Davide Rutigliano

Senior Platform Engineer (AI & GPU Infra) · Italy

(+39) 340 9307386 · davide.ruti@gmail.com · LinkedIn · Github · Medium · Portfolio

Summary

Senior Platform Engineer building GPU-accelerated Kubernetes platforms for AI/HPC workloads. Specialized in inference observability (vLLM, TTFT, OpenTelemetry) and cluster lifecycle operations across cloud and bare-metal environments. Open-source contributor: Kubernetes, Kueue and KubeAI (vLLM operator).

Experience

SUSE

Remote, Italy

Senior Platform Engineer — Dec. 2025 – Present

- **vLLM & GenAI Observability**: Engineered OpenTelemetry (OTel) connectors to instrument vLLM inference (Time-to-First-Token and related KPIs), enabling on-call triage for multi-tenant GPU inference platform.
- **High-performance GPU Monitoring**: Engineered GPU observability solution for Kubernetes and KubeVirt clusters, enabling fine-grained monitoring of NVIDIA A2/H100 with MIG and vGPU segmentation, unlocking 40+% HPC efficiency.
- **Cluster lifecycle & Autoscaling**: Optimized platform efficiency achieving a 62% reduction in infrastructure costs ($100K+ annual savings), by deploying Cluster Autoscaling with Cluster API on multi-cloud (AWS, GCP, On-Prem) platform.
- **AI-driven Observability**: Architected and delivered the SUSE Observability MCP Server from greenfield idea to MVP, embedding LLM-driven analysis into the alerting pipeline; recognized by senior leadership for roadmap inclusion.
- **Infrastructure Automation**: Designed and implemented a Kubernetes operator to orchestrate large-scale virtual machine migrations from KVM to Harvester, enabling the migration of 100+ VMs.

Platform Engineer — Jan. 2025 – Dec. 2025

- **Infrastructure Governance**: Designed and implemented a custom Kubernetes operator to synchronize VLAN and virtual machine assets into NetBox, establishing a single source of truth for infrastructure inventory.
- **Federated Observability**: Architected migration from Prometheus/Grafana stack to SUSE Observability (StackState) for federated multi-cluster observability, cutting troubleshooting time by 25%.
- **Cloud Evolution**: Led cloud architecture evolution across AWS, GCP, and Azure for enterprise migrations, aligning Cloud Landing Zone (CLZ) design with strategic requirements, acting as advisor to internal customers.

Stack: Kubernetes, Rancher, Cluster API, KubeVirt, GPU (NVIDIA MIG/vGPU), vLLM, OpenTelemetry, Prometheus, Grafana, Terraform, GitOps (ArgoCD/Flux), AWS/GCP/Azure

ERICSSON

Pagani, Italy

Devops Engineer Team Lead — Sep. 2023 – Dec. 2024

- **Self-Service Platform**: Led team of 5 engineers in designing Internal Developer Portal (IDP) based on Spotify Backstage, improving developer efficiency by 25% through self-services. Evangelized portal usage and adoption of best practices.

DevOps Engineer — Jun. 2022 – Sep. 2023

- **ML-driven Optimization**: Engineered an ML-driven solution for automated Kubernetes microservice performance optimization, reducing the engineering time required to fine-tune applications configuration from weeks to days.

Cloud Engineer — Mar. 2021 – Jun. 2022

- **Resource Efficiency**: Optimized Ericsson Licensing solution footprint by 25% deploying auto-scaling and fine-tuning resources configuration, leading to a throughput increase of 10%.

Stack: Kubernetes/OpenShift, Docker, Helm, Backstage, Gatekeeper (OPA), GitLab CI/Jenkins, Kafka, PostgreSQL/Cassandra, Go, Python, Java

CISCO

Vimercate, Italy

ML Engineer — Apr. 2020 – Mar. 2021

- Designed and implemented ML-based solution for optical devices troubleshooting, improving defect detection by a 2x factor.

Stack: Python, TensorFlow, Computer Vision

Education

Master Degree in Computer Engineering — Milan, Italy
Politecnico di Milano — 2018 – 2020

Bachelor Degree in Computer Engineering — Fisciano, Italy
Universitá degli Studi di Salerno — 2014 – 2018

Skills

AI & GPU INFRA: NVIDIA MIG/vGPU, GPU-Operator, LLM-Ops, vLLM, Kueue/Slurm, TensorFlow, Pytorch, Computer Vision

OBSERVABILITY: OpenTelemetry (OTel), Prometheus, Grafana, Alertmanager, StackState, Root Cause Analysis (RCA)

RELIABILITY: SLIs/SLOs, alerting strategy, runbooks, incident response, postmortems, capacity planning

CLOUD NATIVE: Kubernetes, Helm, Docker, GitOps (ArgoCD, Flux), Terraform, GCP, AWS, Azure

DEVELOPMENT: Go, Python, Java, Rust, K8s Operators, Event-Driven Architecture, Linux