Run:ai Demo Setup – GPU Job Orchestration

\*\*Author:\*\* Sri Yasaswini

\*\*Date:\*\* {{Add Date}}

\*\*Audience:\*\* AI Engineering, Infrastructure, DevOps

# 🏁 Goal

Set up a functional Run:ai demo environment to demonstrate:  
- GPU job orchestration  
- Fractional GPU allocation  
- AI model training/inference on the cluster  
- Visibility via dashboard and CLI

# ⚙️ 1. Environment Prerequisites

|  |  |
| --- | --- |
| Component | Status / Notes |
| Kubernetes/OpenShift | ✅ Cluster ready |
| GPU Nodes (e.g., H200) | ✅ Present and accessible |
| Helm CLI | ✅ Installed |
| Run:ai Platform Access | ✅ Credentials & API key available |
| CLI Tools | ✅ `runai`, `kubectl`, `helm` installed |

# 📦 2. Setup Instructions

## 2.1. Install NVIDIA GPU Operator

# Use OperatorHub in OpenShift or Helm

## 2.2. Install Run:ai Operator

helm repo add runai https://run-ai-charts.storage.googleapis.com  
helm install runai-operator runai/runai-operator -n runai --create-namespace

## 2.3. Label Namespace

kubectl create namespace sri-gpu-demo  
kubectl label namespace sri-gpu-demo runai-enabled=true

# 🔐 3. Authentication

## 3.1. Run:ai CLI Login

pip install runai  
runai login

# 🚀 4. Submit a Demo Job

## 4.1. Sample PyTorch Job Submission

runai submit sri-demo \  
 -i pytorch/pytorch:latest \  
 --gpu 1 \  
 -v /mnt/data/code:/workspace \  
 --command -- python train.py

## 4.2. Fractional GPU Example

runai submit sri-fractional \  
 -i pytorch/pytorch:latest \  
 --gpu 0.5 \  
 --command -- python train.py

# 📊 5. Monitoring

- ✅ Use `runai list`, `runai logs`  
- ✅ Check the Run:ai dashboard for GPU usage and job tracking

# 🧪 6. Optional Enhancements

|  |  |
| --- | --- |
| Enhancement | Description |
| Harness Pipeline Integration | Automate job submission from GitHub → Run:ai |
| JFrog Artifactory | Pull model weights and dependencies dynamically |
| Run LLM or CV model | Replace `train.py` with real inference model |
| GPU Resource Sharing | Show multiple jobs sharing a single GPU |

# 📁 Attachments

- train.py (basic model training script)  
- runai-config.yaml (optional Helm values)  
- Screenshots of dashboard/jobs