SOLUTION BRIEF

SUPERMICRO GPU SERVERS WITH MANGOBOOST LLMBOOST™ AI ENTERPRISE SOFTWARE: UNLOCKING THE FULL POTENTIAL OF AMD INSTINCT™ GPUS

Combination Delivers a Cost-Effective, Easy-to-Use, High-Performance, Flexible, and Scalable Multiple-Node GenAI Server Solution

AS -8126GS-TNMR

Table of Contents

Executive Summary

Supermicro GPU systems, in conjunction with MangoBoost's LLMBoost™ AI Enterprise software, offer a solution designed to optimize generative AI performance. This collaboration aims to provide a scalable and cost-effective approach for organizations utilizing AMD Instinct™ GPUs. The integrated offering supports the deployment, management, and scaling of various AI workloads, including inference, training, and Retrieval Augmented Generation (RAG) applications. This can help organizations enhance hardware efficiency and streamline their AI development processes.

Motivation: Meeting the Challenges of the GenAI Boom

The rapid rise of Large Language Models (LLMs) has transformed what is possible with generative AI, enabling sophisticated chatbots, advanced visual generation, and productivity tools that augment human capabilities. However, deploying and operating LLMs at scale present unique challenges:

Integration complexity: Productizing an AI model demands integrating it into a larger application stack, adding significant time and engineering overhead.

- Evolving hardware and model demands: Organizations must constantly adapt to new AI model architectures and leverage the latest hardware capabilities to stay competitive.

- Performance and cost trade-offs: AI systems must maintain high throughput, low latency, and cost-efficiency even as user demand and model sizes grow.

- Multiple-node scale: Serving today's largest models increasingly requires multiple-node GPU clusters, which add layers of orchestration and tuning complexity.

Open-source tools offer a solid foundation but often require significant infrastructure expertise to integrate and operate effectively. Achieving peak performance in real-world deployments requires complex, system-wide tuning that spans inference, training, fine-tuning, kernel optimizations, networking, scheduling, and backend engines. For many organizations, delivering a reliable, high-performance, and flexible GenAI service is a challenging and resource-intensive undertaking.

AMD Instinct™ GPUs Optimally Deployed on Supermicro Servers with MangoBoost LLMBoost™

To address these challenges, Supermicro and MangoBoost have collaborated to deliver an optimized end-to-end GenAI stack, combining Supermicro's robust AMD Instinct GPU server portfolio with MangoBoost's LLMBoost AI Enterprise MLOps software.

Supermicro AMD-Based Servers with AMD EPYC and AMD Instinct:

Supermicro's GPU-optimized servers provide a flexible, high-density, and energy-efficient platform for AI workloads. Powered by AMD Instinct MI355X, MI350X, MI325X and MI300X GPUs, these AI servers deliver:

- Competitive compute performance for large model training and inference.

- Industry-leading memory capacity and bandwidth, crucial for hosting massive LLMs with minimal off-chip data movement.

- Enterprise-grade reliability and configurability, meeting the demands of multiple-node GenAI clusters.

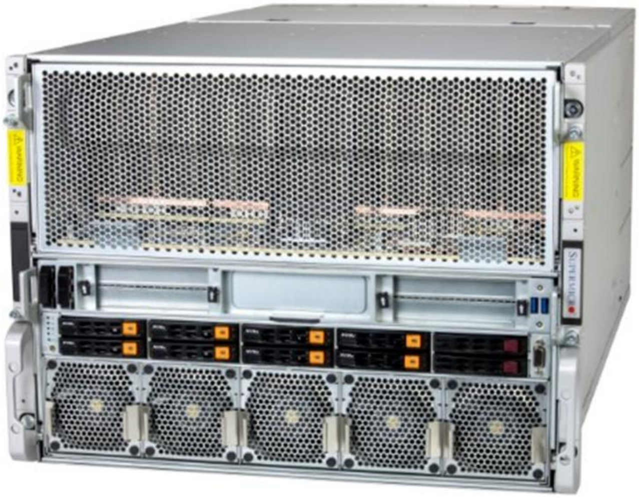


Figure 1 - Supermicro A+ Server AS -8126GS-TNMR

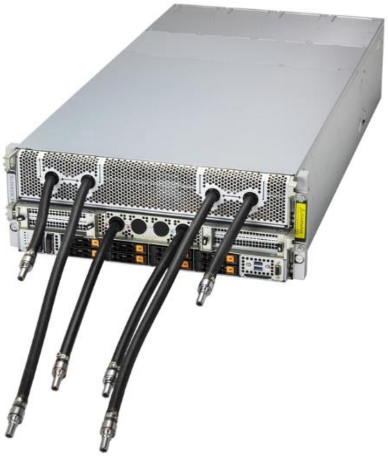
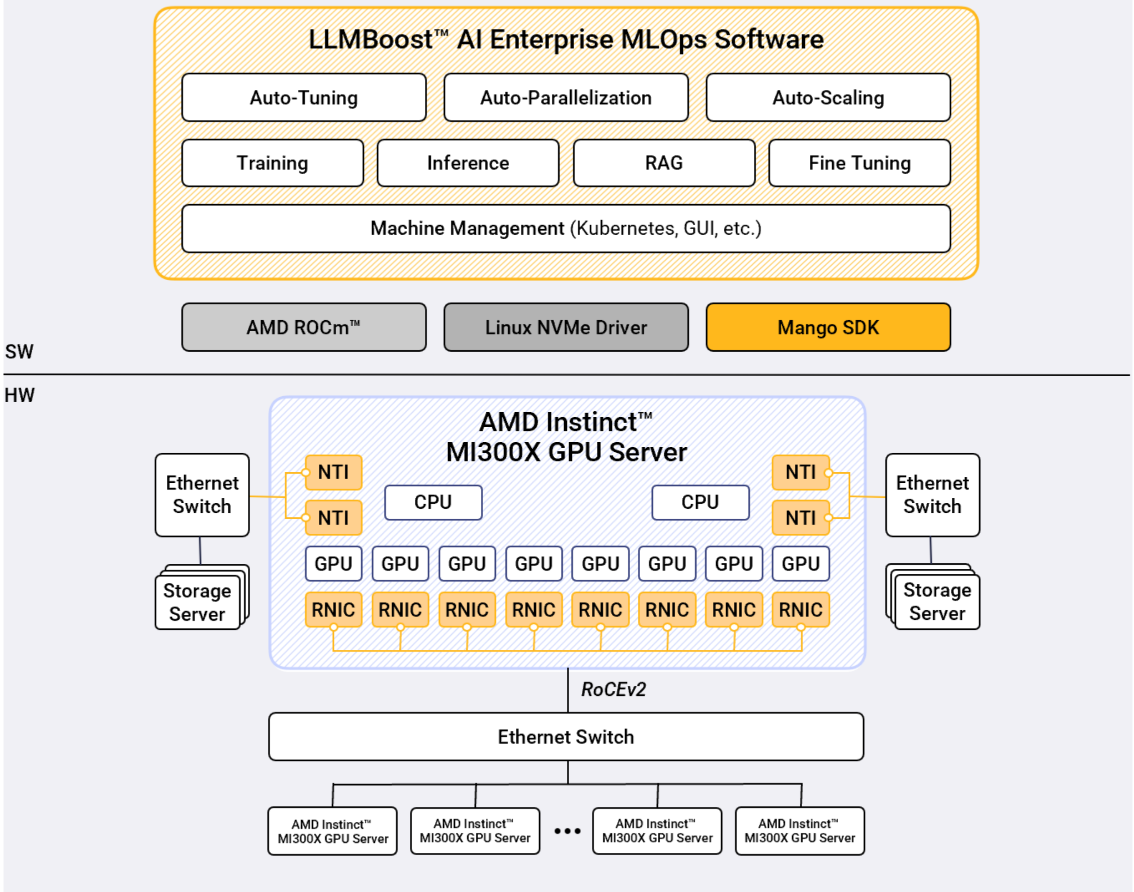


Figure 2 - Supermicro A+ Server AS-4126GS-NMR-LCC

MangoBoost LLMBoost: Enterprise-Grade MLOps for GenAI

MangoBoost's LLMBoost software complements the powerful hardware with a full-stack, production-ready AI MLOps platform:

- Plug-and-play deployment: Pre-built Docker images and an intuitive CLI help developers launch LLM workloads quickly.
- OpenAI-compatible API: Allows developers to integrate LLM endpoints with minimal code changes.
- Kubernetes-native orchestration: Automated deployment and management of autoscaling, load balancing, and job scheduling for seamless operation across single-node or multiple-node clusters.
- Full-stack performance autotuning: Unlike conventional autotuners that only handle model hyperparameters, LLMBoost optimizes every layer, from the inference and training backends to network configurations and GPU runtime parameters. This ensures maximum hardware utilization with no manual tuning required.



** RNIC: Mango GPUBoost^{TM} - RDMA * NTI: Mango StorageBoost^{TM} – NVMe over TCP Initiator * GPU: AMD Instinct^{TM} MI300X*

Figure 3 - Example architecture diagram showing Supermicro AMD Instinct GPU cluster running MangoBoost LLMBoost stack orchestrating AI deployment workflows.

Combined, Supermicro hardware and MangoBoost software provide a turnkey GenAI cluster solution that is high-performance, flexible, and easy to operate, unlocking the full capability of AMD Instinct GPUs with minimal engineering overhead.

Learn more about how Supermicro systems with AMD MI300X GPUs and MangoBoost Software set new records in MLPerf benchmarks, such as:

- The highest-ever llama2-70b performance in MLPerf Inference 5.0 (https://www.mangoboost.io/resources/blog/mangoboost-sets-the-highest-mlperf-inference-v5-0-result-in-history-for-llama2-70b-offline).

- The first-ever AMD multiple-node training on llama2-70b in MLPerf Training 5.0 (https://www.mangoboost.io/resources/blog/mangoboost-sets-a-new-standard-for-multiple-nodes-llama2-70b-lora-on-amd-mi300x-gpu).

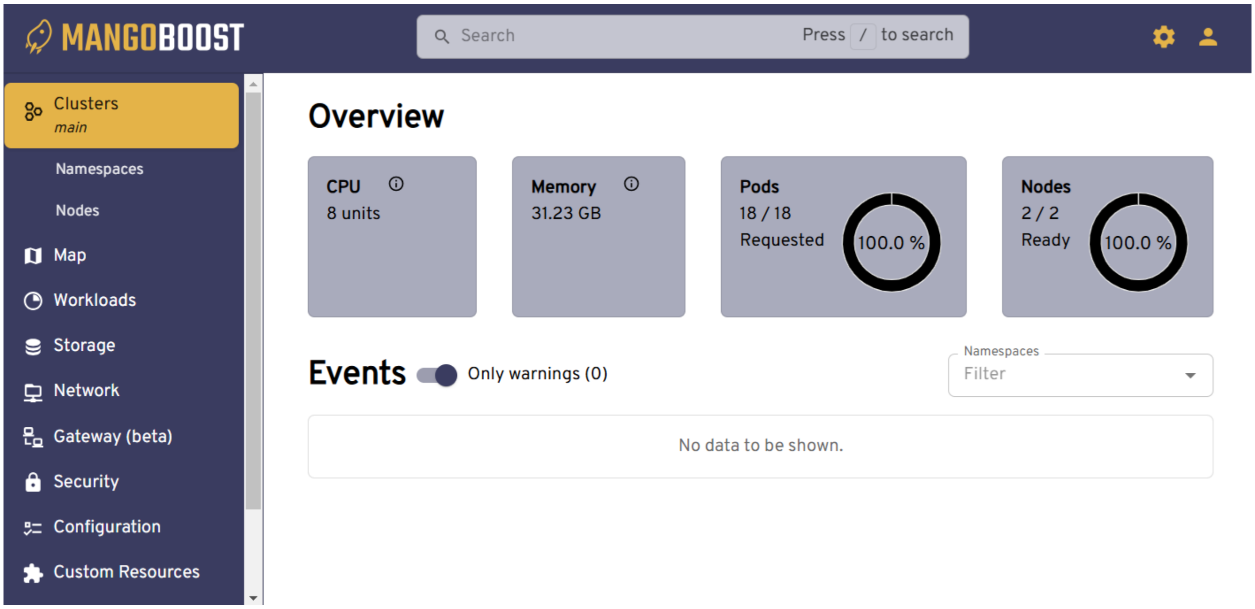
Check out the LLMBoost documentation page to learn more about LLMBoost at https://docs.mangoboost.io.

Evaluation

The combined Supermicro and MangoBoost solution has been tested on real-world GenAI workloads to validate ease of management, scalability, and performance.

Seamless Manageability

LLMBoost offers native Kubernetes integration and an intuitive admin UI for cluster orchestration, workload autoscaling, and resource monitoring, supporting multiple-node deployments.



Overview

CPU
8 units

Memory
31.23 GB

Pods
18 / 18
Requested
100.0 %

Nodes
2 / 2
Ready
100.0 %

Events
Only warnings (0)

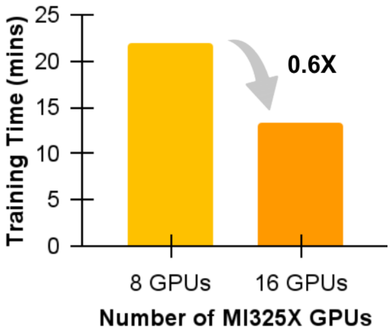
Namespaces
Filter

No data to be shown.

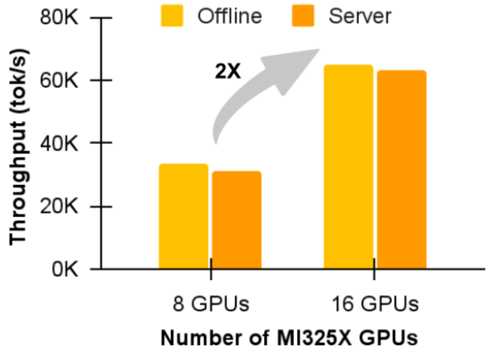
Figure 4 - Kubernetes cluster dashboard and LLMBoost admin interface managing multiple-node deployments.

Proven Multiple-Node Scalability on Industry-Standard MLPerf Benchmarks

LLMBoost delivers impressive multiple-node scalability on Supermicro AMD MI325X servers, as demonstrated in Experiments 1 and 2. These experiments leveraged the industry-standard MLPerf Training 5.0 Llama2-70b-LoRA and Inference v5.0 Llama2-70b benchmarks. The results reveal that LLMBoost reduces training time by 40% for 2-node training and achieves a 1.96X higher throughput for multiple-node inference on these Supermicro AMD servers.



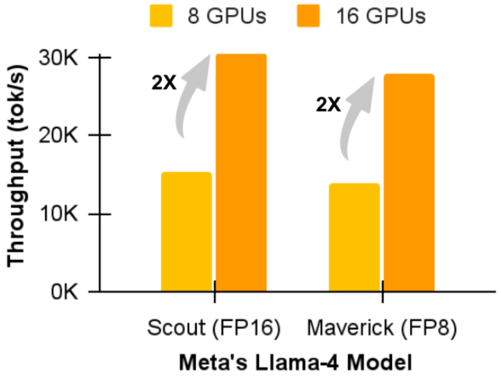
Experiment 1: MLPerf Training on 2×8×MI325 cluster. LLMBoost can train Llama2-70b-LoRA in 21.6 minutes using a single node and in 13.3 minutes using a 2-node MI325X.



Experiment 2: MLPerf Inference on 2×8×MI325 cluster. LLMBoost not only meets the latency constraints but also achieves a high throughput of 31,595.1 tokens/s on a single MI325X node and 61,776.9 tokens/s when scaled to a 2-node MI325X configuration.

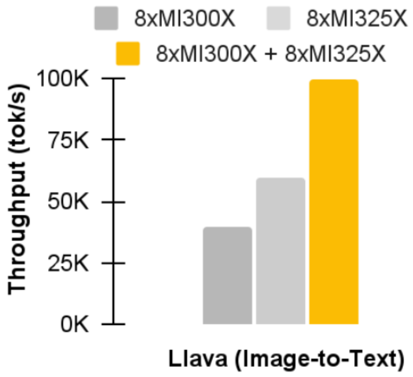
Scaling Performance on any Model

Highlighting the versatility of Supermicro's AMD-based GPU server platforms and LLMBoost software on other open models, Experiment 3 focuses on scaling with Meta's cutting-edge Llama-4 models (Scout and Maverick). The data validates consistent linear performance scaling, indicating robust support for a range of open models.



Experiment 3: Inference throughput for the latest Llama Maverick and Scout models on a 2×8×MI325X cluster.

Finally, experiment 4 showcases LLMBoost's load-balancing capabilities across different generations of Supermicro's AMD-based GPU platforms. This experiment utilized Llava, an image captioning model, on three different setups: a single-node 8xMI300X system, a single-node 8xMI325X system, and a heterogeneous two-node system combining 8xMI300X and 8xMI325X GPUs. The heterogeneous two-node setup achieves 106,770 token/s, within 96% of the sum of individual single-node runs (46,898 + 64,317), demonstrating minimal overhead and high efficiency.



Experiment 4: Inference for image captioning (Llava) on a mixed cluster (8×MI300 + 8×MI325) showing effective load balancing across different GPU types.

Summary of Tests

	Workload	Cluster	Highlights/Performance Summary
Experiment 1	MLPerf Training	2 x 8 x MI325X	Validates multiple-node training scalability, achieving a 40% reduction in training time when scaling from 1 to 2 nodes.
Experiment 2	MLPerf Inference	2×8×MI325X	Demonstrates low-latency, high-throughput inference, delivering 31,595.1 token/s on a single MI325X node and 61,776.9 token/s on a 2-node configuration —a 1.96× throughput improvement
Experiment 3	Llama Maverick & Scout Inference	2×8×MI325X	Highlights near-linear scalability for SOTA LLMs, achieving almost 2× throughput when scaling from 1 to 2 nodes
Experiment 4	Llave Image Captioning (Heterogeneous)	8×MI300X + 8×MI325X	Confirms LLMBoost's flexible deployment and effective load balancing across heterogeneous GPUs. The 2-node setup achieves 106,770.73 tokens/s, within 96% of the combined throughput from separate single-node setups.

Together, these results confirm that Supermicro servers with MangoBoost LLMBoost deliver an easy-to-manage, high-performance, and flexible GenAI solution that scales seamlessly with workload demands.

Conclusion

Supermicro's AS-8126GS-TNMR platforms, powered by AMD Instinct™ MI325X and MI300X GPUs, deliver exceptional performance, scalability, and deployment flexibility for today's most demanding AI workloads. From industry-standard MLPerf™ benchmarks to state-of-the-art large language model (LLM) inference, these systems consistently outperform expectations, achieving near-linear multi-node scalability, low-latency inference, and seamless heterogeneous GPU orchestration. Combined with optimized AI software like LLMBoost™, Supermicro provides a future-ready, cost-effective infrastructure that empowers enterprises to deploy and scale generative AI solutions rapidly.

The experiments demonstrate the strong performance and scalability of Supermicro's MI325X- and MI300X-based platforms across MLPerf Training, MLPerf Inference, and real-world LLM workloads:

- MLPerf Training on LLaMA 2 70B with LoRA showed a 40% reduction in convergence time when scaling from 1 to 2 nodes, highlighting excellent multi-node training efficiency.

- MLPerf Inference throughput doubled from 31,595.1 tokens/s (1 node) to 61,776.9 tokens/s (2 nodes), indicating 1.96× throughput scalability with minimal overhead.

- In-house LLM inference with Maverick and Scout models achieved near-linear scaling on MI325X nodes, proving the readiness of Supermicro systems for real-time generative AI deployment.

- The Llave Image Captioning experiment across heterogeneous 8×MI300X + 8×MI325X configurations showcased Supermicro's ability to balance workloads across diverse GPUs, delivering 106,770.73 tokens/s, within 96% of ideal combined throughput, affirming the effectiveness of LLMBoost's load-balancing mechanisms.

These results validate Supermicro's platform as a robust, high-performance AI infrastructure suitable for both training and inference at scale, including in heterogeneous environments.

For More Information:

Please visit: www.supermicro.com/en/accelerators/amd

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

AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics and visualization technologies. Billions of people, leading Fortune 500 businesses and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible.

Learn more at www.amd.com

MANGOBOOST

MangoBoost is a provider of cutting-edge, full-stack system solutions for maximizing compute efficiency and scalability. At the heart of the solutions is the MangoBoost Data Processing Unit (DPU), which ensures full compatibility with general-purpose GPUs, accelerators, and storage devices, enabling cost-efficient, standardized AI infrastructure. Founded in 2022 on a decade of research, MangoBoost is rapidly expanding its operations in the U.S., Canada, and Korea.

Learn more at www.mangoboost.io