Elastic GPU Service (EGS) -- Workload Automation, Optimization, Cost Reduction, and Observability

Despite advancements in ML scheduling tools like KubeFlow, optimizing GPU and CPU usage remains difficult. Mismatches between resource management and workload orchestration cause idle GPUs: creating delays, and inefficiencies in large-scale setups. Current GPU allocation relies on manual adjustment and lacks dynamic adaptation. Without standardized GPU rating and sharing approaches, advanced ML schedulers still struggle with scheduling, leading to bottlenecks and resource waste.

An Innovative Automated Approach

Elastic GPU Allocator (EGS) dynamically allocates GPUs using performance-based prioritization for efficiency and scalability. It leverages inputs like model size, network topology, and execution path data to create dynamic GPU mappings.

- **ML Automation and Optimization**
  Just-in-time orchestration of pipeline dependencies, task resource needs, and schedulers enables dynamic GPU allocation across heterogeneous GPU pools.
- **Built on KubeSlice (a CNCF sandbox project)**
  Open-source framework enabling seamless ML orchestration, leveraging Avesha's KubeSlice for visibility and tightly coupling tasks in DAG pipelines with scheduling and orchestration.
- **Predictive Resource Allocation**
  Uses historical execution patterns for intelligent GPU allocation, ensuring balanced resource utilization and eliminating manual workload management
- **Advanced Multi-Cluster Orchestration**
  Combines Dynamic Resource Allocation (DRA) with KubeSlice to automate job groups, DAG paths, and data flows, addressing multi-cluster challenges with precise execution and monitoring.

Intelligent GPU Execution

This approach automates manual adjustments and inefficiencies with the following benefits:

Core Benefits

- **Usage Optimization**
  Businesses can align GPU usage with priorities by leveraging predictive allocation, prioritizing critical workloads, minimizing idle resources, and reducing costs through GPU cluster time-slicing for dynamic provisioning and reallocation and improving GPU utilization by up to 50%.
- **Observability with Real-Time Clarity**
  Real-time dashboards provide visibility into GPU status, costs, resource allocation, and workflow efficiency, enabling automated remediation (e.g. on temperature or memory or power). Learn exactly how much each team is allocated and spending.
- **Smart Orchestration**
  Optimized job completion rates result in higher utilization and reduce idle time across GPU workloads. EGS has shown to result in up to 44% improved throughput.
- **Automated Remediation**
  Continuous monitoring of GPU nodes and health checks minimizes manual intervention. Dynamic reconfiguration of nodes ensures adaptability to changing workload conditions.
- **Security**
  EGS offers secure multitenancy with RBAC that ensures data privacy among teams that share the same GPU resources.
- **Scalability and Modularity**
  Features such as observability and cost control are independently deployable, allowing businesses to adopt capabilities incrementally.

Free-tier for Enterprises in Early-stage trials:

- **Separation by Teams**
  EGS ensures isolation of teams while sharing GPU resources. Data is not co-mingled among teams.
- **Ease of Operations**
  Allocation of GPU resources is managed via a central entity ensuring fair and policy-controlled usage of GPUs.
- **Virtual GPU Allocation**
  Teams can schedule GPUs virtually and specify the hyperparameters to match the desired shape and model. In turn, EGS will allocate the resources as it finds them.

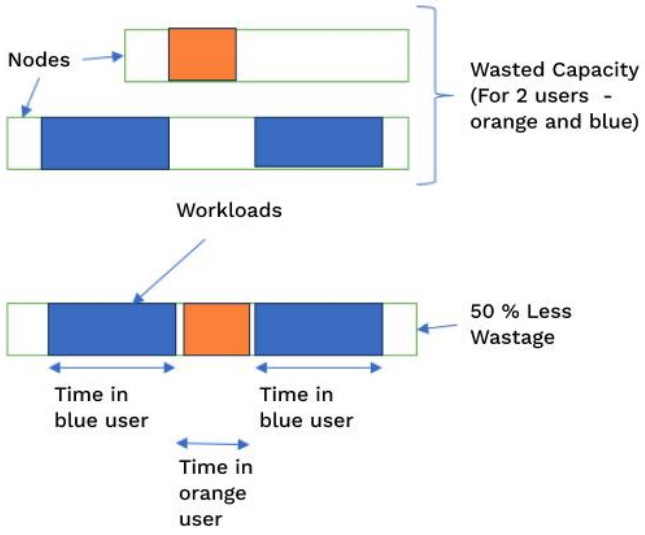

Fig Cluster time-slicing

Advanced Capabilities of Intelligent GPU Execution

Closing the Gap in MLOPs

EGS addresses enterprise needs for multi-user, multi-job group, and multi-cluster orchestration with advanced features:

- **Advanced GPU Allocation**
  Predictive Reinforcement Learning models ensure optimal GPU utilization. The system learns via feedback on GPU utilization and workload behaviors that inform the AI-model on better allocations.
- **Proactive Workload Monitoring**
  Real-time monitoring of model and dataset performance (via integration with MLOPs tools) enables corrective actions, such as auto-checkpointing, model retraining or operator alerts.
- **Priority Preemption**
  Workloads may be evicted from GPUs based on priority.

End-to-End ML Pipeline Optimization & APIs

EGS seamlessly integrates with popular ML frameworks, such as TensorFlow, PyTorch, KubeFlow, and Ray, to optimize every stage of the ML pipeline. For example:

- **Initial Setup:**
  EGS associates slice GPUs with each step in the DAG, mapping resources to models for preprocessing, training, and evaluation stages.
- **Dynamic Adjustments:**
  Real-time metrics trigger preloading configurations, reducing idle time between jobs.
- **Outcome:**
  Enterprises experience faster job completion times, better GPU utilization, and smoother transitions between pipeline stages

Conclusion

No matter what size company, EGS redefines GPU resource management by delivering cost-efficiency, real-time observability, and high throughput. By combining advanced scheduling, dynamic provisioning, and seamless ML framework integration, it empowers enterprises to overcome inefficiencies and unlock the full potential of their GPU infrastructure. This solution addresses the challenges of mixed GPU imbalance and manual workload management, enabling businesses to scale AI-driven operations with confidence and agility. As a result, enterprises can achieve reduced costs, optimized resource utilization, and a competitive edge in innovation.