Avesha

Elastic GPU Service (EGS)

Smart Orchestration for AI Infrastructure

The AI Infrastructure Challenge

Avesha

1. AI workloads are constrained by legacy proprietary frameworks.
2. GPU orchestration inefficiencies lead to high costs and resource waste.
3. Manual allocation lacks adaptability, leading to delays.
4. Fragmented monitoring makes optimization difficult.

The Need?

A seamless, automated, and cost-efficient AI workload orchestration solution.

Introducing Elastic GPU Service (EGS)

Avesha

Kubernetes-native AI-powered orchestration

Dynamic, just-in-time GPU allocation

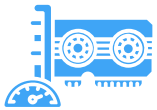
Real-time observability & cost transparency

Seamless multi-cloud & hybrid cloud deployment

Outcome: *Higher efficiency, lower costs, and optimized AI workloads*

Key Capabilities of EGS

Avesha



Smart Orchestration:

Automates GPU allocation based on workloads

Cost Optimization:

Visibility into cost breakdown per DAG, team

Real-time Monitoring:

GPU temperature, utilization, memory

Scalability & Elasticity:

Handles AI training, inference, and real-time processing

Security & Compliance:

Role-based access and secure multi-tenancy

Market Adoption & Trends

Avesha

1 AI compute demand growing exponentially across industries

2 Enterprises struggle with cost, availability, and efficiency of GPUs

EGS - Core Use Cases

Avesha

Automated GPU Allocation

- Eliminates manual GPU provisioning
- SDK/API-driven workload-aware scheduling

Real-Time Observability

- Tracks GPU temperature, utilization, memory & costs
- Proactive alerts for performance anomalies

Cost Optimization & Chargeback

- Per-team, per-project cost transparency
- Reduces idle GPU wastage by up to 40%

Multi-Tenant Secure GPU Sharing

- RBAC ensures teams share GPUs without interference

EGS vs. Traditional GPU Management

Avesha

Feature	Traditional GPU Management	EGS
GPU Allocation	✗ Manual	✓ Fully Automated
Real-time Observability	✗ Limited	✓ Full-stack GPU Insights
Cost Transparency	✗ Opaque	✓ Granular Cost Breakdown
Multi-cloud & Hybrid	✗ Difficult	✓ Seamless Integration
AI-driven Scaling	✗ None	✓ Predictive Scaling

EGS Delivers: *Faster, Smarter and More Cost-Effective GPU Management*

The AI Infrastructure Challenge

Avesha

Real-World Benefits

1. Up to 60% cloud cost savings (Finvi use case)
2. 40% reduction in idle GPU waste (benchmarking results)
3. Faster model training & inference with optimized scheduling
4. Improved developer productivity by reducing manual GPU Ops from days to minutes

Competitive Positioning

Avesha

EGS vs. CoreWeave & Traditional GPU Providers

- More granular workload scheduling with DAG-based orchestration
- Predictive scaling with real-time resource optimization
- Fractional GPU sharing to maximize utilization
- Works across multi-cloud, hybrid, and on-prem

Next Steps

How to Get Started with EGS

Avesha

Pilot Deployment

Identify AI workloads for EGS optimization

Free Tier Trial

Try EGS in a single-cluster environment

https://avesha.io/egs-registration

ROI Analysis

Compare cost savings and performance improvements

Thank You, Questions?

Backup