Dheemanth R Joshi

+1 (979) 575-4570 | dheerj188@gmail.com | LinkedIn | Google Scholar | GitHub

Research Interests

ML Systems & Heterogeneous Computing · **Algorithm–Hardware Co-design for AI** · **Probabilistic ML** (Gaussian Processes, Deep Kernel Learning, Bayesian Optimization) · **Efficient LLM Training & Inference** · **Computer Architecture** (Cache Hierarchies, Memory Systems, Accelerator Design)

Education

Texas A&M University, College Station, TX **GPA: –/–**
Master of Science (Thesis), Computer Engineering *Aug. 2025 – Aug. 2027*

- **Coursework:** Computer Architecture · Parallel Computing · Mathematical Methods for Signal Processing
- **Graduate Merit Scholarship**, Department of Electrical & Computer Engineering
- Thesis focus: Heterogeneous execution frameworks and probabilistic methods for efficient large-scale ML

PES University, Bangalore, India **GPA: 9.35 / 10.00**
Bachelor of Technology, Electronics and Communication Engineering *Dec. 2020 – Jun. 2024*

- **Prof. C. N. R. Rao Scholarship** — awarded to the top 5% of the department (GPA-based selection)
- **Undergraduate Teaching Assistant**, Artificial Neural Networks — conducted lab sessions and guided student projects on network design, backpropagation, and training dynamics
- **Intel Student Ambassador, oneAPI** — promoted heterogeneous programming models and parallel computing tools across the university developer community

Research Experience

Middleware & Runtime Systems Lab, Indian Institute of Science (IISc) Bangalore, India
Project Associate / Pre-doctoral Researcher (Advisor: Dr. Sathish Vadhiyar — Funded by Shell Global) *Jul. 2024 – Jul. 2025*

- **SkipPar — Hybrid CPU-GPU LLM Training Framework** *(HCW @ IPDPS 2026, Accepted, First Author):* Designed a co-execution paradigm where the GPU handles all forward and backward passes while the CPU concurrently runs parameter updates — directly eliminating the dominant idle-time bottleneck found in existing heterogeneous training frameworks.
- Implemented a **4-thread producer-consumer pipeline** coordinating GPU computation with CPU optimizer steps; integrated **PyTorch Distributed Data Parallel (DDP) hooks** to manage gradient aggregation and synchronize weight updates across multiple GPUs with no correctness penalty.
- Engineered a **dual-stage scheduling strategy**: (i) selective layer skipping for CPU updates during the backward pass, and (ii) gradient-norm-based filtering to suppress low-utility updates — reducing CPU overhead and GPU stalls while maintaining convergence behavior.
- Benchmarked on **NVIDIA A100 & H100 GPUs** using **LLaMA-2 (10B)** and **GPT-2 (9B)**; achieved **up to 17% reduction in end-to-end training time** against AAAI- and ICPP-published state-of-the-art heterogeneous training baselines.

Texas A&M Engineering Experiment Station (TEES) College Station, TX
Student Software Developer (Office of the Vice Chancellor of Engineering) *Jan. 2026 – Present*

- Building end-to-end **AI-driven education platforms** at institutional scale — full-stack development spanning data ingestion pipelines, LLM model integration, and scalable backend deployment.
- Designing **intelligent recommendation workflows** for student–advisor matching and academic analytics, incorporating retrieval-augmented generation and embedding-based similarity search.

- Delivering production-ready AI services with robust REST APIs serving large university-wide user bases with reliability and low latency.

Selected Projects

PAC-IPV: Prefetch-Aware Cache Replacement via Insertion & Promotion Vectors | C++, ZSim, PARSEC, SPEC C

- Extended an **RRIP-based LLC replacement policy** in ZSim with prefetch-aware RRPV value assignment, differentiating between demand-fetched and hardware-prefetched blocks to reduce cache pollution under streaming and instruction-intensive workloads. Evaluated across PARSEC and SPEC CPU2006; demonstrated **MPKI reduction and IPC improvement** over LRU, LFU, and SRRIP, with performance competitive to SHiP at lower complexity.
- Derived a **probabilistic Markov-chain analytical model** for the RRIP policy family; validated expected hit-rate predictions against simulation via **Monte Carlo analysis** (<1.5% error) — providing a mathematically grounded framework for reasoning about cache behavior under mixed workload patterns.

VAJRA: Heterogeneous GPU/FPGA Edge Cluster for AI Inference | C, CUDA, FPGA, Embedded Linux

- Architected a heterogeneous edge cluster (Raspberry Pi 5, Intel DE10 SoC FPGAs, NVIDIA Jetson Orin) over a custom Ethernet fabric to study model-parallel inference across heterogeneous memory and compute hierarchies. Developed a **C library for model-parallel DNN inference** with synchronized layer-wise execution and efficient inter-node memory transfers.
- Demonstrated **400M-parameter model inference** using **4 GB collective cluster memory** with no single node holding the full model — validating memory-disaggregated inference at the edge as a practical deployment strategy.

Joint Resource Allocation & Service Migration in Vehicular Edge Networks | Python, PyTorch, DDPG

- Formulated a joint **Markov Decision Process (MDP)** for network-initiated resource allocation and live service migration across MEC servers, incorporating vehicle mobility models, per-link latency constraints, and server load balancing.
- Trained a modified **DDPG agent** over continuous action spaces to learn mobility-aware migration and allocation policies, achieving **26.67% reduction** in service violations over vehicle-initiated baselines. Published at **AIIoT 2024**.

TLCC: Two-Stage Parallel Connected Component Analysis | C++, Parallel Computing, Image Processing

- Designed a two-layer CCA pipeline combining **OR-mask morphological filtering** with region-based parallel CCA, eliminating redundant computation on sparse binary images. Achieved **64.37% reduction in CPU simulation time** over conventional CCA while preserving full component-detection accuracy. Published at **VLSI-SoC 2023**.

Publications — 4 papers · 6 citations

[1] **D. Joshi**, K. Namboori, K. M. Kuriakose, S. Vadhiyar, S. Banerjee, A. Singh, "SkipPar: A Hybrid CPU-GPU Framework for Accelerating LLM Training via Efficient Scheduling of Parameter Updates," *Proc. HCW, International Parallel and Distributed Processing Symposium (IPDPS) 2026.* **(Accepted)**

[2] **D. Joshi**, A. A. Gangotri, S. P. Chennamsetti, G. Bolar, G. Thiagarajan, S. Gurugopinath, "A Two-Layer Connected Component Algorithm for Target Extraction Using K-means and Morphology," *IFIP/IEEE 31st Int'l Conference on VLSI-SoC, 2023.* [IEEE Link]

[3] G. Bolar, **D. Joshi**, S. P. Chennamsetti, V. K. Tumuluru, "DRL Based Service Migration and Resource Allocation in Vehicular Edge Networks," *3rd Int'l Conference on AI for IoT (AIIoT), 2024.* [IEEE Link]

[4] G. Bolar, **D. Joshi**, S. P. Chennamsetti, S. Gurugopinath, "Performance Comparison of Learning Methods for Soil Parameter Estimation using Hyperspectral Data," *8th Int'l Conference on Signal Processing and Communication (ICSC), 2022.* [IEEE Link]

Technical Skills

Systems & HPC	CUDA, OpenMP, OpenMPI, PyTorch (DDP, autograd hooks, backend extensions), ZSim
Programming	C, C++, Python, MATLAB, Verilog, SystemVerilog
ML & Mathematics	PyTorch, OpenCV, Linear Algebra, Probabilistic Modeling, Markov Chains, Monte Carlo Methods
Hardware Platforms	NVIDIA A100, H100, Jetson Orin; Intel DE10 SoC FPGA; Raspberry Pi 5
Tools & Workflow	LaTeX, Git, Linux, Overleaf, System Diagram Modeling