GPU-Scanner:

Extending CNCF Observability for Multi-GPU AI Workloads

Ritika Gupta
Software Engineer, Oracle

Observability Summit NA 2026

github.com/oracle-quickstart/oci-gpu-scanner

The Problem

Observability Summit North America

"Imagine a 25-day training job failing on day 23 because one GPU silently throttled."

Silent GPU Failures

Hardware failures are more frequent than CPU hosts — more moving parts, more ways to break.

Invisible Until Too Late

Most observability stops at the node. GPU health stays invisible until workloads fail or budgets spike.

Onion-Peeling Debugging

Root cause is never the first thing you check — thermal, network, topology issues all compound.

What is GPU-Scanner?

An open-source observability extension for Kubernetes GPU clusters that adds active and passive GPU health checks to your existing CNCF stack.

Cloud-Native

Prometheus & Grafana native. Deploy via Helm.
Fits your existing CNCF observability stack.

Active + Passive Checks

30+ health checks: TFLOPs, memory bandwidth, thermals, NVLink, RDMA, and more.

Vendor Agnostic

Works with NVIDIA and AMD GPUs. Supports H200, MI355X, and other accelerator shapes.

Auto-Remediation

Detects failures, drains nodes, triggers reboots, tags issues — automatically.

How It Works

Observability Summit North America

GPU-Scanner deploys as a DaemonSet on Kubernetes GPU nodes with a Control Plane for management, and a Data Plane for monitoring and health checks.

Control Plane

Portal (React)
Resource pools, health status, on-demand checks

Backend (Django)
REST API, orchestration, remediation triggers

PostgreSQL
Cluster state, health history, config

Deploy via: Helm

Configuration: ConfigMaps & Secrets • Ingress: NGINX Controller

Data Plane (DaemonSet per GPU Node)

GPU Metric Exporters
NVIDIA DCGM Exporter
AMD SMI Exporter
Prometheus Node Exporter

Health Check Engine
Active checks (PyTorch)
Passive checks (every 30s)
Schema converters

NVIDIA: A100, H100, H200, B200 | AMD: MI300X, MI355X

Integrates with Node Problem Detector for auto-tainting failed nodes

Observability Layer — CNCF Stack

Prometheus
Metrics storage & querying. All metrics prefixed oci_lens_ for custom alerting.

Grafana Dashboards
Pre-configured dashboards: cluster overview, node health grid, GPU metrics, RDMA status.

Alerting & Integration
Connect to Slack, PagerDuty, or your existing incident management workflow.

Active vs. Passive Health Checks

Passive Health Checks

Runs every 30 seconds • Non-intrusive

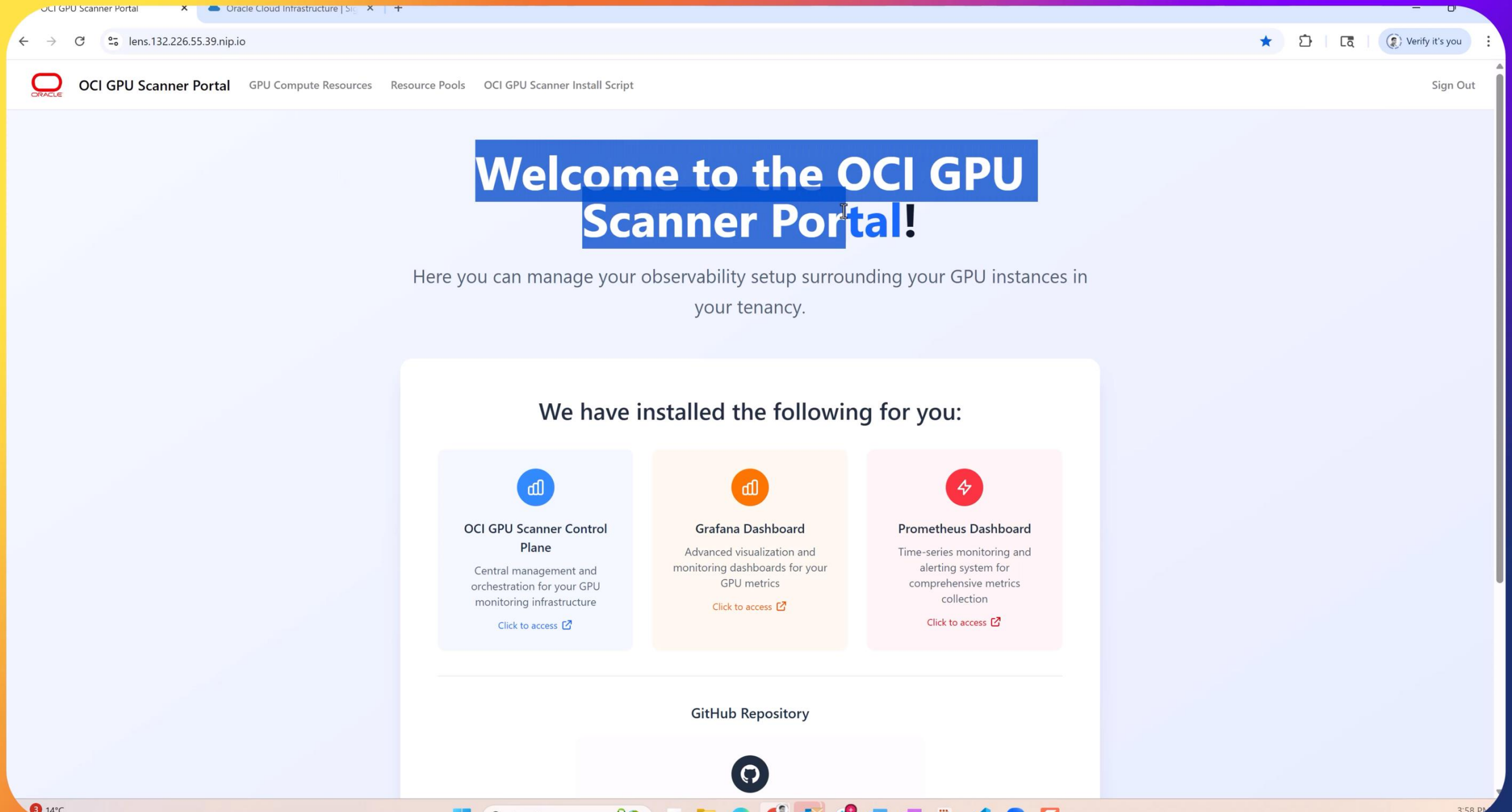
- ✓ GPU Count & Clock Check
- ✓ NVLink & SRAM Status
- ✓ PCIe & XID Error Detection
- ✓ RDMA Link & Flap Check
- ✓ GID Index Verification
- ✓ Kernel Parameter Check
- ✓ Driver & Mode Validation
- ✓ Fabric Manager Check

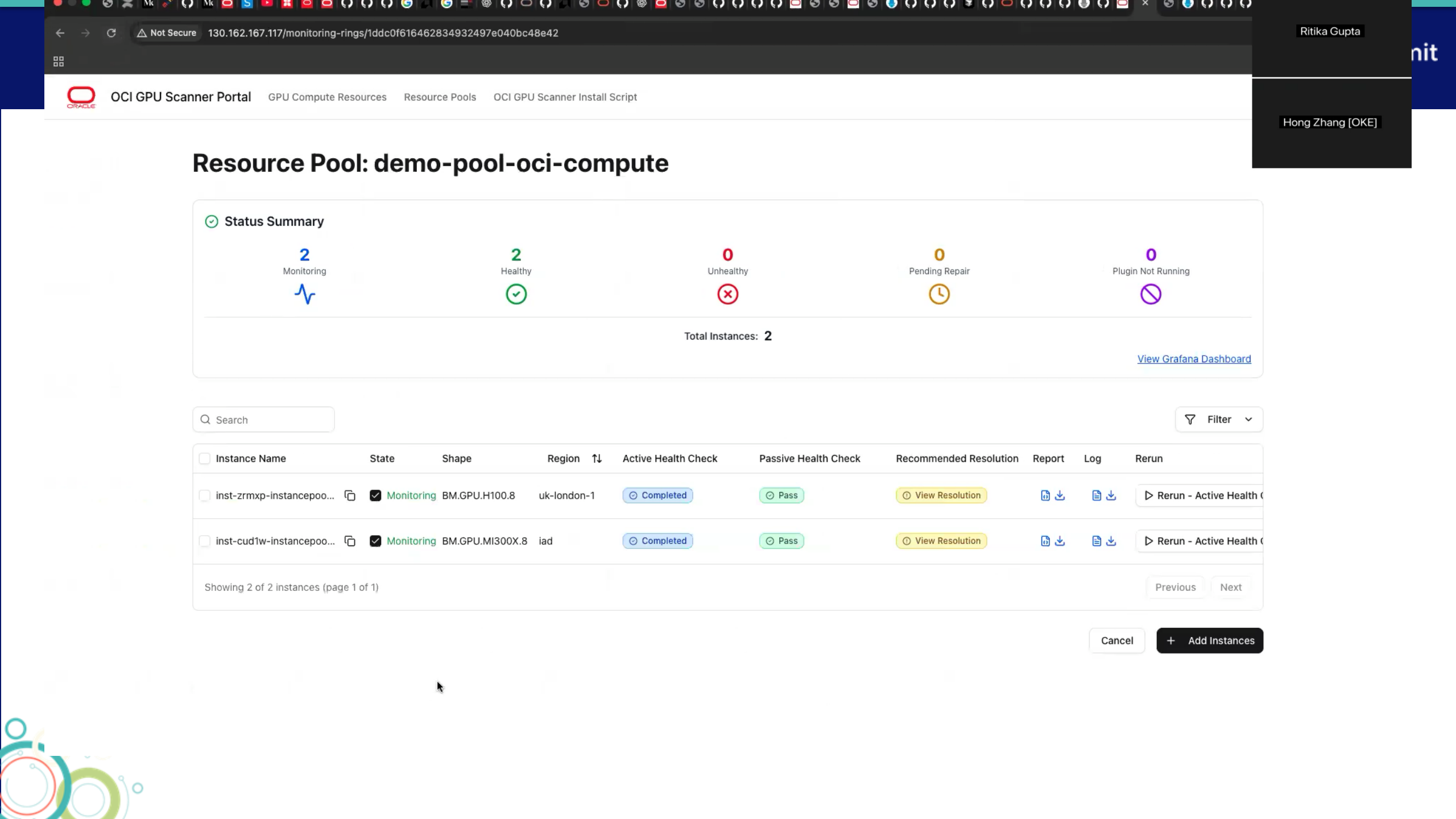
Active Health Checks

On-demand • Occupies GPUs during run

- ▸ Memory Bandwidth Test
- ▸ Tensor Core Utilization
- ▸ Sustained Workload Stability
- ▸ Mixed Precision Validation
- ▸ Power Draw Consistency
- ▸ Temperature Sustained Check
- ▸ GPU Topology Validation
- ▸ MFU: Matmul & Linear Regression
- ▸ Multi-Node MPI (RDMA)

Demo


OCI GPU Scanner Portal
Oracle Cloud Infrastructure | Sig
lens.132.226.55.39.nip.io
Verify it's you
OCI GPU Scanner Portal
GPU Compute Resources
Resource Pools
OCI GPU Scanner Install Script
Sign Out
Welcome to the OCI GPU Scanner Portal!
Here you can manage your observability setup surrounding your GPU instances in your tenancy.
We have installed the following for you:
OCI GPU Scanner Control Plane
Central management and orchestration for your GPU monitoring infrastructure
Click to access
Grafana Dashboard
Advanced visualization and monitoring dashboards for your GPU metrics
Click to access
Prometheus Dashboard
Time-series monitoring and alerting system for comprehensive metrics collection
Click to access
GitHub Repository


Not Secure
130.162.167.117/monitoring-rings/1ddc0f616462834932497e040bc48e42
Ritika Gupta
Hong Zhang [OKE]
OCI GPU Scanner Portal
GPU Compute Resources
Resource Pools
OCI GPU Scanner Install Script
Resource Pool: demo-pool-oci-compute
Status Summary
2
Monitoring
2
Healthy
0
Unhealthy
0
Pending Repair
0
Plugin Not Running
Total Instances: 2
View Grafana Dashboard
Search
Filter
Instance Name
State
Shape
Region
Active Health Check
Passive Health Check
Recommended Resolution
Report
Log
Rerun
inst-zrmxp-instancepoo...
Monitoring
BM.GPU.H100.8
uk-london-1
Completed
Pass
View Resolution
Rerun - Active Health
inst-cud1w-instancepoo...
Monitoring
BM.GPU.MI300X.8
iad
Completed
Pass
View Resolution
Rerun - Active Health
Showing 2 of 2 instances (page 1 of 1)
Previous
Next
Cancel
Add Instances

Your GPU Observability Playbook

Observability Summit North America

What you can do today

1

Deploy GPU-Scanner

Helm install on any OKE or Kubernetes GPU cluster.
Works with NVIDIA and AMD accelerators.

2

Baseline Your Cluster

Run active health checks to establish performance baselines — MFU, bandwidth, thermal profiles.

3

Set Up Alerting

Connect Prometheus alerts to Slack, PagerDuty, or your existing incident management workflow.

4

Enable Auto-Remediation

Let GPU-Scanner drain, reboot, and recover unhealthy nodes before your training jobs notice.

Thank You!

Questions? Let's talk GPU observability!

GitHub: github.com/oracle-quickstart/oci-gpu-scanner

Contributions Welcome!