Ram Reddy

248-894-7875 | ram.n.reddy15@gmail.com | linkedin.com/ram-n-reddy/ | github.com/ramnreddy15

Education

University of Illinois at Urbana-Champaign — Urbana, IL
Bachelor of Science in Computer Engineering — *Aug. 2023 – Dec. 2026*
GPA: 3.91
Honors: Dean's List 2023, 2024, 2025
Selected Coursework: Data Structures, Distributed Systems, Operating Systems, Parallel Programming, AI, Machine Learning, Computer Vision, Robotics, Analog Signal Processing

Experience

Systems Engineering Intern — June 2024 – Aug 2025
Arcfield — *Chantilly, VA*
- Returning intern continuing work on Retrieval-Augmented Generation (RAG) pipelines for LLM applications
- Segmented 100+ page documents into meaningful regions using LlamaIndex and unsupervised clustering
- Built spreadsheet parsers and glossary extractors for production environments and deployed through FastAPI
- Reduced processing time by 90%, accelerating pipelines for system modeling

Live Telemetry Team Lead — Aug 2023 – May 2024
Illini Electric Motorsports — *Urbana, IL*
- Led the design and development of a live telemetry system for a Formula SAE race car, overseeing multiple design reviews and iterations
- Managed integration of Controller Area Network (CAN) bus for in-car data acquisition and coordinated ESP32-based control of XBee Pro 900HP transceivers
- Built a scalable PyQt desktop viewer for visualizing real-time serial data, enabling live performance monitoring
- Achieved a 10x speedup in the viewer by optimizing core data structures and communication logic

Undergraduate Student Researcher — Jan 2025 – May 2025
KIMLAB (Kinetic Intelligent Machine LAB) — *Urbana, IL*
- Researched vision-free robotic grasping and 3D shape estimation using a soft tactile sensor hand driven by air pressure.
- Developed and tested ROS-based scripts to implement dynamic pressure thresholds for grasp detection and various hand poses.
- Used forward kinematics to estimate 3D object geometry from tactile data, improving recognition in occluded/low-light environments.

Projects

Parallelizing GPT Inferencing | *CUDA, Python, NVIDIA Nsight* — Jan 2025 – May 2025
- Coded GPT-2 from scratch and achieved performance of almost 70 tokens/sec which is 14x more tokens generated and 250x faster than baseline performance
- Developed custom matrix multiplication kernel achieving 42% of cuBLAS performance
- Automated kernel configuration tooling for tuning performance
- Implemented KV Caching improving tokens/sec by 57%

ChimpOS | *Operating Systems, RISC-V, C, Debugging, Unit Testing* — Jan 2025 – May 2025
- Built a Unix-like OS from scratch with RISC-V and C in a team of three; 30k+ lines of kernel/test code, capable of running Doom
- Implemented custom multi-level file system, caching, memory management, paging, drivers, processes, threading, and more
- Led development of custom shell with history, autocomplete, and text editor
- Selected by instructors as one of the top projects in the course

Technical Skills

Languages: C++, GO, Rust, CUDA, Python, C, Java, JavaScript, SQL, HTML/CSS, MATLAB
Frameworks: PyTorch, TensorFlow, Django, React, LlamaIndex, OpenCV
Technologies: Git, Linux, Docker, ROS, IoT, Sockets, MongoDB, PostgreSQL, REST API, Ansible, Raspberry Pi, Fusion 360