Akshat Ramachandran

Contact Information

Klaus Advanced Computing Building, Room 3306
266 Ferst Dr NW
Atlanta, GA 30332

akshat.r@gatech.edu
+1 (404)-271-8713
www.akshatramachandran.com

Research Vision

The central focus of my research is in the design and development of efficient and high performance algorithms, architectures and systems for accelerating emergent AI applications. My research breaks down traditional barriers existing between different computing elements and adopts an interdisciplinary approach spanning the entire computing stack.

Research Interests

Computer Architecture, Hardware/Software Co-Design, Networks-on-Chip, Computer Arithmetic, Design Space Exploration, Artificial Intelligence/Machine Learning, Computer Vision, Digital Circuit Design, Simulator Infrastructure

Education

Georgia Institute of Technology, Atlanta, GA — Aug 2023 - Present
M.S./Ph.D. in Electrical and Computer Engineering (ECE) — GPA: 4.0/4.0
Thesis: Synergizing Number Systems, Algorithms, and Hardware for Efficient AI Computing
Advisor: Tushar Krishna

Veermata Jijabai Technological Institute (VJTI), Mumbai — Aug 2018 - May 2022
B.Tech in Electronics Engineering — GPA: 9.80/10.0
Thesis: Next Generation Architecture for Computer Vision
Advisor: John Gustafson (ASU)
Department Rank: 1/83 (Gold Medallist), Institute Rank: 2/480

Work Experience

Intel Corporation, Santa Clara, CA — Aug 2024 - Present
AI Hardware Research Intern, (Manager: Arnab Raha, Souvik Kundu)

- Driving Intel NPU-v7 Compute-in-Memory architecture design and optimization to enable flexible structured sparsity support for LLM applications.

Nvidia Corporation, Santa Clara, CA — May 2024 - Aug 2024
Architecture Energy Modeling Intern, (Manager: Visu Subramanian)

- Developed an AI-based power estimation tool for NVIDIA GPUs.
- Leveraged transfer learning to distill knowledge from parent chip to derivative architecture chips; reducing data collection and training effort by 5×.

Georgia Institute of Technology, Atlanta, GA — Aug 2023 - Present
Graduate Research Assistant, (Advisor: Tushar Krishna)

- Conducting research on efficient and high-performance algorithms, architectures and systems for accelerating emergent AI applications.

Lemurian Labs, Menlo Park, CA — Jun 2023 - Aug 2023
Hardware Intern, (Manager: Ashish Kaul, Theodore Omtzigt)

- Designed and verified a smart NoC to interlink functional units in the Lemurian tensor engine using System Verilog and Python/PyUVM, achieving a 35% increase in operating frequency and 2.5× better performance over existing interconnect.
- Spearheaded cross-functional efforts to deploy Adaptive Logarithmic datatype in Python, C++ and assess mixed precision quantization of NumPy-implemented BERT and GPT, yielding a 10x speedup over floats.

Samsung Research & Development, Bangalore, IN & Suwon, KR — Aug 2022 - Jul 2023
Visual Intelligence Engineer, (Manager: B.V. Vandrotti, Jooyoung Kim)

- Involved in the accelerator/algorithm co-design, optimization and research-to-commercialization of depth perception, processing solutions for XR devices, securing ICIP publication, A2 patent and Samsung Best Paper Awards nomination.
- Co-designed and cross-validated a DL-based ToF Depth Fusion algorithm and SoC architecture in Python and HLS, enabling a 67% reduction in on-device inference latency over previous iterations.
- Designed a systolic dataflow accelerator in System Verilog for ToF acquisition, reducing latency by 60%.

Indian Institute of Technology, Bombay (IIT-B), Mumbai — Jul 2021 – Dec 2022
Research Assistant, (Advisor: Madhav Desai)

- Designed and verified a reconfigurable, programmable, and multi-thread 40-Gb/s NIC and MAC architecture in HLS, Verilog interfaced as a memory-mapped peripheral to the $2 \times 4 \times 32$ Ajit processor. MeiTY funded project.
- Implemented custom packet generation & forwarding (for verification) and NIC support-software libraries in C.

Samsung Research & Development, Bangalore — May 2021 – Jul 2021
Summer Research Intern, (Manager: Ankit Dhiman)

- Researched memory-efficient DL algorithms and architectures for AR applications.
- Remodelled Samsung's 3D reconstruction DL pipeline in Python, C and designed a memory controller that extrapolates the concept of k-means to parameters, achieving a performance boost of nearly 2× over existing pipelines.

Awards, Honors and Scholarships

MICRO student travel grant — 2024

DAC Young Fellow — 2024

Recognized in **top 5%** of advanced research in Samsung Electronics for work on sparse ToF sensors towards the **Samsung Best Paper Award**. — 2023

Samsung Spot Award for outstanding contribution towards design and development of **AR/VR algorithms and acceleration** on Samsung Galaxy devices. — 2022

Best Undergraduate Thesis Award, VJTI — 2022

Professional level SW Competency, Global Samsung Aptitude Test(GSAT) — 2022

Ranked **second** (out of 700 participants) for developing a model for categorisation of post-earthquake satellite imagery at the Vision Beyond Limits Competition. — 2021

Achieved **second** position (out of 1000 participants) in the E-yantra International Robotics Competition for designing a redressal robot on an FPGA to address threats in Industry 4.0. — 2021

Teaching Experience

Microprocessor Systems Lab, VJTI — Fall 2021
Teaching Assistant (Professor: Rohin Daruwala)

Professional Service

Technical Program Committee Member — 2023
Conference on Next-Generation Arithmetic (CoNGA)

Posters, Talks

Algorithm-Hardware Co-Design of Distribution-Aware Logarithmic-Posit Encodings for Efficient DNN Inference, IBM Research, New York, NY — Jun 2024

Deep Learning Model Quantization, NVIDIA GPU Power Architecture Team, Santa Clara, CA — May 2024

Algorithm-Hardware Co-Design of Distribution-Aware Logarithmic-Posit Encodings

for Efficient DNN Inference, Lemurian Labs, Menlo Park, CA May 2024

Data-Free Quantization and Inference using Contrastive Learning and Next Generation Arithmetic, CoCoSyS Apr 2024

Algorithm-Hardware Co-Design of Distribution-Aware Logarithmic-Posit Encodings, poster at CoCoSyS Annual Review Meeting Feb 2024

Publications

MicroScopiQ: Accelerating Foundational Models through Outlier-Aware Microscaling Quantization
Akshat Ramachandran, Souvik Kundu, and Tushar Krishna
Under review

CLAMP-ViT: Contrastive Data-Free Learning for Adaptive Post-Training Quantization of ViTs
Akshat Ramachandran, Souvik Kundu, and Tushar Krishna
European Conference on Computer Vision (ECCV), Oct 2024

Algorithm-Hardware Co-Design of Distribution-Aware Logarithmic-Posit Encodings for Efficient DNN Inference
Akshat Ramachandran, Zishen Wan, Geonhwa Jeong, John Gustafson, and Tushar Krishna
Design Automation Conference (DAC), Jun 2024

GPU based building footprint identification utilising self-attention multiresolution analysis
Rizwan Ahmed Ansari, **Akshat Ramachandran**, and Winnie Thomas
All Earth, Dec 2023

NTrans-Net: A Multi-Scale Neutrosophic-Uncertainty Guided Transformer Network for Indoor Depth Completion
Akshat Ramachandran, Ankit Dhiman, Basavaraja Shanthappa Vandrotti, and Jooyoung Kim
International Conference on Image Processing (ICIP), Oct 2023

Multiresolution based neutrosophic framework for building footprint identification using remote sensing images
Akshat Ramachandran and Rizwan Ahmed Ansari
International Conference on Signal Processing and Integrated Networks (SPIN), Mar 2023

PositIV: A Configurable Posit Processor Architecture for Image and Video Processing
Akshat Ramachandran, John Gustafson, Anusua Roy, Rizwan Ahmed Ansari and Rohin Daruwala
Euromicro Conference on Digital System Design (DSD), Aug 2023

Self-supervised depth enhancement
Akshat Ramachandran and Rizwan Ahmed Ansari
International Conference for Advancement in Technology (ICONAT), Jan 2022

Patents

A Method and System for Acquiring, Processing and Monocular Depth Completion/Estimation of sparse time-of-flight(ToF) sensors
Akshat Ramachandran, Ankit Dhiman, Basavaraja Shanthappa Vandrotti, Jooyoung Kim and Lokesh Rayasandra Boregowda
Provisional patent filed based on work at Samsung Electronics Co. Ltd., Feb 2023