ECI

CASE STUDY

CLOUD HPC CLUSTER FOR AI/ML RESEARCH AT A FAANG COMPANY

Building a scalable, secure, and production-ready HPC environment for next-generation MLOps innovation.

VERTICAL

AI / Machine Learning / HPC

SERVICES

- Cloud HPC & MLOps Enablement
- Multi-Cloud GPU Infrastructure
- Secure Access & Resource Management

WHY ECI

- Expertise in AWS ParallelCluster and Azure Cycle Cloud
- Proven at-scale AI/ML and GPU optimization
- Trusted by global technology leaders for secure HPC innovation

CHALLENGE

A leading FAANG company needed to scale its AI/ML research infrastructure to support hundreds of researchers and thousands of GPUs. On-prem HPC clusters offered control but were costly, slow to expand, and quickly outdated. Cloud HPC provided flexibility but lacked the performance tuning and internal integrations required for enterprise-scale workloads. The company sought a multi-cloud HPC solution that delivered security, scalability, and seamless integration with existing systems.

SOLUTION

ECI deployed a customized HPC Slurm cluster on AWS using ParallelCluster as the base layer—enhanced with secure access, 2FA, Unix user management, and multi-tenant support. The solution integrated S3 pipelines, multiple FSx for Lustre file systems, persistent home directories, and advanced monitoring and accounting to create a fully production-ready environment.

Later, an Azure HPC cluster was added via Cycle Cloud, forming a unified, multi-cloud platform optimized for GPU-intensive research and MLOps workloads.

RESULT

The environment scaled to support 500+ researchers, 20+ clusters, and 5,000+ GPUs across 5+ accounts, managing petabytes of data on S3 and FSx. Researchers gained faster access, improved collaboration, and reduced experiment cycle times. The solution accelerated innovation—so much so that AWS later incorporated concepts from this deployment into ParallelCluster itself.