Avesha

Innovating the Future by automating infrastructure, workflow and GPU/CPU management

Elastic GPU Service

Introduction

EGS (Elastic GPU Service) optimizes GPU infrastructure for AI engineers by providing **usage optimization, observability with real-time clarity, smart orchestration and automation.** It redefines how organizations harness the power of GPU intensive workloads. EGS automation unlocks unparalleled efficiency, scalability, and cost-effectiveness—all tailored for AI, ML, and high-performance computing. EGS usage optimization achieves up to 45% more node allocations and reducing GPU wait times by 32%. This allows engineers to manage both training and inferencing tasks effectively, ensuring scalable and high-performing AI operations.

Key Features

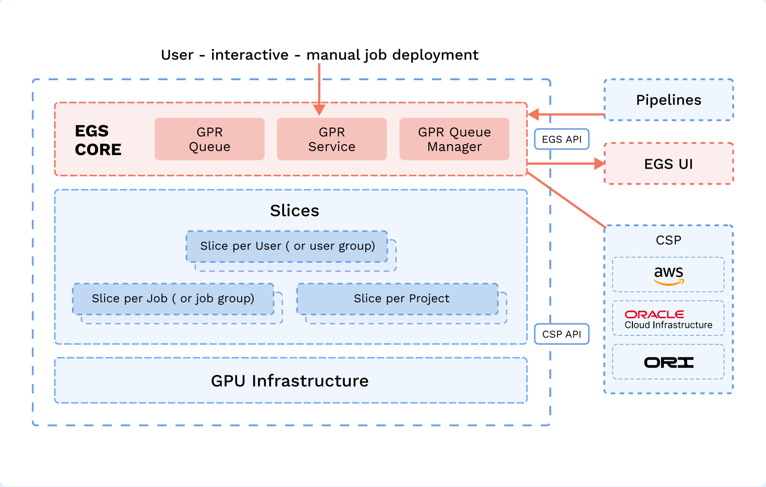
1. Dynamic Resource Allocation
2. Spot Instance Utilization
3. Cost Transparency across workflows, teams, projects and clusters.
4. Predictive Insights
5. Dynamic Scaling
6. Workflow Optimization for DAG pipelines

Benefits

1. Maximizes GPU usage and efficiency
2. Shortens GPU access waiting times
3. Lowers costs by optimizing GPU usage
4. Scalable and flexible for future needs
5. Ensures reliable and predictable GPU allocation

How It Works

EGS uses namespaces for multi-tenancy and data isolation, enabling secure, efficient resource management across teams or projects. It supports automatic GPU provisioning, which increases node allocations by 44% more workload minutes. EGS integrates with MLOps tools like Run and Volcano to optimize workload distribution. Real-time monitoring provides visibility into key metrics, enabling proactive management of GPU resources. With a dual-layered approach combining DAG pipelines and dynamic resource allocation, EGS efficiently manages resources, aligning with current demands and future growth.


User - interactive - manual job deployment
EGS CORE
GPR Queue
GPR Service
GPR Queue Manager
Pipelines
EGS API
EGS UI
Slices
Slice per User (or user group)
Slice per Job (or job group)
Slice per Project
CSP API
CSP
aws
ORACLE Cloud Infrastructure
ORI
GPU Infrastructure

Conclusion

With EGS, you're not just managing GPU resources—you're transforming them into a strategic advantage. By combining usage optimization, observability with real-time clarity, smart orchestration and intelligent automation, EGS empowers your organization to innovate at scale while keeping costs under control. Whether you're training the next LLM, running complex simulations, or managing multi-agent systems, EGS ensures you stay ahead of the curve.