NIKHIL VAIDYAR

Senior Platform Engineer

Bengaluru, KA, India | +91 70422 24751 | nikhilvaidyar1997@gmail.com | LinkedIn | GitHub | Website

SUMMARY

Senior platform engineer with 6+ years building and operating distributed systems in Go on Kubernetes and AWS. Owns the delivery, caching, streaming and observability paths behind peak e-commerce traffic: 120+ microservices migrated onto GitOps, 20K+ RPS sustained through flash-sale events, and MTTR cut ~70% with an AI-assisted SRE layer. Works across the full platform surface, from infrastructure as code and multi-AZ reliability to autoscaling, security hardening and compute cost control.

SKILLS

Cloud & Infra	AWS (EKS, EC2, Lambda, S3, IAM, VPC, RDS, DynamoDB, API Gateway, CloudFront/WAF, CloudWatch, SQS/SNS), GCP, Azure AKS, Terraform, Infrastructure as Code
Kubernetes	Kubernetes (CKAD), Docker, Helm, ArgoCD, Harness, GitHub Actions, Jenkins, GitOps, HPA/KEDA, RBAC, network policies, secrets management, image scanning, progressive rollout and automated rollback
Reliability	Prometheus, Grafana, OpenTelemetry, Loki/ELK, SLOs and error budgets, alerting design, PagerDuty, on-call, multi-AZ and DR, load testing, capacity planning
Languages	Go, Python, Java, SQL, Bash
Data & Messaging	Apache Kafka, Redis, PostgreSQL, MySQL, DynamoDB, MongoDB, Elasticsearch, gRPC, REST, GraphQL
AI & LLM	LangChain, LangGraph, RAG, Claude API, MCP, Pinecone, model routing and inference cost control
Practices	Design review and system design facilitation, mentoring, incident response, TDD, domain-driven design, Agile

PROFESSIONAL EXPERIENCE

Victoria's Secret & Co **Mar 2025 – Present**

Senior Software Engineer *Bengaluru, KA, India*

- Designed and shipped the flash-sale order and inventory platform: inventory, cart, order and payment services in Go on EKS, with Redis-backed reservation (TTL, idempotency keys, atomic operations) to guarantee correctness under extreme concurrency. **Sustained 20K+ RPS and 1,200+ orders/minute through peak sale events, cutting overselling from ~1–2% of peak-event orders to effectively zero.**
- Eliminated cache stampedes on peak read paths by replacing request-time fan-out with a Kafka-driven precompute pipeline partitioned by shop_id, using bounded worker pools, atomic pointer reads and batched MGET. **Stabilised low-latency reads and materially reduced Redis and database contention under peak traffic.**
- Absorbed traffic spikes with HPA/KEDA autoscaling driven by CPU and Kafka consumer lag, backed by rate limiting, circuit breakers, retries and backpressure at service boundaries. **Prevented cascading failure across dependent services during unplanned load.**
- Ran multi-AZ deployments with zero-downtime releases and built the load-testing infrastructure used to validate capacity before every event. **Turned event readiness from a manual judgement call into a repeatable, measured gate.**
- Built an AI-assisted SRE and observability platform: Go platform services and Kubernetes operators on EKS instrumented with Prometheus, Grafana, OpenTelemetry and Loki/ELK to correlate metrics, traces and logs fleet-wide, with an LLM agent doing RAG over runbooks, architecture docs and past incidents to diagnose issues, recommend remediation and auto-execute safe recurring fixes. **Cut MTTR from ~35 minutes to ~10 minutes (~70%) across a fleet serving 15K+ RPS.**
- Added a natural-language layer over Prometheus through MCP integration, translating plain-English questions into PromQL and service-health scores. **Removed the query-writing overhead mid-incident and lowered on-call cognitive load.**
- Hardened the platform with RBAC, network policies, secrets management and image scanning in CI, and routed alerting through Slack and PagerDuty. **Closed the gap between "deployed" and "operable" for every service onboarded to the platform.**
- Architected VogueFlow, a real-time multi-agent AI commerce engine in Go (LangChain, Claude API, Pinecone hybrid RAG, Redis) using tiered semantic and token caching with model cascading from Llama-3.1 to Claude-3.5. **Served 15K RPS at sub-200ms p95 and cut inference cost 42% (~$8K/day), with a projected 22% mobile conversion lift.**
- Lead design reviews and system design sessions for platform and backend changes, and run knowledge-sharing sessions across the Bengaluru engineering team. **Spread platform practice and review standards beyond a single team.**

Victoria's Secret & Co **Apr 2023 – Mar 2025**
Software Engineer *Bengaluru, KA, India*

- Led the migration of 120+ microservices from Consul to Harness and ArgoCD GitOps pipelines, standardising delivery on Helm and GitHub Actions with environment promotion, pod disruption budgets and automated rollback. **Shipped 425 service versions across 2,500 PRs with a cross-functional Bengaluru team and removed manual deployment error from the release path.**
- Designed and built end-to-end observability dashboards in Prometheus and Grafana for service health and regression detection. **Cut on-call troubleshooting time by ~50%.**
- Optimised peak-load performance by restructuring database access patterns and removing redundant Kafka retry loops. **Reduced P95 response times by 30% under peak load.**
- Rebuilt the Add to Bag flow from the product listing page for the Beauty and Sleepwear collections. **Increased annualised sales by $2.3M through reduced purchase friction and improved conversion.**

Biofourmis **May 2020 – Mar 2023**
Software Development Engineer *Bengaluru, KA, India*

- Architected event-driven ingestion pipelines in Go, Kafka and AWS Lambda over PostgreSQL and MongoDB for continuous biometric telemetry from 100K+ remotely monitored patients. **Processed 400B+ data points daily, improved ingestion throughput 5x and reduced clinical alert latency by 40%.**
- Engineered a configurable care-execution workflow engine supporting 50+ clinical protocols. **Held p95 under 250ms across 100K+ concurrent patient sessions.**
- Built and operated the supporting microservices for continuous biometric data collection in a regulated health-tech environment. **Kept high-volume clinical data processing reliable as patient load scaled.**

PROJECTS

Cloud-Native Banking Platform — Go, Kubernetes, Helm, Docker, gRPC, JWT, sqlc, domain-driven design. Applied pessimistic locking on transaction paths to guarantee data consistency under concurrent access.

EDUCATION & CERTIFICATIONS

Rajasthan Technical University — B.Tech, Electronics and Communication Engineering Aug 2015 – May 2019
Scaler — Bootcamp, Computer Science
Languages: English, Hindi, Malayalam
Certified Kubernetes Application Developer (CKAD) | gRPC [Golang] Master Class: Modern APIs and Microservices | Algorithmic Toolbox | Data Structures