Aradhana Mohan Parvathy

Email: amohanpa@purdue.edu | LinkedIn: www.linkedin.com/in/AradhanaMP

RESEARCH SUMMARY: Ph.D. researcher specializing in **efficient deep learning and hardware–software co-design** for LLM and multimodal transformer inference. Experience spans **quantization, post-training optimization, pruning, and structured sparsity**. Developed multiple **novel approximation frameworks** that deliver **significant latency, and energy reductions** across LLM and vision/audio transformers. Recipient of prestigious recognitions including **Qualcomm Innovation Fellowship 2025**.

EDUCATION

Ph.D. in Electrical and Computer Engineering, Purdue University, West Lafayette, IN (2020 – 2026)
- **Advisor:** Prof. Anand Raghunathan
- **Key Achievements:** Won the **Qualcomm Innovation Fellowship North America 2025, DAC Young Fellow 2020 and 2022, Purdue Graduate School Summer Research Grant 2021**

B.Tech. (Hons.) in Electrical and Electronics Engineering, NIT Tiruchirappalli (2016 – 2020)
- **Rank:** Graduated in the top 1% of the EEE Department.
- **Key Achievements:** Won the **DAAD-WISE Scholarship 2019** to pursue an internship at **RWTH Aachen, Germany**

TECHNICAL SKILLS

- **Deep Learning & Efficiency:** Quantization, Pruning, LLM Inference & Finetuning, Multimodal Transformers.
- **Frameworks & Languages:** Python, PyTorch, HuggingFace, TensorFlow, C, C++, CUDA, MATLAB
- **Hardware & Simulation:** SystemVerilog, Verilog, Gem5, SCALE-Sim.

INDUSTRY RESEARCH EXPERIENCE

IBM Research, Yorktown Heights, NY
Research Scientist Intern | *Sept 2025 – Dec 2025*
- Investigated and proposed **approximate computing** techniques for efficient inference of **Large Language Models** (details omitted due to confidentiality)**.**
- **Large-Scale Finetuning:** Implemented LoRA finetuning using DeepSpeed to enable resource-efficient model adaptation and optimization.
- Developed **compiler techniques** to improve the inference efficiency of Deep Neural Networks (DNNs).

Intel Corporation, Remote
Multimodal AI Research Intern | *Oct 2024 – Dec 2024*
- Proposed a novel **approximation technique** and tensor core-like **accelerator architecture** to improve inference efficiency of **ultra-low precision LLMs**.
- Achieved a **performance improvement of up to 25%**.
- Submitted a research paper based on the findings.

EnCharge AI, Santa Clara, CA
Summer Intern | *May 2023 – Aug 2023*
- Explored compression of DNN workloads to reduce off-chip memory accesses for higher performance.
- Mapped diverse DNN workloads onto EnCharge AI's hardware architecture for improved performance.

ACADEMIC RESEARCH EXPERIENCE

Integrated Systems Laboratory, Purdue University, West Lafayette, IN
Graduate Research Assistant | *May 2021 – Present*

- **Efficient Quantized LLM Inference (Under Review)**
  - Proposed hardware-software codesign methods to optimize **Quantized LLM Inference** (Details omitted due to double-blind review).
  - Focused on post-training optimization methods to enable efficient execution on resource-constrained hardware.
- **SoftProx: Efficiency for Multimodal Transformers (Accepted at TCASAI)**
  - Proposed a three-step systematic post-finetuning approximation scheme to mitigate the Softmax bottleneck in emerging **Transformer workloads**.
  - Validated efficacy across **multimodal architectures** (text, vision, audio Transformers), improving inference performance by up to **40.22%**.
  - Presented the work at **TECHCON 2025**.
- **Seprox: Compression for Ultra-Low Precision DNNs (Published at ICCAD 2022)**
  - Developed a lightweight compression method for ultra-low-precision DNNs by encoding weight sequences without auxiliary metadata, achieving up to **35.2% model compression**.
  - Modeled the technique in **SCALE-Sim** (systolic-array simulator) and quantified **up to 14.8% energy** and **18.2% performance** improvements.
  - Presented the work at **IBM AI Hardware Forum 2022**, **TECHCON 2022**, and **ICCAD 2022**.

RELEVANT PROJECTS

Mapping Fine-Grained Sparsity to Tensor Cores | *Purdue University*
- Proposed a tensor-core-like architecture to accelerate sparse GEMM with **N:M sparsity**.
- Designed a cycle-accurate simulator based on the **NVIDIA A100 architecture** to estimate speedup.
- Estimated a speedup of **2x-10x** for **50%-93.75%** sparsity.

Adversarial Training Consistency | *Purdue University*
- Proposed a methodology to ensure consistent accuracy across adversarial perturbations of different attack strengths.
- Achieved a mean accuracy of **92.2%** with a standard deviation of 3.4% on the LeNet-MNIST model.

PUBLICATIONS

- **A. Mohan Parvathy,** S. Krithivasan, S. Venkataramani, A. Raghunathan, V. Srinivasan**.** *Title withheld due to double-blind review. Under review*
- **A. Mohan Parvathy,** SK Ghosh, S. Kundu, A. Raha, S. Kundu, D. Mathaikutty, A. Raghunathan**.** *Title withheld due to double-blind review. Under review*
- **A. Mohan Parvathy**, S. Roy, SK. Ghosh, A. Raha, D. Mathaikutty, A. Raghunathan "Softprox: A systematic methodology to mitigate softmax bottleneck in emerging transformer workloads," *Accepted at IEEE TCASAI.*
- **A. Mohan Parvathy**, S. Krithivasan, S. Sen. A. Raghunathan, "Seprox: Sequence-Based Approximations for Compressing Ultra-Low Precision Deep Neural Networks" *IEEE/ACM ICCAD* 2022 (Acceptance rate: 22.5%)
- S. Du, L. Zheng, **A. Mohan Parvathy**, F. Xie, T. Wei, A. Raghunathan, H. Li, "3D-CIMlet: A Chiplet Co-Design Framework for Heterogeneous In-Memory Acceleration of Edge LLM Inference and Continual Learning" *ACM/IEEE DAC 2025*