Sunil Mehta

Platform, DevSecOps & AI Infrastructure Engineer • Kubernetes • GitOps

Gurugram, India | +91-9983218581 | sunilmehta695@gmail.com | linkedin.com/in/sunilmehta-si
github.com/sunilmehta-si | sunilmehta.si | x.com/sunilmehta_si

Summary

Platform engineer with 9.5 years building and running multi-cloud Kubernetes platforms on GCP, AWS, Azure, DigitalOcean and Hetzner for enterprise SaaS and AI-native products, spanning GitOps delivery, DevSecOps, confidential AI infrastructure and cloud cost governance. Twice the founding DevOps lead; directed 15+ engineers.

- Sustained 99.95–99.99% availability for 200+ microservices on 10+ Kubernetes clusters.
- Scaled confidential AI infrastructure to 300+ TDX-protected GPUs behind gateways handling 100M+ monthly requests.
- Shifted security left: every build scanned, signed and policy-checked before deploy, with runtime threat detection in production.
- Reduced hosting spend 55% and accelerated releases 70% via zero-downtime migrations and keyless GitOps.

Skills

Cloud & Infrastructure	GCP (GKE Autopilot and Standard, Cloud SQL), AWS (EKS), Azure (AKS), DigitalOcean, Hetzner (Talos), Kubernetes, Helm, Terraform (reusable modules), Docker
CI/CD & GitOps	Argo CD, Flux, GitHub Actions, keyless delivery with Workload Identity Federation, Jenkins, Octopus Deploy, multi-cluster release orchestration
Security & DevSecOps	Secret scanning (TruffleHog, git-secrets), SAST (SonarQube), DAST (OWASP ZAP), SCA (Snyk), container scanning (Trivy, Clair), IaC scanning (Checkov, Terrascan), image signing (cosign), admission control (Kyverno, OPA Gatekeeper), runtime security (Falco), vulnerability remediation, OAuth 2.1 / OIDC SSO, External Secrets, Cloudflare, mutual TLS, SOC 2 Type II, ISO 27001
AI Infrastructure & MLOps	GPU node pools, Intel TDX confidential compute and remote attestation, AI sandboxes, LLM gateways (Rust), MCP, Airflow, Prefect, PyTorch
Observability & SRE	OpenTelemetry, Prometheus, Grafana (HA), Loki, Tempo, ClickHouse, Elastic APM, Istio service mesh, SLOs, alert-noise reduction, incident response
FinOps	Right-sizing, scale-to-zero, Spot and Reserved Instances, GCP/AWS Savings Plans, FinOps dashboards
Programming	Python, Go, Rust, TypeScript, Bash
Data & Messaging	PostgreSQL (Cloud SQL), MongoDB, Kafka, Elasticsearch, Redis, RabbitMQ

Experience

Orgn.com — Gurugram, India
Founding Lead DevOps and MLOps Engineer — Feb 2025 – Present

- Architected dev, staging and production GKE across 10+ Kubernetes clusters with Argo CD, Kyverno, Grafana and protected Terraform delivery.
- Ran confidential AI workloads on 300+ Intel TDX GPU nodes: AI sandboxes, Rust LLM gateways serving 100M+ requests per month with usage metering, and remote attestation, all delivered through GitOps.
- Built an end-to-end DevSecOps pipeline that gates every release on secrets, code, dependency, image and IaC checks (TruffleHog, SonarQube, Snyk, Trivy, Checkov), signs images with cosign, enforces Kyverno and OPA admission, and watches runtime with Falco, driving remediation of critical and high vulnerabilities.
- Owned $70K–$90K in monthly cloud spend and the platform behind an engineering org of 50+.
- Migrated AWS workloads to GCP and DigitalOcean with zero downtime, cutting infrastructure cost by 55% through right-sizing and scale-to-zero.
- Automated keyless CI/CD from GitHub Actions into Argo CD, lifting release velocity by 70%.
- Held 99.95% uptime across 200+ services with Istio, Prometheus, Loki and OpenTelemetry, and cut alert noise by 60%.
- Moved the central OAuth 2.1 / OIDC single sign-on service to GKE and Cloud SQL with no downtime.

Infra360.io — Gurugram, India
Founding Lead DevOps Engineer — Sep 2022 – Feb 2025

- Led a team of 15+ DevOps engineers across 8 client accounts, raising delivery efficiency by 45%.
- Designed multi-region EKS and AKS clusters with Terraform, sustaining 99.99% uptime.
- Drove vulnerability management across client accounts with SAST, DAST and SCA, remediating critical and high findings and hardening cloud infrastructure.
- Integrated Istio and mutual TLS for multi-tenant clients handling 50M+ monthly requests.
- Engineered CI/CD with GitHub Actions and Jenkins, cutting rollback incidents by 40%.
- Unified OpenTelemetry and Elastic APM across 100+ microservices, halving debugging time.
- Executed cost-optimization roadmaps with executive leadership, saving $140K in annual overhead.

Zeta.tech Bangalore, India
Senior DevOps Engineer Apr 2021 – Sep 2022

- Managed 100+ microservices on EKS and AKS, raising service availability to 99.9%.
- Enforced OPA Gatekeeper and Kyverno policies, blocking 100% of non-compliant workloads.
- Implemented federated Prometheus and Grafana, reducing mean time to resolution by 35%.
- Accelerated data processing by 60% through Apache Flink and Airflow deployments.

To The New Greater Noida, India
DevOps Engineer Sep 2019 – Apr 2021

- Codified Azure infrastructure with Terraform, cutting environment setup time by 70%.
- Authored CI/CD workflows with Jenkins and Octopus Deploy, delivering 50% faster releases.
- Centralized Prometheus and Elasticsearch monitoring, lowering system downtime by 45%.

CloudInnov8 Technologies Gurugram, India
Cloud & DevOps Engineer Aug 2017 – Sep 2019

- Hardened AWS environments for 15 enterprise clients, passing 3 consecutive security audits.
- Administered 100+ AWS resources with auto-scaling, reducing service interruptions by 40%.
- Introduced self-service CI/CD with Jenkins and Terraform, eliminating 80% of manual steps.

Open Source

- **kubernetes-secure-mlops-pipeline**: every model that reaches production is reproducible, evaluated against the model it replaces, signed and verified at deploy time (Kubeflow Pipelines, MLflow, KServe, Kyverno, Argo CD, cosign, SLSA provenance, ML-BOM); drift alerts start retraining; 24 admission, network, monitoring and GitOps checks run on a real kind cluster in CI.
- **k8s-gpu-scheduler-benchmark**: compares the default kube-scheduler, Kueue, Volcano and NVIDIA KAI Scheduler on identical AI workload traces, in a simulation up to a 1,024-GPU fleet and on the real schedulers in CI. Finding: out of the box only KAI bin-packs GPUs; one setting fixed Volcano and made the same trace 26% faster and 26% cheaper per useful GPU-hour.
- **k8s-inference-routing-benchmark**: a Kubernetes Service against the llm-d Router in front of simulated vLLM pods on kind, eight runs per router. Finding: prefix-aware routing lifts cache hits from 64% to 80% and cuts median time to first token 27%, but load-aware routing keeps the best tail.
- **k8s-inference-autoscaling-benchmark**: the HorizontalPodAutoscaler's algorithm on four signals, checked against the real HPA, Prometheus and prometheus-adapter on kind. Finding: no default setup holds a 2-second first-token objective on a fast ramp; pod startup, not the target, is the constraint.
- **agentic-code-review-tool**: AI code review for the whole pull request: 128 rules across application, infrastructure and AI/agent code (OWASP LLM Top 10, MCP, agent skills), each swept over whole open-source repositories before release, plus existing scanners; an LLM agent confirms each finding and writes the fix. GitHub, GitLab, SARIF and JUnit reports.
- **ai-devsecops-cicd**: reference DevSecOps pipeline for LLM and agent services: OWASP LLM Top 10 tests, SBOM and AI-BOM, Trivy, keyless cosign signing, Kyverno admission.
- **llm-inference-platform**: GPU-backed LLM serving with vLLM-Metal, a Python streaming gateway, Prometheus/Grafana monitoring, reproducible benchmarks and Kubernetes templates.
- **jev-inference-router**: observable routing for local GPU inference with retrieval and Prometheus metrics, evaluated against rule-based and classifier baselines.
- **Upstream**: merged in Argo CD (OCI repositories over HTTP), MLflow (model downloads no longer break signature verification) and NVIDIA KAI Scheduler (GPU fractioning docs); in review at KServe (security fix for compromised root CAs), Tesla vehicle-command (P-256 key storage) and the xAI Python SDK.

Education

Bachelor of Technology, Computer Science – Rajasthan Technical University, Kota 2013 – 2017

Awards

Cost Optimization Champion (Orgn.com) • Spotlight Award for DevOps Excellence (Infra360.io) • Zeta Award for Excellence (Zeta.tech) • Outstanding Team Recognition (To The New) • Commitment to Ownership Award (CloudInnov8)