

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

Staff-level Software Engineer with 8+ years of expertise in High-Performance Computing (HPC), parallel programming, and distributed systems architecture. Proven track record designing scalable cluster architectures, health telemetry, and GPU-accelerated workloads for data center environments. Specialized in System Management Interface (SMI), workload orchestration, Prometheus/Grafana/Slurm observability, and Agentic AI knowledge bases. Led cross-functional teams developing firmware automation frameworks and platform provisioning systems. Strong background in CUDA, ROCm, HIP, and cloud-native infrastructure.

TECHNICAL SKILLS

Programming Languages	Python, C, C++, CUDA, HIP, Shell Scripting, SQL, YAML, LaTeX
Parallel & Distributed Computing	CUDA Programming, ROCm, HIP Programming, Multi-GPU Systems, Cluster Architecture Design, Workload Orchestration, Parallel Algorithm Design
Health Telemetry & Monitoring	System Management Interface (SMI), GPU Telemetry, Prometheus, Grafana, Slurm, Hardware Metrics, Performance Profiling, In-Band Monitoring, Observability Agents, Log Shipping to S3
ML/AI Frameworks & Agents	TensorFlow, PyTorch, LangGraph, FAISS, TensorRT, PyCUDA, OpenAI/Vertex AI, Obsidian Knowledge Base, Agentic Memory Index, Panda
System Software & Firmware	UEFI, EDK2, Firmware Architecture, Boot Optimization, RAS (Reliability, Availability, Serviceability), CPER, GPU Partitioning
DevOps & Infrastructure	Kubernetes, Docker, Docker Compose, Terraform, Jenkins, GitHub Actions, CI/CD Pipelines, Infrastructure as Code, Platform Provisioning, Cluster Valid
Cloud & Services	AWS (EC2, S3, Lambda, SageMaker, ECS, EKS), GCP, Azure, Vertex AI, RESTful APIs, JFrog Artifactory, TeamCity, JIRA

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 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. - Architected in-band health telemetry for GPU clusters using System Management Interface (SMI), monitoring temperature, power, memory, and compute metrics. - Designed dynamic workload orchestration for multi-GPU environments with load balancing based on real-time hardware telemetry and thermal constraints. - Led performance benchmarking for HPC GPUs with test suites covering TensorFlow, PyTorch, and GEMM operations. - Built parallel programming solutions using ROCm and HIP for multi-GPU workload distribution and cluster utilization. - Implemented RAS injection mechanisms for MI2XX and MI3XX platforms with health monitoring and recovery procedures. - Collaborated with cross-functional teams on system architecture and Agentic AI tooling for next-generation HPC platforms.	

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

- Architected platform provisioning infrastructure for large-scale HPC product automation, designing cluster-based deployment systems that reduced manual configuration time by 70% through automated provisioning workflows.
- Led Platform Orchestration Layer (POL) architecture design for Universal Scalable Firmware (USF), establishing scalable firmware configuration management framework supporting heterogeneous computing platforms.
- Designed and implemented Terraform-based infrastructure automation for Kubernetes cluster environments, enabling rapid deployment of development and testing platforms with infrastructure-as-code practices.
- Integrated Azure OpenAI services into firmware development workflow, implementing automated code review, security analysis, and enhancement suggestions that accelerated development cycles.
- Architected CI/CD pipeline infrastructure using GitHub Actions and CloudBees Jenkins, supporting both bare-metal and containerized deployment targets for continuous integration and testing.
- Optimized firmware boot-time performance for client platforms that reduced boot latency%.

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

- Led architecture design for next-generation client platform at USF Platform Orchestration Layer, owning domain responsibilities for Automation Framework, Platform Provisioning, and Generative AI integration for firmware development.
- Managed end-to-end Software Development Lifecycle (SDLC) including licensing, legal compliance, IP planning, code security, and static code analysis, establishing processes that ensured compliance across all firmware releases.
- Architected and implemented automated security scanning infrastructure using Jenkins, integrating Snyk and Coverity for continuous security analysis and vulnerability detection in firmware codebase.
- Developed automated code quality pipeline using GitHub hooks, implementing Python packaging validation, linting with Bandit, and security analysis that improved code quality metrics by 60%.
- Designed and implemented automated firmware validation framework using TeamCity, enabling continuous functional testing and reducing manual testing overhead by 60%.
- Mentored junior engineers and conducted technical knowledge-sharing sessions on firmware architecture, parallel programming concepts, and system design principles.

Firmware Engineer Intern at Intel Technology India Pvt Ltd Jun 2019 - May 2020

- Developed BIOS firmware for client products using EDK2 and UEFI frameworks, contributing to debugging solutions and OpenBoard Package integration.

Teaching Assistantship at Nirma University Aug 2018 - May 2019

- Data Mining and Visualization

Software Developer at Quixom Technology Jul 2017 - Apr 2018

- SaaS-Employee management system (chat, surveys, benchmark, analytics...) and Restful API Development
- Socket programming in python to receive continuous data from IoT device and display chunks of data meaningfully
- Streaming service with Web-Portal, DRF and Kodi across platforms

Freelance Full-stack Developer 2015 - Apr 2019

- Cloud solution for Data gathering and webscraping with selenium and GCP.
- Designing template and banners for High resolution prints

- Inventory management system for Book store with Django.

PROJECTS

Cluster Management and Observability

2025 - Present

Description: Cluster management and observability for large-scale GPU networks.

- Integrate Prometheus, Grafana, and Slurm telemetry for cluster network monitoring.
- Auto-deploy observability runners and agents across large cluster fleets.
- Automate secure log upload and telemetry shipping to AWS S3 with cluster validation CI.

Obsidian Agentic Memory Network

2025 - Present

Description: Agentic design for memory index, knowledge base, and skills sync for humans and agents.

- Deliver Obsidian-based agentic memory network with indexed knowledge base and auto-synced skills.

Agentic AI for Software Engineering Modernization

2025 - Present

Description: Modernized enterprise repository workflows, code reviews, and project management with multi-agent GenAI systems.

- Developed tool-compatible custom skills, automated Git hooks, and multi-modal agents leveraging Claude, Cursor, and OpenAI APIs.
- Engineered security compliance, sandboxed execution verification, and LLM token spend/quota controls.

Context-Aware Task Routing & Triage with LangGraph

2025 - Present

Description: End-to-end context-aware ticket/bug classification and routing engine.

- Built graph-based LLM logic using LangGraph with FAISS vector databases and specialized embeddings.
- Integrated GitHub, Jira, and documentation history with OpenAI, Vertex AI, and local LLMs.
- Formulated a persona-based LLM counselor model behind a quality matrix verification gate.

GPU Cluster Performance Analysis & Health Telemetry System

2024 - Present

Description: Enterprise-grade solution for in-band profiling and health monitoring of GPU clusters in data center environments

- **Cluster Health Telemetry Architecture:** Architected and implemented comprehensive in-band profiling system using System Management Interface (SMI) for real-time GPU cluster health monitoring, capturing telemetry data across temperature, power, memory, and compute utilization metrics.
- **Dynamic Cluster Workload Orchestration:** Designed intelligent workload scheduling system that dynamically adapts to real-time cluster resource availability, implementing load balancing algorithms that optimize multi-GPU workload distribution and improve overall cluster efficiency.
- **Distributed Health Monitoring Infrastructure:** Engineered scalable telemetry collection framework supporting hundreds of GPU nodes, implementing in-band monitoring channels that minimize performance overhead while providing comprehensive hardware metric visibility.
- **Performance Profiling & Analysis:** Developed advanced profiling infrastructure capturing kernel-level performance metrics, call stack timings, and execution patterns to identify bottlenecks and optimize parallel program performance across cluster deployments.
- **Unified Analytics Platform:** Innovated methodology to aggregate and correlate multi-dimensional metrics (thermal, power, utilization, kernel timings) into unified health scores, enabling predictive maintenance and proactive cluster management.

Platform Orchestration & Cluster Provisioning System

2023 - 2024

Description: Automated platform provisioning infrastructure for large-scale HPC cluster deployment

- Architected cluster provisioning framework supporting remote BMC and BIOS firmware flashing across distributed server nodes, enabling zero-touch deployment for pre-production reference boards.

- Designed automated OS deployment pipeline integrated with Jenkins, supporting artifact management through JFrog Artifactory and RedFish API-based system configuration.
- Implemented infrastructure-as-code practices using Terraform for cluster environment provisioning, reducing deployment time from days to hours.

Universal Scalable Firmware (USF) - Platform Orchestration Layer

2022 - 2024

Description: Architecture design of scalable firmware configuration management for heterogeneous computing platforms

- Led architecture design of Platform Orchestration Layer enabling unified firmware configuration across diverse hardware platforms, contributing to UEFI industry standard development.
- Reference: [Firmware Configuration – Past, Present, and Future](#) (UEFI DevCon 2023)

Firmware Automation Framework

2020 - Present

Description: Enterprise automation framework for large-scale firmware development and validation

- Architected end-to-end automation framework supporting thousands of silicon devices, implementing zero-touch BIOS configuration and automated validation workflows.
- Established SDLC processes including security scanning, static code analysis, legal compliance, and licensing management integrated into CI/CD pipelines.
- Designed continuous deployment infrastructure using Jenkins and GitHub Actions, managing release artifacts through JFrog Artifactory with automated testing and validation.
- Open Source: <https://github.com/intel/xml-cli>

EDUCATION

M.Tech Computer Science

Jul 2018 - Jun 2020

Nirma University, GPA: 8.49

B.E. in Computer Science and Engineering

Aug 2013 - Apr 2017 Gujarat Technological University,

CGPA: 8.08

CERTIFICATIONS

2025	AMD ROCm Star - Application Developer by AMD Adaptive Computing Training Team
2025	On-Device AI by Qualcomm & DeepLearning.ai
2025	Essential Terraform in AWS
2021	Product Assurance and Security Yellow Belt - Software
2020	Natural Language Processing with Classification and Vector Spaces Coursera S9DGGMCDBGQ
2020	Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization Coursera SMWQJKJYJW6G
2020	Structuring Machine Learning Projects - Coursera SEJML7H3TN5Z
2020	Neural Networks and Deep Learning - Coursera S2TETZHTLUSL
2020	Machine Learning - Coursera XVNXNPPZ3JMA

ACHIEVEMENTS

2026	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
2023	Firmware Configuration Presentation at UEFI Developers Conference & Plugfest
2023	Tech Talk on Harnessing Multiprogramming across CPU, GPU and QPU (Quantum Processing Unit)
2021	Expert Lecture on CUDA Programming
2019-Present	4000+ reputation on Stackoverflow
2017-Present	Active volunteer on Peer Learning, Cultural and team building activity
2018	Rank #1 for Python challenges on hackerrank
2018	Rank #112 from 2 lacs+ participants in Techgig's National Coding Contest 2018

PUBLICATIONS

• 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 . Presentation, UEFI DevCon 2023	2023