SOLUTION BRIEF

Architecting Scalable Inference on AMD AI Platforms

Supermicro Accelerated Solutions Featuring AMD Instinct™ GPU Clusters

MAY 2026

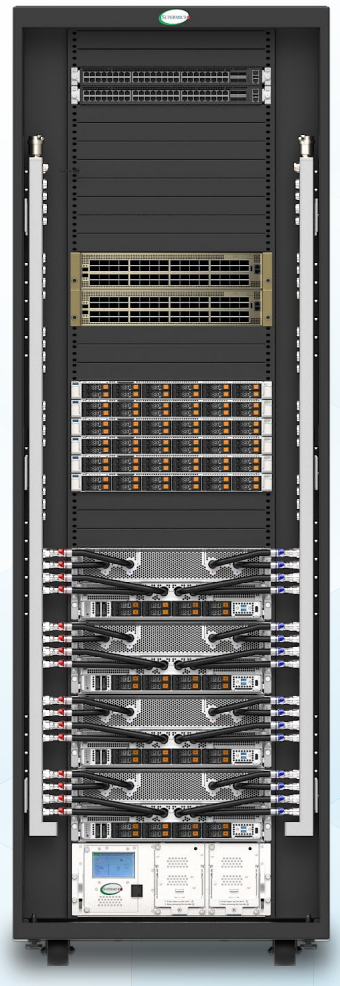


TABLE OF CONTENTS

EXECUTIVE SUMMARY

Enterprises are rapidly shifting from AI experimentation to production-scale deployment, but many are discovering that infrastructure, not models, is the bottleneck, resulting in delayed time-to-market, emergency re-architecture, and high costs of running inference operations on platforms not designed to support them at scale.

Supermicro Accelerated AMD AI Platforms featuring AMD Instinct™ GPUs provide rack-scale infrastructure built for enterprises that need to move AI from proof of concept to production — and build from there. Optimized for AI deployments today and engineered to scale with the full breadth of enterprise AI workloads as they evolve, this platform delivers the production readiness, operational efficiency, and deployment speed required to run AI as a core business capability — at a cost structure that sustains it. The solution delivers:

- **Faster time to production:** Pre-validated, modular systems deploy in weeks, not months.
- **Lower cost per inference:** Memory-dense architecture, liquid cooling, and power-efficient silicon reduce total cost of ownership (TCO).
- **Enterprise AI Software:** AMD Enterprise AI Software Stack provides example blueprint designs and inference microservices to enable rapid time-to-deployment.
- **Vendor independence:** Open software ecosystem (ROCm™) and flexible hardware configurations to support evolving workloads and future accelerators.

Designed for AI Factor Scale and Built for What Comes Next

As inference grows in importance, organizations need infrastructure that can scale efficiently while adapting to new models, deployment patterns, and accelerator technologies. Supermicro optimized this platform for AI environments today while ensuring it remains flexible enough to support a wide range of AI use cases over time.

TODAY'S AI INFRASTRUCTURE CHALLENGE: EXPERIMENTATION TO PRODUCTION

Organizations across every industry have made significant investments in AI pilots, but moving from experimentation to production remains a major hurdle because the infrastructure cannot support the transition. As organizations scale AI into real-world applications, several challenges emerge:

Inference demand: As organizations move from building models to deploying them at scale, the demands on AI infrastructure shift fundamentally — from periodic, high-throughput training runs to continuous, latency-sensitive production workloads spanning inference, fine-tuning, and real-time data pipelines. Infrastructure optimized for training alone cannot absorb this transition without compromising performance, cost efficiency, or operational predictability. Meeting the full scope of production-scale AI demands requires a purpose-built platform for sustained throughput, not just peak compute.

Certainty and performance at scale: Workloads that succeed in a low-volume pilot routinely fail to meet performance and cost expectations at production scale — leaving organizations without the assurance they need that their AI infrastructure will deliver when it matters.

Rigid architectures and lock-in: Closed AI stacks, from silicon to software, limit flexibility, making it difficult and costly for organizations to adapt as AI technologies and workloads evolve.

Power and density constraints: Existing data center environments often cannot support the thermal and power demands of modern AI infrastructure at scale.

THE SOLUTION: SUPERMICRO H14 SYSTEMS WITH AMD AI PLATFORMS FEATURING AMD INSTINCT GPUS AND AMD ENTERPRISE AI STACK

Supermicro Accelerated Solutions featuring AMD Instinct MI350 Series GPUs provide a rack-scale AI infrastructure platform designed for production-scale AI workloads, from high-performance inference to large-scale training, across deployment environments ranging from enterprise data centers to regional AI factories. The platform scales flexibly to match organizational demand, whether teams are standing up initial AI capacity or expanding to full factory-scale operations. The platform reduces storage and data pipeline bottlenecks across inference and training workloads to help sustain throughput, increase GPU utilization, and improve operational efficiency at scale.

The Supermicro AI Platforms offer several cluster size options, with preconfigured L11 bill of materials (BOM) with full rack configurations including servers, switches, cables, and cooling solution.

GPU Family	MI350X	MI355X	MI355X	MI355X
Supermicro Server	**AS -8126GS-TNMR**	**AS -A126GS-TNMR-T**	**A S -4126GS-NMR-LCC**	**AS -4126GS-NMR-LCC**
Server Cluster Size	4, 16, 64	4, 16, 64	4, 32, 64	8, 48, 128
Number fo GPUs	32, 128, 512	32, 128, 512	32, 256, 512	64, 384, 1024
GPU	MI350X AC	MI350X AC	MI355X DLC	MI355X DLC
Number of AI NICs	40, 160, 640	40, 160, 640	40, 320, 640	80, 480, 1280
Supermicro L11 BOM SKU	SRS-48UAC-MI350-8U4N-R0	SRS-48UAC-MI355-10U4N-R0	SRS-48ULC-MI355-4U8N-R0	SRS-52ULC-MI355-4U12N-R0

Supermicro's Building Block approach enables organizations to scale infrastructure without re-architecting or over-provisioning. The same platform supports a range of AI workloads and deployment models, allowing teams to expand capacity, introduce new accelerators, and adapt to changing requirements without disrupting operations.

Supermicro AMD AI Platforms are available through major CSPs and leading neocloud providers, enabling AI builders to deploy AMD Instinct-accelerated workloads across on-premises, hybrid, and cloud environments.

By combining Supermicro servers with AMD high-performance compute and AMD Enterprise AI Software Suite, the platform delivers a faster, more flexible path from AI pilot to production.

The Supermicro + AMD Difference

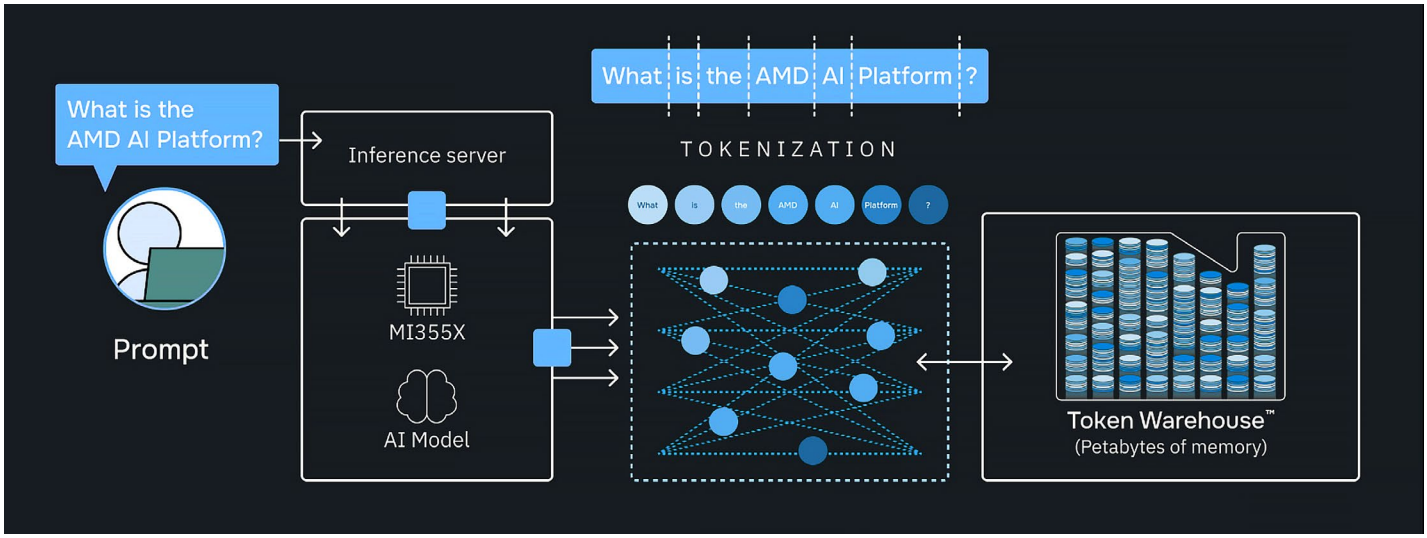
- **Faster time to deployment:** Modular, pre-validated system designs integrate the latest AMD EPYC™ Series CPUs, Instinct GPUs, and Pollara 400G NICs, reducing time from procurement to production.
- **Flexible, right-sized infrastructure:** Supermicro provides the industry's broadest portfolio of AMD-based AI Platform configurations, enabling teams to precisely match infrastructure to workload requirements from edge inference to large-scale data center deployments.
- **Proven Enterprise AI Software:** AMD Enterprise AI Software Stack, verified on the MI350 series, enables confident deployment and faster time-to-launch.
- **Open architecture:** Open software ecosystem and standards-based infrastructure provide flexibility to evolve as AI workloads and technologies change.

Key Features of Supermicro AMD AI Platforms

- Modular, pre-validated GPU 8U and 10U air-cooled and 4U liquid-cooled systems, deployable in weeks through Supermicro's Building Block architecture.
- AMD Instinct MI350X, MI355XGPUs feature 288 GB HBM3e per GPU (2.3 TB per node), 8 TB/s memory bandwidth, and hardware-native MXFP4/MXFP6/MXFP8/FP16 precision support.
- Dual-socket AMD EPYC 9005 Series CPUs (up to 192 cores/CPU) and AMD Infinity Fabric Link interconnect (1.12 TB/s aggregate), eliminating both CPU and GPU-to-GPU bottlenecks.
- AMD Pensando™ Pollara 400 AI NICs provide high performance, scale-out Ethernet switched GPU-to-GPU back side networking.
- Direct Liquid Cooling (DLC-2), reducing power consumption by up to 40%[i], physical footprint by up to 60%[ii], and TCO by up to 20%.
- AMD ROCm open software with over 2M+ models from Hugging Face and Day-0 model open framework support across PyTorch, JAX, TensorFlow, vLLM, SGLang, and Triton—plus the AMD Enterprise AI Suite delivers a fully supported, enterprise-ready AI stack with optimized frameworks, validated tools, and standards-based runtimes, with zero licensing fees, reducing integration effort and accelerating deployment of both training and inference workloads.
- Supermicro SuperCloud Composer is a unified, single pane of glass which enables centralized fleet management with Redfish API access.
- Consistent deployment from edge pilots to cloud-scale clusters, with identical software and operational playbooks across tiers.

The Supermicro network cluster guide provides details on various Ethernet switches and network topologies to optimize architectures for providers and customers building infrastructures with AMD Instinct™ MI350X-MI355X Accelerators.

DATA FLOW BLOCK DIAGRAM



ARCHITECTURE SNAPSHOT

Supermicro's GPU-optimized systems share a common processor platform, networking architecture, and software stack across air-cooled and liquid-cooled configurations, allowing teams to standardize deployment and operations regardless of environment.

Feature	Air-Cooled 8U System Supermicro AS -8126GS-TNMR	Air-Cooled 10U System Supermicro AS -A126GS-TNMR	Liquid-Cooled 4U System Supermicro AS -4126GS-NMR-LCC
GPU Platform	AMD Instinct MI350X (8 GPUs)	AMD Instinct MI350X (8 GPUs)	AMD Instinct MI355X (8 GPUs)
Cooling	Dual-zone air cooling	Air cooling (up to 19x 8cm fans)	Direct-to-chip liquid cooling (DLC)
Processor	Dual AMD EPYC 9575F	Dual AMD EPYC 9575F	Dual AMD EPYC 9575F
GPU Memory	288 GB HBM3e per GPU (2.3 TB/node)	288 GB HBM3e per GPU (2.3 TB/node)	288 GB HBM3e per GPU (2.3 TB/node)
Precision	OCP-FP8, MXFP8, MXFP6, MXFP4, FP16, BF16* (hardware-native)	OCP-FP8, MXFP8, MXFP6, MXFP4, FP16, BF16* (hardware-native)	OCP-FP8, MXFP8, MXFP6, MXFP4, FP16, BF16* (hardware-native)
Networking	Pollara 400 Gbps AI NIC	Pollara 400 Gbps AI NIC	Pollara 400 Gbps AI NIC
Ethernet Interface	PCIe® Gen5.0x16; OCP® 3.0	PCIe® Gen5.0x16; OCP® 3.0	PCIe® Gen5.0x16; OCP® 3.0
Power	6x 5,250W Titanium Redundant	6x 6,600W Titanium Redundant	4x 6,600W Titanium Redundant

* *MXFP8, MXFP6, and MXFP4 are OCP Microscaling (MX) formats natively supported in hardware on AMD CDNA4 architecture (MI350X and MI355X). OCP-FP8 (E4M3/E5M2) is the standard OpenCompute FP8 format supported across CDNA3 and CDNA4. FP32, TF32, and FP64 are also supported.*

SOFTWARE STACK

Supermicro built the platform using an open, production-ready software ecosystem that enables rapid deployment, simplifies operations, and allows customer choice. Key components:

- **AMD ROCm:** Open-source software stack including drivers, development tools, and APIs that enable GPU programming from low-level kernel to end-user applications, with leading AI frameworks support - including PyTorch, JAX, TensorFlow, vLLM, SGLang, and Triton. ROCm offers a suite of optimizations for AI workloads and supports the broader AI software ecosystem, including open frameworks, models, and tools.
- **AMD Enterprise AI Suite:** Helps organizations deploy AI faster and with less complexity. Built on AMD ROCm software and optimized for AMD Instinct, the suite delivers a fully supported, enterprise-ready AI stack with optimized frameworks, validated tools, and standards-based runtimes that reduce integration effort and accelerate deployment of both training and inference workloads, streamlining the entire AI lifecycle with no licensing fees. Installation details for on-premises deployments are now available.
- **Open Ecosystem by Design:** Over 2 million models from Hugging Face for containerized deployment reduce migration effort, enabling flexibility across environments and future hardware.

PROVEN PERFORMANCE

The AMD Instinct MI355X, running on the Supermicro AS -4126GS-NMR-LCC liquid-cooled 4U platform, delivers independently validated, production-representative AI inference performance — the operational priority that dominates enterprise AI infrastructure decisions at scale. Benchmarked using the industry-standard MLPerf Inference 6.0 suite, the MI355X not only redefines generational performance but sets a new production milestone: for the first time in MLPerf history, an AMD platform surpassed 1 million tokens per second at multi-node scale. On Llama 2 70B, a single 8-GPU node delivers 100,282 tokens per second — 3.1x the throughput of the prior-generation MI325X — while remaining competitive across Offline, Server, and Interactive scenarios. At 11 nodes and 87 GPUs, the MI355X scales to 1,042,110 tokens per second at 93% efficiency, validating that performance does not erode as deployments grow. These results are not theoretical peak figures — they are MLCommons-submitted, Supermicro-reproduced configurations, within 4% of AMD's own reference numbers, giving enterprises the confidence to plan production deployments around them.[iii]

Workload Category	Benchmark & Model	Configuration	Result*	Business Impact
Large Model Inference	MLPerf 6.0 — Llama 2 70B Server	1 node, 8X MI355X	100,282 tokens/sec	3.1x higher throughput vs. MI325X; 97% of B200 Server, tied B200 Offline, 119% of B200 Interactive — more requests served per dollar of infrastructure
Inference at Scale	MLPerf 6.0 — Llama 2 70B	11 nodes, 87x MI355X	1,042,110 tokens/sec Offline; 1,016,380 Server; 785,522 Interactive	First AMD submission to exceed 1M tokens/sec; 93% Offline/Server scale-out efficiency and 98% Interactive efficiency validate fabric performance and investment predictability
Emerging Model Inference	MLPerf 6.0 — GPT-OSS-120B	1 node, 8x MI355X	111% of competitor Offline; 115% of competitor Server	First-time model enablement with day-one competitive performance demonstrates ROCm software readiness for emerging 120B+ model families
Ecosystem Reproducibility	MLPerf 6.0 — Llama 2 70B	Supermicro + 8 ecosystem partners	Within 4% of AMD reference results	Partner-validated results confirm that benchmark performance translates directly to production deployments, reducing implementation risk
Fine-Tuning	MLPerf 5.1 — Llama2-70B LoRA	1 node, 8X MI355X vs. 4 nodes, 32X MI300X	9.96 min vs. 10.92 min**	One MI355X node outperforms four MI300X nodes — 75% reduction in cluster footprint and associated power, cooling, and licensing costs

* *MLPerf Inference 6.0 results from MLCommons, April 1, 2026. Single-node test platform: Supermicro AS -4126GS-NMR-LCC, 2x AMD EPYC 9575F, 3 TB DDR5-6400, 8x AMD Instinct MI355X 288GB HBM3e. Fine-tuning result from MLPerf v5.1; MLPerf Inference 6.0 does not include a fine-tuning benchmark.*[iii]

** *MLCommons MLPerf v5.1 results. Test platform: Supermicro AS -4126GS-NMR-LCC, 2X AMD EPYC 9575F, 3 TB DDR5-6400, 8X AMD Instinct MI355X 288GB HBM3e.*[iv]

The platform scales without re-architecture. The same software stack, management plane, and operational model apply across all deployment sizes, so what you build or pilot today can scale seamlessly

IMPACT AT SCALE

- More output per dollar: Higher inference throughput increases the number of requests served without scaling infrastructure.
- Proven AI solution at scale: AMD Instinct GPUs, EPYC CPUs, Pollara NICs and ROCm are used by 8 of the top 10 AI organizations in the world with large clusters running numerous production workloads.
- Faster iteration cycles: Improved fine-tuning efficiency enables faster model updates and deployments.

SEE IT IN ACTION: THE JUMPSTART PROGRAM

Supermicro + AMD JumpStart provides hands-on access to bare-metal systems so you can benchmark real workloads, validate the software stack against your existing toolchain, and assess real-world economics against your specific deployment profile.

Get Ready for Tomorrow's AI

Contact your Supermicro or AMD account team to enroll in the **JumpStart program**

A+ page: **Broad Range A+ Servers for Data Center. Cloud. AI | Supermicro**

---

i. Supermicro's DLC-2, the Next Generation Direct Liquid-Cooling Solutions, Aims to Reduce Data Center Power, Water, Noise, and Space, Saving on Electricity Cost by up to 40%, and Lowering TCO by up to 20%

ii. Data Center Building Block Solutions® (DCBBS) | Supermicro

iii. MLPerf 6.0: AMD Instinct™ MI355X GPUs Surpass 1M Tokens/Sec, Power New Workloads and Demonstrate Distributed Inference

iv. Optimizing Enterprise AI: Accelerating Time to Value with Supermicro Platforms with AMD

SUPER MICRO COMPUTER, INC.

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 on your requirements.

www.supermicro.com

Printed in USA ♻ Please Recycle 53_2026-ArchitectingScalableInference_AMD_R06 **MAY 2026**