# **Supermicro Edge Al** Edge Video Transcoding, Edge Inference, Edge Training

Across industries, businesses whose employees and customers engage at edge locations - in cities, factories, retail stores, bospitals, and many more - are increasingly investing in deploying AL at the edge. By processing data and utilizing AL and

hospitals, and many more - are increasingly investing in deploying AI at the edge. By processing data and utilizing AI and ML algorithms at the edge, businesses overcome bandwidth and latency limitations, enabling real-time analytics for timely decision making, predictive care and personalized services, and streamlined business operations.

Purpose-built, environment-optimized Supermicro Edge AI servers with various compact form factors deliver the performance needed for low-latency, open architecture with pre-integrated components, diverse hardware and software stack compatibility, and privacy and security feature set required for complex edge deployments out of the box.

## **Systems**

## Short-Depth 5G/Edge & Hyper E

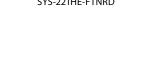
Compute and AI Performance at the Edge

#### Extra Large Workload: 2U Hyper-E

- 3 NVIDIA H100 PCIe
- 6 NVMe drives
- 32 DIMMs DDR5-4800



- Medium Workload: Short-Depth Multi-GPU Edge Server
- 1U Compact Edge/5G Server
- 2 NVIDIA L4 2 Internal Drive Bays
- 8 DIMMs DDR5-4800



SYS-111E-FWTR

## Fanless and Wallmount Edge

Compact Systems for the Intelligent Edge

#### Large Workload: Compact System

- Powerful expandable server for the Edge
- 1 NVIDIA L40S or 2 L4
- 8 DIMM slots DDR5-4800
- 4 NVMe Drives



SYS-E403-13E

#### Small Workload: Embedded System

- Ultra-compact Fanless Edge Server CPU (or ASIC) based Inference
  Up to 64GB DDR5
- M.2 M/B/E-Key with Nano SIM Card Slot



SYS-E100-13AD

## **Recommended NVIDIA GPUs**



L4

- HHHL SW
- PCle 4.0 x16
- 72W24GB GDDR6



L40S

- FHFL DW
- PCle 4.0 x16350W
- 48GB GDDR6



L40

- FHFL DW
- PCIe 4.0 x16
- 300W 48GB GDDR6

## **Accelerate Edge Al Workloads**

Edge Video Transcoding, Edge Inference, Edge Training

#### **Opportunities and Challenges:**

- · Space and weight limitation, power constraints
- Balancing data throughput for video and audio requirements with cost of storage and bandwidth constraints
- · Latency impacting response time and service quality
- Data privacy and security, regulatory compliance
- Resiliency in face of network outages
- · Long product lifecycle requirements

## **Key Technologies:**

- CPU or GPU-based Edge AI Inferencing, GPU-based Edge AI training, and video transcoding/encoding/decoding
- NVIDIA L4, L40S, L40, A30, A40, T4, A2 GPUs
- Short-depth chassis design for edge locations with AC or DC power supply options
- Front I/O with broad range of expansion and I/O port for flexibility and serviceability
- Ruggedized systems designed to be placed outside of the data center

#### **Solution Stack:**

- NVIDIA<sup>®</sup> TensorRT<sup>™</sup> and Triton Inference Server
- NVIDIA DeepStream, Clara, Merlin, Metropolis, Morpheus, Omniverse, and Riva
- NVIDIA Fleet Command
- Intel<sup>®</sup> OpenVINO

## Use Cases:

- Video processing: decode, encode, and transcode
- Edge inference: vision, speech, anomaly detection, etc.
- · Markets: security and surveillance, retail, manufacturing, healthcare, and medical devices

## GPU Acceleration for Complete Range of Workloads



Go to www.supermicro.com/ai or scan the QR code to download the AI Workload Solution Brochure:



© 2023 Copyright Super Micro Computer, Inc. Specifications subject to change without notice. All other brands and names are the property of their respective owners. All logos, brand names, campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part, without express written permission by Supermicro Corporate Marketing.