Harshith Kantamneni

kantamneniharshith@gmail.com | +1 (414) 916-5799 | linkedin.com/in/hk4231 | github.com/Drogon4231
Madison, WI

M.S. ECE student specializing in CUDA Kernel Optimization, GPU Memory Systems, and Performance Engineering

Objective

- Seeking roles in **GPU Performance Engineering**, **CUDA Optimization**, or **HPC Software**, focusing on kernel tuning, memory hierarchy analysis, and large-scale benchmarking.

Skills

- **Languages:** C, C++, Python, CUDA
- **Tools: CUDA Toolkit**, **Nsight Compute (CLI)**, PyTorch, Slurm, **gem5**, McPAT, Git
- **Domains: GPU Performance Optimization**, **High-Performance Computing**, **Computer Architecture**, ML for Systems

Education

University of Wisconsin-Madison — *Madison, USA*
M.S. Electrical and Computer Engineering — Sep 2024 – Dec 2025

- Coursework: High Performance Computing, Advanced Computer Architecture, Machine Learning, Fault-Tolerant Computing

Vellore Institute of Technology — *Amaravati, India*
B.Tech in Electronics and Communication Engineering — 2020 – 2024

Projects

OpenGPUPerf: CUDA Kernel Performance Workbench — Sep 2025 – Present
*C++, **CUDA**, Python*

- Implemented optimized **CUDA** kernels for **GEMM** and **reductions** using **shared-memory tiling**, **warp-shuffle primitives**, and **WMMA tensor cores**.
- Profiled kernels using **Nsight Compute** to analyze **warp stalls**, **memory coalescing**, **achieved occupancy**, and **DRAM throughput**.
- Developing an auto-tuning engine to sweep **block sizes**, **tile sizes**, and **unroll factors** to maximize SM utilization.
- Built reusable GPU benchmarking harness for throughput tracking and roofline-style performance comparison.

ML-Guided CUDA Kernel Optimizer — Jan 2025 – May 2025
*Python, PyTorch, **CUDA**, Slurm* — GitHub

- Trained a model to predict **CUDA grid/block sizes** from workload features, reducing manual tuning by **>95%**.
- Achieved up to **30%** runtime improvement on matrix workloads using learned configurations.
- Benchmarked workloads on GPU clusters with **Slurm-driven pipelines** and CUDA event timing.

MI300X Memory-System Reverse Engineering & gem5 Calibration — Sep 2025 – Present
***CUDA**, **gem5**, Python*

- Developed **CUDA microbenchmarks** (pointer-chasing, stride sweeps) to measure **L2/L3/HBM latency and bandwidth** on MI300X.
- Analyzing GPU performance counters to model **chiplet-level cache behavior** and **memory access patterns**.
- Tuning **gem5 MI300X memory hierarchy** (cache sizes, line sizes, latencies) to reduce simulation–hardware mismatch.
- Created repeatable methodology for reverse-engineering **modern GPU memory systems**.

ML-Assisted Task Graph Partitioning — Jan 2025 – May 2025
Python, XGBoost, Slurm — GitHub

- Predicted optimal **TGD partitioning** strategies with **<5% error** using graph-structural features.
- Reduced design-space exploration time by **25%**.

ABFT-GEMM Reliability Study — Sep 2025 – present
Python, Slurm

- Performed **Monte-Carlo fault injection** on **GEMM kernels** to evaluate resilience vs execution-time overhead.
- Measured SDC behavior across matrix scales using HPC batch simulation.