SOLUTION BRIEF

SUPERMICRO TURNKEY AI WORKLOAD SOLUTION

Pre-configured, ready-to-deploy platform for AI & Generative AI

Supermicro Superworksation SYS-531A-I

Table of Contents

Executive Summary

Adoption of AI and Generative AI is now critical across all industries, yet effectively building and managing secure AI environments remains a challenge for IT and DevOps teams. To successfully adopt AI and accelerate innovation, companies must address:

- Infrastructure Readiness: Ensure that existing IT environments can support AI workloads, including storage, networking, and compute resources.
- Data & Workload Management: Effectively managing the associated data and image workloads is crucial.
- Operational Costs: Factor in hardware, software, power consumption, and maintenance expenses when deploying AI solutions.
- Orchestration layer: An orchestration layer is needed to enable seamless communication and coordination between the infrastructure and the workload, yielding desired results.
- Performance Optimization: Choose hardware and software configurations that maximize AI training and inference efficiency.
- Skill & Expertise: Ensure that IT teams and developers have the necessary expertise to manage and optimize AI workloads effectively.

Supermicro and RAVEL have collaborated to bring to market an industry-leading turnkey AI workload solution that delivers seamless, ready-to-deploy AI infrastructure designed to eliminate the complexity of AI adoption. By integrating the latest NVIDIA GPUs and Intel® Xeon® processors with RAVEL Orchestrate software, this pre-configured system enables businesses

and developers to implement AI and GenAI workloads efficiently with minimal setup. Offering optimized performance, enterprise-grade reliability, and scalable architecture, this solution empowers organizations to accelerate innovation while reducing time-to-value. Whether for machine learning, generative AI, or edge AI applications, Supermicro and RAVEL's AI solution provides the necessary hardware, software, and tools to drive success in AI-powered initiatives.

This solution addresses the needs of several key market segments:

1. Enterprises Scaling AI Operations – Companies looking to integrate AI for automation, predictive analytics, and decision-making without the burden of infrastructure complexity.

2. AI & Machine Learning Developers – Startups and research teams needing high-performance computing for training and deploying AI models efficiently.

3. Data Centers & Cloud Providers – Organizations offering AI-as-a-Service (AIaaS) requiring scalable, high-performance AI infrastructure.

4. Healthcare & Life Sciences – Businesses utilizing AI for medical imaging, drug discovery, and diagnostics that need reliable AI processing power.

5. Manufacturing & Industry 4.0 – Companies adopting AI for quality control, predictive maintenance, and intelligent automation in production lines.

6. Financial Services & FinTech – Firms leveraging AI for fraud detection, algorithmic trading, and risk analysis.

7. Retail & eCommerce – Businesses enhancing customer experiences with AI-driven recommendation systems and demand forecasting.

8. Media & Entertainment – Organizations using AI for content generation, video analytics, and personalization.

9. Government & Defense – Agencies requiring AI-powered cybersecurity, surveillance, and intelligence analysis tools.

Solution Highlights

Supermicro's turnkey AI Workload Solution simplifies AI deployment by integrating cutting-edge hardware and software into a pre-configured, ready-to-deploy system. Powered by RAVEL Orchestrate software, Intel Xeon processors, and the latest generation of NVIDIA GPUs, this solution seamlessly enables businesses and developers to integrate AI capabilities into their operations with minimal effort.

Key Benefits

● Plug-and-Play AI Deployment: Eliminate the complexity of setting up AI infrastructure with a pre-configured solution that is ready to use out of the box.
● Optimized Performance: Leverage NVIDIA’ s latest GPUs to accelerate AI and GenAI workloads for superior efficiency and scalability.
● Comprehensive Software Integration: Utilize RAVEL Orchestrate software to simplify AI workload management and orchestration.

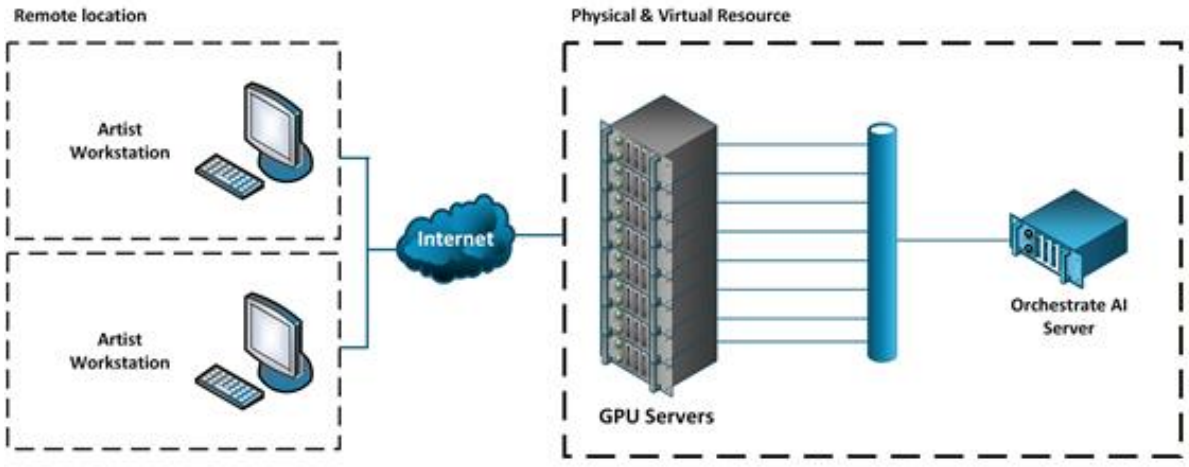
● Enterprise-Grade Reliability: Built with Supermicro's industry-leading hardware to ensure high performance, reliability, and scalability.
● Reduced Time to Value: Quickly implement AI-driven applications without the need for extensive configuration or technical expertise.

Technical Specifications

● Hardware: Latest-generation Supermicro AI workstations with Intel Xeon processors and NVIDIA GPUs for high-speed AI computation.
● Software: RAVEL Orchestrate AI for streamlined AI workload management, deployment, load balancing, and optimization.
● Scalability: Flexible architecture that supports growing AI demands and can be customized for specific workloads.
● Security: Enterprise-grade security measures to protect sensitive AI data and workloads.

Use Cases

● Generative AI (GenAI) & AI Model Deployment: Deploy and run generative AI and AI models efficiently for applications in content creation, image generation, and language modeling.



Orchestrate AI - AI Batch Jobs & Load Balancing

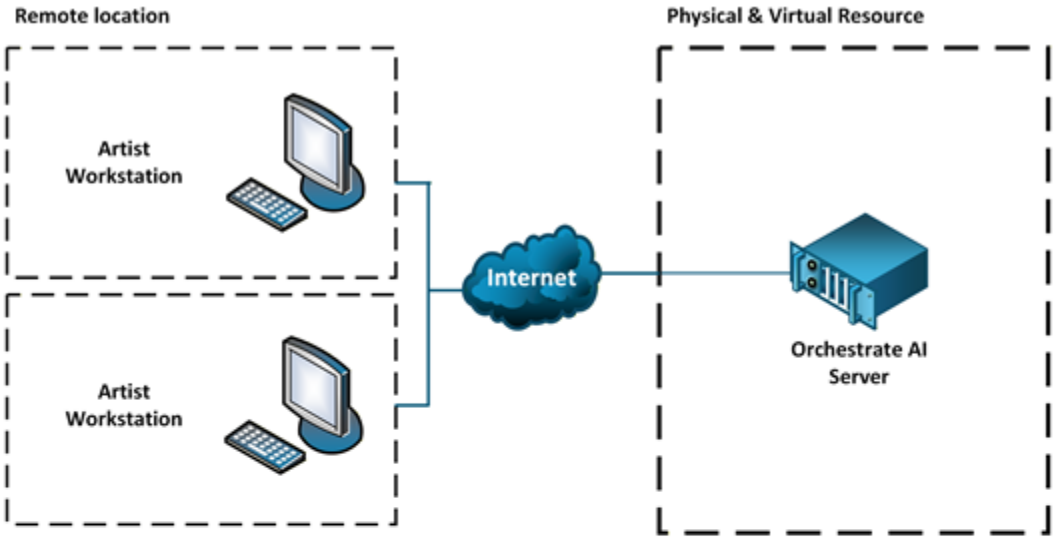
Generative AI (Image & Video):
ComfyUI and Stable Diffusion - advanced distributed workflows for high-fidelity image and video (WAN 2.2) synthesis, optimized for GPU-accelerated rendering pipelines and multi-node orchestration.

Inferencing Engines:
Ollama and KoboldCPP - high-performance local LLM inferencing stacks enabling low-latency conversational AI, offline deployment, and full CUDA optimization for scalable multi-GPU workloads.

Training Infrastructure:
Custom CUDA-optimized environments with distributed dataset prefetching, gradient checkpointing, and mixed-precision optimization to accelerate fine-tuning and reinforcement learning across heterogeneous GPU clusters.

AI Driven Analytics:
TensorBoard and LangFuse - supported by Orchestrate AI for real-time training and inference visibility. They track metrics, prompts, and GPU performance across distributed nodes, providing synchronized insight and seamless observability across every active workload.

- AI Batch Jobs & Load Balancing: Accelerate model training and inference for a wide range of AI applications.



Orchestrate AI - Generative AI (GenAI) & AI Model Deployment

Generative AI (Image & Video):
ComfyUI and Stable Diffusion - advanced distributed workflows for high-fidelity image and video (WAN 2.2) synthesis, optimized for GPU-accelerated rendering pipelines and multi-node orchestration.

Inferencing Engines:
Ollama and KoboldCPP - high-performance local LLM inferencing stacks enabling low-latency conversational AI, offline deployment, and full CUDA optimization for scalable multi-GPU workloads.

Training Infrastructure:
Custom CUDA-optimized environments with distributed dataset prefetching, gradient checkpointing, and mixed-precision optimization to accelerate fine-tuning and reinforcement learning across heterogeneous GPU clusters.

AI Driven Analytics:
TensorBoard and LangFuse - supported by Orchestrate AI for real-time training and inference visibility. They track metrics, prompts, and GPU performance across distributed nodes, providing synchronized insight and seamless observability across every active workload.

System Hardware Configuration

			
Model	SYS-532AW-C	SYS-531A-I	SYS-551A-T
Processor	Intel Core Ultra 9 285K	Intel Xeon W7-2595X	Intel Xeon W7-3565X
GPU	Integrated	NV RTX PRO 6000 Blackwell Max-Q (300W)	NV RTX PRO 6000 Blackwell Workstation Edition (600W)
System Memory	64GB DDR5, UDIMM	128GB DDR5 5600, ECC	256GB DDR5 5600, ECC
Storage	1TB NVMe SSD	2TB NVMe SSD	2TB NVMe SSD
Networking	1 x 1GbE 1 x 2.5GbE	1 x 1GbE 1 x 10GbE	1 x 1GbE 1 x 10GbE
Power Supply	1000W (Gold)	1200W (Platinum)	2000W (Platinum)
Form Factor	Mid-ATX Tower	Mid-ATX Tower	Full ATX Tower
Dimensions	17.7 x 8.1 x 18.5” 450 x 205 x 470mm	16.7 x 7.6 x 20.7” 424 x 193 x 525mm	21.6 x 8.7 x 22.6” 535 x 222 x 573mm

Figure 1 - Recommended configurations for Supermicro Intel Workstations SYS-532AW-C, SYS-531A-I, and SYS-551A-T with RAVEL Orchestrate AI

The Supermicro SYS-532AW-C, SYS-531A-I, and SYS-551A-T workstations with NVIDIA RTX PRO™ 6000 Blackwell Workstation Edition GPUs and Intel Xeon processors. Specifically, the layout is detailed below:

Get Started

Supermicro is a trusted leader in high-performance computing solutions, delivering industry-leading hardware optimized for AI workloads. By combining Supermicro’s expertise with NVIDIA’s cutting-edge GPU technology, Intel Xeon processors, and RAVEL Orchestrate AI software, this turnkey solution delivers a reliable, efficient, and scalable AI infrastructure that accelerates innovation. Discover how Supermicro’s X RAVEL turnkey AI workload solution can revolutionize your AI initiatives. Contact us today to learn more or request a demo.

For Integrators & Resellers Interested in partnering to host your own AI Workload POC, please get in touch with us at https://ravelinc.com/partner/# .

Further Information

https://www.supermicro.com/en/

https://www.ravelinc.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

RAVEL

RAVEL is a leading provider of DevOps tools for the orchestration of cloud, on-prem, and data center infrastructure for complex workloads and workflows. Our flagship product RAVEL Orchestrate™ provides teams with a no-code, DevOps & IT solution that automates & rapidly assembles complex software and hardware environments, including managing infrastructure, virtualization technology, identity management, and customized software images. From digital content creation studios to architecture and engineering environments, to large enterprises in the energy, computing, and healthcare sectors, our solutions help our customers elevate their productivity and improve their workflow.

Learn more at: www.ravelinc.com