

Maximum GPU Density, Familiar 2U Rack Simplicity



ThinkSystem SR650a V4

Lenovo

× intel

Up to 8 GPUs, Dual Intel® Xeon® 6, One 2U Rack Server for AI Inference

The Lenovo ThinkSystem SR650a V4 is an ideal 2-socket 2U rack server for customers who want to maximize GPU compute power while retaining a traditional 2U rack form factor. It supports up to four double-wide GPUs or eight single-wide GPUs at the front of the chassis, paired with one or two Intel® Xeon® 6 processors (up to 86 cores, 350W TDP) featuring Intel® Advanced Matrix Extensions (Intel® AMX) acceleration for deep learning training and inference.

The purpose-built SR650i V4 Inference Model pairs dual Intel® Xeon® 6530P processors with accelerator support and 512GB memory, delivering a validated, ready-to-deploy configuration for LLM and VLM inference workloads. Customers avoid re-architecting racks or cooling for high-density GPU chassis, instead scaling AI compute within familiar 2U infrastructure. Support for high-bandwidth accelerator configurations, PCIe 5.0 throughout, and optional Neptune Core liquid cooling make it well-suited to AI lifecycle, training, and inference at data center scale.

Scale AI training and inference in a familiar 2U platform powered by Intel® Xeon® 6.

Intel Products and Technologies

Accelerating AI and Analytics at the Edge



Intel® Xeon® 6 Processors



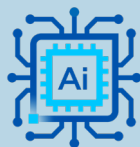
OpenVINO™ Toolkit

Intel® Advanced Matrix Extensions

Intel® QuickAssist Technology

Key Features

Empowered by Intel.



Built-In AI
Acceleration



Purpose-Built
Inference
Configuration



High-Density GPU
Compute



Enterprise-
Ready
Scalability

Lenovo ThinkSystem SR650a V4



intel ai

Intel® AI Edge Systems

Product Specifications

CPU	Up to 2x Intel® Xeon® 6 P-core processors (6500-series or 6700-series); up to 86 cores / 172 threads, up to 4.0 GHz, up to 350W TDP. Inference Model: 2x Intel® Xeon® 6530P (32C, 225W, 2.3GHz)
Graphics	Up to 4x double-wide GPUs or 8x single-wide GPUs
Memory	Up to 32 DIMMs (16 per CPU), 8 channels/CPU; up to 8TB using 256GB 3DS RDIMMs; MRDIMMs supported up to 8000 MHz. Inference Model: 512GB (8x 64GB TruDDR5 6400MHz)
Storage	Up to 8x 2.5" SAS/SATA/NVMe or 8x E3.S NVMe hot-swap bays; up to 491.52TB max (E3.S NVMe). Inference Model: 2x 3.84TB NVMe + 2x 960GB NVMe M.2 (RAID)
LAN	Up to 2x OCP 3.0 slots (PCIe 5.0, up to 400GbE). Inference Model: Broadcom 57414 10/25GbE SFP28 2-port
Operating / Storage Temp.	5 °C ~ 25 °C / -40°C ~ 70°C
Dimension (W x D x H)	445 mm x 924 mm x 87 mm
Weight	Up to 38.8 kg
OS	Microsoft Windows Server, Red Hat Enterprise Linux, SUSE Linux Enterprise Server, VMware ESXi, Ubuntu Server
I/O Options (Input/Output)	Front: 2x USB 3, optional Mini DisplayPort; Rear: 2x USB 3, 1x VGA, 1x RJ-45 1GbE (XCC), optional DB-9 serial

Learn More

[Lenovo ThinkSystem SR650a V4](#)

Intel technologies may require enabled hardware, software or service activation. // No product or component can be absolutely secure. // Your costs and results may vary. // Performance varies by use, configuration and other factors. // See our complete legal [Notices and Disclaimers](#). // Intel is committed to respecting human rights and avoiding causing or contributing to adverse impacts on human rights. See Intel's [Global Human Rights Principles](#). Intel's products and software are intended only to be used in applications that do not cause or contribute to adverse impacts on human rights.

© Intel Corporation. Intel, the Intel logo and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.