

LENOVO'S AMD-BASED AI PORTFOLIO UPDATE

STEVE MCDOWELL, CHIEF ANALYST
APRIL 25, 2024

CONTEXT

Lenovo [announced](#) a comprehensive set of AMD-based updates to its AI infrastructure portfolio, which include GPU-rich and thermal-efficient systems designed for compute-intensive workloads in various industries, including financial services and healthcare.


The new offerings, designed in partnership with AMD, address the growing demand for compute-intensive workloads across industries, providing the flexibility and scalability required for AI deployments.

THINKSYSTEM SR685A V3

The Lenovo ThinkSystem SR685a V3 is a high-performance GPU server designed for compute-heavy AI workloads like generative AI and LLMs. Developed in collaboration with AMD, it targets applications requiring extreme I/O bandwidth and large memory capacity.

Introducing ThinkSystem SR685a V3

A 2-socket, 8U system workload optimized for performance with AMD processors and GPUs



ThinkSystem

Ideal for both AI and HPC

- **Industries:** Financial services, healthcare, energy, climate science, and transportation
- **AI Workloads:** Model development, training and retraining, Machine Learning, Deep Learning, LLMs
- **HPC workloads:** Modeling and simulation, rendering, financial tech, scientific research

Performance

- Industry-leading 4th Gen AMD EPYC™ processors
- Maximum GPU Performance w/ AMD Instinct™ MI300x
- Interconnects for an ultra-fast transfer rate with AMD Infinity Fabric™ at 800GB/sec

Flexible

- Support for AMD MI300X, NVIDIA H100 / H200/B100 GPUs
- Drop-in upgrade to AMD next-gen processor
- Easy to manage and simplified deployment with Lenovo XClarity System Management Software
- Fits into industry-standard 19-inch racks

Advanced Thermals

- Designed with additional thermal headroom to accommodate future, high powered, GPUs
- Advanced Air-Cooled Design

The key features of the new ThinkSystem SR685a V3 include:


- **High-Performance Processors:** The server comes equipped with leading-edge 4th-generation AMD EPYC Processors.
- **GPU-Rich Architecture:** The new server is equipped with 8 AMD Instinct MI300X GPUs, offering the acceleration necessary for large AI models. It also supports NVIDIA's latest HGXTM GPUs (H100/H200/B100) for added flexibility.
- **High Memory Capacity:** The server features 1.5TB of high-bandwidth (HBM3) memory and substantial I/O bandwidth of up to 1TB/s.
- **Efficient Cooling:** The ThinkSystem SR685a V3 has substantial thermal headroom via air-cooling, allowing it to sustain maximum performance even under high-power CPUs and GPUs.
- **Flexible Deployment Options:** The server is suitable for enterprise on-premises AI and public AI cloud service providers, offering flexibility to support a range of deployment environments.
- **Advanced Interconnect:** Lenovo's ThinkSystem SR685a V3 uses AMD's Infinity Fabric interconnect, providing high-speed connections between CPUs and GPUs.

THINKSYSTEM SD535 V3

The new Lenovo ThinkSystem SD535 V3 is a high-performance, multi-node server for intensive transaction processing, cloud computing, and large-scale data analytics. Its architecture focuses on maximizing processing power while maintaining thermal efficiency.

ThinkSystem SD535 V3

High performance Multi-node Rack Server



Cross Industry, Targeted Workloads:

- Cloud Native Applications
- Virtualization
- High Performance Workloads
- Transaction Processing
- Big Data Analytics

Performance

- 1U1S multi-node half-width server powered by a single 4th Generation AMD EPYC™ series processor
- 2x the rack density of a standard 1U rack server
- 6x SSD drives adding storage and density

Flexible

- Mix ThinkSystem node types(1U/2U) and CPU Types (AMD/Intel) within the same D3 chassis
- Added agility with up to 4 nodes installed in a 2U ThinkSystem D3 Chassis

Energy Efficient

- Up to 30% less power than a standard rack server
- 1U optimized thermals with up to four nodes sharing 2 or 3 power supplies

Key features of the ThinkSystem SD535 V3 include:


- **Compact Multi-Node Design:** The server uses a 1S/1U half-width node, allowing it to fit more computing power into a smaller footprint. The 2U chassis can house up to four nodes, providing a scalable and dense computing solution for data centers.
- **Flexible CPU Options:** The ThinkSystem SD535 V3 is powered by a single 4th Gen AMD EPYC processor per node, allowing customers to choose their preferred CPU architecture. The server also supports mixing nodes based on AMD and Intel CPUs within the same 2U Lenovo ThinkSystem D3 chassis, giving customers greater flexibility in optimizing workloads.
- **Scalability and Agility:** The new ThinkSystem SD535 V3 can be configured with up to four nodes.
- **Unified Power and Cooling:** The server's unified power and cooling implementation helps reduce energy consumption by up to 30% compared to standard 1U rack servers.
- **Simplified Management:** Lenovo's XClarity system management software facilitates automation, orchestration, and deployment, streamlining large-scale operations.

THINKAGILE MX455 V3 EDGE PREMIERE SOLUTION

Lenovo's new ThinkAgile MX455 V3 Edge Premier Solution brings AI and real-time data analysis capabilities to the edge. The solution integrates with Azure Stack HCI, offering a versatile platform that delivers enhanced AI and compute performance while maintaining strong power efficiency. This is ideal for use in distributed edge environments, such as retail, manufacturing, and healthcare, where on-premises AI is crucial.

Introducing ThinkAgile MX455 V3 Edge Premier Solution

A GPU-rich and CPU-rich platform with Hybrid Edge to Cloud AI-optimized industry leading AMD processors



Ideal for
AI Workloads: Real-time data inferencing and analytics at the edge. Predictive maintenance, secure video analytics, streaming, personalized retail assistants, and healthcare analysis for faster diagnosis
Industries: Retail, Manufacturing, SmartCity and Healthcare
Retail Use Case for AI workloads, analytics, and consolidating critical data in retail Chains with multiple stores

Customer Experience

- Increased reliability and reduced downtime with continuous validation & testing by Lenovo and Microsoft allows customers to leverage technology to innovate faster, stay competitive and be more resilient
- Faster problem resolution with single point of contact for all hardware and software issues
- Data protection through self-encrypted storage and robust physical security

Flexible

- GPU and storage configurations meets specific needs of each environment
- Simplified deployment and seamless cloud based (fleet) management across thousands of nodes and geographies with LOC-A, XClarity Systems Management, Azure Arc and Cloud Deploy

Performance

- Maximum Performance with 4th Gen AMD EPYC™ processors up to 6 GPUs (AMD and NVIDIA GPUs) delivering
 - 32% less power consumption for the same workload¹
 - 2x better CPU performance²
- Edge optimized, intensive compute performance for AI inference & real-time data analysis directly at the edge
- Ruggedized design, with shock and vibration resistance, dust filtration and extended operating temperature range -5 to 55°C

Here are the key features of the ThinkAgile MX455 V3 Edge Premier Solution:

- **Edge Computing Capabilities:** The platform enhances AI inferencing and real-time data analysis at the edge, providing rapid insights and low-latency responses for edge-based applications.
- **Integration with Azure:** The new ThinkAgile system integrates seamlessly with Azure Stack HCI, enabling turnkey integration with both on-premises and Azure cloud environments. This flexibility allows customers to extend their cloud-based applications to the edge.
- **High Power Efficiency:** Powered by AMD EPYC 8004 processors, the ThinkAgile MX455 V3 Edge Premier Solution offers high performance with lower power consumption. According to Lenovo, this efficiency makes it one of the market's most power-efficient Azure Stack HCI solutions.
- **Simplified Management:** The solution offers unique Lenovo Open Cloud Automation (LOC-A) features, providing near zero-touch provisioning and automated deployment. It also supports centralized, cloud-based fleet management, reducing the time and effort needed to maintain the infrastructure.

ANALYSIS

Lenovo's new suite of AI-centric infrastructure systems and solutions, developed in collaboration with AMD, represents a significant step forward in advancing hybrid AI innovation. These offerings, from the ThinkSystem SR685a V3 to the ThinkAgile MX455 V3 Edge Premier Solution, offer the performance, flexibility, and scalability needed to support the growing demands of AI workloads.

Lenovo's announcements are a solid step in continuing the company's efforts to establish itself as a leader in the AI infrastructure space. By offering a range of GPU-rich, thermally efficient systems integrated with flexible as-a-service options and backed by robust professional services, Lenovo hopes to position itself to capitalize on the growing demand for scalable AI solutions across industries.

While the solutions are solid additions to Lenovo's portfolio, the real winner is AMD. Lenovo gives the chipmaker another OEM outlet for its recently introduced MI300x accelerators, following Dell Technologies' December 2023 announcement that its PowerEdge XE9680 will support AMD's accelerator technology.

Lenovo hasn't yet announced support for Intel's new Gaudi 3 accelerators, which competes with AMD's MI300X and Nvidia's latest generation GPUs, though Intel [indicated](#) at its recent Intel Vision event that its accelerators will be available from Lenovo, HPE, Dell and Supermicro.

The availability of non-Nvidia accelerators, such as AMD's MI300x and Intel's Gaudi 3, over the coming quarters will be a substantial test of Nvidia's grip on the market. If OEMs such as Lenovo and Dell Technologies find success in shipping these accelerators, it will demonstrate that the Nvidia's dominance isn't unbreakable and that the market is ready for choice. We'll all be watching.