Ayush Gala

Raleigh, NC | (619) 382-0493 | ayushgala2@gmail.com | github.com/Ayush-Gala | linkedin.com/in/ayush-gala

EDUCATION

North Carolina State University **Raleigh, NC**
Master of Computer Science **May 2026**
Courses: Parallel Systems, Parallel Architecture, Distributed Systems, Operating Systems, Cloud Computing **GPA: 4.0**

Savitribai Phule Pune University **Pune, India**
Bachelor of Engineering (Honors) in Computer Engineering **May 2024**
Courses: Algorithms, Computer Networks and Security, HPC, System Programming, Software Engineering **GPA: 3.76**

EXPERIENCE

Software Engineering Intern | FlexGen - Durham, NC **May 2025 - Aug 2025**

- Built fleet-scale hardware-agnostic BESS simulations for electric grids in C++. Designed ephemeral cloud environments using Terraform and AWS to create 'one-click' sandboxes for sales demos and test teams, unlocking $100,000 in annual savings.
- Architected a parallel testing framework (FlexTest) using task decomposition and dynamic load balancing to synchronize concurrent validation workloads. Optimized loop structures for CPU cache locality, speeding up software release time by 87%.

Research and Development Intern | Reverie Language Technologies - Bengaluru, India **Aug 2023 - Mar 2024**

- Accelerated Deep Learning inference pipelines by optimizing CUDA kernels to reduce GPU memory footprint and minimize PCIe transfer latency, enabling real-time speech processing on resource-constrained GPU hardware.
- Conducted performance analysis using NVIDIA Nsight Systems and Compute profiling tools, identifying and resolving system-level bottlenecks in memory throughput and kernel warp occupancy.
- Built a high-throughput data ingestion pipeline to process and validate 280+ hours of 16kHz audio data for model training.

Software Engineering Intern | MasterCard - Pune, India **May 2023 - Jul 2023**

- Built a data-parsing engine in Python to extract regression test analytics from Jenkins CI/CD logs to a centralized team dashboard.
- Streamlined DevOps lifecycle for 40 distributed teams across 6 countries, lowering Mean Time to Resolution (MTTR) by 26%.

PROJECTS

High-performance Sparse Matrix Kernels | C, CUDA, OpenMP, MPI, Slurm, Apptainer, Nsight (Systems/Compute)

- Built and benchmarked sparse matrix multiplication kernels (COO SpMV) in C using OpenMP and distributed MPI (Bcast, Scatterv, Reduce) on Slurm/PBS clusters, achieving **9-13× speedup** on 8 nodes while profiling GFLOP/s and memory bandwidth.
- Implemented GPU SpMV in CUDA (atomicAdd accumulation, tuned 256-thread blocks, CUDA-event profiling) and extended to multi-node MPI+CUDA, delivering up to **127× end-to-end speedup** vs serial and sustaining 415 GB/s device bandwidth.
- Designed a reproducible HPC benchmarking framework (warm-ups, adaptive iteration counts, multi-trial averaging, correctness validation via MPI reductions), reducing timing variance from ~2× cold-start swings to <5% CV for reliable scaling analysis.
- Implemented communication-aware SUMMA GEMM in MPI (Stationary-A/B/C) using 2D Cartesian topologies, MPI subarray datatypes, and Reduce_scatter; strong-scaled to 64 ranks, and quantified compute vs communication limits across grid sizes.

Cellular Automata Accelerators | C++, CUDA, OpenMP, Slurm, Apptainer, Nsight (Systems/Compute)

- Designed a HPC simulation and benchmarking suite for Conway's Game of Life, comparing CPU (serial, OpenMP) and GPU (CUDA) implementations on large grids (10000×10000, 557 generations) to study memory-bound cellular automata workloads.
- Implemented a CUDA kernel with double buffering, ghost-cell boundary handling, branchless update logic, and coalesced memory access; achieved **111× speedup** over serial and **86.4% parallel efficiency** on an NVIDIA A100 GPU.
- Optimized data movement using pinned host memory, CUDA streams, and shared-memory caching. Profiled scaling behavior across serial, OpenMP, and CUDA variants, identifying and solving for memory bandwidth as the primary bottleneck.
- Orchestrated workloads via Slurm batch jobs on a cluster equipped with Mellanox ConnectX-6 SmartNICs, utilizing Apptainer (Singularity) for building reproducible containerized runtime environments.

Distributed Key-Value Store | Go, Raft, gRPC, Docker, Prometheus, Makefile

- Engineered a fault-tolerant storage backend implementing Log-Structured Merge (LSM) trees (WAL, MemTables, SSTables) and Raft consensus to ensure linearizable consistency and crash-safe recovery across sharded replica groups.
- Utilized Prometheus instrumentation to benchmark replication lag, I/O throughput, and latency under high-concurrency stress.

Unix Filesystem Defragmenter | C, Linux System Calls, Qemu, Makefile, Bash, GitHub Actions

- Engineered a userspace utility to parse Superblock and Inode structures, optimizing disk layout by restructuring non-contiguous data blocks into contiguous clusters to reduce seek latency. Interfaced directly with block devices using low-level Linux system calls, validating data integrity against reference images in a QEMU-emulated environment.

SKILLS

Programming: C, C++, Python, CUDA, Verilog, Bash, MPI (OpenMPI, MVAPICH2), OpenMP, SIMD, Pthreads
HPC and Systems: Slurm, Lustre, GPFS, BeeGFS, RDMA, InfiniBand, Apptainer, Kubernetes, Linux (RHEL, Rocky)
Infrastructure and Tools: GDB, Valgrind, Nsight Compute/Systems, AWS, Terraform (IaC), Prometheus, Make, Git, Perf

RESEARCH EXPERIENCE

- Contextual flow of information using BLE proximity detection to enhance the tourism experience. Link
- Advancing CUDA Kernels for High Performance and Memory-Efficient Cellular Automata Processing on GPUs. Link
- Impact of Hybrid Sampling on ML-based Network Intrusion Detection Systems. Link
- Exploring Speech Enhancement Models for Indic Language in Noisy Environments: A Benchmarking Study. Link
- A survey on enhancing deep learning based speech recognition systems using noise suppression techniques. Link
- Automating Her Own Job: An Ethics Case Study in Software Engineering. Link

ACHIEVEMENTS

- **Runner Up at Gaming Frontier Challenge:** A national level Game Jam hosted by IIT Madras for my game ‘Morphogen’.
- **State Karate Gold Medalist:** Champion of Mumbai. The only one to win a medal for my school. 4th place at Nationals.
- **Best Speaker - Debate:** Defended the motion ‘Gandhian principles are still valid in today’s competitive environment.’
- **Best Editor at Pictoreal:** Awarded for my contributions to the annual college magazine Pictoreal Volume 24.
- **Runner Up at Xenathon:** Awarded for building a proximity ticketing IoT platform recognized by the Computer Society of India.

VOLUNTEERING

- **Disability Resource Office -** Promoting academic inclusivity by delivering prompt and courteous support to students and faculty.
- **Hack_NCState Planning Committee -** Sponsorship coordinator responsible for raising funds for the event. Led successful partnerships with sponsors like Fidelity, AWS, Cisco, NetApp, NordVPN, and Monster.
- **ACM Student Chapter (PICT) -** Promoted student engagement by conducting technical workshops in DevOps, C++, Machine Learning, Linux Development, etc. We won the ‘ACM Best International Chapter Website Award’.
- **University Magazine Secretary (Pictoreal) -** Led a team of 150+ members to publish the yearbook and conduct a plethora of events under the ’Tech for Good’ initiative. Organized campus-wide blood donation camps with 580+ total blood donors.
- **Marketing Lead (GameDevUtopia) -** I taught students game development using Lua and Unity. GDU won 4 National Game Jams during my tenure. Managed sponsor relations for the annual event ‘Glitched’.