

SUPERMICRO AND SPECTRO CLOUD PALETTE: THE POWERFUL PLATFORM FOR EDGE AI MADE EASY



Supermicro E403 Edge System

Table of Contents		Executive Summary
Executive Summary	1	<p>Edge AI is quickly becoming a critical priority across various industries, including healthcare, retail, manufacturing, and transportation. It connects cloud computing with on-device intelligence, allowing for real-time processing, improved security, and more efficient utilization of network resources. Supermicro and Spectro Cloud deliver a comprehensive AI edge solution through pre-validated, optimized hardware and Spectro Cloud’s Palette EdgeAI platform, designed to simplify and scale the deployment and management of AI workloads across diverse edge environments.</p> <p>Common Use Cases</p> <p>Everyday use cases include AI-driven diagnostics on portable medical devices, smart checkout systems, real-time customer insights in retail, predictive maintenance in manufacturing, and intelligent traffic and surveillance systems for smart cities.</p>
Common Use Cases	1	
SuperMicro + Spectro Cloud Edge-in-a-Box Solution . . .	2	
Kubernetes	2	
Supermicro + Spectro Cloud: A Complete Solution for Edge AI	3	
How it Works: Setting Up Your Edge in a Box	5	
Benefits	6	
Why Spectro Cloud	6	
Conclusion	7	
For More Information	7	

By putting AI workloads at the edge, organizations benefit from:

- **Low Latency & Real-Time Processing** – AI models running on edge devices can make decisions instantly without relying on cloud connectivity, which is crucial for applications like autonomous vehicles, industrial automation, and Healthcare.
- **Reduced Bandwidth Costs** – Processing data locally reduces the need to transmit large amounts of information to the cloud, saving bandwidth and operational costs.
- **Enhanced Privacy & Security** – Sensitive data can be processed on-device, minimizing exposure to external threats and compliance risks related to data sharing.
- **Offline Functionality** – Edge AI enables AI applications to function without an internet connection, ensuring reliability in remote areas or mission-critical systems.
- **Energy Efficiency**—AI models optimized for edge devices can run efficiently with lower power consumption, which benefits battery-powered IoT devices and wearables.

Features of the Supermicro + Spectro Cloud Edge-in-a-Box Solution

- **Purpose-Built Edge Systems Utilizing Nvidia GPUs** – Supermicro's servers are compact and edge-ready, providing unmatched performance for your AI-powered applications.
- **Flexible Configurations for AI at the Edge** – Supermicro accommodates a variety of GPUs, which can support everything from predictive analytics to powerful generative AI for media creation and editing.
- **Integrated Hardware and Kubernetes Management for Edge AI** – The integration of Supermicro's edge systems with Spectro Cloud's advanced Kubernetes management platform offers a complete, secure, and scalable solution for edge AI.
- **Seamless Day-Two operations with reusable cluster profiles** – Automate patching, upgrades, certificate rotation, and consistency using Cluster Profiles to ensure uniform, repeatable operations across thousands of edge locations.
- **Run Legacy VMs and Kubernetes Containers Concurrently** – Support legacy and modern container-based edge applications through integrated Virtual Machine Orchestration from Spectro Cloud.

Kubernetes

Unsurprisingly, organizations aiming to implement AI at the edge are turning to Kubernetes. Kubernetes provides several key advantages for managing these complex projects, including portability for consistent deployment across environments, automation for scaling and self-healing, and ecosystem integration for seamless toolchain compatibility.

However, running AI and Kubernetes stacks at the edge at the enterprise scale means IT teams need to overcome several challenges:


- Limited Computational Resources for Model Optimization—Edge devices often have lower processing power, memory, and storage than cloud-based systems. This constraint requires AI models to be optimized via quantizing and pruning, which allows them to fit edge hardware but possibly affects accuracy.
- Scalability & Maintenance – Deploying and updating AI models across many distributed edge devices can be complex and resource-intensive.
- Security Risks—Edge devices are more vulnerable to physical and cyber attacks due to their decentralized nature. Protecting them requires robust security protocols.
- Interoperability & Standardization – Different hardware architectures and AI frameworks create integration challenges, making it harder to build universal solutions.
- Limited Stack Flexibility - Many edge platforms lock teams into hardware, OS, and software choices, limiting flexibility across edge environments, workloads, and critical AI applications.
- Ongoing K8s Infrastructure Maintenance—Remote edge clusters need continual patching, upgrades, cert rotation, and drift remediation—difficult to manage at scale.
- Running VMs and containers in parallel—Many edge environments still rely on legacy VM-based apps. Still, they lack the infrastructure to run VMs and containers side by side, slowing the transition to cloud-native architectures.

Supermicro + Spectro Cloud: A Complete Solution for Edge AI

Supermicro and Spectro Cloud have partnered to create an “edge in a box” solution that streamlines the edge lifecycle.

Supermicro, known for its innovative infrastructure solutions, offers a diverse range of servers optimized for AI edge computing. Supermicro’s modular architecture, designed to meet the demands of edge deployments, ensures reduced latency and minimal network traffic.

Paired with Spectro Cloud’s Kubernetes management platform, the solution simplifies, scales, and accelerates edge AI adoption. It supports both containers and VMs, enforces security compliance, standardizes cluster deployment, and enables full Day 0 to Day 2 management—even in disconnected or low-connectivity environments.



SYS-E300-13AD

Dimensions (mm) H × W × D

43.0 × 264.8 × 225.8

Key Features for Predictive

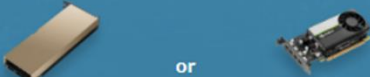
AI computer vision for up to 8 streams¹

Up to ~2,000 automatic speech recognition (ASR) samples per second²


Key Features for Generative

LLM up to 8 billion parameters

Image and video generation with Stable Diffusion at ~1 image every 4-5 seconds



Single A2 or T1000



SYS-E403-14B-FRN2T

Dimensions (mm) H × W × D

117.3 × 266.7 × 406.4

Key Features for Predictive


AI computer vision for up to 32 streams¹

Up to ~22,000 automatic speech recognition (ASR) samples per second²

Key Features for Generative

LLM up to 48 billion parameters

Image and video generation with Stable Diffusion at ~1-2 images per second



H100 or L40S or RTX 6000 Ada

1. Based on using an image classification model similar to EfficientNet-B4, dependent on video stream compression and other workloads on the system

2. Based on using an ASR model like QuartzNet

Spectro Cloud's Palette Edge management platform enables you to:

- Design and build custom edge software images with your choice of OS, Kubernetes distribution, and other software integrations — including simplified deployment of the NVIDIA GPU Operator.
- Deploy and onboard edge Kubernetes clusters at scale, whether in disconnected / airgapped or connected environments. Validated to 10,000+ edge clusters under management with no performance slowdown.
- Automate cluster lifecycle management, including certificate rotation, patching, and security scans.
- Secure your sensitive AI Kubernetes stacks. Palette is highly secure and compliant for use in regulated environments: trusted boot, immutable images, full disk encryption, and more.

- Maintain critical application availability. Palette supports zero-downtime rolling upgrades and is tolerant of dynamic networks.
- Unify all your Kubernetes compute environments with a single pane of glass for complete lifecycle management of edge Kubernetes clusters, as well as your other Kubernetes clusters in clouds and data centers.

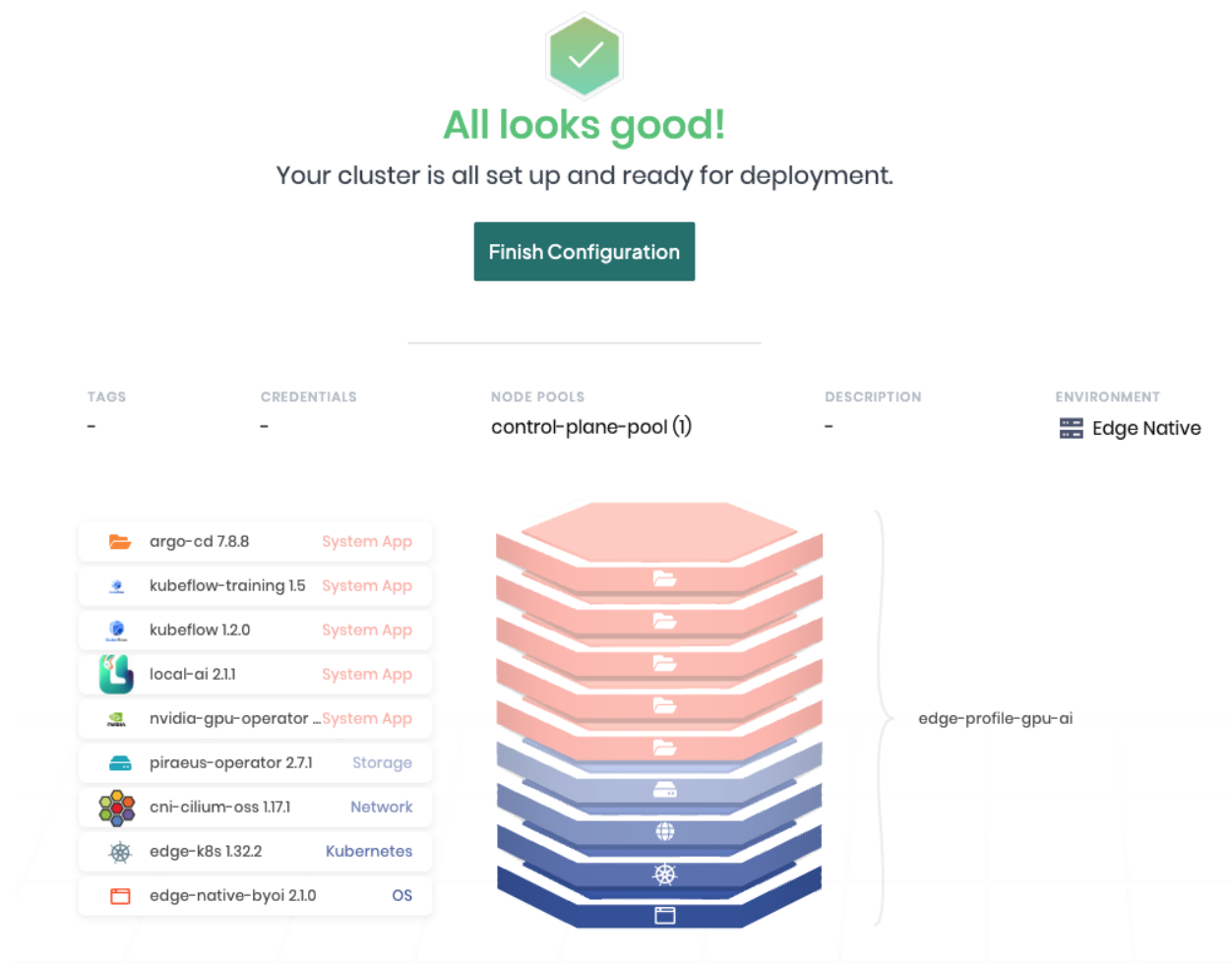


Figure 1 - Spectro Cluster Deployment Configuration

How it Works: Setting Up Your Edge in a Box

Deploying Edge AI solutions is simple with Supermicro and Spectro Cloud, which is relatively easy. Below are the workflow steps for getting up and running:

- Build Your Custom Image with Requirements – Use Spectro Cloud’s Kubernetes packs to build your Edge AI stack. Choose your OS, Kubernetes Distribution, and the AI tools you need, and create your image that will be loaded onto the hardware.
- Images are preloaded by Supermicro at the factory – Custom images are loaded directly onto your Supermicro edge devices, ensuring consistency and saving time when installation starts.

- Devices are shipped to the facility and powered on – Pre-integrated systems arrive ready to deploy. Simply unbox, rack (if needed), and power them on—no manual software setup required, thanks to the pre-install images.
- Devices phone home to the Palette platform and autoregister for management – When the device powers on, it will connect to Spectro Cloud’s Palette management system to register itself for management through the Palette UX.
- Full lifecycle management purpose-built for the edge – Easily manage the entire edge footprint—alongside cloud and on-prem clusters—from a single platform. Maintain visibility, enforce consistency, and streamline operations, even in low/no connectivity environments.

Benefits

Together, Supermicro and Spectro Cloud provide a seamlessly integrated solution that simplifies the deployment, scaling, and operation of edge AI applications. This “edge-in-a-box” approach reduces complexity and enhances team productivity across industries, helping you get up and running more quickly.

- Access to top-tier edge-ready hardware configurable with NVIDIA GPUs - capable of handling your edge AI needs.
- Faster deployments of new edge hardware to the site without the need for skilled resources: zero-touch onboarding from power-on, saving you up to 90% in field engineering costs.
- Reduced business risk through complete software consistency across sites.
- Choice and flexibility to design and deploy the right software images for each business requirement.
- Simplified procurement through Supermicro and your preferred global systems integrator.

Why Spectro Cloud

Spectro Cloud is the world’s leading provider of edge management and orchestration (EMO) solutions for Kubernetes. It is:

- Ranked as the leader for two years running in GigaOm’s Radar for Edge Kubernetes.
- An STL top 100 edge computing company for three years running.
- A Gartner Cool Vendor in Edge Computing.
- Trusted to provide edge Kubernetes solutions by the US military and federal government, as well as Fortune 1000 leaders across retail, healthcare, manufacturing, and other sectors.
- A leader in security, with the world’s only Kubernetes platform, which has FIPS (Federal Information Processing Standards) validated from platform to cluster, plus ISO 27001 and SOC 2.
- Committed to customer choice and open source. We sponsor the CNCF Sandbox project Kairos, the factory for immutable edge OS images.

Conclusion

Supermicro and Spectro Cloud bridge the gap between AI innovation and practical, reliable edge deployment, making it a strategic tool for modern AI-driven enterprises, enabling organizations to deploy and scale AI at the edge with speed, security, and simplicity.

Further Information

Supermicro Edge Website: <https://www.supermicro.com/en/solutions/edge-ai>

Spectro Cloud Website: www.spectrocloud.com

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

SPECTRO CLOUD

Spectro Cloud delivers simplicity and control to organizations running Kubernetes at any scale. With its Palette platform, Spectro Cloud empowers businesses to deploy, manage, and scale Kubernetes clusters effortlessly — from edge to data center to cloud — while maintaining the freedom to build their perfect stack. Trusted by leading organizations worldwide, Spectro Cloud transforms Kubernetes complexity into elegant, scalable solutions, enabling customers to master their cloud-native journey with confidence. Spectro Cloud is a Gartner Cool Vendor, CRN Tech Innovator, and a ‘leader’ and ‘outperformer’ in GigaOm’s 2025 Radars for Kubernetes for Edge Computing, and Managed Kubernetes.

For more information, visit <https://www.spectrocloud.com>