Secure Knowledge Synthesis

We helped GrowthEngine AI set a new standard for intelligent GPU scaling.

Clients:
GrowthEngine AI
U.S. Air Force

Sector:
Defense / Education

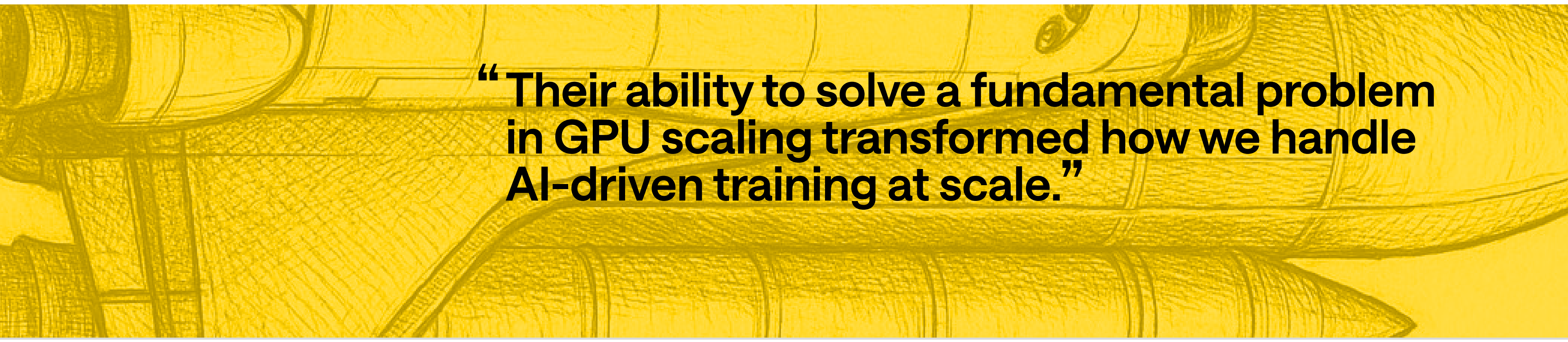
Technologies:
GPUs
Kubernetes
AWS
Go
PyTorch
BERT
Hugging Face
Transformers
High-Performance Computing (HPC)

Resources:
https://growthengineai-private.netlify.app/

Challenge: The U.S. Air Force required a knowledge synthesis tool capable of transforming vast amounts of enterprise knowledge into structured educational materials. Due to the sensitive nature of the data, off-the-shelf AI models were not an option, necessitating the deployment of secure, custom-built models. Additionally, the project faced a major infrastructure challenge—highly burstable traffic during training workshops. A single weekend session could require an immense surge in GPU resources, which then needed to be scaled down immediately after, making conventional cloud scaling solutions inadequate due to the complexity and cost of GPU-based workloads.

Solution: We developed a custom Kubernetes-based GPU controller in Go, designed specifically for dynamic GPU scaling—an unsolved problem at the time. Unlike traditional scaling methods that adjust CPU and RAM allocations, GPU scaling is significantly more complex due to the need for intelligent model allocation and memory management. Our system not only scaled GPU resources dynamically but also optimized VRAM usage by intelligently loading and unloading models based on real-time demand. This required an AI-driven orchestration layer capable of predicting load, preemptively managing deployments, and efficiently shuffling models across available hardware. By leveraging cutting-edge machine learning techniques, we created a system that not only met the Air Force's strict security and performance requirements but also drastically reduced unnecessary GPU overhead costs.

Result: Our scalable, AI-driven GPU orchestra-tion system enabled the Air Force to run highly efficient knowledge synthesis operations without over-provisioning expensive GPU resources. GrowthEngine AI continues to refine the system for broader military and enterprise applications, setting a new standard for intelligent GPU scaling.



> **"Their ability to solve a fundamental problem in GPU scaling transformed how we handle AI-driven training at scale."**