Release notes for EGS 0.5.0

Release date: 8/21/2024

Elastic GPU Service (EGS) platform provides a system and workflows for effective resource management of GPUs across one or more kubernetes clusters.

EGS documents

- For Admin guide please see the documentation on the website
- For User guide please see the documentation on the website
- For Installation guide please see the documentation on github repo

We continue to add new features and enhancements to EGS.

What's New

This is the first beta release of EGS.

For a full list of features, overview and guides please see documentation on the website.

Note: EGS built around core components of KubeSlice Enterprise.
For information of KubeSlice Enterprise please see the documentation on website

These are some of the main features in this release. For details please refer to the documentation on the website.

EGS Slice

EGS slice (workspace) is Slice with one or more namespaces associated with it to provide a Slice workspace for a User or a Team. EGS Slice will be a no-network slice and will be associated with a Slice VPC (GPU VPC). Slice will be associated with a User. Slice will be associated with an RBAC and a service account that will be used by the User to access the EGS UI portal and Kubernetes cluster namespaces.

EGS Slice VPC

EGS Slice VPC is a logical boundary for a User (or a team) workspace. The Slice can be viewed as a VPC that spans one or more Kubernetes clusters. The Slice VPC workspace will be associated with one or more namespaces where user(s) can deploy their AI workloads. Users can deploy non AI workloads (CPU workloads) any time.
AI workloads that need GPUs need to have GPU nodes provisioned in the Slice VPC. Without the GPU nodes provisioned in the Slice VPC the AI workloads (Pods/etc) will go into the pending state - waiting for the GPU resources. Users need to create GPU request(s) for GPU nodes to be provisioned to the slice.

EGS supports two different personas: Admin and User

Admin
Admin is responsible for the installation and administration of EGS platform. EGS provides an Admin portal to perform the Day 0/1/2 operations. EGS also supports YAML (manifests) based admin workflows for these operations so that these workflows can be integrated with CI/CD or MLOps pipelines.

User
Users (can be a Data Scientist, Researcher or a ML engineer) uses EGS User portal to create and manage the life-cycle of GPU provisioning requests for User’s Slice(s) Workspace(s). Uses the portal to get deep visualization of the AI workloads and associated GPUs metrics and other data.

EGS GPU Provision Requests

Users create GPU requests for the Slice using the EGS User portal. The EGS control plane manages the queue of the GPU requests across the Slices/Users (teams) and Clusters. During the GPU request creation time User will be given an estimated provisioning time by EGS control plane (backend).

Dynamic GPU Provisioning in a Slice

The GPR manager periodically checks with Inventory and Queue managers to get the next GPR allocation for provisioning. Once the GPR is allocated the resources needed to provision the request will be identified. The GPR manager works with the worker cluster EGS component - aiOps Operator - to complete the provisioning. The GPR manager creates appropriate CRs for the aiOps operator to complete the provisioning of the GPU nodes into the worker cluster Slice VPC.

AI Workload/GPU Observability

EGS Admin and User portal offers a detailed view of AI workloads that are running in the User workspace (namespaces). In addition, a user-focused dashboard shows key metrics across Users Slices, workloads, and GPUs. It provides detailed events and notifications information for various alerts for GPUs, workloads, GPU requests, and so on.

Admin and User Workflows

- For Admin guide please see the documentation on website

- For User guide please see the documentation on website

Installation

- For Installation guide please see the documentation on github repo