Avesha

Elastic GPU Service (EGS)

Making MLOPs Easier

EGS integrates observability, orchestration, and cost optimization for GPUs, seamlessly combining these capabilities through automation to deliver significant business value.

Challenges in GPU Resource Management

1 Inefficiency

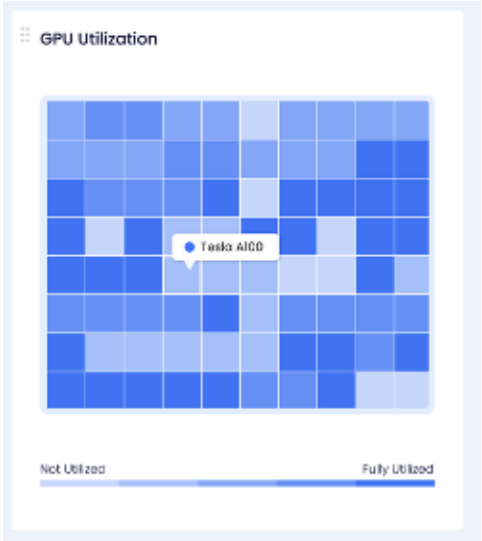
Optimizing GPU and CPU usage remains difficult. Sharing GPUs and managing chargebacks is hard.

2 Delays

Resource mismatches cause idle GPUs, delays, and inefficiencies.

3 Manual

Manual GPU allocation lacks adaptability and precision. Lack of visibility and automation makes allocation very difficult.



Optimized

Inefficient

EGS – Enterprise

1 **Scalability - Usage Optimization**

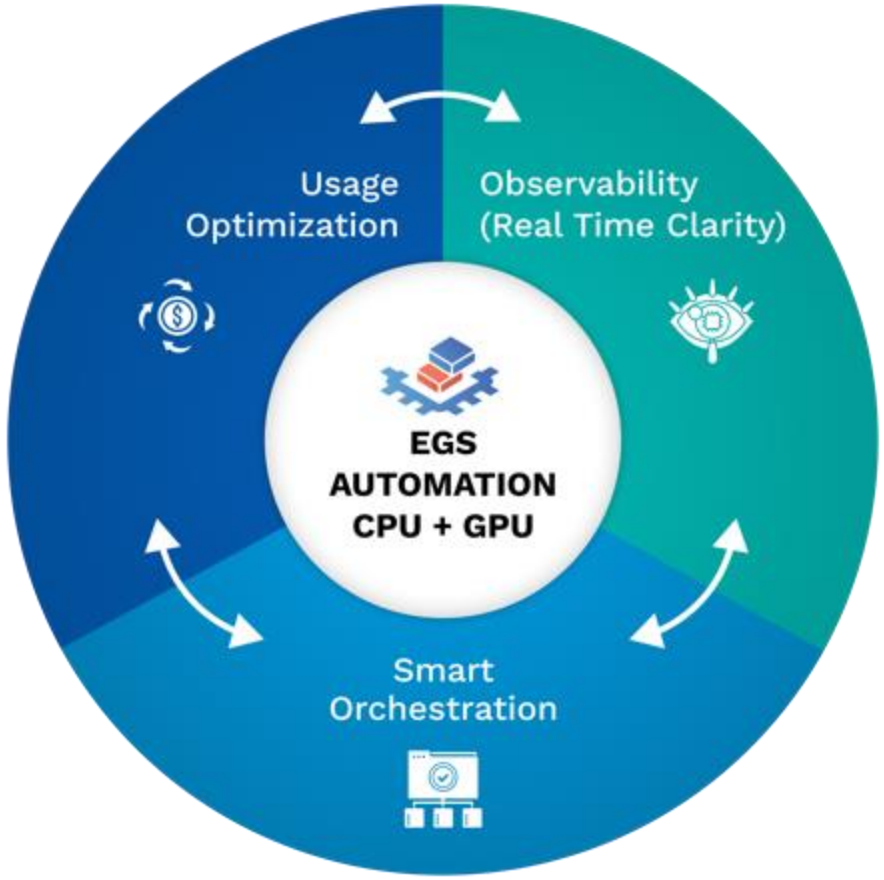
Cost efficiency is woven into every layer of EGS, making it a game-changer for organizations balancing innovation with budgets.

2 **Observability Real Time Clarity (RTC) at Every Layer**

EGS provide a 360-degree view of your GPU utilization, workload performance, and costs.

3 **Smart Orchestration & Isolation**

Seamlessly coordinating workflows and dynamically allocating resources by teams.



EGS – Innovations

1 Automation

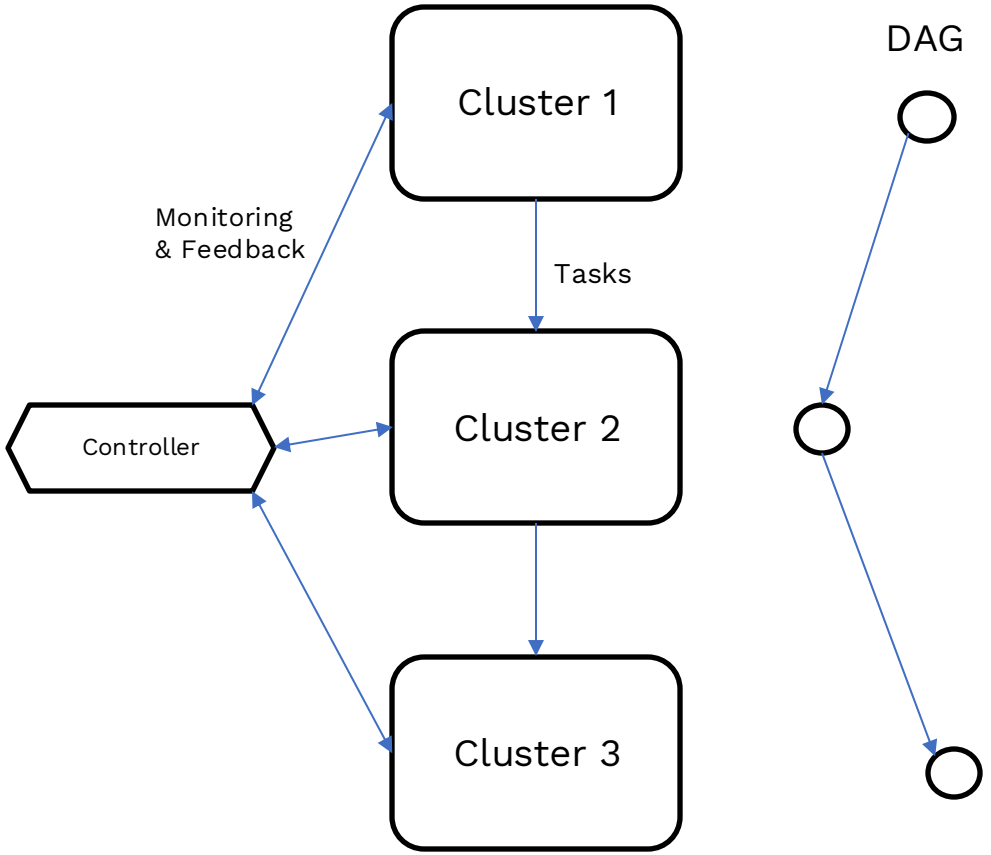
- Orchestrates GPU resources dynamically across pipelines.
- Leverages DAG pipelines for workload efficiency.
- Self-healing -- auto-remediation & auto-checkpointing

2 Predictive Allocation

- Historical data patterns inform GPU allocations.
- Ensures balanced resource utilization.

3 Multicluster Orchestration

- Dynamic Resource Allocation (DRA).
- Automates tasks across clusters -- GPUs & CPUs



Key Benefits of EGS

Smart Orchestration

- **Dynamic Scaling:** Our system provisions GPUs precisely when and where they're needed and decommissions them when they're not.
- **Security:** RBAC ensures secure multi-tenant GPU sharing. Multiple teams can share resources without fear of data privacy violation.
- **Workflow Optimization:** Whether it's training an LLM or running real-time inference, EGS ensures every task in DAG pipelines are optimized for speed and reliability.
- ***Outcome:*** *Run large-scale, distributed workflows without worrying about resource bottlenecks or overspending.*

EGS Automation

- **Auto-Provisioning:** Automatically allocate GPUs based on workload demands, ensuring zero idle resources.
- **Self-Healing Systems:** Detect and resolve workflow failures without manual intervention, keeping your operations running smoothly.
- **Cost-Aware Automation:** Schedule non-critical tasks during off-peak hours or prioritize cost-efficient resources like spot instances.
- ***Outcome:*** *Save time, reduce errors, and scale operations effortlessly with automation that works around the clock*

Key Benefits of EGS

Usage Optimization

- **Predictive Cost Management:** Leverage data-driven insights to forecast resource needs and align costs with usage patterns.
- **Spot Instance Utilization**: Prioritize cost-effective spot GPUs for batch jobs and non-critical tasks, cutting compute costs significantly.
- **Dynamic Resource Allocation**: Automatic GPU allocation based with EGS' time-slice feature. Unused GPU capacity automatically reallocated for efficient use.
- ***Outcome**: Achieve up to 40% savings in GPU costs while maintaining peak performance for your workloads*

Observability RTC

- **Monitoring:** Get a 360-degree view of your GPU utilization, workload performance, and costs, empowering you to make informed decisions in real time.
- **Proactive Insights:** Get instant alerts on anomalies like underutilized GPUs or sudden cost spikes.
- **Cost Transparency**: Dashboards break down resource usage and spending across workflows, teams, projects and clusters.
- ***Outcome:** Eliminate inefficiencies before they become bottlenecks, ensuring every GPU cycle delivers value.*

Comparison to Coreweave

	Coreweave	Avesha + Infra Provider
SW-Defined Orchestration	Automation, Preemption, Fine-grained Sharing	Automation, Preemption, Fine-grained Sharing
API, SDK, Integration	Tensorflow, Pytorch, etc	Frameworks and MLOPs tools
Elastic Scaling	Spin up 1000s of GPUs	Subject to Infra Provider Network
Intelligent Job Placement	Mapping of jobs	Predictive
Distributed Storage	VAST integration	Predictive Pre-fetch (coming soon) for distributed FS
Efficient execution	Diverse workloads	Diverse workloads with DAG-awareness

EGS – Free Tier (Single Cluster)

1 Team Separation

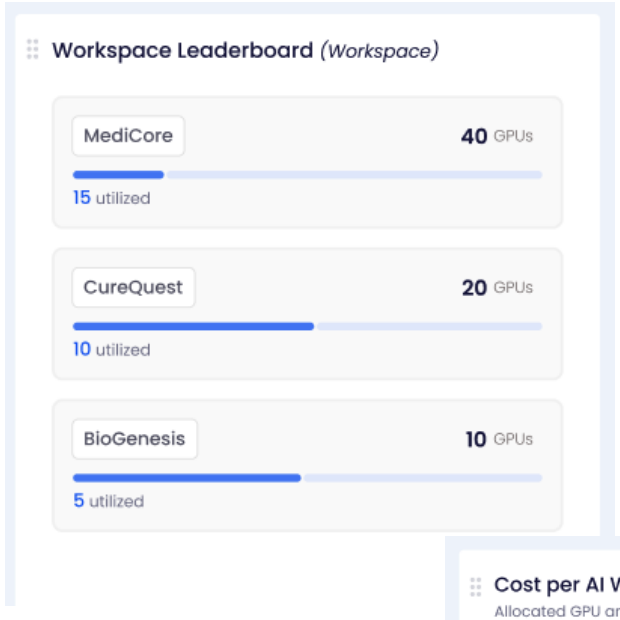
- Isolated Data & GPU resources for different teams.

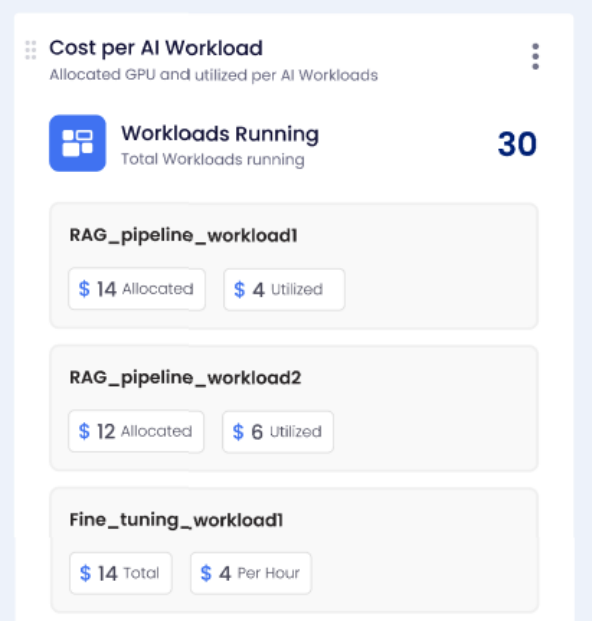
2 Ease of Operations

- Centralized management ensures fair resource allocation.

3 Virtual GPU Allocation

- Teams can specify virtual GPU requirements dynamically.
- Hyperparameters align with workload needs.





EGS – Premium Features

Priority Preemption

- Intelligent workload eviction based on priority.

Predictive Allocation

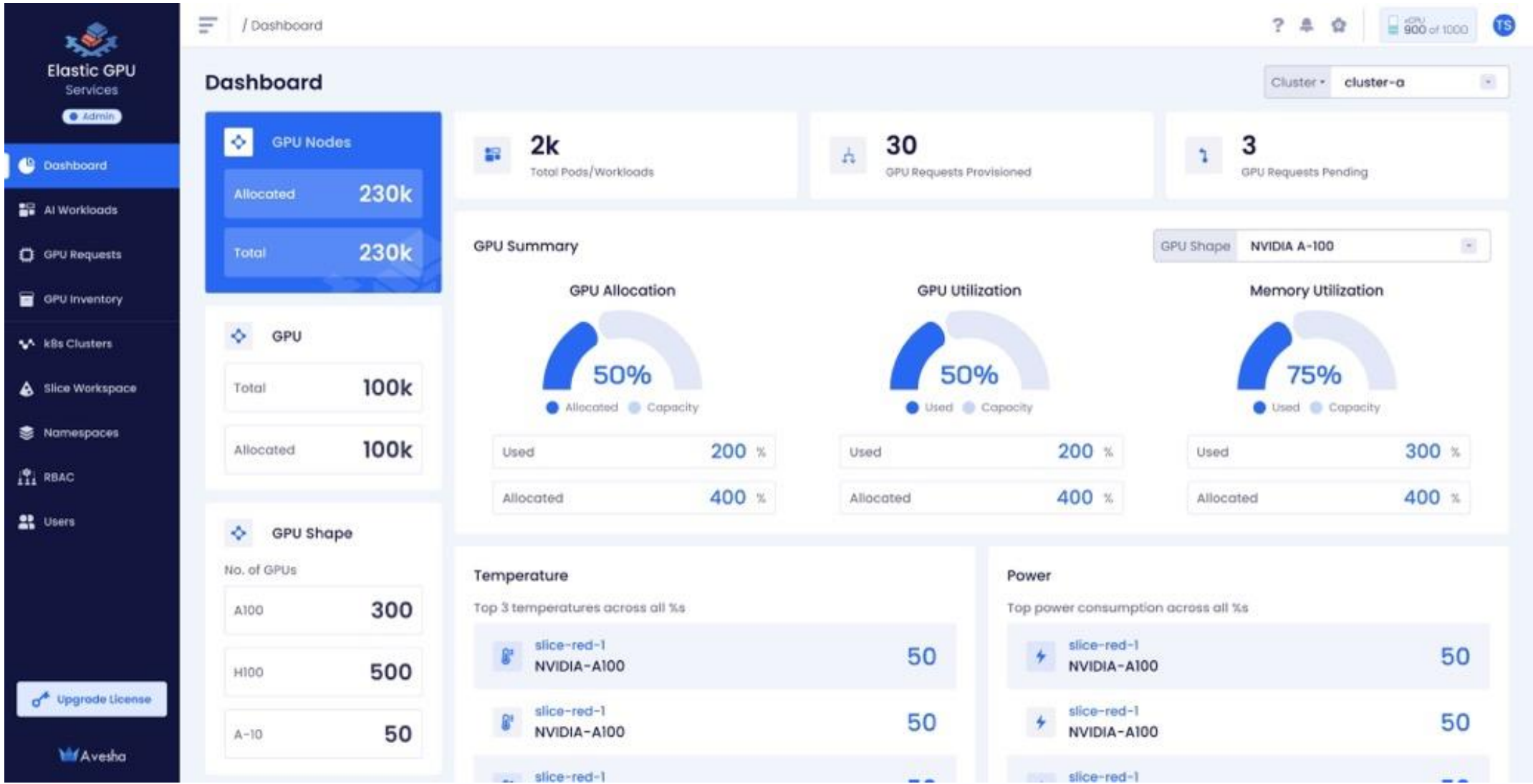
- Reinforcement Learning optimizes GPU usage.

Workload Monitoring

- Real-time dataset and model performance checks.
- Automated retraining and alerts.

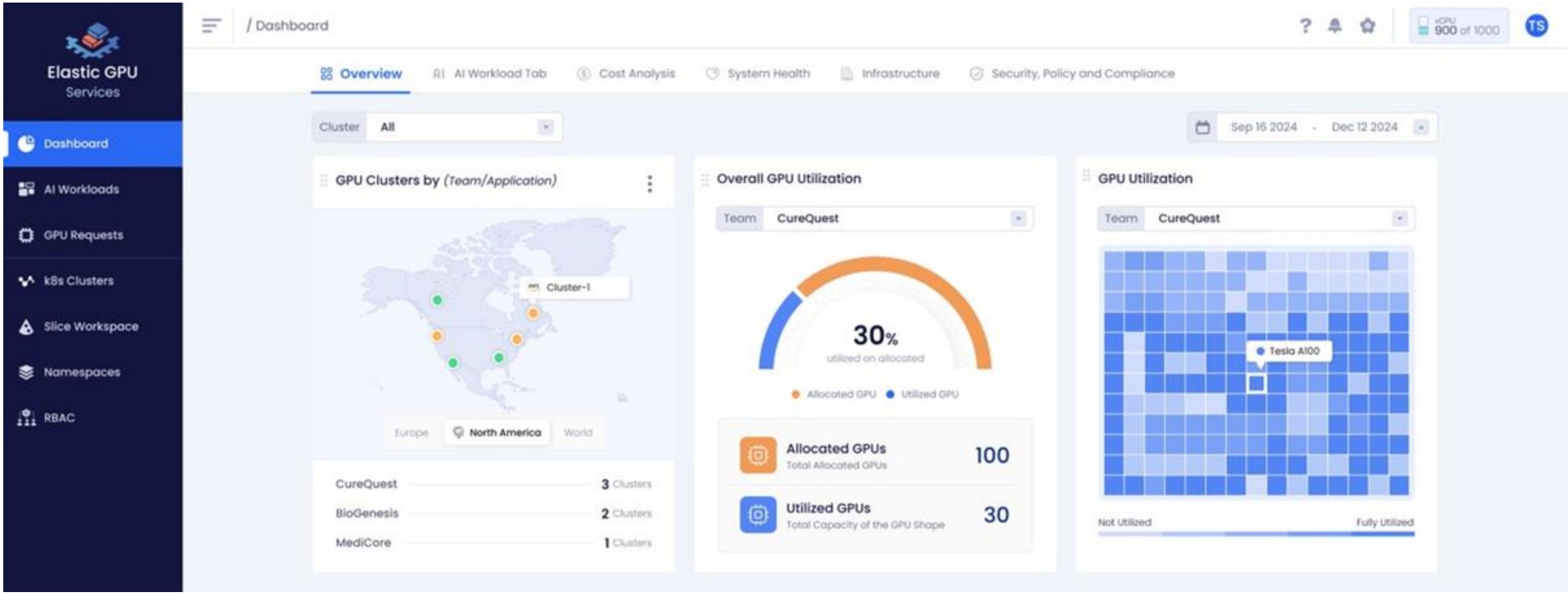
EGS Dashboard

Overall visibility to Allocation & Utilization



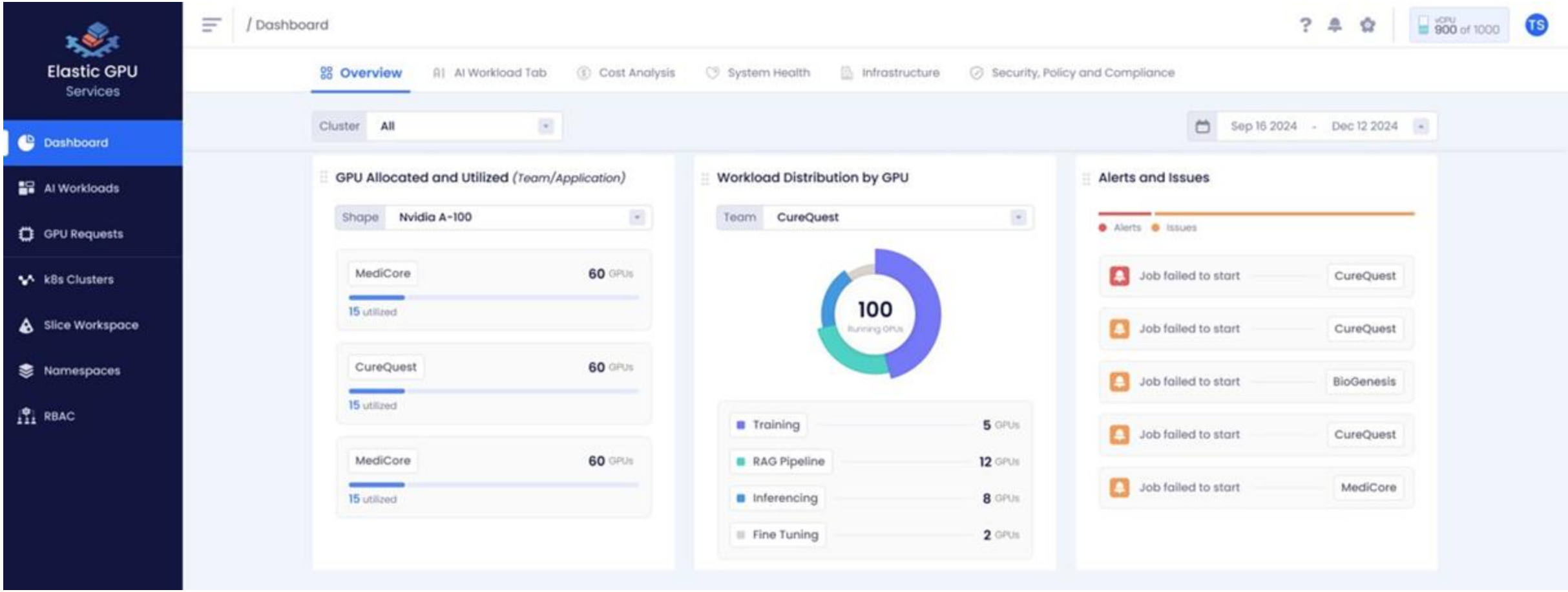
EGS Dashboard

Heatmap to highlight Issues



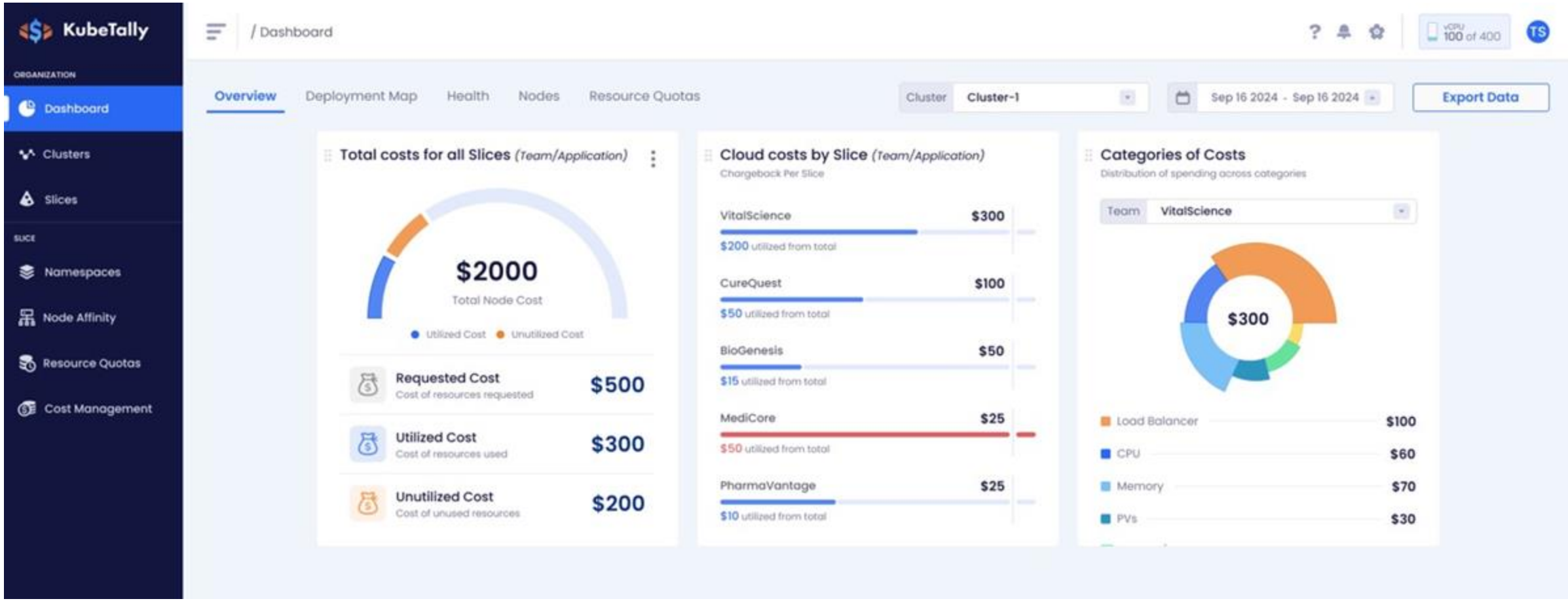
EGS Dashboard

Workload visibility by Team



EGS Dashboard

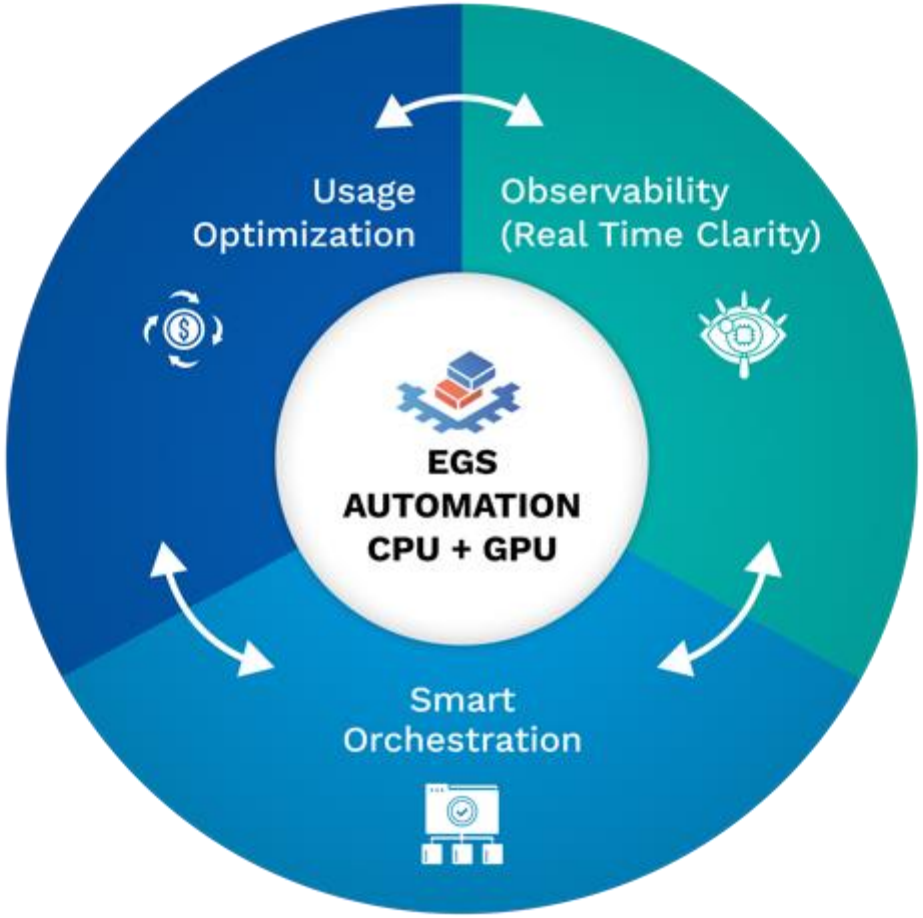
Cost Visibility by Team at the Cluster Level



EGS Summary

EGS transforms your GPU infrastructure into a competitive advantage—empowering your business to scale smarter, innovate faster, and save more.

The future of GPU workloads isn’t just about power—it’s about precision. And EGS delivers both.

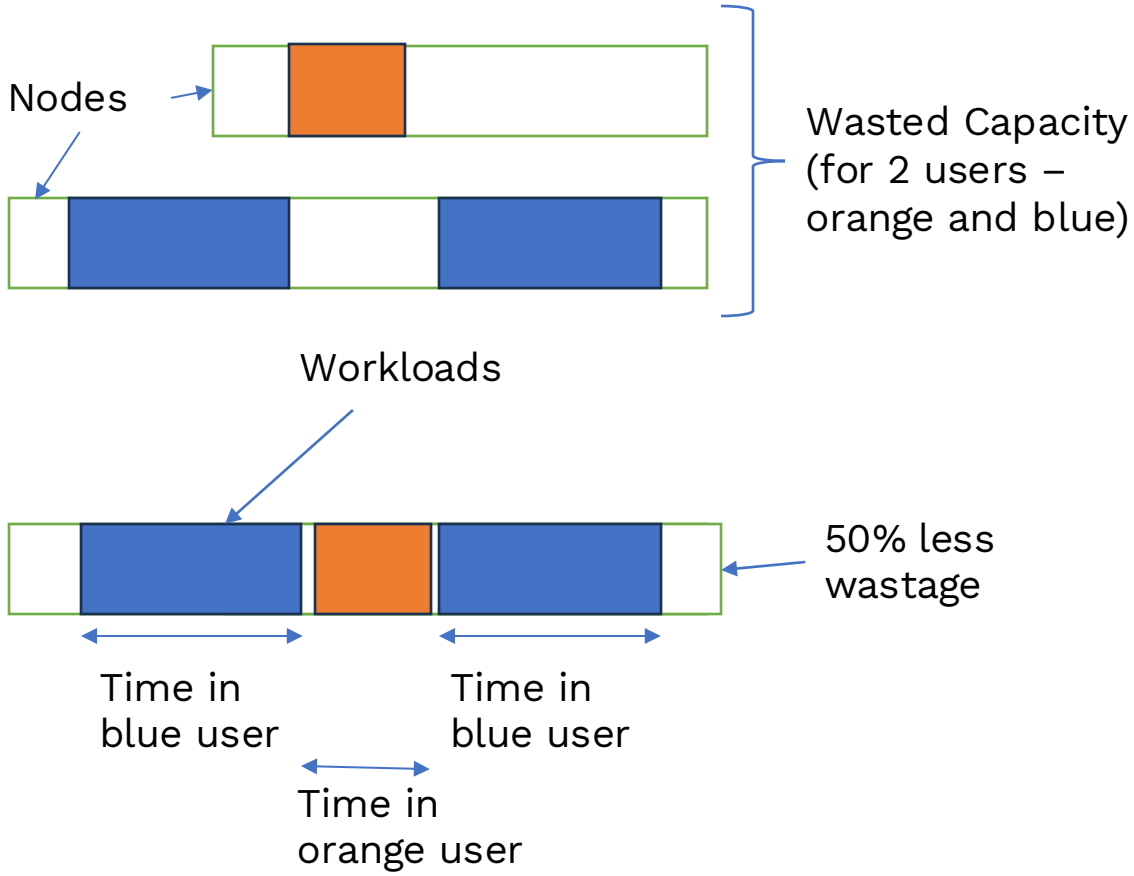


Backup

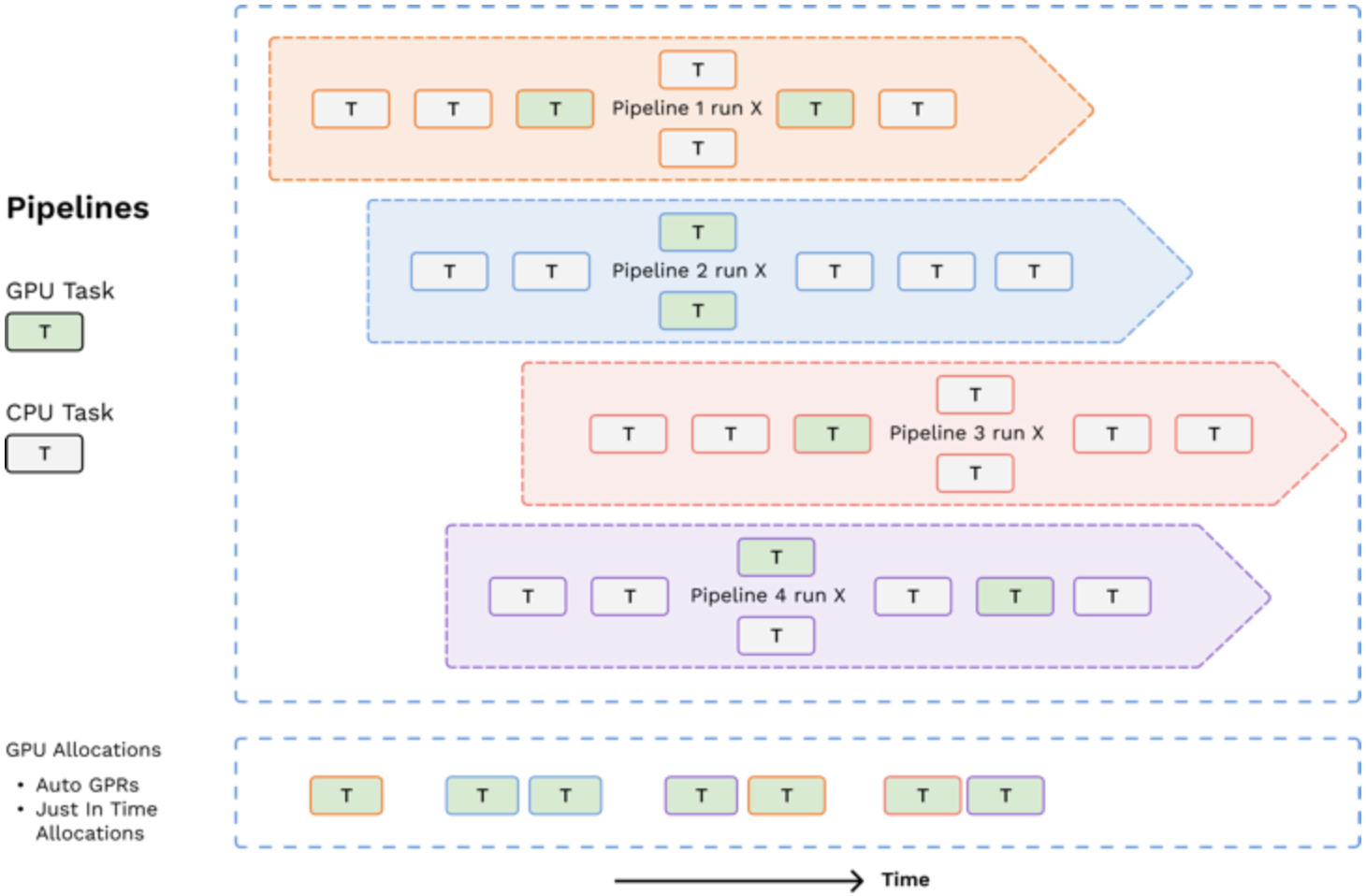
Smart Time-Slicing for Max Utilization

Powerful Features for Optimal GPU Management

- Automatic GPU provisioning via SDK
- Model data management with tenancy
- GPU cluster time slicing
- Priority Queuing
- Real-time GPU performance metrics
- Dynamic GPU virtualization
- Single cluster, multi-cluster & multi-cloud environments
- Diverse cloud providers offering flexibility and resilience
- Built on KubeSlice opensource (CNCF sandbox project)



GPU Allocations for DAG Pipelines



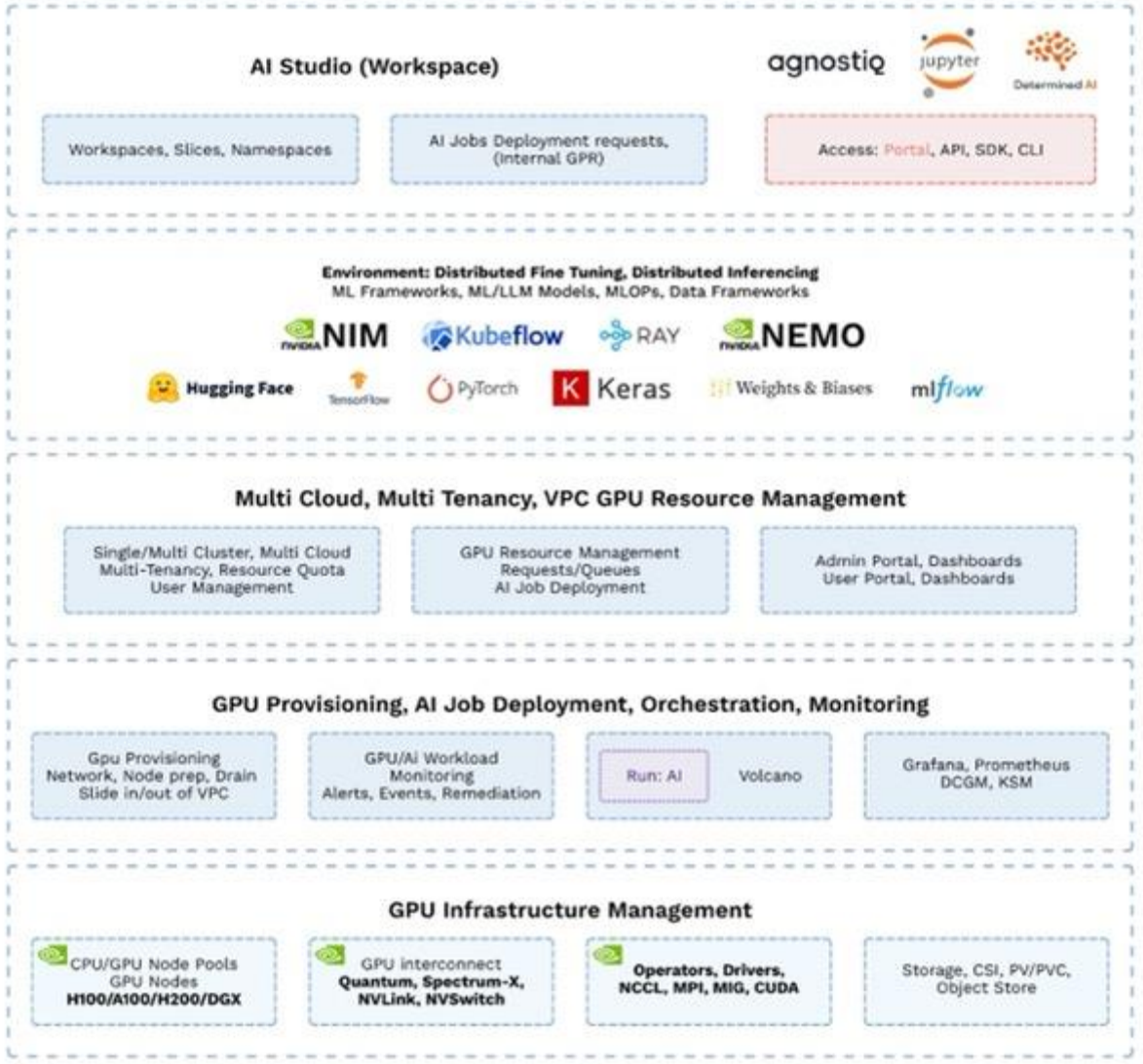
1. GPUs can be shared when nodes of the DAG are not actively using them
2. DAGs of different priorities are simultaneously active
3. Unused but allocated GPUs for high priority DAG can be made available to a lower priority DAG that needs it

Where EGS fits in the Ecosystem

Kubernetes

Where EGS works

Infrastructure



Thank You,
Questions?