Altos aiWorks

Smarter Resource Management for AI & LLM Computing

Why Altos aiWorks

Comprehensive Software - Hardware Integration Platform

Adaptive and Scalable Computing Resource Allocation

Streamlined Deployment of Development Environments

Secure Remote Access via Web-Based Management Console

Intelligent Resource Management

- Boost Team Productivity Instantly: Clear, role-based access ensures Admins, Team Leaders, and Developers work efficiently with zero friction.
- Maximize Every GPU Investment: Real-time cost and usage dashboards give full visibility into GPU spending and utilization for smarter budgeting.
- 24/7 Performance Optimization: MIG/MPS scheduling delivers multi-task parallelization and dynamic GPU allocation to keep workloads running at peak efficiency.

Rapid Development Environment

- Build AI Faster: Pre-configured tools - JupyterLab, TensorFlow, PyTorch, and Ollama - help teams start developing in minutes, not days.
- Scale Without Complexity: Native support for Docker and Kubernetes enables highly scalable, reproducible environments with minimal overhead.
- One Platform for All AI Workloads: Seamlessly supports diverse frameworks, accelerating research, prototyping, and enterprise deployment.

One-Click AI Applications

- Immediate Access to Leading LLMs: Supports OpenAI, DeepSeek, LLaMA, and other models so enterprises can instantly deploy private AI assistants and knowledge systems.
- Powerful Image Generation: Integrated Stable Diffusion (txt2img, img2img, inpaint) delivers high-quality visual outputs for creative and automation workflows.
- Break Language Barriers: Built-in Whisper provides fast, multilingual transcription and subtitling for global operations.
- AI for Everyone: A no-code, visual interface lowers the barrier to AI adoption across all departments.

Fully Private Deployment

- Your Data Never Leaves Your Premises: All computation and storage stay on-site - ideal for healthcare, finance, government, and sensitive environments.
- Enterprise-Grade Security: TLS/SSL encryption and One-Time Password (OTP)-based authentication ensure end-to-end protection against unauthorized access.
- Compliance from Day One: Meets GDPR, HIPAA, and other enterprise security standards to simplify audits and regulatory approval.

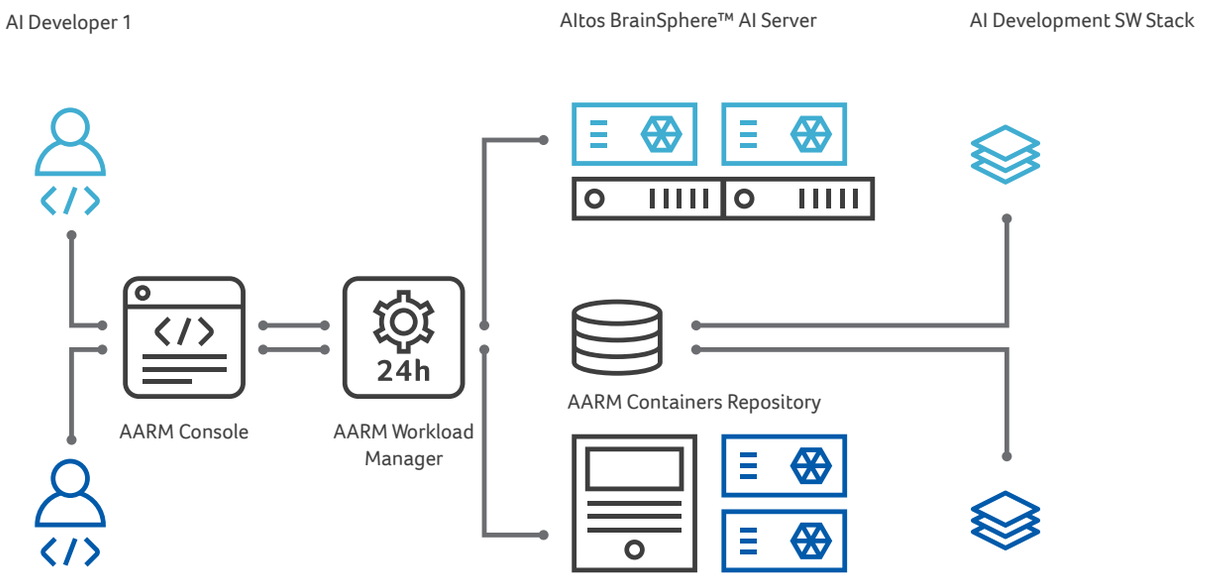
On-Premises vs. Cloud Deployment

Category	Altos aiWorks On-Premises	Cloud Service
Security & Compliance	Your data stays 100% on-site. Enterprise-grade TLS/SSL + OTP, full governance control.	Requires data upload and external storage; security and compliance depend on the vendor, not your organization.
Cost Model	One-time investment with predictable, lower long-term TCO. No surprise bills; ideal for continuous, heavy GPU workloads.	Pay-as-you-go but becomes significantly more expensive over time, especially for AI training or high-usage teams.
Performance & Latency	Local execution delivers the fastest possible performance, ultra-low latency, and guaranteed 24×7 stability.	Performance varies with network conditions; potential lag and instability for large models or real-time workloads.
Flexibility & Control	Full ownership of all compute resources. Supports Bring Your Own Model (BYOM), hybrid cloud, and complete customization to meet enterprise workflows.	Quick setup but limited control and restricted customization due to vendor-locked environments.

Altos aiWorks

Smarter Resource Management for AI & LLM Computing

Start Your Journey in AI/Digital Transformation



Industry Use Cases

Altos aiWorks integrates NVIDIA Triton, Dynamo, and NIM to accelerate generative AI and inference. With disaggregated serving, it supports parallel multi-model processing. The GPU Planner optimizes resource allocation, while the Smart Router intelligently directs workloads for maximum efficiency and performance.

Healthcare

- Medical imaging with MONAI framework
- Voice-to-text clinical documentation
- Federated learning via NVIDIA FLARE for cross-institution collaboration
- HIPAA/GDPR compliance for sensitive data handling

Education & Research

- Centralized GPU management for classrooms and labs
- One-click JupyterLab and AI course environments
- Role-based access for professors, TAs, and students
- Support for Bring-Your-Own-Model (BYOM) to accelerate collaboration

Enterprise & Public Sector

- Private LLM assistants for internal knowledge search
- RAG-based document retrieval and regulation lookup
- Intelligent contact centers with voice transcription and emotion analysis
- Automated document summarization and translation for digital governance