Shakya Jayakody

Orlando, FL 407-580-5009 shakya@ucf.edu LinkedIn https://shakyajayakody.github.io/

RESEARCH FOCUS

- High-Performance GPU Computational Kernels
- Hardware-Software Co-Design in Machine Learning Systems
- GCN training in a Hardware-Contained environment
- Fault tolerance in LLMs
- Computer Architecture

EDUCATION

University of Central Florida, Orlando, FL Aug. 2020 – Dec. 2025
Ph.D. in Electrical Engineering
Dissertation Title: "High-Performance Computational Kernels for Algorithm-System Co-Design in Machine Learning: Enabling Efficient GCN Training and LLM Inference Serving"
Supervisor: Dr. Jun Wang, FIEEE

University of Central Florida, Orlando, FL Jan. 2024 – Dec. 2024
M.S. in Computer Engineering
Specialization: Machine Learning and Computer Architecture
Supervisor: Dr. Mingjie Lin

Louisiana Tech University, Ruston, LA Dec. 2018 – May 2020
M.S. in Electrical Engineering
Specialization: Robotics, Automation Control Systems and Microsystems
Supervisor: Dr. Chester G. Wilson

Louisiana Tech University, Ruston, LA Sep. 2013 – May 2016
B.S. Electrical Engineering, Minor in Mathematics

SKILLS

Programming Languages:	CUDA, C/C++, Python, and MIPS Assembly
Miscellaneous:	PyTorch, PyCUDA, and Verilog
Architectural Simulator:	Gem5, and Ramulator
EDA Tools:	Xilinx IDE, and Vivado HLx
Circuit Design Tools:	Fusion 360, and Eagle

PUBLICATION

[1] **[ICMLA'25]** Pinnock, A.*, **Jayakody, S.***, Roxy, K. A., & Ahmed, Md. R. EdgeProfiler: A Fast Profiling Framework for Lightweight LLMs on Edge Using Analytical Model. 2025 IEEE 24th International Conference on Machine Learning and Applications. (* Equal contribution.) (Accepted, publication pending.)

[2] **[ASAP'25]** **Jayakody, S.**, Zhao Y., & Wang, J. AIRES: Accelerating Out-of-Core GCNs via Algorithm-System Co-Design. 2025 IEEE 36th International Conference on Application-specific Systems, Architectures and Processors, pp. 1-8, doi: 10.1109/ASAP65064.2025.00011.

[3] **[MLArchSys'24]** **Jayakody, S.**, & Wang, J. (2024). Peridot: Accelerating Out-of-Core GCN Data Reuse Pattern and Co-Design on GPU. In Machine Learning for Computer Architecture and Systems 2024.

[4] [**CIC'23**] **Jayakody, S.**, & Wang, J. EMBARK: Memory Bounded Architectural Improvement in CSR-CSC Sparse Matrix Multiplication. 2023 IEEE 9th International Conference on Collaboration and Internet Computing, pp. 8-17, doi: 10.1109/CIC58953.2023.00012.

EXPERIENCE

Postdoctoral Researcher Dec. 2025 – May 2026
Computer Systems Architecture and Data Science Lab, University of Central Florida, Orlando, FL

- Working on DiT-AR, a video generation framework that combines diffusion transformers with autoregressive modeling to produce high-fidelity, temporally coherent video from text and multimodal prompts. My focus is on designing and optimizing the model and inference pipeline so it can scale efficiently across multiple GPUs.
- Developing competitive funding proposals that align with sponsor guidelines and institutional requirements.

Graduate Research Assistant Aug. 2020 – Dec. 2025
Computer Systems Architecture and Data Science Lab, University of Central Florida, Orlando, FL
Supervisor: Dr. Jun Wang

Research related to Machine Learning and Computer Architecture. Engaged in research focused on custom CUDA kernel acceleration for machine learning. Collaborated closely with the academic advisor through weekly meetings to discuss research progress and evaluate algorithms. Conducted research using advanced NVIDIA GPUs (A100, H100, H200) with access provided by the Delta AI Lab, funded by the National Science Foundation (NSF).

- Developed and enhanced SpGMM algorithm and introduced memory controller architecture to optimize SpGMM performance.
- Optimized and extended Sparse Matrix Multiplication for Memory-Constrained GCN Training on GPU through Algorithm-System Co-Design.
- Enhancing Fault-Tolerant Inference for Large Language Models in Multi-GPU Setups through Algorithm-System Co-Design.
- Compiled research technical reports and communicated research findings to the advisor.
- Conducted simulations, gathered results, and undertook data analysis to support research objectives.

Instructor and Teaching Assistant Aug. 2020 – May. 2025
Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL

Instructor for undergraduate courses:
- Digital Systems Lab (Fall 2020 and Summer 2021)
- Electronics I Lab (Fall 2021)
- Embedded Systems Lab (Spring 2022, Fall 2022, Spring 2023, Summer 2023, and Spring 2024)
- Computer Organization Lab (Fall 2023, Spring 2024, and Summer 2024)
- Electrical and Computer Engineering Design Lab (Summer 2024, Fall 2024, and Spring 2025)

Teaching assistant for undergraduate courses:
- Engineering Analysis and Computation (Spring 2021, Summer 2021, and Fall 2022)
- Computer Architecture (Fall 2021)
- Massive Storage and Big Data (Spring 2022, Fall 2022 and Fall 2024)
- Digital Signal Processing Fundamentals (Summer 2023)
- Computer Organization (Fall 2023)
- Topics in Machine Learning (Fall 2024)
- Introduction to Deep Learning (Spring 2025)
- Engineering Applications of Intelligent Systems Electrical and Computer Engineering (Spring 2025)

Student Ambassador (Graduate Assistant) May 2019 – Aug. 2020
International Student Office, Louisiana Tech University, Ruston, LA

Collaborated closely with the International Student Director to achieve key objectives set by the senior leadership. Engaged in weekly strategy meetings with the director and fellow ambassadors to devise and refine student recruitment marketing strategies. Played a pivotal role in crafting innovative marketing approaches to enhance recruitment efforts.

- Led a team of international student ambassadors in planning and implementing outreach programs, fostering a wider and more inclusive student network.
- Conducted a detailed analysis of graduate application statistics to inform responsive strategies during the SARS-CoV-2 pandemic, aiding in maintaining a steady influx of applications.
- Actively recruited international students for undergraduate and graduate programs, contributing to a diverse and vibrant student body.

Technical Director Dec. 2017 – Nov. 2018
Lanka Telephone Co. (Pvt) Ltd., Colombo, Sri Lanka

Oversaw and optimized financial budgets in collaboration with directors, focusing on forecasting and enhancing efficiency for cost-saving measures. Led business development discussions with managers and spearheaded the preparation of bids and tender procurement processes. Fostered technical growth among staff through coaching and promoting career development in technical fields.

- Convened bi-weekly status meetings with lead engineers, sales managers, and crews to strategize on installation processes.
- Managed work and educational event assignments for engineers while maintaining a comprehensive staff calendar.
- Evaluated and improved logistics, supply, and maintenance management operations through detailed analysis, recommendations, and corrective actions.

Lead Electrical and Telecommunication Engineer May 2017 – Dec. 2017
Lanka Telephone Co. (Pvt) Ltd., Colombo, Sri Lanka

Collaborated with multi-disciplinary engineering teams to address design issues and adhere to project schedules.

- Formulated and implemented test procedures for each installation site, including onsite event briefings and guidance to installation crews.
- Ensured compliance with local and state regulatory and institutional policies through diligent monitoring and analysis of activities.
- Oversaw productivity, quality assurance, and contract adherence to maintain high standards of operation.

Electrical and Telecommunication Engineer Jan. 2017 – May 2017
Lanka Telephone Co. (Pvt) Ltd., Colombo, Sri Lanka

Compiled and reviewed technical reports, presenting findings to the Technical Director. Engaged with customers to gauge satisfaction levels post-installation and configuration, and addressed project roadblocks effectively.

- Coordinated installation and maintenance projects, ensuring safety and compliance with customer requirements and legal standards.
- Supervised a team of electrical engineers on multi-million dollar projects, ensuring timely and quality delivery.
- Developed and presented project budgets encompassing materials, construction, and labor costs to the director board.

Software Engineer July 2016 – Jan. 2017
Lanka Telephone Co. (Pvt) Ltd., Colombo, Sri Lanka

- Developed billing software for the telephone system utilizing C/C++.
- Created company inventory software using C#.NET.
- Upgraded and maintained the company website, leveraging skills in WordPress, HTML, CSS, and PHP.

MAJOR PROJECT HIGHLIGHTS

Fault-Tolerant Inference for Large Language Models 2024 – Current
We identify the fault tolerance challenges in prior LLM systems and propose an algorithm-system Co-design solution for efficient data recovery in the event of multiple simultaneous GPU failures.

GCNs training in resource-constrained systems 2023 – 2024
We introduce the AIRES framework, a novel algorithm-system co-design approach dedicated to enhancing the performance of GCNs training in resource-constrained systems, such as single GPU-CPU systems. On the algorithm level, AIRES adopts a row block-wise (RoBW) alignment partitioning method for matrices in compressed format and develops a CUDA tiling algorithm for compressed matrix multiplication to facilitate the RoBW. On the system level, AIRES employs a three-phase dynamic scheduling protocol that features a dual-way data transfer strategy within a tiered memory system to maximize GPU utilization. This research project is funded by the National Science Foundation (NSF).

Memory bounded architectural improvement in CSR-CSC Sparse Matrix Multiplication 2021 – 2023
We introduce EMBARK, is a memory-bounded architectural improvement designed for Compressed Sparse Row (CSR) - Compressed Sparse Column (CSC) sparse matrix multiplication. This approach aims to reduce matrix decompression/compression overhead and optimize storage allocation. The method employs memory partitioning techniques, where values and row IDs are stored together, while column pointers are kept separately in the main memory. Additionally, hot data is stored in main memory, while non-volatile memory (NVM) stores partial hot data using a rank-based page replacement strategy. This research project is funded by the National Science Foundation (NSF).

PREPRINTS/UNDER REVIEW

[**U1**] [**ISCA'26**] Zhao, Y., **Jayakody, S.**, Price, D., Wu, D., & Wang, J. Compute More, Save More: Reliable and Sustainable LLM Inference. 2026 IEEE International Symposium on Computer Architecture.

[**U2**] [**MLSys'26**] **Jayakody, S.**, Zhao, Y., Nehate, C., & Wang, J. GhostServe: A Lightweight Checkpointing System for Fault-Tolerant LLM Serving. 2026 Conference on Machine Learning and Systems.

LABORATORY COMPUTING FACILITIES

Personal laboratory resources supporting graduate student research.

• Local GPU workstation equipped with an NVIDIA RTX 5090, enabling deep learning model development, machine learning research, and performance benchmarking.

• Secondary workstation powered by an NVIDIA RTX 3070, dedicated to resource-constrained and algorithmic efficiency research in machine learning.

• Digilent Nexys 3 FPGA board utilized for hardware acceleration experiments, digital system prototyping, and computer architecture research.

• Planned Expansion: Intend to acquire an NVIDIA DGX Spark system to establish a scalable infrastructure for multimodal and large language model (LLM) training, fostering future collaborations and competitive external grant proposals.

FUNDING PROPOSAL DEVELOPMENT

Assisted in the development of a funding proposal with the University of Central Florida.

AWARDS/ACKNOWLEDGMENTS

- Acknowledged for "Outstanding Analytical Skills" by the Graduate School Dean, Louisiana Tech University 2020
- Acknowledged for "Top Student Ambassador" by the International Student Director, Louisiana Tech University 2020
- Endowed Superior Graduate Student Scholarship, Louisiana Tech University 2019 – 2020
- Bulldog Out-of-State Scholarship, Louisiana Tech University. 2013 – 2016

PROFESSIONAL TRAINING

- Information Security Awareness Training, (UCF) 2025
- Responsible Conduct of Research for Engineers (CITI) 2020
- Doing the Right Thing: Know About Research Misconduct (UCF) 2020

SUB-ADVISING/MENTORING

University of Central Florida Jan. 2025 – Sep. 2025
Ashvik Vijai (Undegrad Student)