SRI HARSHA MUDUMBA

(515)-916-3011 ◇ srim@iastate.edu ◇ LinkedIn • GitHub • Iowa, IA

EDUCATION

Master's in Computer Engineering, Iowa State University, Iowa, IA — Aug 2023 - Aug 2025
GPA: 3.77/4.0
Relevant Coursework: Deep Learning, High-Performance Computing, Computer System Architecture, Hardware Design for ML, Advanced HPC for AI, Distributed Systems, Computational Perception.

B.Tech in Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, — July 2016 - Aug 2020
GPA: 7.5/10.0
Relevant Coursework: Data Structures and Algorithms, Computer Architecture, Computer Networks

TECHNICAL SKILLS

Programming Languages	Python, C++, C, SQL, Shell scripting
AI Frameworks	TensorFlow, PyTorch, ONNX, scikit-learn
HPC	OpenMP, MPI, Slurm
Databases	Oracle 11g/12c/19c, Oracle EBS
Other	Benchmarking, ML, Distributed Systems, LLMs, RAG Systems, RPC, Gem5, Linux

RELEVANT PROJECTS

LEXA: Lightweight Local Retrieval-Augmented Generation GitHub
Technologies: Python, FastAPI, TinyLlama, HuggingFace, SentenceTransformers, LLM

- Built a fully-local Retrieval-Augmented Generation (RAG) system using TinyLlama 1.1B with FastAPI and MiniLM embeddings.
- Introduced a cosine-similarity-based early exit strategy, reducing mean response time by 40% and GPU power usage by 30%.
- Benchmarked query latency for knowledge documents up to 2000 tokens using RTX 4060 GPU.

Benchmarking 1BitLLM with ONNX and IREE GitHub
Technologies: Python, ONNX, IREE, PyTorch

- Designed a performance benchmarking framework for 1-bit quantized LLMs across ONNX Runtime and IREE backends.
- Achieved 2.3x speedup over PyTorch baseline and reduced memory usage by 44% with IREE static compilation.
- Visualized inference trends across 4 batch sizes, showcasing consistent latency improvements across configurations.

SnaPEA Neural Network Inference Optimization GitHub
Technologies: Python, TensorFlow, PyTorch

- Implemented Sparse Neuron Activation Pruning and Early-Exit Architecture (SnaPEA) for lightweight inference.
- Reduced compute complexity by 68% using percentile thresholding of neuron activations.
- Combined sparse masking and early termination to achieve 74% reduction in FLOPs with just 1.5% accuracy drop on CIFAR-100.

Google File System (GFS) - Fault-Tolerant Distributed Storage GitHub
Technologies: C++, gRPC, Protocol Buffers, POSIX, Linux

- Designed and implemented a GFS-like distributed file system with Master, Chunkserver, and Client nodes communicating over gRPC.
- Achieved fault tolerance through periodic chunk heartbeat reports and master metadata checkpointing.
- Enabled high concurrency and consistent replication across 3 chunkservers using write-ahead logs and lease-based consistency.

Vehicle Number Plate Detection System GitHub
Technologies: C++, OpenCV, Python, MPI

- Developed a real-time vehicle license plate detection system using OpenCV, EasyOCR, and MPI, optimized for execution on Nova cluster nodes.
- Enabled parallel frame processing using 16 MPI processes, achieving a speedup of over 18x on 2.45-minute videos.
- Compiled processed frames into output videos and recorded vehicle counts with frame-by-frame annotations for traffic analytics.

PROFESSIONAL EXPERIENCE

Associate Database Administrator — Aug 2020 - Jul 2023
Cognizant Technology Solutions, Bengaluru, India

- Administered Oracle 11g/12c/19c databases, supporting Intuit Finance APP and BioMarin infrastructure.
- Automated ADOP patch processes, reducing patch time by 50% and optimizing Oracle concurrent manager performance.
- Migrated and cloned environments using RMAN, integrated Active Directory for role-based access, enforced audit logging for compliance, and developed Linux scripts for Oracle database cloning.