ACCELERATING EDGE INNOVATION WITH SUPERMICRO EDGE AI SERVERS

Closer to Data, Ahead of Tomorrow's Intelligence

Supermicro's Edge AI Portfolio

TABLE OF CONTENTS

Executive Summary

Artificial Intelligence (AI) is transforming industries, with edge computing playing a crucial role. Rather than relying solely on data centers, computing resources are moving closer to where data is generated, which provides significant benefits for distributed use cases. From hospitals and smart cities to retail and manufacturing, organizations are adopting edge systems to process AI workloads locally.

This white paper outlines Supermicro's edge AI servers, which are engineered for enterprise AI with high performance, reliability, and seamless integration. It examines key market drivers, product architecture, and real-world applications, including digital avatars, medical imaging, synthetic data generation, intelligent stores, and Battery Energy Storage System (BESS) monitoring.

It also highlights Supermicro's strengths - compact air-cooled designs, high GPU density, modular storage, and rugged durability. A valuable resource for IT leaders and architects, this guide demonstrates why Supermicro is a leading choice for AI at the edge.

The Edge Computing Landscape

Why Edge, Why Now?

With the increased amount of data generated at the edge, local processing enables faster responses and reduces latency. Applications like AI chatbots and factory automation require real-time actions that cloud-based systems cannot adequately support. Edge computing also enhances resilience, enabling systems in remote areas to operate independently of cloud access. It lowers communication costs and improves bandwidth usage, helping IoT-heavy networks remain efficient. In summary, edge delivers speed, reliability, and reduced costs—essential for today's connected world.

Use Cases Driving Demand

Edge computing is transforming industries. It powers digital assistants, customer service chatbots, and analytics in enterprise settings. Retailers use it for smart checkouts, inventory tracking, and loss prevention. In healthcare, it supports imaging and real-time video processing. Manufacturing benefits from robotics, quality assurance, and predictive maintenance. Telecommunications rely on the edge for 5G RAN, private networks, and Multi-access Edge Computing (MEC) infrastructure.

Edge-Ready AI Performance with Supermicro Edge AI Servers

Deploying AI at the edge presents challenges like limited space, heat, harsh environments, and cyber threats. Supermicro's edge AI servers address these needs with rugged, compact designs, advanced thermal controls, and strengthened security. They provide enterprise-grade AI exactly where it is required most, even under extreme conditions.
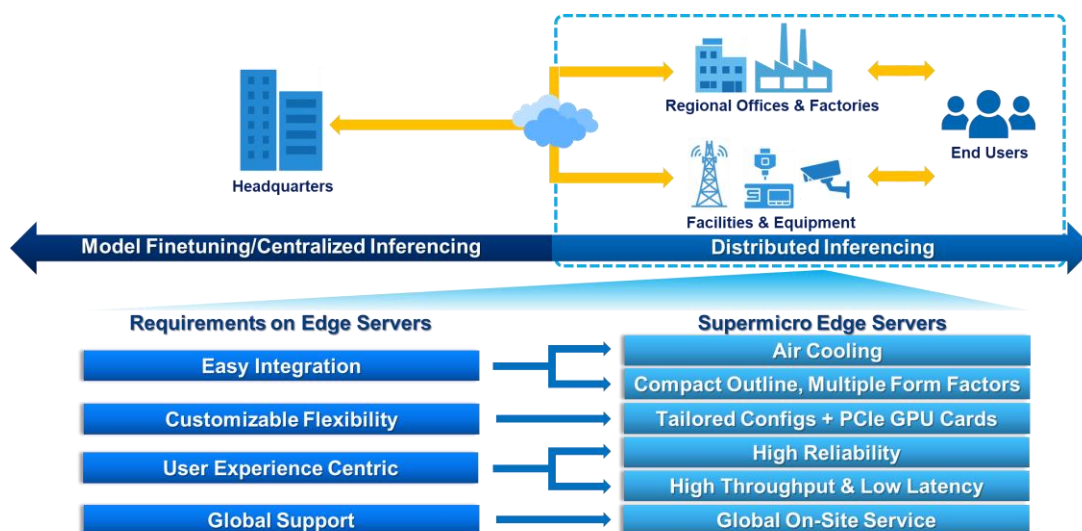

Figure 1 – Distributed Edge Deployment

Enterprise Edge Applications in Action

Use Case 1: Digital Concierge System for Hospitality and Smart Venues

Overview:

Modern customer service requires real-time, interactive engagement. Digital concierge kiosks with lifelike 3D avatars are now used in hotels, airports, malls, and smart buildings. Imagine a hotel kiosk where a realistic 3D avatar greets guests. The kiosk requires a compact, power-efficient system, such as the Supermicro SYS-E102 box-like PC. Rich I/O (COM for card readers/keypads, USB for cameras, mics, barcode scanners) captures user input and sends voice and image data to backend servers. In the server room, the short-depth AS-2115HE GPU server (574mm) processes data using AI to detect gender, emotion, and transcribe speech. This text is sent to a second AS-2115HE, which hosts a vector database and LLM-based RAG system. It retrieves relevant info and generates a response. The first server converts this text to speech, creates avatar animations with lip sync, and displays them in real-time, enabling natural and expressive conversations.

Why Supermicro:

Supermicro offers an end-to-end solution: compact PCs for kiosk control and scalable GPU servers for AI and media processing.

- SYS-E102: Small, dependable, and integration-ready, ideal for 24/7 kiosk use.

- AS- 2115HE: Short-depth 2U server with hot-swappable fans, redundant power, six NVMe bays, and support for up to 3 NVIDIA RTX 6000 Ada or L40S GPUs—ideal for scalable AI workloads.

Value to Customers:

The system reduces staffing needs while delivering rich, multilingual service. Its modular design enables easy expansion—from single kiosks to full-scale, multi-avatar deployments.
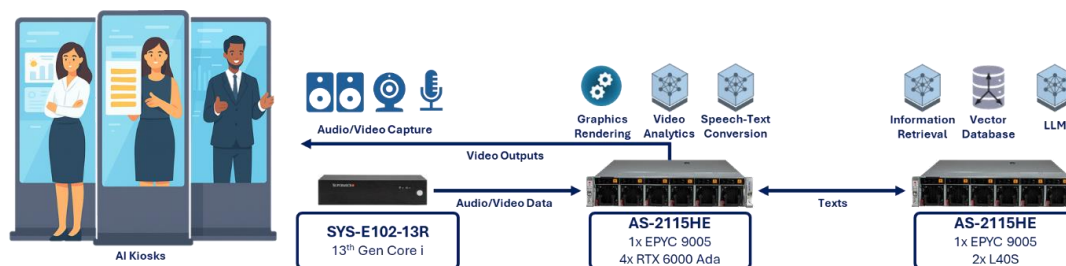

Figure 2 - AI Concierge Workflow

Use Case 2: AI-Enhanced Laparoscopic Surgery System

Overview:

Modern operating rooms depend on precision, real-time imaging, and AI guidance, especially in minimally invasive laparoscopic surgery. AI overlays can track instruments, highlight tissues, and assist surgeons during procedures. In the OR, the Supermicro SYS-212B-FLN2T, a short-depth (450mm), front-I/O server, processes laparoscopic video. A frame grabber card captures surgical video, while the NVIDIA RTX 4000 SFF Ada GPU runs AI models for tool detection, tip localization, and image segmentation—all synchronized with the surgeon's actions. When abnormalities are detected, the surgeon can capture and encrypt images using onboard TPM-generated credentials and temporarily store them in the U.2 SSDs. After surgery, these images are uploaded to the hospital's PACS system via onboard 10GbE ports.

Why Supermicro:

The SYS-212B-FLN2T offers a compact 450mm depth and front I/O for easy integration into cabinets. Its hot-swappable fans, drive bays, and power supplies simplify maintenance. Its versatile PCIe slot layout supports one full-height and six low-profile cards for video capture and AI acceleration. Supermicro also provides tailored support, including fan curve optimization to reduce noise, which is essential for hospital environments.

Value to Customers:

This solution enhances surgical accuracy through real-time AI guidance, ensuring secure and traceable image storage. It provides a reliable, turnkey platform for medical OEMs focused on safety, efficiency, and compliance.
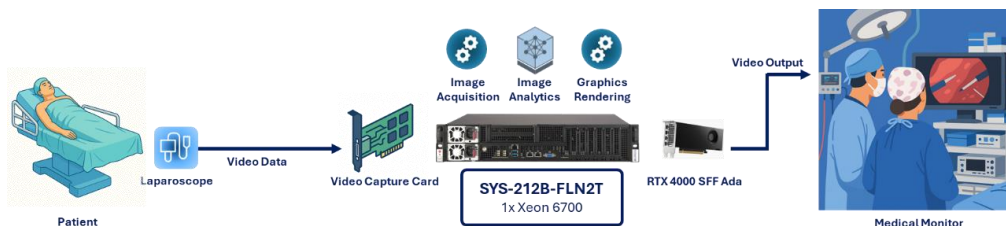

Figure 3 - AI-Assisted Surgery

Use Case 3: Synthetic Data Generation for Financial Services

Overview:

Financial institutions rely on large datasets to develop and test algorithms for trading, fraud detection, and risk modeling. However, access to real data is often limited due to privacy and regulatory concerns, necessitating the generation of synthetic data in secure, low-latency environments. Using the Supermicro SYS-322GA-NR, a financial firm utilizes an AI interpreter to analyze user queries, retrieves relevant data from a vector database hosted on the AS-2126HS-TN, and activates the appropriate generative AI model (VAE, Transformer, or DDM-based) to produce synthetic datasets customized to the user's needs.

September, 2025

Why Supermicro:

Supermicro delivers more than just high-performance hardware. In addition to the SYS-322GA-NR —a 3U server supporting dual Intel® Xeon® 6900-series processors and up to eight 350W GPUs (e.g. NVIDIA L40S) for demanding AI workloads— the AS -2126HS-TN complements this setup as an AI-enabled storage server with 24 U.2 NVMe SSDs and a built-in NVIDIA L40S GPU to accelerate vector-based data retrieval and embedding model performance. Supermicro also offers expert technical consultancy. This helps customers select the optimal server configuration and GPU quantity for a cost-efficient, high-performance AI infrastructure.

Value to Customers:

This solution enables secure, compliant synthetic data generation, accelerating the development of financial algorithms while minimizing reliance on real or third-party datasets. It offers a scalable, in-house platform tailored for sensitive financial environments.
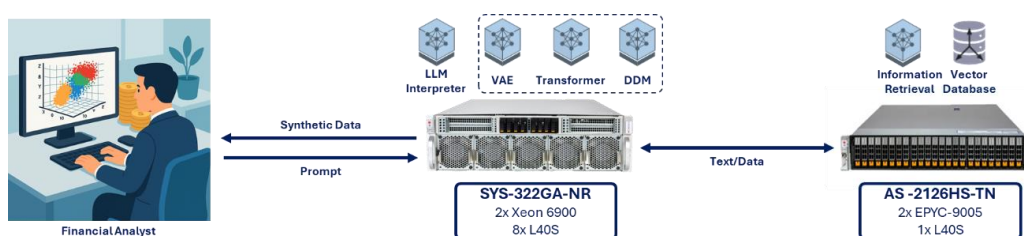
Figure 4 - Synthetic Data Platform

Use Case 4: Unified Retail Infrastructure for Convenience Stores

Overview:

Efficiency is critical in convenience retail, where space is tight and operations run 24/7. Retailers must modernize without expanding their physical footprint. A unified edge platform can consolidate applications such as POS, digital signage, inventory tracking, and AI analytics into one system, instead of separate proprietary devices. The compact Supermicro SYS-E403-14B-FRN2T fits under the counter and runs virtualized retail applications. Cameras stream customer behavior into the server, where an NVIDIA RTX 6000 Ada GPU analyzes movement and links it to sales trends. Shelf sensors and ERP data enable real-time inventory updates and automated restocking. Previously, each system ran on a separate workstation, making installation and maintenance complex. Supermicro's unified platform simplifies everything.

Why Supermicro:

The SYS-E403-14B-FRN2T is a uniquely compact edge server with enterprise-class performance. Its efficient cooling, flexible mounting, and support for operating temperatures of up to 45°C make it suitable for various store environments, including non-air-conditioned areas such as backroom warehouses.

Value to Customers:

This versatile yet powerful solution reduces hardware footprint, lowers costs, and simplifies deployment and management. Retailers achieve higher uptime, better service personalization, and improved operational efficiency, resulting in increased profitability.
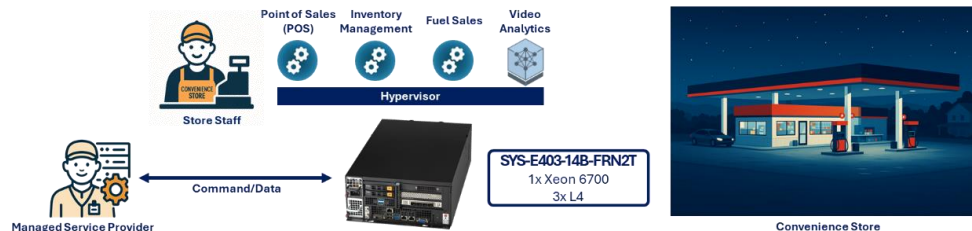

Figure 5 - Retail Workload Consolidation

Use Case 5: AI-Powered Loss Prevention in Retail Environments

Overview:

Shrinkage due to theft, mis-scans, and checkout fraud remains a significant issue in retail. Traditional methods—such as manual monitoring and audits—are reactive and resource-intensive. Supermicro's AI-powered edge platform facilitates proactive, real-time loss prevention. In a busy self-checkout area, overhead IP cameras monitor every transaction. Feeds are analyzed by a Supermicro SYS-212B-FN2T server in the back office, powered by an Intel Xeon 6700/6500-series processor and NVIDIA RTX 6000 Ada GPUs. AI models identify items selected by customers and generate "forecasted barcodes" based on visual recognition and behavior analysis. When a customer scans an item, the system compares the actual barcode to the predicted one. Mismatches trigger real-time alerts to store staff. All events, including false alarms, are stored with video and prediction data for future review and AI model improvement. This enables immediate response and ongoing system learning.

Why Supermicro:

The SYS-212B-FN2T is a compact 2U, 450 mm-deep server that supports two 350W GPUs and operates at up to 40°C—ideal for space- and cooling-limited store environments. It features dual 10GbE ports for high-bandwidth camera input and four hot-swappable U.2 NVMe SSDs for video storage. Supermicro also provides configure-to-order services, helping retailers select the right GPU type and quantity for stores of different sizes.

Value to Customers:

This integrated solution reduces shrinkage, cuts operational overhead, and increases staff efficiency by combining AI analytics, alerts, and storage on a single platform, delivering clear, scalable ROI.
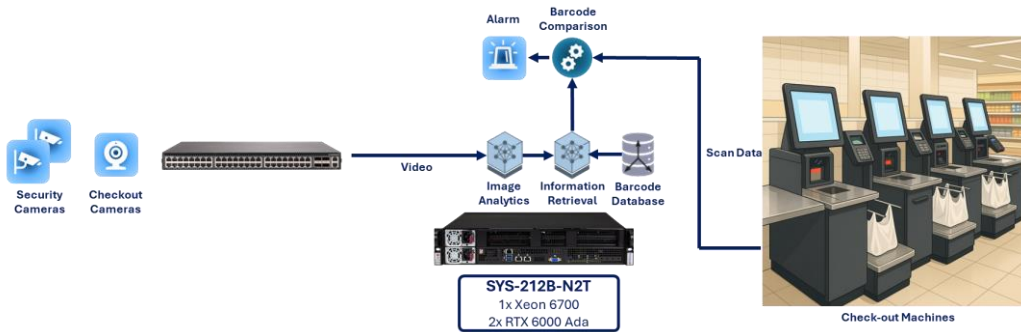
Figure 6 - AI Loss Prevention

Use Case 6: Battery Energy Storage System (BESS) Monitoring in the Energy Sector

Overview:

As renewable energy adoption increases, Battery Energy Storage Systems (BESS) play a crucial role in ensuring grid stability and managing peak loads. Often situated in remote locations, these systems require real-time monitoring and control. Supermicro's edge computing solutions deliver datacenter-level performance to energy sites. For example, at a single facility, a compact Supermicro SYS-E403-14B-FRN2T server collects and processes battery data, including voltage, temperature, and charge status, from the Battery Management System (BMS). This data is transmitted to a centralized SYS-212B-FN4TP server for analytics, visualization, archiving, and access via a web-based dashboard. Commands from the operator interface are routed back to the SYS-E403 to manage battery operations locally.

Why Supermicro:

The SYS-E403-14B-FRN2T is a rugged, compact edge server designed for harsh environments, supporting temperatures up to 45°C and allowing for flexible installation near battery enclosures. The SYS-212B-FN4TP, with its 300mm depth and front I/O, is ideal for tight control rooms with limited cooling, operating reliably up to 40°C. It also supports up to three NVIDIA L4 GPUs to accelerate video processing, graphics rendering, and natural language tasks. Both servers feature IPMI out-of-band management for remote configuration, monitoring, and recovery, reducing on-site maintenance and operational costs.

Value to Customers:

This edge solution cuts risk, boosts reliability, and lowers costs. Customers gain real-time control, secure operations, and improved energy ROI in a scalable, rugged package.

September, 2025
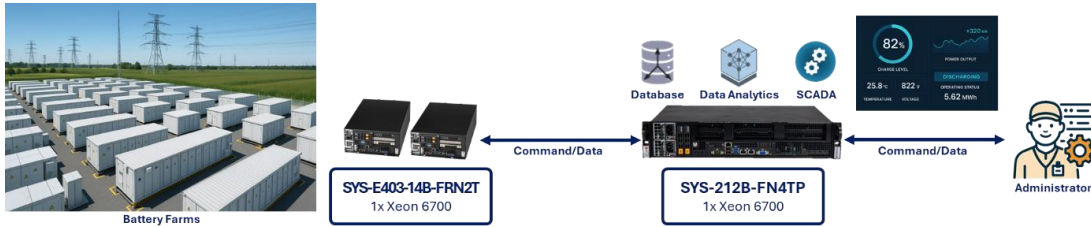
Figure 7 - BESS Edge Monitoring

Summary and Portfolio Overview

Edge AI is accelerating digital transformation by bringing intelligence closer to where data is generated. Supermicro's edge AI servers offer industry-leading performance, rugged reliability, and GPU-rich configurations across various environments. With a broad product portfolio and customized technical support, Supermicro enables enterprises to deploy scalable, real-time AI solutions at the edge. As the complexity of AI increases, Supermicro stands at the forefront of edge innovation. The table below outlines recommended Supermicro systems for various deployment environments, helping decision-makers select the ideal edge platform for their specific use case.

Deployment Environment	Edge Data Center
Typical Edge AI Scenario	High-performance edge AI servers for LLM inference, synthetic data generation, and big data analytics. Ideal for dense compute with powerful CPU and GPU support.

Recommended Models			
SYS-322GA-NR - Dual Intel® Xeon® 6900-series processors - Up to 6TB DDR5 8800 - 10 PCIe x16 or 20 PCIe x8 slots - Support up to 4 600W or 8 350W GPUs	**SYS-212GB-NR** - Single Intel® Xeon® 6700/6500-series processor - Up to 2TB DDR5 5200 - 7 PCIe x16 slots - Support up to 2 600W or 4 350W GPUs	**AS -2115HE-(F)TNR** - Single AMD EPYC™ 9004/9005-series processor - Up to 9TB DDR5 4400 - 4 PCIe x16 or 8 PCIe x8 slots, 2 AIOMs - Support up to 4 350W or 2 600W GPUs	**SYS-222HE-(F)TN** - Dual Intel® Xeon® 6700/6500-series processors - Up to 8TB DDR5 5200 - 4 PCIe x16 or 8 PCIe x8 slots, 2 AIOMs - Support up to 3 350W or 2 600W GPUs

September, 2025

Deployment Environment	Control Room / Edge Node
Typical Edge AI Scenario	Short-depth 2U edge AI servers for on-site AI workloads such as video analytics, visualization, and industrial data processing.

Recommended Models

SYS-212B-(F)N2T
- Single Intel® Xeon® 6700/6500-series processor
- Up to 1TB DDR5 6400
- Max 4 PCIe x16, 1 PCIe x8 slots
- Support up to 2 350W GPUs

SYS-212B-(F)LN2T
- Single Intel® Xeon® 6700/6500-series processor
- Up to 1TB DDR5 6400
- Max 4 PCIe x16, 3 PCIe x8 slots
- Support up to 1 600W or 7 L4 GPUs

AS -2116S-(F)NTRT
- Single AMD EPYC™ 9004/9005-series processor
- Up to 4.5TB DDR5 6400
- Max 4 PCIe x16, 2 PCIe x8 slots
- Support up to 1 600W or 6 L4 Tensor Core GPUs

SYS-212B-FN4TP
- Single Intel® Xeon® 6700/6500-series processor
- Up to 2TB DDR5 6400
- Max 3 PCIe x16, 2 PCIe x8 slots
- Support up to 3 L4 Tensor Core GPUs

Deployment Environment	Field Cabinet or Enclosure
Typical Edge AI Scenario	Compact boxes and 1U front I/O edge AI servers for real-time signal processing, protocol conversion, and edge control near sensors and PLCs.

Recommended Models

SYS-E403-14B-FRN2T
- Single Intel® Xeon® 6700/6500-series processor
- Up to 2TB DDR5 6400
- 3 PCIe x16 slots
- Support up to 1 350W or 3 L4 GPUs

SYS-112B-FWT (FDWR)
- Single Intel® Xeon® 6700/6500-series processor
- Up to 1TB DDR5 6400
- 3 PCIe x16 slots
- Support 1 L4 Tensor Core GPU cards

AS-1115S-F(D)WTRT
- Single AMD EPYC™ 8004-series processor
- Up to 768GB DDR5 4800
- 3 PCIe x16 slots
- Support 1 L4 Tensor Core GPU cards

ARS-E103-JONX
- NVIDIA Jetson Orin NX Series
- Up to 16GB LPDDR5 3200
- 3 M.2 sockets
- 5 LAN ports, 2 COM ports, 1 CAN port

SYS-E300-14AR
- Intel® Core™ Ultra Series 2
- Up to 96GB DDR4 3200
- 1 PCIe x16 slot, 2 M.2 sockets
- 5 LAN ports

SYS-E102-13R
- 13th Gen Intel® Core™ i
- Up to 64GB DDR5 4800
- 3 M.2 sockets
- 2 LAN ports

SYS-E302-13AD
- 12th Gen Intel® Core™ i
- Up to 32GB DDR5 4800
- 3 M.2 sockets
- 2 LAN ports

SYS-E100-13AD
- 12th Gen Intel® Core™ i
- Up to 64GB DDR5 4800
- 3 M.2 sockets
- 2 LAN ports, 4 COM ports
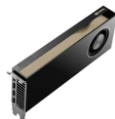
NVIDIA GPU Cards for Edge Computing
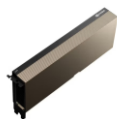
NVIDIA RTX PRO™ 6000 Blackwell Server Edition

- Blackwell
- 96 GB GDDR7 with ECC
- 24,064 CUDA Cores
- 752 Tensor Cores
- 188 RT Cores
- Power consumption 400W - 600W

NVIDIA RTX PRO™ 6000 Blackwell Max-Q Workstation Edition

- Blackwell
- 96 GB GDDR7 with ECC
- 24,064 CUDA Cores
- 752 Tensor Cores
- 188 RT Cores
- Power consumption 300W

NVIDIA RTX PRO™ 5000 Blackwell Server Edition

- Blackwell
- 48 GB GDDR7 with ECC
- 14,080 CUDA Cores
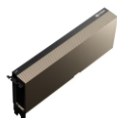- 440 Tensor Cores
- 110 RT Cores
- Power consumption 300W

NVIDIA RTX PRO™ 4500 Blackwell Server Edition

- Blackwell
- 32 GB GDDR7 with ECC
- 10,496 CUDA Cores
- 328 Tensor Cores
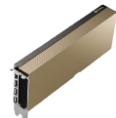- 82 RT Cores
- Power consumption 200W

NVIDIA RTX PRO™ 4000 Blackwell Server Edition

- Blackwell
- 24 GB GDDR7 with ECC
- 8,960 CUDA Cores
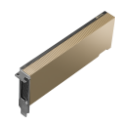- 280 Tensor Cores
- 70 RT Cores
- Power consumption 140W

NVIDIA H200 NVL
- Hopper
- 141GB HBM3e
- 16,896 CUDA Cores
- 528 Tensor Cores
- 2/4-Way NVLink (900 GBps)
- Power consumption up to 600W

NVIDIA L40S
- Ada-Lovelace
- 48 GB GDDR6 with ECC
- 18,176 CUDA Cores
- 568 Tensor Cores
- 142 RT Cores
- Power consumption 350W

NVIDIA L4 Tensor Core
- Ada-Lovelace
- 24 GB GDDR6 with ECC
- 7,680 CUDA Cores
- 240 Tensor Cores
- 60 RT Cores
- Power consumption 72W

For More Information

For more information about Supermicro's Edge AI portfolio, visit www.supermicro.com/edge-ai

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Learn more at www.supermicro.com