Scaling On-Premises Hardware Solutions for Generative AI

Solutions for the Expanding GPU Market

Every company will be an AI company in the future, but building the compute infrastructure to make this possible can be complex. Generative AI (Gen AI) requires extensive GPU power to effectively train massive Large Language Models (LLMs) that start at billions of parameters in size.

In partnership with **AMAX Engineering**, a growing Gen AI software company recently transitioned their cloud-based operations to customized on-premises solutions to scale their AI model training. This shift drastically reduced their Total Cost of Ownership (TCO) and allowed the exclusive compute availability they needed to train their proprietary model.

AMAX On-Premises Vs. Cloud

Choosing between cloud services and on-premises hardware depends on specific business needs. Cloud services offer flexible scaling and low initial costs, making it suitable for projects with fluctuating demands. On-premises hardware provides significantly better long-term cost efficiency, unrestricted access, and security. It allows for specialized builds optimized for high-volume AI tasks and offers enhanced security, crucial for managing sensitive data.

For our customer, who consistently trained, retrained and fine-tuned their models, moving to on-premises hardware made more sense than continuing to pay expensive cloud fees. Our solution allowed them to efficiently meet their intensive, ongoing AI demands.

Customer Timeline

Aug 2022 — Workstation

Jan 2023 — NVIDIA H100 GPU Server

Apr 2023 — 8x NVIDIA HGX™ H100 Cluster

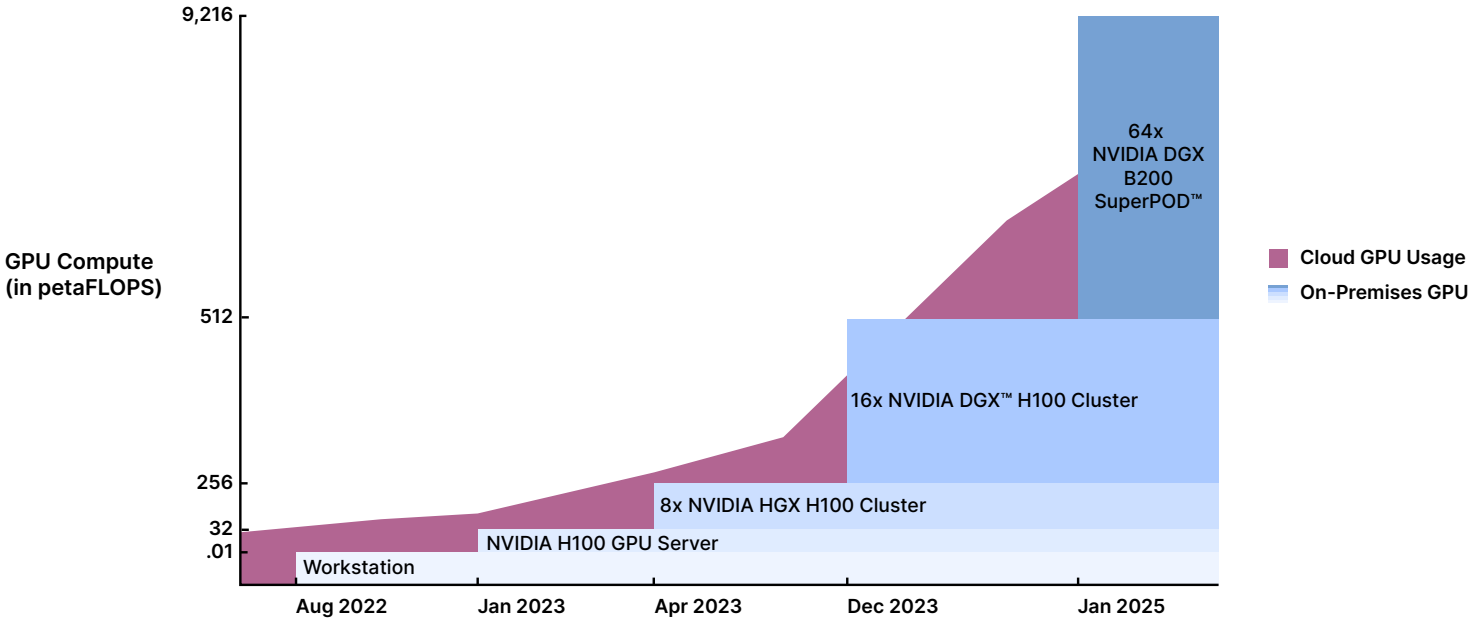
Dec 2023 — 16x NVIDIA DGX™ H100 Cluster

Jan 2025 — 64x NVIDIA DGX™ B200 SuperPOD

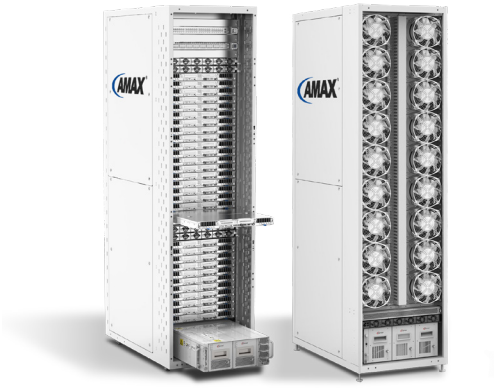
AMAX | 1565 Reliance Way, Fremont, CA 94539 | 1 (408) 505-45988 | www.amax.com | info@amax.com

AMAX Deployment Timeline

Cloud Costs Can Be 5X Higher Than On-Premises Hardware



Startup Growth Pattern with AI Hardware



IntelliRack A45 + Sidecar

Moving to the Data Center - Price, Performance, and Access

Transitioning from cloud to on-prem systems improves performance, cost management, and gives exclusive access to hardware. Our customer began with minimal hardware and supplemented GPU demands with cloud services while testing their applications. This approach is very expedient when you are rapidly scaling, but once you have a clear idea of your trajectory, it is much more affordable to build extra capacity to meet demand during growth periods.

After increasing their investment in their GPU Cluster, they briefly experienced an oversupply of GPU resources. However, as their operations expanded, this surplus was quickly absorbed, bringing their GPU capacity back to just meeting demand. This scenario highlights the importance of strategic planning to ensure a steady balance of GPU resources, optimizing availability and cost efficiency during the growth process.

AMAX Complete AI Solutions

As your company looks to scale its Gen AI capabilities, consider the comprehensive technical expertise that **AMAX Engineering** offers. Our services encompass data center layout design, cluster architecture, network topology design, cluster bring-up, performance tuning, engineering and liquid or air-cooled facility retrofit solutions. We also provide co-location services for site hosting. Partnering with AMAX ensures access to top-tier infrastructure solutions that are essential for advanced AI deployments.

Start Planning your Next AI Deployment with AMAX

AMAX | 1565 Reliance Way, Fremont, CA 94539 | 1 (408) 505-45988 | www.amax.com | info@amax.com