SOLUTION BRIEF

OUT OF THE BOX AGENT-AI AT ANY SCALE

Innovative Solutions Deliver Optimized TCO with Options

Table of Contents

Executive Summary

As Generative AI (GenAI) becomes a mainstream technology, it is rapidly being adopted for use cases such as code generation, chatbots, marketing content creation, product design, and more. Business leaders, however, are eager to move beyond experimentation and see real impact, fast. Many organizations have limited tolerance for extended development cycles and are seeking immediate value, whether through enhanced customer experiences, innovative product offerings, streamlined operations, or optimized total cost of ownership (TCO). To meet these demands, Supermicro, AMD, and Piovation have jointly developed a solution that delivers on all.

Solution Overview

This joint solution addresses the intrinsic complexities associated with deploying large language models (LLMs) and AI agent frameworks by offering a pre-validated, turnkey infrastructure that reduces deployment overhead, enhances observability, and ensures sovereign data control. Designed to scale from compact on-premises clusters to large-scale, multi-tenant cloud environments, the architecture integrates Supermicro's turnkey rack-level systems, AMD Instinct™ GPUs, and Piovation's agentic AI platform—PioSphere—to deliver out-of-the-box agentic AI at any scale.

This out-of-the-box AI solution is a full-stack solution where an autonomous microservice chains LLM prompts, invokes domain-specific tools, and integrates seamlessly with your existing systems via REST, gRPC, or event streams running on the pre-validated Supermicro server powered by AMD Instinct GPUs. The Model Context Protocol (MCP) underpins every Agent's ability to interact with external tools in a modular, composable fashion. It governs how tools are registered, discovered, invoked, and composed dynamically at runtime, handling input/output serialization, maintaining execution context, and enforcing consistency across toolchains. MCP enables context-aware tool usage, making every Agent interoperable, auditable, and enterprise-ready from the start.

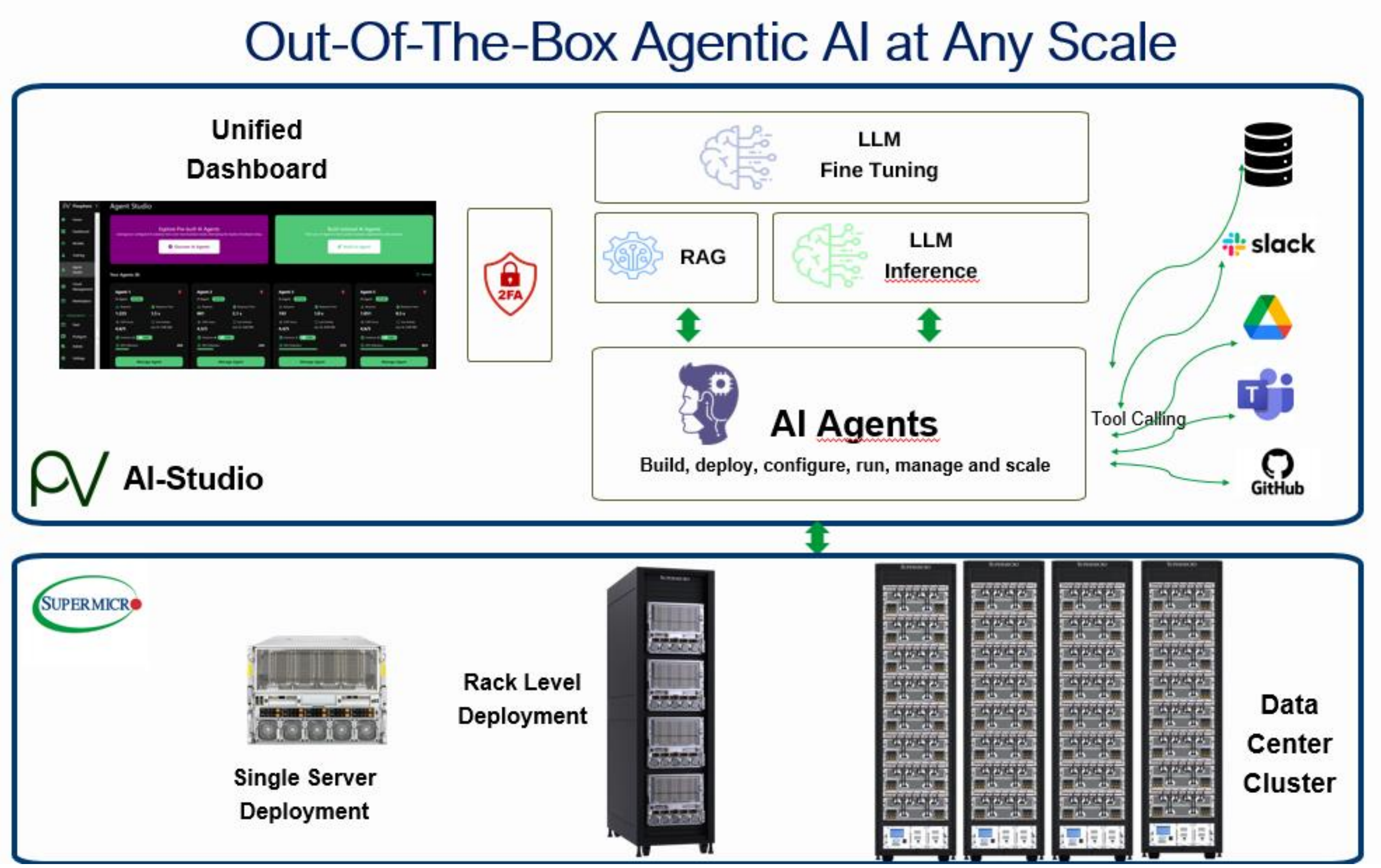


Figure 1 – Turn-Key Solution

Key Solution Components

Hardware – AMD Instinct™ GPU (AS -8126GS-TNMR - Front & Rear)

Figure 2 - Supermicro Air Cooled AS -8126GS-TNMR

Hardware – AMD Instinct™ GPU (AS -4126GS-NMR-LCC - Tray, Front & Real)

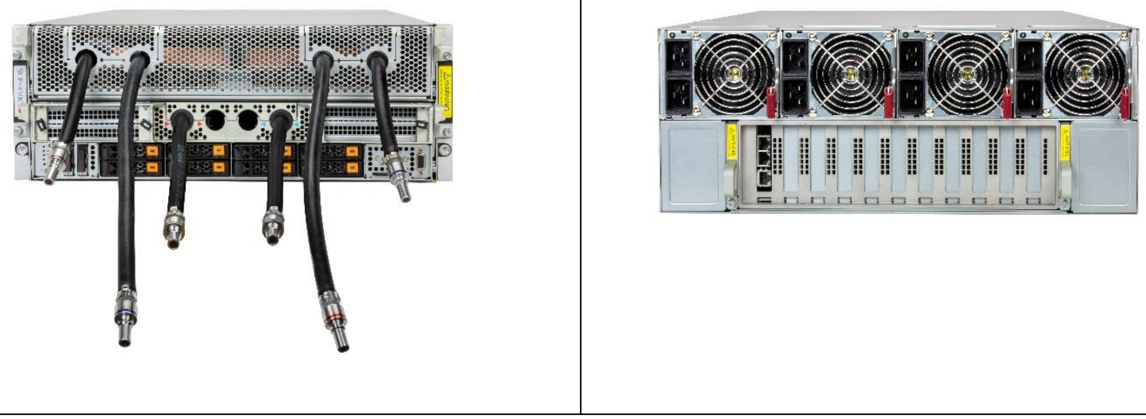
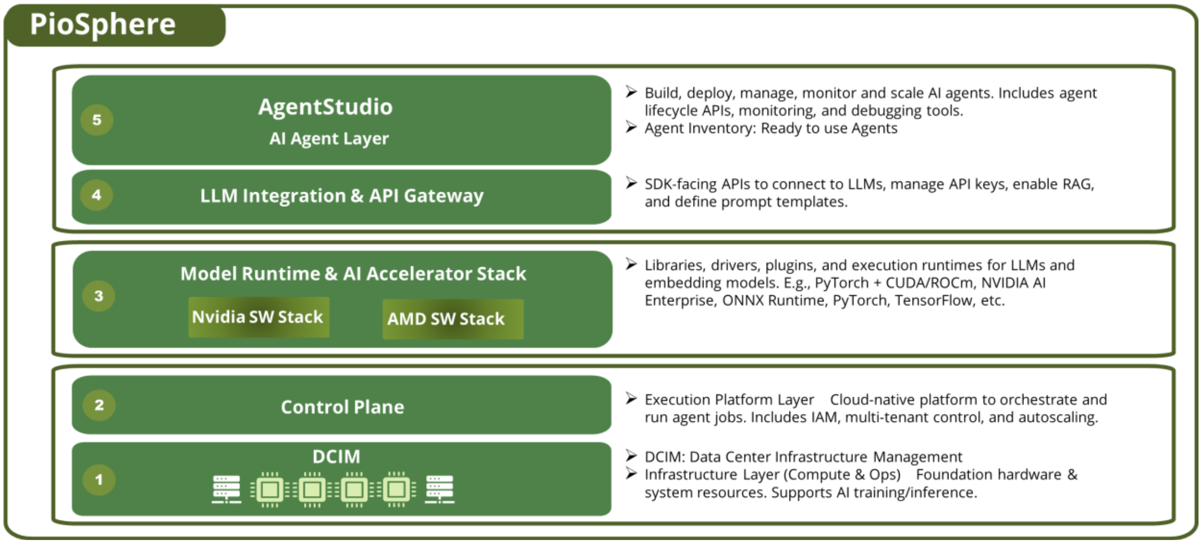


Figure 3 - Supermicro Liquid Cooled AS -4126GS-TNMR

Software - PioSphere



- ➤ Build, deploy, manage, monitor and scale AI agents. Includes agent lifecycle APIs, monitoring, and debugging tools.
- ➤ Agent Inventory: Ready to use Agents
- ➤ SDK-facing APIs to connect to LLMs, manage API keys, enable RAG, and define prompt templates.
- ➤ Libraries, drivers, plugins, and execution runtimes for LLMs and embedding models. E.g., PyTorch + CUDA/ROCm, NVIDIA AI Enterprise, ONNX Runtime, PyTorch, TensorFlow, etc.
- ➤ Execution Platform Layer Cloud-native platform to orchestrate and run agent jobs. Includes IAM, multi-tenant control, and autoscaling.
- ➤ DCIM: Data Center Infrastructure Management
- ➤ Infrastructure Layer (Compute & Ops) Foundation hardware & system resources. Supports AI training/inference.

Figure 4 – The Holistic AI Value Chain

Solution Capabilities and Differentiators

This section highlights the total solution Capabilities and Differentiators—breaks down the key technical and operational strengths that not only power this total solution but also position Supermicro as the technology leader.

- **Turnkey Deployment**

PioSphere COS has been pre-validated on the Supermicro MI350X platform powered by AMD Instinct GPUs, enabling rapid deployment without complex integration or setup.

- **Unified Operations Stack**

Eliminates fragmented AI tooling by providing a tightly integrated environment that combines:

  - Agent orchestration
  - Tool Use via MCP
  - Agent and LLM workflow management
  - Access control
  - Monitoring and governance

- **No-Code Agent Development**

PioSphere's AgentStudio™ empowers non-technical users to design, deploy, and iterate AI agents via a no-code interface, dramatically expanding who can contribute to AI-driven innovation.

- **Sovereign Data Compliance**

The solution includes built-in controls to support GDPR, HIPAA, and other compliance frameworks, with robust API-key governance and RBAC to maintain secure data boundaries.

- **Multi-Tenant Scalability**

Architected for enterprise scale, PioSphere COS enables secure, isolated environments for different business units or clients — all within a shared infrastructure footprint.

- **Integrated LLMOps and Agent Lifecycle Management**

The platform includes:

  - Integrate any LLM published on Hugging Face or Kaggle via one-click connectors, while still supporting proprietary in-house models.
  - Built-in RAG (Retrieval-Augmented Generation) pipelines and external Tool Calling via MCP.
  - Built-in pipelines for fine-tuning and continued pre-training across single or multi-GPU nodes.
  - Full agent lifecycle tools — from development and deployment to audit and optimization

- **Intelligent Autoscaling**

PioSphere supports dynamic autoscaling for both compute (e.g., GPU workloads) and agent environments based on real-time demand. This ensures optimal resource utilization, cost efficiency, and seamless performance during workload spikes.

Key Governance Features

- **Unified Agent Dashboard**: Centralized control for agent creation, deployment, monitoring, and auditing
- **Comprehensive Observability**: Real-time tracking of token usage, agent behavior, performance metrics, and response quality
- **Transparent Operations**: Full visibility into prompt logic, decision paths, data access patterns, with built-in testing tools
- **Policy Enforcement**: Enforce fine-grained operational rules, including integration restrictions, model access policies, and RAG source controls
- **Role-Based Access Control (RBAC)**: Multi-level access ensures clear separation of responsibilities for administrators, developers, and reviewers

The solution provides three principal deployment topologies, each catering to distinct operational scales and use cases:

Deployment Options for PioSphere COS

Deployment	Target Use	Specs & Capacity	Key Advantages
MiniStack	• SMBs • Pilots • Research • Edge	• Single Supermicro CPU control node • 128–256GB RAM • Local NVMe, SSD Storage • Up to 16 Supermicro MI350X GPUs	✓ Fast rollout, ✓ Compact ✓ Local AI ✓ No cloud needed ✓ Low hardware footprint
EdgeCluster	• Regulated sites • Branches • HA required	• 1–4 Supermicro CPU control nodes for HA • 256GB+ RAM/node • Local NVMe, SSD Storage • Up to 128 Supermicro MI350X GPUs	✓ High availability ✓ Edge-ready ✓ Regulatory fit ✓ On-prem scalability

Deployment	Target Use	Specs & Capacity	Key Advantages
Cloud Deployment	• CSPs • Large Orgs • AI Providers	• 12+ Supermicro CPU control nodes • Enterprise network/storage infrastructure • Unlimited Supermicro MI350X GPU scalability	✓ Multi-tenant cloud ✓ Modular extendable ✓ Compliance ✓ Integrated billing ✓ Simple Orchestration

Note: Both deployment options support rapid provisioning and are validated for Ubuntu-based environments, with setup times as short as 5 minutes for the base OS.

Learn More

Need more information? Reach out to a_plus_server_taskforce@supermicro.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

PIOVATION

Founded by Mazda Sabony in Munich, Piovation is at the forefront of democratizing AI technology while maintaining the highest standards of data sovereignty and compliance. Our mission is to make powerful AI accessible to every organization, regardless of size or technical expertise. With strategic partnerships, we deliver enterprise-grade solutions that are both powerful and practical.

Learn more at: www.piosphere.de

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