ACCELERATE AI VALUE FROM DAY ZERO

Supermicro servers with AMD ROCm vLLM

Let's Get Moving!

THE AI CHALLENGE: CUSTOMERS NEED A CLEAR PATH TO VALUE

AI transformation begins with a bold vision – but turning that vision into reality is complex. As organizations invest in building platforms to unlock AI's full potential, many encounter roadblocks – especially those with deep investments in legacy infrastructure.

Choice and diversity are critical but so is getting to production and getting AI open for business.

CHOICE

Customers want maximum vendor choice across HW and SW. But many have also already made infrastructure choices, and everything must elegantly integrate to make AI work.

Many customers don't understand how to integrate all their choices.

COMPLEXITY

Between hardware accelerators (GPUs, DPUs), ML frameworks (PyTorch, TensorFlow), plus orchestration and deployment platforms, customers don't have the time to evaluate all pieces of the stack.

Even more customers are worried about switching or moving.

COSTS

All those moving pieces, from hardware and software to data center infrastructure, are large investment strategies. Customers want to feel confident in the bets they're making.

Nearly every customer is worried about high solution costs and expensive uncertainty.

THE SOLUTION: SIMPLIFIED DEVELOPMENT AND DEPLOYMENT VIA PRE-CONFIGURED, READY-TO-USE ENVIRONMENTS

Customers want to build AI – at their pace, on their path. **Supermicro servers with AMD ROCm vLLM Prebuilt Docker Image for MI350 Series GPUs** quickly move AI builders from proof of concept to production, no matter where they start.

Customers get to try AMD GPUs with zero stress and maximum AI capacity..

High-Performance Large Language Model (LLM) Inference

Enables rapid, efficient inference for advanced LLMs:

- Llama 3.1 (405B and 70B)
- Mixtral 8x22B
- DeepSeek-R1 (671B)

Delivers optimized throughput and low latency by leveraging AMD's hardware accelerations and containerized performance tuning, ideal for applications like chatbots and AI assistants.

Performance Benchmarking & Validation

Provides standardized, reproducible benchmarks for inference and training workloads.

Supplies prebuilt containers and scripts validated by AMD, ensuring users can easily verify system performance against expected ranges and troubleshoot deviations and optimize performance.

AI Application Development & Deployment

Offers out-of-the-box Docker environments accelerating the development and deployment of AI applications without complex setup.

Simplifies integration of LLMs into production workflows, enabling developers to quickly build, test, and launch interactive AI solutions optimized for AMD GPUs.

FOUR SIMPLE STEPS TO A BETTER WAY TO BUILD AI

Experiencing the MI350X is now easier than ever.

1 **Download** the prebuilt Docker image.

2 **Launch** the container with GPU access.

3 **Configure** the environment and run benchmarks or inference workloads.

4 **Validate and tune** performance as needed.

AS-8126GS-TNMR

THREE GIANT BENEFITS FOR TECHNOLOGY AND THE BUSINESS

No matter where their AI journey starts, Supermicro and AMD are ready to help you build for momentum that keeps you moving towards what's next.

1 MORE (AND BETTER) AI CHOICE. AVOID LOCK-IN, ACCELERATE GROWTH

Industry-leading AI performance: 288GB HBM3e, 8TB/s peak bandwidth.

Supports top LLMs with optimizations for high throughput, low latency.

2 BUILD, MIGRATE, AND MOVE AI WITH MAXIMUM SIMPLICITY

Fully prebuilt Docker image; no setup needed, start immediately.

Pre-configured, validated systems reduce setup time and guesswork.

3 MINIMIZE TCO ACROSS HW, SW, & INFRASTRUCTURE

Run larger models on fewer GPUs; lower hardware and power costs.

Containerization cuts overhead, accelerates deployment, enables rapid scaling.

Let's Get Moving!