Rik Kisnah

Lead Principal Systems Software Engineer, AI/ML Infrastructure, Oracle Cloud Infrastructure
Seattle, Washington, United States • +1 206 724 5771 • rikesh@gmail.com • linkedin.com/in/rikkisnah • rik-kisnah.ai • github.com/rikkisnah

SUMMARY

Infrastructure engineer with 20+ years building and operating production platforms, the last seven on GPU/HPC compute at fleet scale. Leads the OCI GPU Infrastructure team, owning rack-scale cluster design, the data plane, validation, fleet health, and repair automation for a fleet spanning NVIDIA H100, GB200/GB300 NVL72, and AMD accelerators across dozens of production regions.

Co-authored the engineering behind OCI's dedicated GB200 NVL72 rack-scale APIs. Hands-on in Go and Python; comfortable across C/C++. Technical advisor to senior executives, public speaker, and author of a widely read body of writing on GPU infrastructure, NCCL, AI data centers, and AI-assisted engineering.

CORE SKILLS

GPU cluster design & rack-scale bring-up (NVL72) • NVLink / NVSwitch fabric • RDMA / InfiniBand / RoCE, NCCL & collective performance • Validation & acceptance criteria (multi-vendor, topology-aware) • Fault classification & repair orchestration • Fleet telemetry & observability • NVIDIA / AMD / OEM technical interface • Go, Python, C/C++ • Terraform / IaC, CI/CD • Cross-org technical leadership & public speaking

EXPERIENCE

Oracle Cloud Infrastructure

Seattle, WA | Full-time | 7 yrs 8 mos

Lead Principal Systems Software Engineer — June 2026 – Present

- Leads the OCI GPU Infrastructure team, owning rack-scale cluster design, bring-up, validation, and fleet health for a heterogeneous NVIDIA H100, GB200/GB300 NVL72, and AMD fleet across dozens of production regions.
- Technical interface with NVIDIA and OEM hardware partners for new GPU generations: firmware and driver interoperability, NVLink/NVSwitch provisioning, and RDMA fabric validation for GB200/GB300 bring-up.
- Owns the validation and acceptance-criteria framework that gates every host before release: ~190 test definitions across three silicon vendors, topology-aware from single host through pair, full rack (18 trays / 72 GPUs), and multi-rack collectives, with performance thresholds set per shape and per fabric.
- Owns the fault-classification and remediation path that routes production hardware failures to the right recipe automatically (rerun, soft reset, deep power cycle, repair script, ticket), fed by fleet-wide telemetry for hosts and NVLink/NVSwitch fabric (health, port state, error counters, thermals, power).
- Built the diagnostics and repair-orchestration path that drains and repairs an individual host while jobs keep running on the rest of the NVL72 rack; added an LLM-assisted triage tier behind deterministic rules with confidence-gated escalation to humans.
- Technical advisor to senior executives on architecture and roadmap; drove the organization's AI-assisted engineering adoption program. Multiple patent and invention disclosures in GPU fault management and topology-aware AI systems.

Consulting Member of Technical Staff — August 2024 – June 2026

- Tech lead for AI/ML infrastructure across OCI's HPC/GPU product line: control-plane APIs, host software, data-plane validation.
- Co-authored the engineering behind OCI's dedicated GB200 NVL72 rack-scale APIs, which launch, resize, monitor, and repair a full 72-GPU rack as a single unit, including firmware-version consistency enforcement across hosts, GPUs, and NVLink switches.

Principal Engineer — January 2021 – August 2024

- Tech lead for the OCI Compute HPC/GPU organization; shipped the OCI HPC/GPU host plugins that run across all OCI GPU instances: RDMA configuration, GPU metrics collection, and fault detection, integrated with cluster networking up to 400 Gbps.

Principal Engineer (DevOps/SRE), OCI Compute HPC and Bare Metal January 2019 – December 2020

Amazon Web Services

Seattle, WA | 6 yrs 1 mo

Senior System Development Engineer December 2012 – December 2018

- Elastic Load Balancing: built and operated automation for one of AWS's highest-traffic distributed services, where every change ships to a large, globally distributed, always-on fleet.
- Previously delivered AWS WorkSpaces, EC2 Windows, and Amazon Kindle.

Aeroflex

2 yrs 7 mos

Principal Staff Development Engineer, PXI RF test for Qualcomm chipsets June 2010 – December 2012

Motorola

7 yrs 6 mos

Senior Staff Software Engineer, mobile devices (RAZR V3, ROKR E1, V70) January 2002 – June 2009

PUBLICATIONS & SPEAKING

GPU Burn-In at Scale: Commands, Gates, and Troubleshooting for NVIDIA and AMD Clusters
rik-kisnah.ai, September 2026. Diagnostics, network validation, and acceptance gates for cluster commissioning.
www.rik-kisnah.ai/posts/gpu-burn-in-at-scale-reference-guide

The Complete NCCL Reference Guide for OCI GPU Infrastructure
rik-kisnah.ai, November 2025. NCCL tuning and errors for GB200, H100/H200, A100, MI300X. Widely shared reference.
www.rik-kisnah.ai/posts/nccl-complete-reference-guide

Attention Is All You Need: An Infrastructure Engineer's Guide
rik-kisnah.ai, February 2026. Attention cost, KV-cache sizing, and interconnect design. Well received across the industry.
www.rik-kisnah.ai/posts/attention-is-all-you-need-and-all-you-need-to-know

Zettascale Performance: OSU and NCCL Benchmarks on H100 AI Workloads
Oracle Cloud Infrastructure Blog, November 2025. H100 performance analysis at scale.
blogs.oracle.com/cloud-infrastructure/zettascale-osu-nccl-benchmark-h100-ai-workloads

Behind the Scenes: Scale Your NVIDIA GB200 NVL72 Deployments with Dedicated OCI APIs
Oracle Cloud Infrastructure Blog, September 2025 (updated January 2026). Co-authored; companion Oracle video.
blogs.oracle.com/cloud-infrastructure/behind-the-scenes-scale-nvidia-gb200-nvl72-deployments

Mauritius Doesn't Need to Build Fable. It Needs an AI Pass.
Le Mauricien (Forum), September 2026. Latest of three AI opinion pieces. Talks: GB200 for OCI and AI Workloads (NUS HPC, Aug 2025); HPC Cloud Builders (Europe HPC, Sep 2021). Full list: rik-kisnah.ai/publications
www.lemauricien.com/opinions/mauritius-doesnt-need-to-build-fable-it-needs-an-ai-pass

EDUCATION

Massachusetts Institute of Technology 2026
Postgraduate Certification, AI/ML

The University of Texas at Austin 2019
Postgraduate Certificate, AI/ML

University of Wales, Aberystwyth 2007
MSc, Computer Science (Neural Networks)

Nanyang Technological University, Singapore 2001
BASc, Computer Engineering (Machine Learning)

CERTIFICATIONS & LANGUAGES

Java Developer • MCPD • AWS Solutions Architect • PMP • Languages: English, French, French-based Creole (native/bilingual); Mandarin, Bahasa Melayu (conversational)