Rahul Surya

Edinburgh, UK · +44 7742 977485 · rahulsurya021@outlook.com · LinkedIn · GitHub · Portfolio

SUMMARY

HPC and Research Software Engineer with hands-on experience on the UK's ARCHER2 national supercomputer, currently completing an MSc in High-Performance Computing with Data Science at EPCC, University of Edinburgh. Proficient in MPI, OpenMP, CUDA, C/C++, and Python for developing and optimising parallel scientific applications. Additional ML research experience at ISRO on large-scale geospatial datasets. Seeking HPC Engineer and Research Software Engineer roles in the UK from August 2026.

SKILLS

- **Parallel Programming:** MPI (point-to-point, collectives, domain decomposition), OpenMP, CUDA, Pthreads
- **Languages:** C, C++, Python, Go, SQL, Fortran basics
- **HPC Tools:** ARCHER2 (Cray EX), Slurm, Performance Profiling (gprof, Valgrind), CMake
- **ML & Data:** PyTorch, TensorFlow, Apache Spark, PostgreSQL, TimescaleDB
- **DevOps:** Docker, Kubernetes, Prometheus, Git, CI/CD, Microsoft Azure

PROFESSIONAL EXPERIENCE

Machine Learning Engineer Intern · ISRO – National Remote Sensing Centre Oct 2023 – Jan 2024

- Optimised deep learning pipelines processing **500GB+ of WRF simulation data**, reducing model training time by 35% through efficient parallel data loading and VAE-based compression
- Developed a **ConvLSTM-Seq2Seq** geospatial prediction research system in PyTorch achieving 92% forecast accuracy
- Integrated inference models with REST APIs serving 15+ researchers on a daily basis

Data Science Intern · Clustor Computing Jun 2023 – Sep 2023

- Built predictive ML models for quantitative finance on large historical datasets and produced Tableau dashboards for client reporting

Software Engineer Intern · Develoscope Software Solutions Jun 2023 – Sep 2023 *(concurrent)*

- Delivered four production web applications, contributing to the full software development lifecycle

PROJECTS

Dissertation: ML-Based Memory Leak Detection in Containerised Environments · Python, Prometheus, eBPF, TimescaleDB, Kubernetes

Building an early-detection system for memory leaks using ML on time-series metrics collected via Prometheus and eBPF from containerised workloads on an EIDF HPC cluster.

Massively Parallel Cellular Automata Simulation · C, MPI, OpenMP

Designed and implemented a **2D domain-decomposed** parallel simulation on the ARCHER2 supercomputer, achieving **85%+ parallel efficiency** across 64 CPU cores. Applied halo-exchange communication patterns and cache-optimised memory access to minimise MPI synchronisation overhead.

LLM Inference Engine · Python, CUDA, PyTorch, Prometheus

Engineered GPU-accelerated inference server with CUDA memory management and continuous batching to maximise throughput. Applied quantisation-aware techniques and kernel fusion to reduce inference latency.

Distributed ETL & Caching Pipeline · Python, PostgreSQL, Redis

Architected a high-throughput data pipeline with Redis caching, reducing query latency significantly on large financial datasets.

EDUCATION

University of Edinburgh – *MSc High-Performance Computing with Data Science* Sep 2025 – Aug 2026

Coursework: Message Passing Programming (MPI), Threaded Programming (OpenMP), GPU Programming (CUDA), High Performance Data Analytics, Reinforcement Learning

G H Raisoni Institute of Engineering – *B.Tech Artificial Intelligence (First Class with Distinction)* Aug 2020 – Jul 2024

CERTIFICATIONS

Microsoft: Azure DevOps Engineer Expert · Azure Developer Associate · Fabric Analytics Engineer · AI Fundamentals · Data Fundamentals **Oracle:** OCI Generative AI Professional **Coursera:** Machine Learning – Stanford Online · Google Cybersecurity Professional