

GAHAN SARAIYA

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

OBJECTIVE

High Performance Computing (HPC) Development Engineer with 9+ years of Research and Development experience in system software, GPU computing, and distributed systems architecture. Currently at AMD India Pvt Ltd., specializing in performance optimization, AI/ML workload execution on HPC accelerators (GPUs), and cluster observability (Prometheus, Grafana, Slurm). Deep expertise in Python, modern C++, and Bash for large-scale on-prem and cloud deployments. Strong background in operating systems internals, computer networks, Agentic AI knowledge bases, and high-performance applications across heterogeneous environments (CPUs, GPUs, accelerators). Proven track record deploying and optimizing TensorFlow and PyTorch on enterprise HPC clusters. Open-source contributor advancing parallel computing and distributed systems.

TECHNICAL SKILLS

- **Programming Languages:** C++, C, Python, Bash/Shell Scripting, SQL
- **HPC & GPU Computing:** ROCm, HIP Programming, CUDA, PyCUDA, OpenCL, MPI, OpenMP, GPU Profiling, Memory Optimization, Kernel Development
- **Deep Learning Frameworks:** TensorFlow, PyTorch, TensorRT, ONNX Runtime, Large-Scale Model Training, Distributed Training
- **Distributed Systems:** CPU/GPU/Accelerator Architecture, Parallel Programming, High-Performance Networks, InfiniBand, RDMA, MPI
- **Operating Systems:** Linux Internals, Kernel Development, UEFI/BIOS Firmware, System Performance Tuning, Memory Management
- **Cloud & Infrastructure:** AWS (EC2, S3, SageMaker), GCP, Azure, Kubernetes, Docker, Terraform, Large-Scale On-Prem HPC Clusters, Cloud Security Policies
- **CI/CD & DevOps:** Jenkins, GitHub Actions, TeamCity, JFrog Artifactory, Cluster Validation CI
- **Observability & Cluster Ops:** Prometheus, Grafana, Slurm, Telemetry, Log Shipping to S3, Observability Agents, Cluster Monitoring
- **Agentic AI:** Obsidian Knowledge Base, Agentic Memory Index, LangGraph, FAISS, LLM APIs (OpenAI, Vertex AI)
- **System Level IP:** RAS (Reliability, Availability, Serviceability), Compute Partitioning, Boot-time Optimization, CPER, SMI

EXPERIENCE

Senior System Software Designer at AMD India Pvt Ltd

Mar 2024 - Present

- **Cluster Observability:** Manage cluster network observability and telemetry using Prometheus, Grafana, and Slurm; auto-deploy observability runners and agents for large GPU fleets.
- **Agentic Memory Network:** Deliver Obsidian agentic design for agent memory index, knowledge base, and skills auto-sync (human-readable and agent-readable).
- **Telemetry Pipeline:** Automate secure log upload and cluster validation telemetry shipping to AWS S3.
- **R&D Leadership:** Research and development of high-performance computing solutions for next-generation AMD GPU accelerators (MI2XX, MI3XX series).
- **Performance Engineering:** GPU benchmarking frameworks using **modern C++** and **Python** for functional and performance validation of HPC devices.
- **Deep Learning Framework Optimization:** Deploy and optimize **PyTorch** workloads on large-scale GPU clusters, including GEMM operations and distributed training pipelines.
- **System-Level Optimization:** Load-balanced GPU performance criteria with thermal management, power profiling, memory optimization, and core utilization metrics.
- **HIP/ROCm Development:** Firmware capabilities including compute partitioning, HIP kernel profiling, RAS injection, and parallel execution strategies.

- Collaborate with cross-functional R&D teams on HPC solutions and Agentic AI tooling.

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

- **HPC Infrastructure R&D:** Platform provisioning and infrastructure design for mass automation of high-performance computing products.
- **System Architecture Design:** Led Platform Orchestration Layer architecture for Universal Scalable Firmware (USF) using **modern C++** and **Python**.
- **AI/ML Integration:** Implemented Azure OpenAI Generative AI solutions for automated code review and security analysis.
- **Large-Scale Infrastructure:** Designed CI/CD pipelines with GitHub Actions and Jenkins for distributed build systems across baremetal and containerized environments.
- **Operating System Optimization:** Enhanced firmware-based boot-time performance through deep OS-level optimizations.

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

- **R&D Technical Lead:** Spearheaded next-generation platform architecture design, owning domains including Automation Framework, Platform Provisioning, and Quantum Firmware Design.
- **C++ and Python Development:** Developed high-performance firmware validation tools using **Python scripting** for large-scale automated testing.
- **Distributed Systems:** Managed end-to-end SDLC for distributed firmware systems including security analysis (Snyk, Coverity) and static code analysis.
- **Scripting & Automation:** Developed **Bash** and **Python** scripts for automated testing pipelines, GitHub hooks, and CI/CD integration with Jenkins/TeamCity.
- **Operating Systems Expertise:** Deep involvement in UEFI/BIOS firmware development and system-level debugging.
- Mentoring junior engineers on HPC concepts, parallel programming, and modern C++ best practices.

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

- **Low-Level Systems Development:** BIOS/UEFI Firmware development using EDK2 framework with **C** and **C++** for client products.
- **Operating Systems:** Deep understanding of boot processes, memory management, and hardware abstraction layers.

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 2026 - Present

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

Grafana, and Shurm telemetry; auto-deploy observability runners and agents; automate secure log shipping to AWS S3 for cluster validation CI.

Obsidian Agentic Memory Network 2026 - Present

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

Agentic AI for Software Engineering Modernization 2025 - Present

Description: Modernized software engineering, repository management, and code review processes using Agentic AI and GenAI (Claude, Cursor). Developed multi-tool compatible skills, agents, and hooks with security compliance and token spend restrictions.

Context-Aware Task Routing & Resolution with LangGraph 2025 - Present

Description: Led a team and managed sprints for context-aware task/issue disposal using LangGraph. Implemented smart routing via LLMs and semantic FAISS, integrating OpenAI, Vertex AI, and local LLMs. Built data source connectors for GitHub, Jira, and historical data, alongside a persona-based counselor model.

Prompt Engineering and Generative AI for firmware stack 2023 - 2024

Description: Solution design leveraging Azure LLM OpenAI services (GPT3.5 & GPT4) with automated task summarization and advanced security analytics.

Platform Provisioning with Pre-Production Reference Board 2023 - 2024

Description: Platform Provisioning with automating most common features for pre-production server processor such as:

- Remote BMC and BIOS Firmware Flashing
- OS Deployment from artifactory through Jenkins Pipeline
- Boot to EFI Shell and OS
- Configuration of Initial system setup with Jenkins pipeline steps with Baseboard management controller Red-Fish API

Universal Scalable Firmware 2022 - 2024

Description: Architecture design of Platform orchestration layer for Firmware Configuration

- Motivation of workgroup - [Firmware Configuration – Past, Present, and Future](#)
- [Firmware Configuration](#)

Firmware Automation Framework 2020 - Present

Description: Lead SDL for an automation foundation framework for high scaled volume of silicon devices with edk2 UEFI firmware;

- Seamless configuration of system without any manual intervention for BIOS configuration.
- SDL activities like legal compliance, security, licensing, static code analysis
- Continuous deployment and test with jenkins and tox configurations
- Release Artifact management to JFrog Artifactory using GitHub Actions
- <https://github.com/intel/xml-cli>

Multiview Video Summarization 2019 - 2020

Description: Summarizing video of multiple view angles of egocentric location in to one to generate effective video summary, minimizing overhead to explore multiple cam feed to discover accidents or anomalies.

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 |