RYA SANOVAR

(+1) 404-400-9490 ◇ rsanovar3@gatech.edu ◇ LinkedIn ◇ Website

EDUCATION

Georgia Institute of Technology *Aug 2025 - Present*
Ph.D. in Computer Science
Advised by Prof. Moinuddin Qureshi, Future Architectures and Systems Lab

Birla Institute of Technology and Science, Pilani *Oct 2020 - Jun 2024*
B.E. in Electronics and Communication GPA: 8.91/10.0

WORK EXPERIENCE

Microsoft Research Bangalore, India
Research Fellow (Previously Research Intern) *Jan 2024 - Jul 2025*

· Developed **LeanAttention**: A hardware-aware scalable exact-attention execution mechanism that accelerates attention by 2.6x over FlashAttention-2.
  – Integrated into **ONNXRuntime**: ONNXRT LeanAttention
· Published and filed patents in key novel techniques to improve efficiency of LLM inference.

Central Electronics Engineering Research Institute Chennai, India
Research Intern *May 2022 - Jul 2022*

· Implemented video-based physiological signal extraction via eulerian video magnification to estimate respiratory rates from facial recordings, enabling low-cost health monitoring.

PUBLICATIONS & POSTERS

Rya Sanovar, Srikant Bharadwaj, Renee St. Amant, Victor Rühle, Saravan Rajmohan. *"Lean Attention: Hardware-Aware Scalable Attention Mechanism for the Decode-Phase of Transformers"*
*The Eighth Annual Conference on Machine Learning and Systems **(MLSys)**, Santa Clara, May 2025*
https://arxiv.org/abs/2405.10480

Rya Sanovar. *"Horses for Courses: Unique Hardware Efficiency Challenges and Solutions for Prefill and Decode Inference."*
Poster presented at the *Young Professionals Symposium @ MLSys, Santa Clara, May 2025.*

PATENTS

Rya Sanovar, Srikant Bharadwaj, Victor Rühle. *"Hardware-aware attention mechanism with dynamic workload distribution for transformer models."*

RELEVANT COURSEWORK

Computer Science: Advanced Computer Architecture, Operating Systems, Machine Learning
Mathematics: Probability and Statistics, Multivariate Calculus, Differential Equations
Electrical and Electronics: Microelectronic Circuit Design, Microprocessors and Interfacing

TECHNICAL SKILLS

Languages: CUDA, C/C++, Python
Libraries & Frameworks: CUTLASS, TensorRT-LLM, NSight Compute, Vivado HLS
Hardware: Modern GPU Micro-architectures and FPGAs