End-to-End On-Prem Solutions

Design to Deployment Services

For over 40 years, AMAX has been at the forefront of engineering advanced computing solutions, specializing in AI infrastructure, liquid cooling, and advanced manufacturing. Our expertise in designing, building, and deploying scalable compute environments gives businesses the power to take control of their AI and HPC workloads with unmatched efficiency, security, and scalability.

The Shift to On-Prem

As AI training and inference needs accelerate, businesses are increasingly moving away from cloud-based infrastructure. Rising costs, data security concerns, and the need for predictable performance are driving a shift toward on-prem solutions, enabling organizations to scale AI workloads more efficiently while optimizing long-term investments.

On-Prem Solution Options

Turnkey Data Center

- Fully integrated compute, cooling, and cluster monitoring
- Designed for new builds or retrofitting existing environments
- Full-service deployment by AMAX, from design to implementation

Colocation

- Deploy with your chosen colocation provider or one of our trusted partners
- Infrastructure setup, optimization, and ongoing support
- Full-service deployment by AMAX

AMAX On-Site Hosting

- Rapid bring-up in our facilities for immediate compute access
- Ideal for bridging gaps while permanent infrastructure is completed
- Fully managed environment with enterprise-grade reliability

GPU Cluster Design

AMAX designs GPU cluster architectures based on NVIDIA HGX™ and NVIDIA DGX™ reference designs, ensuring optimized performance for AI and HPC workloads. Whether scaling a single-node system or deploying an NVIDIA BasePOD™ or DGX SuperPOD™, we build infrastructure that meets your compute, networking, and storage demands.

NVIDIA DGX SuperPOD with B200 Architecture

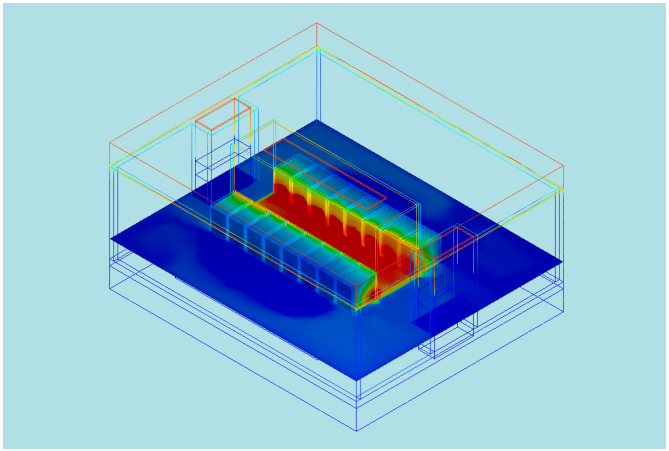
Building AI Infrastructure On-Prem

Moving AI workloads on-prem unlocks greater control, lower latency, and long-term cost efficiency. AMAX customers transitioning off the cloud often see significant savings within months by eliminating usage-based fees and optimizing resource allocation. With dedicated infrastructure, AI models run with predictable performance and scalability.

Optimizing Compute, Networking, and Storage

AI workloads depend on high-performance compute, efficient storage, and networking designed for scale. AMAX builds GPU-optimized systems, integrates high-speed storage for rapid data access, and designs low-latency network topologies to eliminate bottlenecks. Every system is validated to ensure sustained AI performance.

AceleMax® POD with NVIDIA HGX B200



Thermal Assessment of Data Center

Power, Space, and Cooling Considerations

AI clusters push the limits of power and thermal management, consuming significantly more energy than traditional workloads and generating more heat. AMAX optimizes power distribution for high-density deployments, implements liquid cooling to maintain GPU efficiency, and optimizes rack design to maximize scalability within existing data centers.

Expert Deployment, Installation, and Management

AMAX manages the full deployment process, ensuring AI infrastructure is installed, configured, and optimized for production. Our team handles rack integration, power distribution, and network cabling, bringing systems online, burn-in, and full validation. We oversee every stage, coordinating logistics, managing timelines, and ensuring a smooth transition from cloud to on-prem with minimal disruption and maximum efficiency.

Go On-Prem with AMAX

AMAX | 1565 Reliance Way, Fremont, CA 94539 | 1 (800) 800-6328 | www.amax.com | info@amax.com |
v1-2025