Optimized Solutions For Edge AI Inference

Enterprise AI at the edge

Enterprise AI is transforming organizations by embedding advanced AI into core business processes and workflows. This transformation is particularly evident at the network edge, where workloads such as real-time data analysis, predictive maintenance, and autonomous operations demand rapid processing, low latency, and localized decision-making.

Edge AI inference matters because it speeds up decision-making, minimizes bandwidth requirements, and reduces potential security issues by keeping and processing sensitive data on-site rather than transmitting it to centralized servers. By deploying systems with AI-optimized GPU compute and connectivity, enterprises can run workloads across distributed environments—even in locations with constrained or outdated infrastructure—by scaling the right systems or updating as needed, accelerating insights and operational efficiency without major overhauls.

Supermicro and NVIDIA accelerating enterprise AI

Supermicro and NVIDIA are bringing enterprise-ready systems to the edge, enabling industries like retail, manufacturing, telecommunications, and smart spaces to run powerful AI inference closer to their data sources. With compact, efficient architectures and latest-generation GPU acceleration, these solutions transform how enterprises process and act on real-time information.

Key technical advantages

- **Edge-optimized systems:** Supermicro platforms such as the E403 compact edge platform and 2U single-socket NVIDIA RTX PRO™ Server feature compact, thermally optimized, and low-power designs for optimized for edge AI inference.
- **CPU replacement:** Purpose built to replace traditional CPU-only servers, these systems deliver orders-of-magnitude improvements in computing power and efficiency for edge AI workloads.
- **GPU acceleration:** NVIDIA GPUs—including the NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU and NVIDIA L4 GPU—enable high-performance edge inference. The NVIDIA RTX PRO 6000 Blackwell Server Edition GPU accelerates workloads by up to 45× faster than CPU-only systems, ideal for demanding AI workloads at the edge, while the NVIDIA L4 GPU provides efficient acceleration in space-constrained environments, ensuring low-latency inference and high throughput.
- **Enhanced software stack:** NVIDIA's virtualization platform extends the power of GPUs to employees throughout the organization, with VDI and virtual machines expanding capabilities, allowing enterprises to fully leverage AI and graphics at the edge.

Business outcomes

- Real-time decision-making powered by low-latency inference closer to data source and user.
- Efficient deployment in space- and power-constrained environments.
- Process more data faster with significantly higher compute power and scalability compared to CPU-only infrastructure.
- On-site data processing negates the need to transmit across networks or to the cloud, enhancing security and data sovereignty.

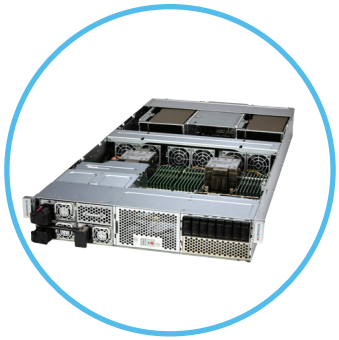
Supermicro Systems Optimized For Edge AI Inference

Key points of value

- Range of compact form factors that allow placement close to edge data source
- Thermally optimized systems architectures able to operate in non-traditional IT environments
- Single-socket architectures designed for maximum power efficiency while also reducing initial capital outlays
- GPU-accelerated systems that can replace existing CPU-only infrastructure for significant performance increases

Relevant verticals/industries

- Retail
- Manufacturing
- Healthcare
- Telecom
- Smart spaces



2U UP NVIDIA RTX PRO Server

Single-socket solution with up to 4-GPU support
Based on NVIDIA MGX™ architecture
E1.S drive support

SYS-212GB-NR

- **CPU** - Single Intel® Xeon® 6700 series processor with P-cores
- **Maximum GPU Quantity** - 4 double-width
- **Memory** - 216 DIMM slots; up to DDR5-6400
- **Storage** - 4 front hot-swap E1.S NVMe drive bays
- **Power** - 3x Redundant 2000W Titanium Level power supplies

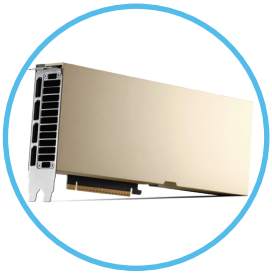
E403 Edge AI Platform

Single-socket solution for maximum efficiency
Up to 3 PCIe 5.0 x16 FHFL slots
Compact 2.5U design for front or back-of-store

SYS-212GB-NR

- **CPU** - Single Intel® Xeon® 6700/6500 series processor with P-cores or 6700 series processor with E-cores
- **Maximum GPU Quantity** - 1 double-width
- **Memory** - 8 DIMM slots; up to DDR5-6400
- **Storage** - 2 front hot-swap 2.5" NVMe drive bays + 2 internal fixed 2.5" SATA drive bays
- **Power** - Single 800W Redundant Platinum Level power supply

NVIDIA GPUs optimized for edge AI inference



NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU

- Universal GPU performance - 5th Gen Tensor Cores + 4th Gen Ray Tracing Cores
- 96GB GDDR7
- Up to 600W
- Accelerates workloads from CPU only systems (45x performance vs CPU-only)
- Multi Instance GPU (MIG) for secure GPU sharing in VDI and virtual machine environments
- Best for: Agentic & generative AI, industrial & physical AI applications, media & entertainment, AI-driven rendering & graphics

NVIDIA L4 GPU

- 4th Gen Tensor Cores + 3rd Gen Ray Tracing Cores
- 24GB GDDR6
- 72W
- Single-width form factor
- Highly efficient edge acceleration
- Best for: Generative AI, virtualization, video streaming & transcoding