

Ganish Deepak N

+91 88256 48984 • Bengaluru, India • gdthe03n@gmail.com • gd03.me • github.com/gd03champ • linkedin.com/in/gd03n

SUMMARY

Infrastructure & Platform Engineer with 2+ years of experience operating production Kubernetes, AWS, Terraform, and CI/CD platforms. Own and lead organization-wide AI/LLM infrastructure while building developer platforms as products for internal engineering teams. Experienced across production reliability, incident response, developer tooling, autoscaling, cost optimization, security, and AI-assisted operations.

CORE SKILLS

Cloud & Infrastructure: AWS (EKS, EC2, ECS, ECR, IAM, ALB, EventBridge), Terraform, Azure, GCP

Kubernetes & Delivery: Kubernetes, Helm, Karpenter, KEDA, Istio/Envoy, Docker, GitOps, canary & progressive delivery

CI/CD & Developer Experience: GitHub Actions, self-service deployment tooling, internal developer platforms, ArgoCD

Reliability & Observability: Prometheus, Grafana, Loki, incident response, production debugging, cost optimization, CPU right-sizing

AI/LLM Infrastructure: vLLM, llm-d, LiteLLM, LangChain, MCP, LLM inference, autonomous agent systems

Languages & Backend: Go, Python, Node.js, C++, FastAPI, gRPC, REST

Data Stores: PostgreSQL/Aurora, OpenSearch, Redis, MongoDB

EXPERIENCE

SWIGGY | SDE 2 | AI Infrastructure & Platform Engineering

May 2026 – Present

Bengaluru, India

- Own and lead organization-wide AI/LLM infrastructure spanning the inference platform, LLM gateway, and internal MCP tooling.
- Built and deployed an autonomous LangChain-based agent for independent production outage detection and mitigation, alongside a self-service MCP debugging server exposing 270+ Kubernetes tools - reducing mean time to first actionable debug insight from 24 hours to under 2 minutes.
- Built and operated a production LLM inference stack on EKS using vLLM, llm-d, Karpenter, KEDA, and Istio, achieving 48.6 tok/s on Qwen3.8-27B in L40s with speculative decoding; [published](#) the architecture and operational runbook as open_source.

SWIGGY | SDE 1 | AI Infrastructure & Platform Engineering

Apr 2025 – Apr 2026

Bengaluru, India

- Owned a new scheduler trigger for Shuttle, Swiggy's internal CI/CD platform, across 4 repositories, enabling teams to schedule jobs declaratively through app.yaml without Terraform or Lambda; added 36 unit tests and drove UAT through rollout.
- Led multi-cloud security incident response across AWS, Azure, and GCP following an exposed-credential and supply-chain incident; rotated 50+ compromised keys and coordinated organization-wide remediation end-to-end within a single weekend.
- Led an organization-wide OpenSearch FGAC access-control migration using Terraform and IAM; authored the canonical reference implementation and rollout SOP, closing a cluster-wide authentication gap.
- Designed and drove a zero-downtime OpenSearch cross-region replication (CCR) strategy, achieving 100% replication success with 1.52s average lag, and mentored two contractors through the rollout.
- Migrated a single-instance production service to highly available ECS Fargate with canary rollout via Route 53, reducing infrastructure cost by 48.6%.
- Built organization-wide CPU right-sizing automation across production Kubernetes clusters, saving approximately \$3,000/day, and converted it into an ongoing Terraform-managed infrastructure pipeline.

SWIGGY | Software Development Intern

May 2024 – Mar 2025

Bengaluru, India

- Built a custom Prometheus exporter for SupervisorD in Go and [open-sourced](#) the project.
- Designed and built a [distributed log-processing system](#) - RabbitMQ/NATS Streaming for message routing across worker and drainer pods on Kubernetes - reducing ~900K log lines to 730 (99.92% reduction) in <2mins via RAG-based analysis.

EDUCATION

Bannari Amman Institute of Technology | Bachelor of Engineering, Computer Science and Engineering

2021 – 2025

- 2× Smart India Hackathon National Winner
- 1 Granted Patent (202241060687)