AMD ENTERPRISE AI SUITE: UNIFIED, OPEN, AND BUILT TO ACCELERATE AI AT SCALE

EXECUTIVE SUMMARY

AI adoption is accelerating across industries, but enterprises face significant hurdles in operationalizing models at scale. Fragmented tools, proprietary systems and governance challenges often hinder progress, making it difficult to achieve efficiency and scalability.

The AMD Enterprise AI Suite offers a game-changing solution: a unified, open, Kubernetes-native software stack that simplifies the entire AI lifecycle on AMD Instinct™ accelerators. By embracing openness and modularity, the suite empowers enterprises to innovate freely, scale efficiently and avoid vendor lock-in.

Purpose-built for enterprise platforms, infrastructure teams and AI practitioners, the suite delivers consistent performance, observability and governance across development, deployment and production operations.

AMD ENTERPRISE AI SUITE

In combination with the AMD GPU Operator, the AMD Enterprise AI Suite offers a comprehensive set of tools purpose-built to simplify and accelerate the deployment of enterprise AI solutions on AMD-based hardware. These tools, designed to address the complexities of modern AI workflows, enable seamless integration, optimized performance and scalable operations.

SCALE AI FASTER WITH THE AMD OPEN, UNIFIED PLATFORM

The AMD Enterprise AI Suite delivers a unified, open and performance-optimized platform that enables enterprises to securely and efficiently scale AI and GenAI from experimentation to production.

UNIFIED AI INFRASTRUCTURE
Simplifies deployment, reduces operational friction and accelerates time-to-value for enterprise AI initiatives.

OPEN AND MODULAR
Freedom to innovate, integrate and evolve without vendor lock-in. Protects investments, reduces dependency on proprietary ecosystems, and provides long-term flexibility.

PERFORMANCE-OPTIMIZED AND GOVERNED AT SCALE
Harness the full power of AMD Instinct™ GPUs – efficiently and securely. Maximizes infrastructure efficiency, enforces governance and compliance, and delivers measurable performance gains.

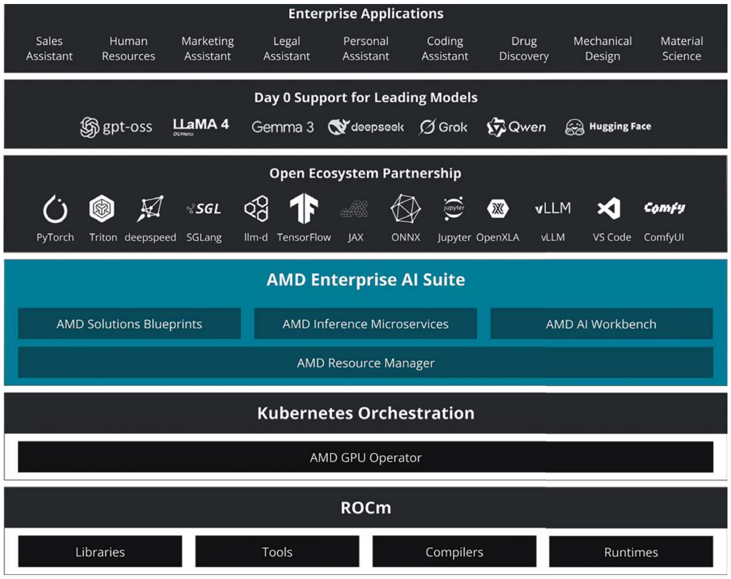
Let's explore each component in detail to see how they empower enterprises to innovate and grow with confidence:

AMD Inference Microservices (AIM)

A scalable, production-ready inference runtime platform for deploying and scaling AI models on AMD Instinct™ GPUs. AIMs abstracts away the complexities of configuring and serving AI models by providing pre-built, optimized containers that bundle the model, inference engine, and necessary configurations.

- **Optimized execution:** Hardware-tuned runtimes for AMD GPUs—including multi-GPU and multi-node configurations, with one-command deployment of validated LLM models.
- **Hardware-aware tuning:** Automatically selects optimal precision and parallelism settings based on GPU characteristics. Provides local caching and advanced memory management.
- **Flexible engines:** Supports open-source inference engines like vLLM and SGLang.
- **Standard interfaces:** Provides OpenAI-compatible APIs for fast integration with existing applications.
- **Scalable orchestration:** Uses Kubernetes and KServe for intelligent routing, scheduling, and autoscaling.
- **Built-in observability:** Prometheus and integrated metrics and health checks for monitoring and lifecycle management.

FULL STACK SOLUTION



Enterprise Applications

Sales Assistant | Human Resources | Marketing Assistant | Legal Assistant | Personal Assistant | Coding Assistant | Drug Discovery | Mechanical Design | Material Science

Day 0 Support for Leading Models

gpt-oss | LLaMA 4 | Gemma 3 | deepseek | Grok | Qwen | Hugging Face

Open Ecosystem Partnership

PyTorch | Triton | deepspeed | SGLang | llm-d | TensorFlow | JAX | ONNX | Jupyter | OpenXLA | vLLM | VS Code | ComfyUI

AMD Enterprise AI Suite

AMD Solutions Blueprints | AMD Inference Microservices | AMD AI Workbench

AMD Resource Manager

Kubernetes Orchestration

AMD GPU Operator

ROCm

Libraries | Tools | Compilers | Runtimes

AMD Solution Blueprints

Ready-to-use reference architectures for generative and agentic AI deployments integrate AIMs with data pipelines, retrieval systems and orchestration frameworks. These architectures provide a blueprint for enterprises to accelerate AI innovation, streamline workflows and deploy scalable, production-ready solutions with ease.

- **Agentic Translation Pipeline:** document-aware, instruction-controlled translation.
- **LLM Chat Sandbox:** prompt experimentation with plug-in retrieval modules.
- **Local LLM IDE Assistant:** developer assistance with code refactoring and contextual search.
- **Financial Stock Intelligence:** a proof-of-concept showcasing how AMD hardware and software can combine LLMs and real-time data and news to provide comprehensive financial insights.
- **AutoGen Studio Platform:** configurable multi-agent workflows with real-time debugging and AIMs integration.
- **Continue.dev Coding Assistant:** An AI pair programmer that integrates with code editors to suggest and fix code using a local LLM.

AMD Resource Manager

A centralized control plane designed for multi-tenant GPU infrastructure enables administrators to efficiently manage clusters, users and workload policies. By streamlining resource allocation and governance, it provides fair usage, enhances operational efficiency and simplifies the complexities of managing large-scale AI environments.

- **Cluster Management:** Provides a single dashboard to view and manage on-premise and cloud-based AMD GPU clusters.
- **Workload Orchestration:** Intelligently schedules AI workloads (like training, inference) across clusters, optimizing GPU utilization and reducing idle time.
- **User & Project Management:** Sets up organizations, teams, and projects, assigning quotas and access rights to provide fair resource distribution.
- **Resource Allocation:** Maps users and groups to compute, data and storage resources, managing usage limits (quotas).
- **Monitoring & Visibility:** Offers real-time insights into cluster health, utilization and capacity.
- **Security:** Manages secrets (API keys, credentials) for secure access to data stores.

AMD AI Workbench

A unified developer interface that simplifies AIMs deployments, accelerates model building and streamlines AI workflow orchestration—empowering teams to innovate faster and operate more efficiently.

- **AIMs Catalog:** deploy and manage inference services, share APIs with internal users.
- **AI Workspaces:** pre-configured GPU development environments (VS Code, JupyterLab).
- **Fine-Tuning:** UI-driven or CLI-driven workflows with expert-tuned hyperparameter “recipes”.
- **Chat & Compare:** side-by-side evaluation of models and AIMs deployments.
- **GPU-as-a-Service:** self-service provisioning with governance and quotas.

- **AMD Enterprise AI Suite Documentation**
- **AMD Inference Microservices (AIMs)**
- **AMD Solution Blueprints Catalog**
- **AMD Resource Manager and AI Workbench**

FOR MORE INFORMATION, VISIT AMD ENTERPRISE AI SUITE

PID: 254162850