIe 3.0 x4 NHS SSD (Industrial)	1
BS5M	ThinkEdge SE360 V2 M.2 Cabled Adapter Module	1
BYLN	ThinkSystem M.2 N-30m2 1.92TB Read Intensive NVMe PCIe 3.0 x4 NHS SSD (Industrial)	1
BS2C	ThinkSystem NVIDIA L4 24GB PCIe Gen4 Passive GPU	2
BS5E	ThinkEdge SE360 V2 Riser Assembly (PCIe Riser + PCIe Riser) w/ Geotracking	1
BUGU	ThinkEdge SE360 V2 AC Power Input Board	1
BW8U	ThinkEdge SE360 V2 500W 230V/115V Non-Hot Swap Power Supply	1
6313	2.8m, 10A/120V, C13 to NEMA 5-15P (US) Line Cord	1
BS5W	ThinkEdge SE360 V2 Fan Assembly (Front to Rear)	1
BS4E	ThinkEdge 130mm USB-C to VGA Display Cable	1
B6Q3	ThinkEdge Rubber Feet	1
BRPJ	XCC Platinum	1
BS6A	ThinkEdge SE360 V2 Top Cover for Geotracking	1
BTJK	ThinkEdge SE360 V2 Air Baffle for Processor	1
BS66	ThinkEdge SE360 V2 IO Cover Assembly for 1GbE I/O Module	1
BS64	ThinkEdge SE360 V2 Rear Operational Panel Module	1
BS63	ThinkEdge SE360 V2 Operational Panel Module	1
BUGV	ThinkEdge SE360 V2 AC Power Module Board Air Baffle	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
7S0YCTO1WW	Lenovo Open Cloud Automation w/Support	1
SD2S	Lenovo Open Cloud Automation - nZTP with Device Management platform onboarding for 1-socket ThinkEdge server with 1 year support. Price per node	1

Large SE455 V2

7DBYCTOAWW	SE455 V3 Large : ThinkEdge SE455 V3 - 3Yr Warranty with Controlled GPU	1
BVTK	ThinkEdge SE455 V3 Chassis	1
BW2T	ThinkEdge SE455 V3 AMD EPYC 8324P 32C 180W 2.65GHz Processor	1
BQ39	ThinkSystem 32GB TruDDR5 4800MHz (1Rx4) 10x4 RDIMM-A	2
C18N	ThinkSystem 2.5" U.2 VA 1.92TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	1
BVUU	ThinkEdge SE455 V3 2.5" NVMe Backplane	1
BVUY	ThinkEdge SE455 V3 M.2 SATA/x4 NVMe Adapter with Carrier	1
BXMG	ThinkSystem M.2 PM9A3 1.92TB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	1
BS2C	ThinkSystem NVIDIA L4 24GB PCIe Gen4 Passive GPU	3
BVUR	ThinkEdge SE455 V3 Riser1	1
BMH8	ThinkEdge 1100W 230V/115V Platinum Hot-Swap Gen2 Power Supply	2
BMH2	ThinkEdge 600mm Ball Bearing Rail Kit	1
BS4E	ThinkEdge 130mm USB-C to VGA Display Cable	1
BVV6	ThinkEdge SE455 V3 Intrusion Switch	1
BVTX	ThinkEdge SE455 V3 Standard EIA Bracket	1
BVTL	ThinkEdge SE455 V3 Motherboard	1
BRPJ	XCC Platinum	1
BW38	ThinkEdge SE455 V3 Supercap Holder	1
BW37	ThinkEdge SE455 V3 M.2 Air Baffle Extension	1
BW36	ThinkEdge SE455 V3 M.2 Air Baffle	1
BVVY	ThinkEdge SE455 V3 2.5" Drive Bay Filler	3
BVUT	ThinkEdge SE455 V3 Riser2 Filler	1
BY8T	ThinkEdge SE455 V3 OCP Filler	1
BVTP	ThinkEdge SE455 V3 Fan	5
BVV1	ThinkEdge SE455 V3 Fan Board	1
BVTM	ThinkEdge SE455 V3 Root of Trust	1
BVUK	ThinkEdge SE455 V3 Power Distribution Board	1
BVVF	ThinkEdge SE455 V3 Riser Side Support	3
BW3A	ThinkEdge SE455 V3 CPU Air Baffle for 2U Heatsink	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
7S0YCTO1WW	Lenovo Open Cloud Automation w/Support	1
SD2S	Lenovo Open Cloud Automation - nZTP with Device Management platform onboarding for 1-socket ThinkEdge server with 1 year support. Price per node	1
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1

X-Large SE455 V2

7DBYCTOAWW	SE455 V3 XLarge : ThinkEdge SE455 V3 - 3Yr Warranty with Controlled GPU	1
BVTK	ThinkEdge SE455 V3 Chassis	1
BY8X	ThinkEdge SE455 V3 AMD EPYC 8324PN 32C 130W 2.05GHz Processor	1
BUVV	ThinkSystem 96GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM-A	6
C18M	ThinkSystem 2.5" U.2 VA 3.84TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	1
BVUU	ThinkEdge SE455 V3 2.5" NVMe Backplane	1
BVUY	ThinkEdge SE455 V3 M.2 SATA/x4 NVMe Adapter with Carrier	1
BXMF	ThinkSystem M.2 PM9A3 3.84TB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	1
BYFH	ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU	2
BVUR	ThinkEdge SE455 V3 Riser1	1
BVUS	ThinkEdge SE455 V3 Riser2	1
BMH9	ThinkEdge 1800W 230V Platinum Hot-Swap Gen2 Power Supply	2
BMH2	ThinkEdge 600mm Ball Bearing Rail Kit	1
BS4E	ThinkEdge 130mm USB-C to VGA Display Cable	1
BVV6	ThinkEdge SE455 V3 Intrusion Switch	1
BVTX	ThinkEdge SE455 V3 Standard EIA Bracket	1
BVTL	ThinkEdge SE455 V3 Motherboard	1
BRPJ	XCC Platinum	1
BW39	ThinkEdge SE455 V3 CPU Air Baffle for 1U Heatsink	1
BW38	ThinkEdge SE455 V3 Supercap Holder	1
BW37	ThinkEdge SE455 V3 M.2 Air Baffle Extension	1

BW36	ThinkEdge SE455 V3 M.2 Air Baffle	1
BVVY	ThinkEdge SE455 V3 2.5" Drive Bay Filler	3
BY8T	ThinkEdge SE455 V3 OCP Filler	1
BVTP	ThinkEdge SE455 V3 Fan	5
BVV1	ThinkEdge SE455 V3 Fan Board	1
BVTM	ThinkEdge SE455 V3 Root of Trust	1
BVUK	ThinkEdge SE455 V3 Power Distribution Board	1
BVVJ	ThinkEdge SE455 V3 Riser2 Rear Support	1
BVVH	ThinkEdge SE455 V3 Riser1 Rear Support	1
BVVF	ThinkEdge SE455 V3 Riser Side Support	2
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
7S0YCTO1WW	Lenovo Open Cloud Automation w/Support	1
SD2S	Lenovo Open Cloud Automation - nZTP with Device Management platform onboarding for 1-socket ThinkEdge server with 1 year support. Price per node	1
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1

Resources

Lenovo ThinkEdge SE100 Edge Server Product Guide

Lenovo ThinkEdge SE350 V2 Server Product Guide

Lenovo ThinkEdge SE360 V2 Server Product Guide

Lenovo ThinkEdge SE455 V3 Server Product Guide

https://lenovopress.lenovo.com/lp1992-lenovo-xclarity-one

Change history

Version 1.0	June 27, 2025	Initial release with 9 Edge Configurations and sizing models

Trademarks and special notices

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkAgile®
ThinkEdge®
ThinkShield®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

Intel®, Intel Core®, OpenVINO®, and Xeon® are trademarks of Intel Corporation or its subsidiaries.

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

Microsoft®, Windows Server®, and Windows® are trademarks of Microsoft Corporation in the United States, other countries, or both.

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 serve 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 is at your own risk