Rasool Sharifi

434-249-3650
as3mx@virginia.edu
rasool-sharifi.github.io/
github.com/danteisalive

Overview

My interests lie broadly in Computer Architecture and Domain-Specific Compilers. I am interested in exploring new optimization techniques to enhance the performance, energy efficiency, and programmability of AI accelerators.

•**Highlights of my work experience:** As a Machine Learning Performance Architect at d-Matrix.ai, my primary responsibilities involved optimizing and enhancing the performance of machine learning models on Corsair hardware. I collaborated closely with the compiler team to comprehend both the model algorithms and the hardware pipeline, which helped in identifying new optimization opportunities. My duties included developing Speed of Light (SOL) simulation models to profile the performance of Large Language Models on Corsair systems, identifying performance bottlenecks, and conducting experiments to test various optimization techniques.

•**Highlights of my research project:** Graph Neural Networks (GNNs) represent a cutting-edge synergy in advancing compiler optimization. GNNs are adept at capturing the intricate dependencies and relationships inherent in graph-like data structures, which are fundamental in compiler design. By representing various program constructs and their interactions as a graph, GNNs can effectively analyze and predict optimal layouts assignment and tiling. Integrating LLMs into this framework adds a layer of sophisticated decision-making and heuristic understanding. LLMs, trained on vast amounts of optimization patterns, can provide high-level insights and recommendations, guiding the GNN in exploring the vast solution space more intelligently. This combination can significantly enhance the compiler's ability to determine the best layout assignments, taking into account not only the structural aspects captured by the GNN but also the nuanced patterns and best practices distilled by the LLM. Such a collaborative approach could lead to more efficient, faster, and energy-saving compiled programs, pushing the boundaries of current compiler optimization techniques.

Work Experience

d-Matrix.ai, California, USA, May. 2023 - August. 2023
ML Performance Architect - Intern
Responsibilities:
- Developed a python-based pipeline for modeling collective operations
- Developed a python-based pipeline for roofline performance analysis of Large Language Model workloads

d-Matrix.ai, California, USA, June. 2022 - Dec. 2022
ML Performance Architect - Intern
Responsibilities:
- Performance modeling, correlation, and projection
- Maintaining tools for high-level system simulation of Corsair systems
- Developed a MLIR compiler pipeline for direct performance evaluation of MLIR workloads

Notable Projects

Optimizing Layout Assignments: A Collaborative Approach with Graph Neural Networks and Large Language Models,
Language: *Python*, **Frameworks**: *Pytorch, JAX, Torch Geometrics (PyG)*
Links: https://github.com/danteisalive/tpugraph (will be publicly available after publication)

Design and Implementation of MLIR-Based Pipeline for d-Matrix.ai Corsair Performance and Power Projection,
Language: *Python, C++*, **Platform**: *MLIR, Lark*
Links: Part of the internship at d-Matrix.ai

Automatic Resource-aware Generation of Energy-efficient CNN Inference Accelerator for Edge Embedded FPGAs,
Language: *C, C++, Chisel, Verilog*, **Platform**: *Xilinx Zynq SoC*
Links: https://bitbucket.org/alijahan/bnn-pynq/src/master/

Speculation-Driven Dynamic Binary Optimization,
Language: *C++, Python*, **Framework**: *GEM5*
Links: https://github.com/logangregorym/gem5-changes

Compiler for Tiny AVR Microcontrollers,
Language: *Java*, **Platform**: *Atmel AVR*
Links: https://github.com/danteisalive/AVRCompiler

Architecture Support for Memory and Type Safety,
Language: *C++, Python*, **Framework**: *LLVM, GEM5, Pin*
Links: https://github.com/danteisalive/llvm-typecheking
Links: https://github.com/danteisalive/gem5-tc
Links: https://github.com/danteisalive/typetracking_pin

Processing in Memory for Bioinformatics Applications,
Language: *C++, Python*, **Framework**: *DRAMSim3*
Links: https://github.com/umd-memsys/DRAMsim3

Education

University of South Carolina, Columbia, USA — Jan. 2023 - Present
Ph.D., Computer Science

University of Virginia, Charlottesville, USA — Sep. 2018 - Dec. 2022
Ph.D., Computer Science (Transferred to UoSC on Dec. 2022)

University of Tehran, Tehran, Iran — Sep. 2013 - June 2016
M.Sc., Electronics-Circuit and Systems

Isfahan University of Technology, Isfahan, Iran — Sep. 2009 - June 2013
B.Sc., Electrical Engineering

Research Experience

University of South Carolina, Columbia, USA, — Jan. 2023 - Present
Graduate Research Assistant

University of Virginia, Charlottesville, USA, — Sep. 2018 - Dec. 2022
Graduate Research Assistant

University of Tehran, Tehran, Iran, — Sep. 2013 - June 2016
Graduate Research Assistant

Publications

- K. Skadron, M. Lenjani, **Rasool Sharift** , and L. Wu, "Scalable in situ dram-based accelerators and methods of operating the same," *US Patent*, 2023
- Lingxi Wu, Rahul Sreekumar, **Rasool Sharifi**, Mircea Stan , Kevin Skadron, and Ashish Venkat, "Hardware Trojans in eNVM Neuromorphic Devices," in *Design, Automation and Test in Europe Conference*, 2023. **Nominated for Best Paper Award**
- Logan Moody, Wei Qi, **Rasool Sharifi**, Layne Berry, Joey Rudek, Sreenivas Subramoney, Jayesh Gaur, Jeff Parkhurst, Kevin Skadron, and Ashish Venkat, "Speculative Code Compaction: Eliminating Dead Code via Speculative Microcode Transformations," in *IEEE/ACM 53rd International Symposium on Microarchitecture (MICRO)*, 2022
- L. Wu, **Rasool Sharifi**, A. Venkat, and K. Skadron, "Dram-cam: General-purpose bit-serial exact pattern matching," *IEEE Computer Architecture Letters (CAL)*, 2022
- Lingxi Wu, **Rasool Sharifi**, Marzieh Lenjani, Kevin Skadron, and Ashish Venkat, "Sieve: A Scalable In-Situ DRAM-based Accelerator for Massively Parallel K-mer Matching," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, 2021, **Acceptance Rate: 18%**
- **Rasool Sharifi** and Ashish Venkat, "CHEx86: Context-Sensitive Enforcement of Memory Safety via Microcode-Enabled Capabilities," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020, **Acceptance Rate: 18%**, **Top Pick in Hardware and Embedded Security – Selected from Architecture/Security/VLSI Design Conferences Held Between 2015-2020!**
- Lingxi Wu, **Rasool Sharifi**, Kevin Skadron, and Ashish Venkat, "DRAM-CAM: General-Purpose Bit-Serial Exact Pattern Matching," in *IEEE Computer Architecture Letters (CAL)*, 2022
- Ali Jahanshahi, **Rasool Sharifi**, Mohammadreza Rezvani, and Hadi Zamani, "Inf4Edge: Automatic Resource-aware Generation of Energy-efficient CNN Inference Accelerator for Edge Embedded FPGAs," in *2021 IEEE Workshop on Energy-Efficient Machine Learning (E2ML)*, 2021
- **Rasool Sharifi** and Zain Navabi, "Online Profiling for cluster-specific variable rate refreshing in high-density DRAM systems," in *2017 22nd IEEE European Test Symposium (ETS)*, 2017

Awards

CHEx86: Context-Sensitive Enforcement of Memory Safety via Microcode-Enabled Capabilities — Dec. 2021
Top Pick in Hardware and Embedded Security
Selected from Architecture/Security/VLSI Design Conferences Held Between 2015-2020

Skills

Programming Languages/Frameworks:
- C/C++, LLVM, MLIR, Python, PyTroch, JAX

References

Two references will be made available upon request.