

GAHAN SARAIYA

☎ +91 7359508173 ◊ ✉ gahansaraiya@gmail.com ◊ in [linkedin.com/in/gahan-saraiya](https://www.linkedin.com/in/gahan-saraiya) ◊ 🐙 github.com/gahan9

PROFESSIONAL SUMMARY

Senior DevOps and MLOps Engineer with 8+ years of experience designing and managing scalable cloud infrastructure, CI/CD pipelines, and machine learning operations. Proven expertise in containerization, Infrastructure as Code (IaC), multi-cloud platforms (AWS, Azure, GCP), Kubernetes, and automated ML model deployment. Strong background in Python automation, Docker, Terraform, Jenkins/GitHub Actions, and observability (Prometheus, Grafana, Slurm). Experience managing AI/ML workloads on HPC GPUs, cluster telemetry, log shipping to S3, and Agentic AI knowledge-base systems at AMD India Pvt Ltd.

TECHNICAL SKILLS

Programming Languages	Python, Bash Shell Scripting, SQL, YAML, JSON, PowerShell
DevOps & CI/CD Tools	Jenkins, GitHub Actions, GitLab CI/CD, Terraform, Ansible, Docker, Docker Compose, Kubernetes (K8s), Helm, Git, TeamCity, JFrog Artifactory
Cloud Platforms	AWS (EC2, S3, CloudWatch, VPC, Lambda, CloudFront, ECR, ECS, EKS, SageMaker, IAM, Route53), Azure (DevOps, OpenAI, Container Registry), Cloud Storage, Cloud Functions), Digital Ocean
MLOps & ML Frameworks	PyTorch, TensorFlow, LangGraph, FAISS, MLflow, Model Serving, Model Deployment, Model Monitoring, AWS SageMaker, MLOps Pipelines, Model Versioning
Infrastructure & Monitoring	Infrastructure as Code (IaC), Container Orchestration, Prometheus, Grafana, Slurm, System Monitoring, Logging, Observability, Telemetry, Log Shipping to S3, Observability Agents, Auto-scaling, Load Balancing
Agentic AI & Knowledge Ops	Obsidian Knowledge Base, Agentic Memory Index, Skills Auto-Sync, LangGraph, FAISS, Claude, Cursor, LLM APIs
APIs & Services	RESTful APIs, Microservices Architecture, API Gateway, Service Mesh, Redfish API, Selenium Automation

PROFESSIONAL EXPERIENCE

Senior System Software Designer at AMD India Pvt Ltd	Mar 2024 - Present
- Manage cluster network observability and telemetry using Prometheus, Grafana, and Slurm; auto-deploy observability runners and agents across large GPU cluster fleets - Design Obsidian-based agentic memory network for agent memory index, knowledge base, and skills auto-sync (human-readable and agent-readable) - Automate secure telemetry log shipping to AWS S3 for cluster validation CI with IAM and network security controls - Architected and managed distributed workload execution pools across geographic regions with Jenkins CI/CD for build, test, and deployment - Led release management for ROCm builds across production, staging, and development environments with OS and toolchain matrix configurations - Designed Docker containerization with enterprise registry (JFrog Artifactory), optimizing image builds and deployment pipelines - Migrated platform provisioning to Infrastructure as Code (IaC) using Terraform for automated infrastructure deployment - Developed MLOps pipelines for AI/ML workload execution on HPC GPUs, integrating automated testing and deployment workflows	

- Conducted functional and performance benchmarking of HPC GPUs with test suites aligned to industry standards

Senior Firmware Development Engineer at Intel Technology India Pvt Ltd Oct 2023 - Mar 2024

- Designed and implemented platform provisioning and infrastructure automation for large-scale HPC product deployments, reducing manual setup time by 70%
- Architected Platform Orchestration Layer for Universal Scalable Firmware (USF), enabling automated configuration management and deployment workflows
- Built complete Infrastructure as Code (IaC) solution using Terraform for on-demand development environments on Kubernetes clusters, including node provisioning from scratch
- Implemented Kubernetes-based container orchestration for development environments, enabling rapid scaling and resource optimization with automated node management, pod scheduling, and resource allocation policies
- Integrated Azure OpenAI services for automated code review, security vulnerability scanning, and code quality assessment, improving development productivity
- Integrated GitHub Actions and JIRA for end-to-end CI/CD pipeline management, implementing continuous deployment and automated testing through CloudBees Jenkins
- Managed hybrid Jenkins infrastructure with bare-metal and Docker containerized nodes, optimizing build performance and resource utilization

Firmware Development Engineer at Intel Technology India Pvt Ltd Jun 2020 - Sept 2023

- Led next-generation client platform architecture design at USF POL, owning domains for Automation Framework, Platform Provisioning, Applied Generative AI, and Quantum Firmware Design
- Managed end-to-end Software Development Life Cycle (SDLC), including licensing, legal compliance, IP planning, code security, and static code analysis integration
- Configured and maintained Jenkins CI/CD pipelines with security scanning tools (Snyk, Coverity), implementing automated security vulnerability detection in build processes
- Enhanced GitHub code review automation with Python packaging validation, linting (Bandit), and Flask-based Jenkins pipeline integration for improved code quality
- Implemented automated testing frameworks for firmware functional validation using TeamCity, reducing manual testing effort by 60%
- Built and maintained continuous integration pipelines for automated build, test, and deployment workflows
- Established DevSecOps practices by integrating security scanning, code analysis, and compliance checks into CI/CD pipelines
- Mentored team members on DevOps best practices, automation frameworks, and CI/CD pipeline development

Software Developer at Quixom Technology Jul 2017 - Apr 2018

- Developed SaaS-based employee management system with RESTful API architecture, implementing microservices design patterns
- Built real-time data processing pipelines using Python socket programming for continuous IoT device data ingestion and analytics
- Designed and deployed cross-platform streaming service with Django REST Framework (DRF) backend and web portal frontend
- Implemented automated data processing workflows for handling large-scale data streams from IoT devices

- Designed and deployed cloud-based data gathering and web scraping solutions on Google Cloud Platform (GCP) using Selenium automation
- Built scalable data processing pipelines for automated data collection, storage, and analysis
- Developed inventory management system using Django framework with automated workflows for stock management and reporting
- Implemented cloud infrastructure solutions for high-availability web applications

KEY PROJECTS

Cluster Management and Observability

2026 - Present

Cluster management and observability for large-scale GPU networks. Integrate Prometheus, Grafana, and Slurm telemetry stacks; auto-deploy observability runners and agents; automate secure log upload to AWS S3 for cluster validation CI.

Obsidian Agentic Memory Network

2026 - Present

Agentic design for agent memory index, knowledge base, and skills auto-sync optimized for human-readable and agent-readable consumption.

Agentic AI for Software Engineering Modernization

2025 - Present

Designed GenAI engineering workflows and automated repository management agents. Built custom tool integrations and Git hooks compatible with Claude and Cursor. Engineered security compliance layers and token spend/quota restriction mechanisms.

Automated Context-Aware Task Routing & Triage with LangGraph

2025 - Present

Developed context-aware ticket triage and resolution pipelines. Built graph-based LLM routing with LangGraph, semantic classification using FAISS, and connectors to GitHub, Jira, and historical system data. Integrated Vertex AI, OpenAI, and local LLM endpoints.

GPU Performance Analysis and MLOps Pipeline for Data Center Workloads

2024 - Present

Designed and implemented end-to-end MLOps pipelines for automated ML model training, deployment, and monitoring on GPU clusters. Built Docker-based containerization strategies and Terraform IaC solutions for automated GPU cluster provisioning. Engineered comprehensive monitoring solutions for GPU telemetry using System Management Interface (SMI). Integrated automated testing, model validation, and deployment workflows into Jenkins pipelines for continuous ML model delivery.

MLOps Pipeline for Generative AI Code Review Automation

2023 - 2024

Built automated MLOps pipelines integrating Azure OpenAI (GPT-3.5, GPT-4) for intelligent code analysis and review automation. Integrated AI-powered code review into Jenkins CI/CD pipelines, enabling automated code quality checks and security vulnerability detection. Enhanced security posture by automating code security scanning, vulnerability assessment, and compliance checking, reducing manual code review time by 50%.

Infrastructure Automation and Platform Provisioning

2023 - 2024

Developed Terraform-based IaC solutions for automated infrastructure provisioning and configuration management. Built end-to-end Jenkins pipelines for automated remote BMC and BIOS firmware flashing, reducing deployment time by 75%. Integrated JFrog Artifactory with Jenkins pipelines for automated OS image deployment and version management. Implemented automated system configuration using Redfish API for Baseboard Management Controller (BMC) operations.

DevOps Automation Framework for Firmware Development

2020 - Present

Designed and implemented fully automated system configuration framework, eliminating manual BIOS configuration intervention for thousands of devices. Established comprehensive DevSecOps practices including legal compliance automation, security scanning, licensing validation, and static code analysis in CI/CD pipelines. Built end-to-end continuous deployment and testing pipelines using Jenkins with tox configurations. Developed and maintained open-source automation tools: <https://github.com/intel/xml-cli>

EDUCATION

M.Tech Computer Science
Nirma University, GPA: 8.49

Jul 2018 - Jun 2020

B.E. in Computer Science and Engineering
Gujarat Technological University, CGPA: 8.08

Aug 2013 - Apr 2017

CERTIFICATIONS

2025	Essential Terraform in AWS - LinkedIn Learning
2025	AMD ROCm Star - Application Developer - AMD Adaptive Computing Training Team
2025	On-Device AI - Qualcomm & DeepLearning.ai
2021	Product Assurance and Security Yellow Belt - Software - Intel
2020	Machine Learning Specialization (5 courses) - Coursera (DeepLearning.ai)

ACHIEVEMENTS & CONTRIBUTIONS

- Top 1% recognition for AI use-case development (2026)
- Top 10 in AI Hackathon: Hack2Skill Prompt Wars by Google (In-person, Bangalore) (2026)
- Rank #3 in Virtual Prompt Wars Hackathon (2026)
- Presented Firmware Configuration at UEFI Developers Conference & Plugfest (2023)
- Tech Talk on Harnessing Multiprogramming across CPU, GPU and QPU (2023)
- Expert Lecture on CUDA Programming (2021)
- 4000+ reputation on Stack Overflow (2019-Present)
- Rank #1 for Python challenges on HackerRank (2018)
- Rank #112 from 200,000+ participants in Techgig's National Coding Contest (2018)

PUBLICATIONS & PRESENTATIONS

- [HIP Parallel Programming with ROCm](#) Lecture Series, YouTube (2025)
- [Beyond the Power — Data Center & BMC](#) Lecture Series, YouTube (2025)
- [System Design for Parallel Programming](#) Lecture Series, YouTube (2025)
- [Firmware Configuration – Past, Present, and Future](#), UEFI DevCon 2023