

SUMMARY

Infrastructure and Inference Engineer with **5+ years** of experience building and scaling **high-traffic AWS infrastructure**, **GPU-based inference systems**, and production **Kubernetes** platforms. Proven track record in **load testing at 10M+ TPM scale**, self-hosting ML/LLM models (vLLM) serving **30% of production traffic**, shipping **real-time voice-AI over WebRTC**, and driving **68% inference cost reductions**. Strong in **Terraform IaC**, **CI/CD automation**, and **cloud security hardening**.

EXPERIENCE

Senior DevOps Engineer, Primetrace

Dec 2024 — Present

- Owned **end-to-end load testing** for Crafto's peak traffic events, preparing the platform to handle **~10M TPM**; system scaled to **1.6M concurrent users** and **6M DAU**, supporting **INR 4Cr revenue in 36 hours**.
- Deployed **self-hosted inference models** (**flux2-dev-32b 4-bit quantised**, **flux2-klein-4b**) and LLMs (**Gemma 26B via vLLM**) now serving **30% of production traffic**, reducing per-inference cost from **INR 3.5–4 to INR 0.8**.
- Optimized **GPU infrastructure** for Pixora upscaling from **2×A100 80GB (\$11/hr)** to **1×L40S (\$3.5/hr)** — **68% cost reduction**. Moved Sundar AI & Looks AI upscaling fully self-hosted, eliminating **\$3,200/mo** Replicate spend.
- Built the production **WebRTC media infrastructure** on **EKS** powering real-time voice-AI calls — stood up a **coturn** TURN/STUN relay behind an internet-facing **NLB** to carry **SRTP/UDP** audio into private-subnet pods, enabling **sub-second-latency** Android↔AI voice.
- Built **Android CI/CD** end-to-end via **GitHub Actions + Slack**, replacing manual builds — now critical infrastructure for the entire Crafto Android team.
- Completed **zero-downtime MySQL upgrade** (**8.0.42 → 8.4.7**) on major production DB; created reusable **Terraform module** for **EKS Pod Identity** reducing setup from 3 resources to 1 declaration (**3× productivity**).
- Diagnosed and resolved a **critical InnoDB history-length runaway** on a production MySQL DB — traced to **long-running transactions** under **REPEATABLE READ** blocking **undo-log purge**, restoring purge throughput.
- Rolled out **KEDA** event-driven autoscaling across most production workers with tuned **Marbot** alerting, right-sizing capacity to queue depth.
- Rolled out **AWS WAF** to production for Kutumb/Labs — analysed multi-day **WAF logs** across iterations to eliminate false positives, then cut over to **block mode** on prod with **zero downtime**.

DevOps Engineer, Ollion

Aug 2022 — Nov 2024

- Saved **\$230,000/year** for Indonesian digital bank through **EC2 rightsizing** and **VPC endpoint consolidation** (**90% reduction**).
- Led migration of Indonesian OTT platform from **GKE to AWS EKS**, managing **Terraform**, ingress controllers, and monitoring.
- Hardened security across **80+ AWS accounts** for UK-based media organization with automated compliance scripts.
- Executed cross-cloud migration from **S3-Redshift** to **GCS-BigTable/BigQuery** with secure **VPN** connectivity.

Senior Systems Engineer, Infosys

Nov 2020 — Aug 2022

- Migrated on-prem **Redis** to **Azure Cache** and **Go** apps to **AKS** via **Terraform** for the world's largest private bank and a Fortune 500 fintech.

SKILLS

- Cloud & Infra:** AWS (EKS, EC2, RDS, S3, Route53, CloudWatch, WAF), Azure (AKS), GCP (GKE, BigQuery)
- Orchestration & IaC:** Kubernetes, Terraform, ArgoCD, Argo Workflows, Docker, Helm
- AI/ML Inference:** GPU Inference (A100, L40S, H100), Diffusion Models, Codeformer, GFPGAN, vLLM, Gemma, LoRA
- Scaling & Real-time:** KEDA, Locust, WebRTC, coturn, SRTP, MySQL/InnoDB internals, CloudWatch, Marbot
- Tools & Languages:** GitHub Actions, Vault, Pritunl VPN, MySQL, Redis, Python, Bash

AWARDS

- 1st place — AWS GameDay** (Acquisition Meltdown · AWS DevOps Agent) — led team to the top score (99,270) among competing startups; teammates placed 2nd and 3rd.

EDUCATION

B.Tech, Electronics & Communication Engineering

Oct 2020

Maharaja Agrasen Institute of Technology, Delhi

Class 12, CBSE

March 2016

Apeejay School, Pitampura, Delhi