Harshith Kantamneni

kantamneniharshith@gmail.com | +1 (414) 916-5799 | linkedin.com/in/hk4231 | Madison, WI 53726

M.S. ECE student specializing in GPU Performance, ML-Accelerated Systems, and Parallel Computing
Seeking full-time roles in: GPU Performance Engineering, Deep Learning Systems, ML + HW Co-Design

Skills

- **Languages:** C, C++, Python, CUDA
- **Tools:** PyTorch, Nsight, TensorRT, gem5, ModelSim, Git, MATLAB
- **Domains:** GPU Performance Optimization, Parallel Computing, ML for Hardware

Education

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

- Relevant Coursework: High Performance Computing, Machine Learning, Advanced Computer Architecture, Fault-Tolerant Computing, Introduction to Artificial Neural Networks

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

Certifications

- **NVIDIA Deep Learning Institute (DLI) – Getting Started with Accelerated Computing using CUDA C++**, 2025

Projects

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

- Built PyTorch model to predict CUDA launch parameters (grid/block sizes) from workload dimensions.
- Achieved **30% runtime speedup** through improved **SM utilization, warp occupancy, and memory coalescing**.
- Benchmarked kernels across varied matrix sizes, **evaluating runtime trends and scaling behavior.**.
- Validated prediction stability across GPU architectures ensuring performance portability.

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

- Engineered workload features from Task Dependency Graphs (TDGs) for partition prediction.
- Reduced configuration search time by **25% with <5% error**, accelerating HPC scheduling.
- Validated scalability across diverse matrix workloads and runtime environments.

5-Stage Pipelined RISC CPU (Course Project) — Aug 2024 – Dec 2024
Verilog, ModelSim

- Designed WISC-F24 ISA CPU with hazard detection, forwarding logic, and branch prediction.
- Verified with cycle-accurate testbench achieving full instruction coverage.
- Applied microarchitecture concepts relevant to GPU/CPU design pipelines.

Experience

Society for Space Education, Research and Development — *Bengaluru, India*
Research Intern — Nov 2021 – Jan 2022

- Developed ESP32 firmware for telemetry and GPS acquisition.
- Optimized boot latency by **30%**, improving startup performance in real-time environments.