Project #7: GPU Testing for LLM Blueprints

Project Page: GPU Testing for LLM Blueprints

Issue: https://github.com/kubeflow/trainer/issues/2432

Mentors: @andreyvelich, @varodrig

Project Size: 350 hrs

Summary

This project aims to use self-hosted runners to run GPU-intensive tasks like LLM blueprint or (planned) AI Playground. The necessary infra is provided by Oracle, plan is to use Oracle Kubernetes Engine (OKE) with NVIDIA GPUs for this task. Any code or sample that requires GPU-intensive resources will be transferred to OKE infra instead of generic GitHub infra for faster and more efficient execution.

For now, the idea is to have a specific policy that whenever any Jupyter Notebook code will be added to a `trainer/examples/pytorch/**` folder (e.g., in trainer/example/pytorch/image-generation/sample.ipynb), that action is transferred to OKE infra by the GitHub self runner. For security reasons, this process will require manual approval from one of the maintainers to trigger the self-runner build. I will set up the GitHub workflow to monitor changes in the respective folder. Once approved, the CI action will execute the code using the GitHub self-runner on the OKE infrastructure. Additionally, we will set up a dashboard for monitoring and metrics to understand usage patterns and identify bottlenecks.

The scope of this project is set up on OKE, but theoretically, this is platform-agnostic; it can be deployed on any Kubernetes cluster with sufficient GPU resources.

Personal Information

- **Full Name**: Akash Jaiswal
- **Email Address**: akashjaiswal3846@gmail.com
- **GitHub**: https://github.com/jaiakash
- **LinkedIn**: https://www.linkedin.com/in/akashjaiswal03/
- **CNCF Slack**: https://cloud-native.slack.com/team/U02JS2NFSBF
- **University/College**: NIT Trichy, India
- **Degree Program**: B.Tech, Civil Engineering
- **Year of Study**: Graduated in May 2024
- **Work:** Software Engineer 1 at Oracle India
- **Work Experience:** 10 months
- **Country of Residence**: India
- **Timezone**: IST, UTC+05:30

Qualifications; Motivation?

- **Experience**
  - Software Engineer 1 at **Oracle India** Linkedin (June 2024 - Present)
  - Software Engineer Intern at **Oracle India** Linkedin (May 2023 - July 2023)
  - **GSoC 2022 contributor** at CC Extractor Linkedin (May 2022 - Oct 2022)
  - Internationalization Contributor for the **p5.js (Processing Foundation)** Website
  - Founding Engineer in 3x startups
  - Contributor to **KubeFlow**, **Volcano**, **OWASP**, **Processing Foundation**, **Oppia**, **Keploy**, **Google** and **HackerRank** Github

- **Pre GSoC Contribution**
  - KubeFlow https://github.com/kubeflow/website/issues/4045 https://github.com/kubeflow/website/issues/4043 https://github.com/kubeflow/website/pull/4044 https://github.com/kubeflow/website/pull/3998 https://github.com/kubeflow/website/pull/3963
  - Volcano https://github.com/volcano-sh/website/pull/367 https://github.com/volcano-sh/website/issues/361 https://github.com/volcano-sh/website/issues/366
  - OWASP - https://github.com/OWASP/Nest/issues/950 https://github.com/OWASP/Nest/issues/888 https://github.com/OWASP/Nest/issues/945 https://github.com/OWASP/Nest/issues/946 https://github.com/OWASP/Nest/issues/886 https://github.com/OWASP/Nest/issues/887 https://github.com/OWASP/Nest/pull/951 https://github.com/OWASP/Nest/pull/1033 https://github.com/OWASP/Nest/pull/964 https://github.com/OWASP/Nest/pull/941 https://github.com/OWASP/Nest/pull/948 https://github.com/OWASP/Nest/pull/902 https://github.com/OWASP/Nest/pull/903

- **Community/Leadership/Certification**
    - Community
        - KubeCon India Blog on CNCF Website
        - Microsoft Student Summit, Bangalore
        - KubeCon India 2024
        - Hack This Fall Community meetup in Hyderabad
        - WikiMedia and IIIT Hyderabad Scholarship
    - University Leadership
        - Lead TeCOS, the open-source community in college with 1000+ participants, to help college juniors with open-source and related programs.
        - Technical Secretary, 2023 - 2024 NIT Trichy (Managing 27 technical clubs, south-India largest techno-managerial fest Pragyan and helping 7000+ student community)
    - Certification
        - Introduction to KubeFlow Introduction to AI/ML Toolkits with Kubeflow LFS147
        - Oracle Cloud Infrastructure 2025 Foundations Associate
        - Certified Kubernetes Application Developer (CKAD)
        - Jenkins - From Zero to Hero Specialization
- **Availability**

I am committed to dedicating **30-35 hours per week** during the GSoC period. My working hours will be: Weekdays (Mon-Fri **4 hours per day**) and Weekends (Sat-Sun **6.5 hours per day**) Regular working time slots:

- 07:00 - 09:00 IST
- 19:30 - 24:00 IST

I will have limited availability from August 5th to August 8th due to KubeCon India. *I have also volunteered to help KubeFlow set up their booth in KubeCon India.* ☝️☝️ I have kept a buffer time to cover any backlog and blockers. Besides this, I have no other prior commitments. If any emergency commitments arise, I will promptly inform mentors to find a suitable workaround.

- **Motivation**

I have been active in open source since my college days, and I like to experiment and contribute to open source in my free time. I have been past GSoCer as well. In my work at Oracle, internally, I was tasked to develop a PoC to leverage k8s to run CI/CD pipeline. We were looking to efficiently configure/allocate resources in OKE to run CI/CD pipeline, and that's where I got to know about Volcano's batch scheduling. I started reading about Volcano and in turn, KubeFlow. I started reading and contributing to MLOps and KubeFlow.

Then I started joining community calls of KubeFlow, that's where I told **Chase Christensen** that I work at Oracle. He motivated me to learn and contribute to Oracle distro of KubeFlow. I also had meet with Andrey, Francisco and Victor about the plan for Oracle to donate GPU infra to KubeFlow.

After my little experimentation and research on trainer/issues/2432, I am confident to contribute to this project as a GSoC contributor.

Goals

- [ ] Set up a sample LLM Blueprint
- [ ] Configure GPU nodes on OKE
- [ ] Establish ACR on the OKE Cluster for deploying the LLM
- [ ] Create a GitHub Action for manual triggers and runners on the OKE cluster
- [ ] Implement metrics and analytics for the GPU Cluster
- [ ] Develop an AI Playground on OKE

Non-Goals

1. The GPU cluster for production deployment should be provided by Oracle. For testing purposes, I have a sufficiently powerful personal machine (Ryzen 7 8600G, 32GB RAM, Nvidia RTX 4060) to conduct tests.
2. Once the infrastructure for the self-runner is set up, running the AI Playground will require minimal setup. The primary focus of this project is to establish the infrastructure for running the LLM blueprint on OKE. The AI Playground is a secondary priority for this GSoC project, but I will continue working on it if it is not completed within the GSoC period.

Estimation of Deliverables

1. **Milestone 0 (May 8 - June 1)**: Community Bonding Period
2. **Milestone 1 (June 2 - June 10): Setup sample LLM Blueprint**
3. **Milestone 2 (June 11 - June 20): Setup Github Action**
4. **Milestone 3 (June 21 - July 6): Setup OKE with GPU nodes**
5. **Buffer Period (July 7 - July 13)**
6. **Midterm Evaluation (July 14 - July 18)**
7. **Milestone 4 (July 19 - July 27): Setup ACR on OKE Cluster**
8. **Milestone 5 (July 28 - August 10): Metrics and analytics**
9. **Milestone 6 (August 11 - August 24): AI Playground on OKE**
10. **Final Submission (Aug 25 - Sept 1)**

Technical Details

TechStack

GitHub Actions (and ARC), Kubernetes, Oracle Cloud, PyTorch, Python, Linux

Setup LLM Blueprint (Milestone 1 (June 2 - June 10))

To set up the same LLM blueprint that can be triggered based on admin approval. We have already one sample on in trainer repo, here I have tested a sample project for running on my local system. I will be adding 2-3 base more samples with different requirements for our testing.

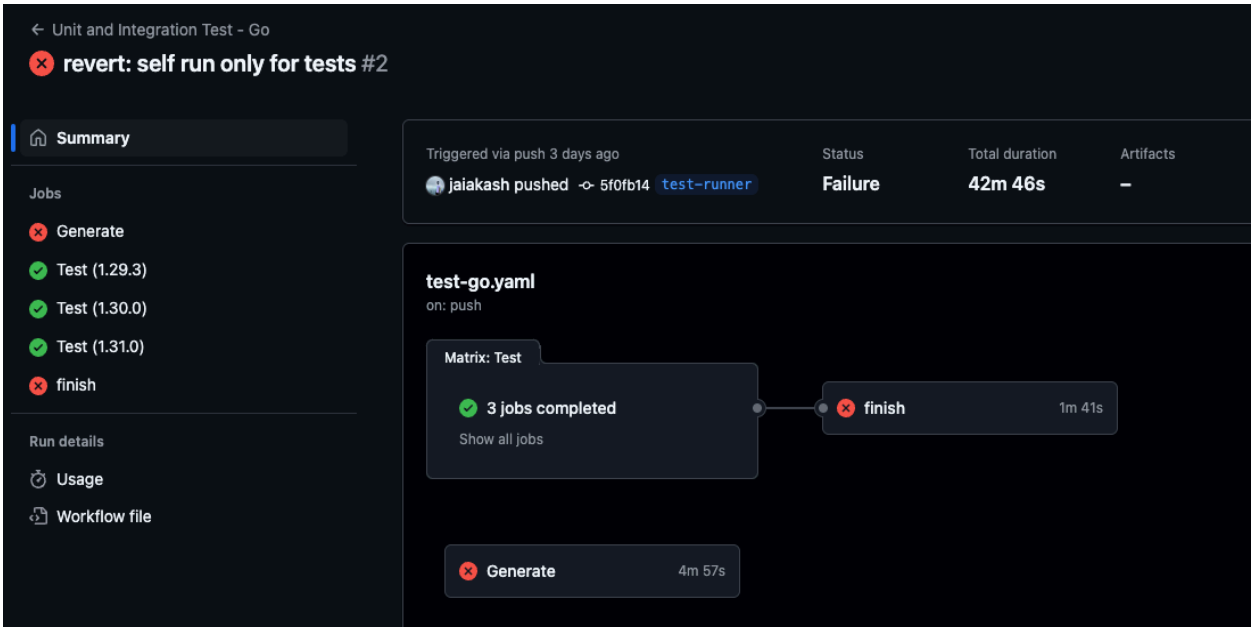
These additional samples will be designed to cover a range of scenarios and configurations, thereby enhancing the versatility and applicability of the LLM blueprint. This approach will not only facilitate thorough testing but also provide valuable insights into optimizing the deployment and execution of LLMs on the OKE infrastructure.

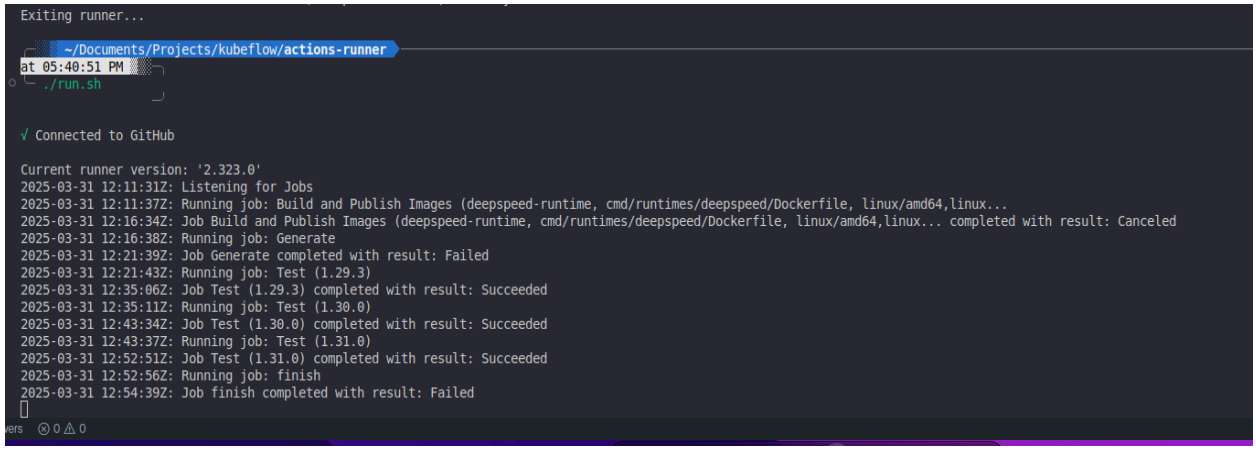
Github Action (Milestone 2 (June 11 - June 20))

Create a GitHub action for checking changes in files in `trainer/example/self-runner` and wait to trigger the self-runner after approval from the maintainers. Once the maintainer approves the scan, the code is executed in the self-runner (that is OKE infra). Assuming it takes some time and resources, we will implement queuing so that resources don't get flooded with requests. We will maintain a queue for requests, and report the result back to CI accordingly.

Here is the branch - test-self-runner

Screenshot:





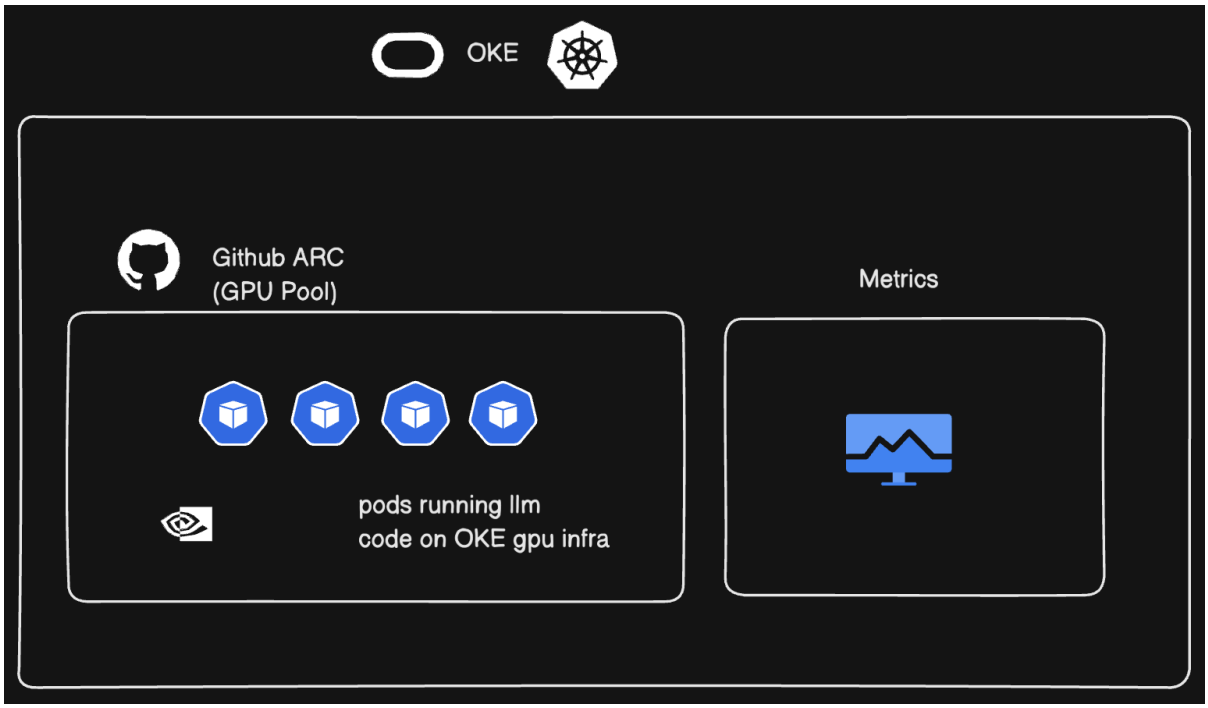

```yaml
name: Check llm changes and run on self-runner infra

on:
  pull_request:
    paths:
      - 'examples/self-runner/**'

jobs:
  request-approval:
    runs-on: ubuntu-latest
    steps:
      - name: Request maintainer approval
        uses: hmarr/auto-approve-action@v3
        if: github.event.pull_request.user.login == 'jaiakash'
        with:
          github-token: ${{ secrets.GITHUB_TOKEN }}

  wait-for-approval:
    needs: request-approval
    runs-on: ubuntu-latest
    steps:
      - name: Wait for maintainer approval
        uses: hmarr/auto-approve-action@v3
        with:
          github-token: ${{ secrets.GITHUB_TOKEN }}

  run-llm-code-on-oke-infra:
    needs: wait-for-approval
    runs-on: self-runner
    steps:
      - name: Checkout code
        uses: actions/checkout@v4

      - name: Run LLM code on OKE Infra
        run: |
          echo "Running on OKE infra"
```

Setup and access control of OKE Cluster with GPU (Milestone 3 (June 21 - July 6))

In this milestone, the aim is to setup an OKE Cluster with GPU node. System: Ubuntu 22.04 LTS The GPU image has the GPU drivers pre-installed.

Cluster Architecture: The OKE cluster will be configured with:

- GPU Node Pool utilizing NVIDIA GPUs with pre-installed drivers
- Standard Node Pool for regular workloads (Metrics and Queue)
- Managed Control Plane by OCI

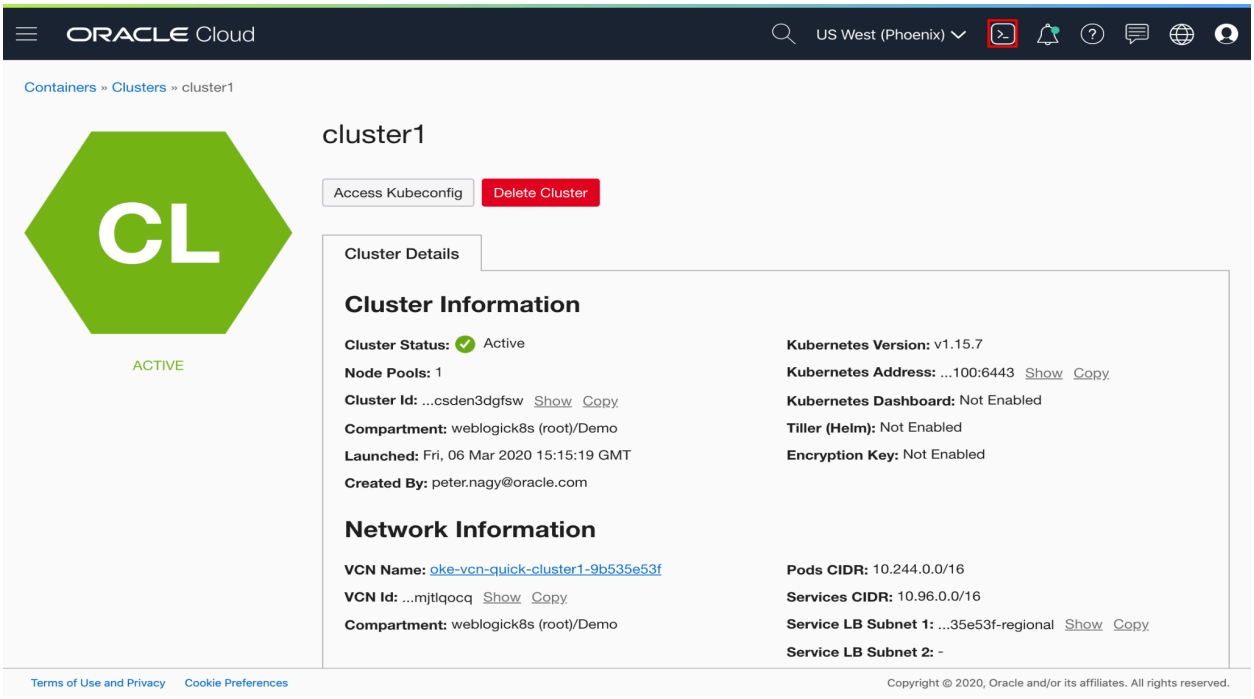


Accessing Cluster

1. OKE can be accessed with `kubeconfig` file
2. Or via Bastian

Access control for the OKE cluster

- Cluster Administrators: Full cluster management rights
- Maintainers: Limited administrative access
- CI/CD Systems: Restricted service account access



Images for NVIDIA shapes

- GPU driver 570 & CUDA 12.8
- GPU driver 560 & CUDA 12.6
- GPU driver 550 & CUDA 12.4

Reference

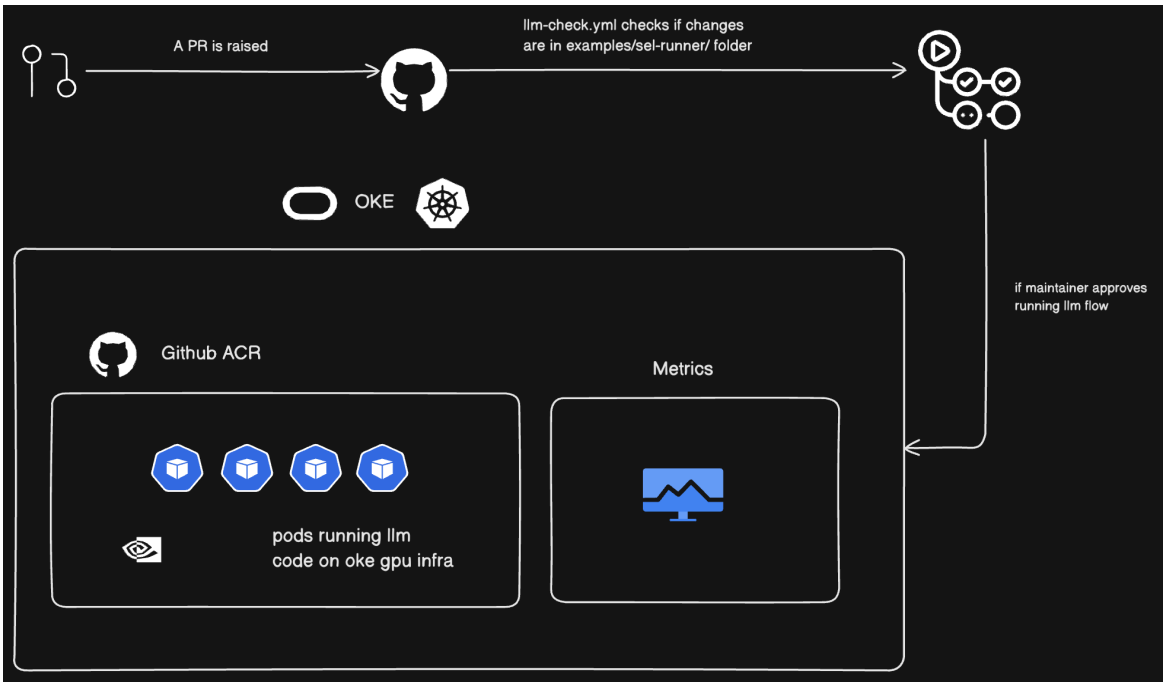
- https://github.com/oracle-quickstart/oci-hpc-oke
- https://blogs.oracle.com/java/post/create-k8s-clusters-and-deply-to-oci-from-vscode

Buffer period and midterm evaluation (Buffer period (July 7 - July 13))

This buffer period is for covering any backlogs and blockers. This time, I will use it to cover any pending changes and fixes. I will also try to demo in a community call. One of the agendas in this buffer period is to write a blog about the progress and status of the current project. By this time, the main project would have been completed.

Setup GitHub Actions Runner Controller (ARC) (Milestone 4 (July 19 - July 27))

Actions Runner Controller (ARC) is a Kubernetes operator that orchestrates and scales self-hosted runners for GitHub Actions. This is advanced phase of our project where we use k8s operator that is useful to scale and orchestrate pods based on the action CI.



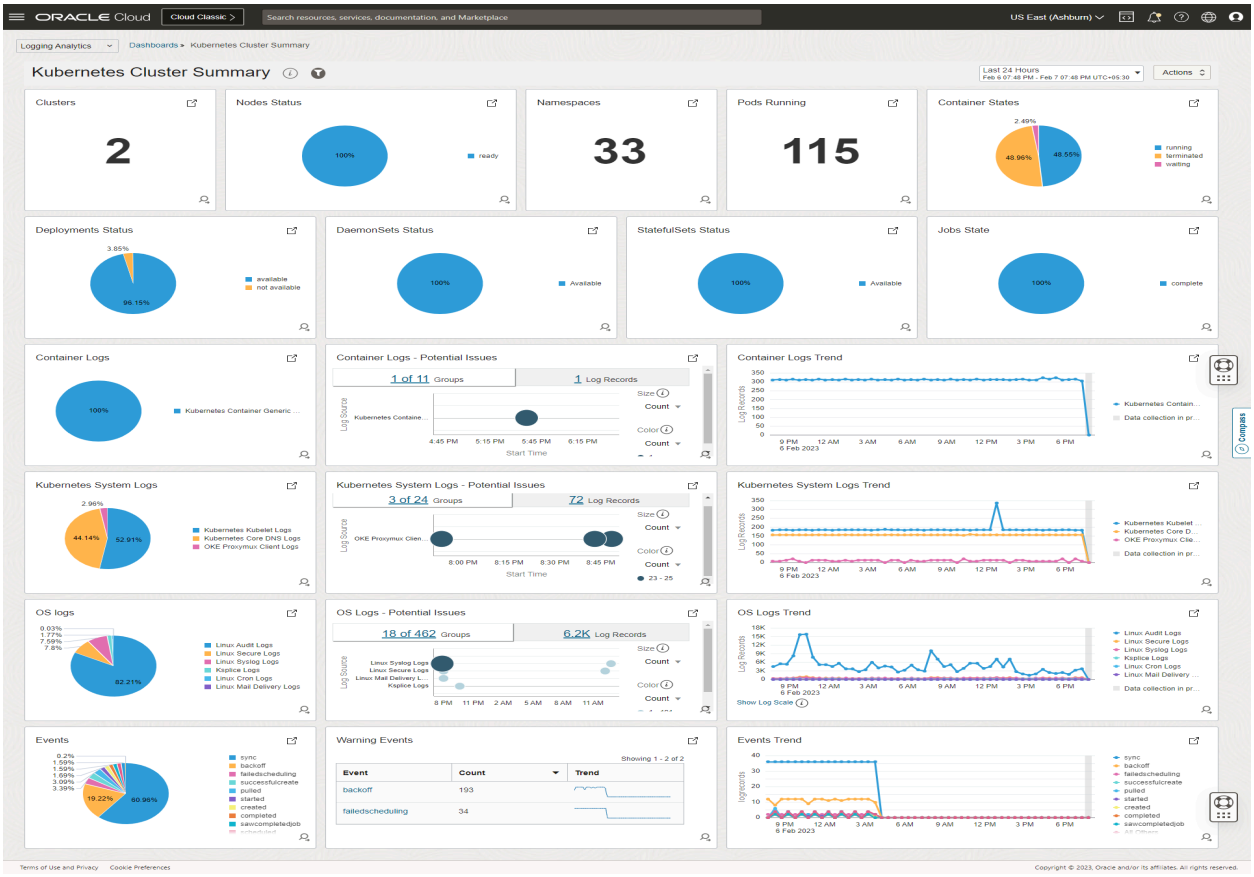
OKE Monitoring (Milestone 5 (July 28 - August 10))

For admins, we also need to maintain monitoring to see the metrics and resource utilisation of the OKE infra. Oracle already provides an open-source sample for OKE Monitoring, so we can leverage that. Out of various options, installation via Helm is sufficient for our basic needs.

Estimated monthly cost: **$0/month**

Metrics needed

Avg CPU Usage	Avg Queue timing
Avg GPU Usage	Avg Build timing
Peak GPU Usage	



AI Playground (Milestone 6 (August 11 - August 24))

This is the final phase of the project, with LLM CI deployment, as in various KubeCons various users wanted to deploy a model quickly to test and run KubeFlow. The idea is to setup sample models where user can Open Kubeflow Jupyter Notebook -> select Kubeflow LLM blueprint -> fine-tune model with Kubeflow Trainer -> serve it with Kubeflow KServe. We will set up similar GitHub action for it during events. Also we need to implement safeguards to prevent misuse, given our limited GPU infrastructure. One option is to implement GitHub OAuth and provide access on an as-needed basis during KubeCon.

Test Plan

During the GSoC period, until OKE infra is donated to KubeFlow. I will be testing local machine with 32GB RAM, Nvidia RTX 4060 GPU, Ryzen 7 8700G. I will have a demo during mid term evaluation in community call, once that is finalized by mentors.

I will be using OKE to deploy after production. To make sure there is no unnecessary usage of infra while testing, i will be putting certain guardrails on prod OKE.

Reference

Oracle Docs https://oracle.github.io/fmw-kubernetes/wccontent-domains/oracle-cloud/prepare-oke-environment/ https://github.com/oracle/weblogic-kubernetes-operator/blob/main/kubernetes/hands-on-lab/tutorials/setup.oke.ocishell.md

KubeFlow Docs https://www.kubeflow.org/docs/components/trainer/getting-started/ https://www.kubeflow.org/docs/started/architecture/

GitHub ACR Docs https://docs.github.com/en/actions/hosting-your-own-runners/managing-self-hosted-runners-with-actions-runner-controller/about-actions-runner-controller

GitHub Self Runner Docs https://docs.github.com/en/actions/hosting-your-own-runners/managing-self-hosted-runners/adding-self-hosted-runners

Cloude 3.5 Sonnet - For formatting text and improving this proposal

Diagram - https://app.eraser.io/

Thanks for helping and guidance @andreyvelich, @varodrig @thesuperzapper, @chasecadet, @varodrig