Mahmoud Khairy A. Abdallah — Curriculum Vitae, December 2024

Contact Information

Personal Email: khairy2011@gmail.com
Personal Website: https://mkhairy.github.io/
Other Websites: Linkedin - Google Scholar - GitHub - Medium

Interests

GPU Architecture, Multi-GPU/Multi-Chiplet Scaling, Deep Learning Acceleration, Performance Modeling and Simulation, Hardware-Software Co-design, Data Center Microservices

Education

Purdue University, West Lafyatte, IN, US

Ph.D., Electrical and Computer Engineering, January 2017 - August 2022

- Thesis Title: “Scalable and Energy-Efficient SIMT Systems for Deep Learning and Data Center Microservices”
- Advisor: Prof. Timothy Rogers

Cairo University, Giza, Egypt

M.Sc., Computer Engineering, Sep 2011 - May 2015

- Thesis Title: “Efficient Utilization of GPGPU Cache Hierarchy”

Cairo University, Giza, Egypt

B.Sc., Computer Engineering, Sep 2006 - May 2011

- Distinction with degree of honor (90.19%) – GPA: 3.88/4.0
- Ranked 3rd in group of 45 students.

Peer-Reviewed conference & Journal Publications

1. Ahmad Alawneh, Ni Kang, **Mahmoud Khairy**, Timothy G. Rogers. “ThreadFuser: A SIMT Analysis Framework for MIMD Programs”, 2024 International Symposium on Microarchitecture (**MICRO 2024**).

2. **Mahmoud Khairy**, Ahmad Alawneh, Aaron Barnes, and Timothy G. Rogers, “SIMR: Single Instruction Multiple Request Processing for Energy-Efficient Data Center Microservices”, 2022 International Symposium on Microarchitecture (**MICRO 2022**), acceptance rate=22%.

3. Cesar Avalos, **Mahmoud Khairy**, Roland N. Green, Mathias Payer, Timothy G. Rogers, “Principal Kernel Analysis: A Tractable Methodology to Simulate Scaled GPU Workloads” 2021 International Symposium on Microarchitecture (**MICRO 2021**), acceptance rate=22%.

4. Vijay Kandiah, Scott Peverelle, **Mahmoud Khairy**, Amogh Manjunath, Junrui Pan, Timothy G. Rogers, Tor Aamodt, Nikos Hardavellas, “AccelWattch: A Power Modeling Framework for Modern GPUs” 2021 International Symposium on Microarchitecture (**MICRO 2021**), acceptance rate=22%.

5. **Mahmoud Khairy**, Vadim Nikiforov, David Nellans, and Timothy G. Rogers,“Locality-Centric Data and Threadblock Management for Massive GPUs” 2020 International Symposium on Microarchitecture (**MICRO 2020**), acceptance rate=18%.

6. **Mahmoud Khairy**, Jason Shen, Tor M. Aamodt, and Timothy G. Rogers, “Accel-Sim: An Extensible Simulation Framework for Validated GPU Modeling” 2020 International Symposium on Computer Architecture (**ISCA 2020**), acceptance rate=18%. **Selected In ISCA@50 Retrospective: 1996-2020, only 98 out of 1,077 papers were selected.**

7. **Mahmoud Khairy**, Amr G. Wassal, and Mohamed Zahran. “A Survey of Architectural Approaches for Improving GPGPU Performance, Programmability and Heterogeneity.” Journal of Parallel and Distributed Computing 127 (2019): 65-88 (**JPDC 2019**).

8. Jain Akshay*, **Mahmoud Khairy***, Timothy G. Rogers, “A Quantitative Evaluation of Contemporary GPU Simulation Methodology”, ACM SIGMETRICS, 2018, (**SIGMETRICS 2018**), *First Co-authors.

9. **Mahmoud Khairy**, Mohamed Zahran, and Amr G. Wassal, “SACAT: Streaming-Aware Conflict-Avoiding Thrashing-Resistant GPGPU Cache Management Scheme”, IEEE Transcation on Parallel and Distributed Systems (**TPDS 2016**)

Workshops

1. **Mahmoud Khairy**, Mohamed Zahran, and Amr G. Wassal, “Efficient Utilization of GPGPU Cache Hierarchy”, 8th Workshop on General Purpose Computing using GPUs (GPGPU8), 2015, (**GPGPU 2015**)

Posters

1. Christin David Bose, Cesar Avalos, Junrui Pan, **Mahmoud Khairy**, Timothy Rogers ”MAccel-sim: A Multi-GPU Simulator for Architectural Exploration.” 2024 IEEE International Symposium on Workload Characterization (**IISWC 2024**).

2. Ahmad Alawneh, **Mahmoud Khairy**, Timothy G. Rogers, “A SIMT Analyzer for Multi-Threaded CPU Applications.” 2022 IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS 2022**).

3. **Mahmoud Khairy**, Jain Akshay, Tor M. Aamodt, and Timothy G. Rogers, “A Detailed Model for Contemporary GPU Memory Systems.” 2019 IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS 2019**).

Technical Blogs

1. Tim Rogers and **Mahmoud Khairy** “An Academic’s Attempt to Clear the Fog of the Machine Learning Accelerator War.” **SigArch blog**, August 2021
   Blog Link: https://www.sigarch.org/an-academics-attempt-to-clear-the-fog-of-the-machine-learning-accelerator-war/

2. **Mahmoud Khairy** “TPU vs GPU vs Cerebras vs Graphcore: A Fair Comparison between ML Hardware,” **Medium article**, July 2020
   Article Link: https://khairy2011.medium.com/tpu-vs-gpu-vs-cerebras-vs-graphcore-a-fair-comparison-between-ml-hardware-3f5a19d89e38

Technical Reports

1. **Mahmoud Khairy**, Jain Akshay, Tor Aamodt, and Timothy G. Rogers. “Exploring Modern GPU Memory System Design Challenges through Accurate Modeling.” arXiv preprint arXiv:1810.07269 (2018).

2. Hughes, Clayton, Simon David Hammond, **Mahmoud Khairy**, Mengchi Zhang, Roland Green, Timothy Rogers, and Robert J. Hoekstra. “Balar: A SST GPU Component for Performance Modeling and Profiling”. No. SAND2019-10389. Sandia National Lab., 2019.

3. **Mahmoud Khairy**, Mengchi Zhang, Roland Green, Simon David Hammond, Robert J. Hoekstra, Timothy Rogers, and Clayton Hughes. “SST_GPU: An Execution-Driven CUDA Kernel Scheduler and Streaming-Multiprocessor Compute Model.” No. SAND2019-1967. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2019

Pending Patents

1. **Mahmoud Khairy Abdallah** et al.“A Method and Apparatus for Heterogeneous-aware Work Group Scheduling to Reduce Traffic in Multi-Chiplet Designs” US Pending Patent, app# 18612505, filed by AMD Systems

2. **Mahmoud Khairy Abdallah** et al. “A Method and Apparatus for Profiling Based Work Group Scheduling on Multiple GPU Chiplets” US Pending Patent, app# 18744231, filed by AMD Systems

3. Timothy Rogers, **Mahmoud Khairy** “System and Methods for Single Instruction Multiple Request Processing”, US Pending Patent, app# 63307853, filed by Purdue University

Dissertations

1. **Mahmoud Khairy** “Scalable and Energy-Efficient SIMT Systems for Deep Learning and Data Center Microservices”, August 2022, Ph.D. dissertation, Purdue University.

2. **Mahmoud Khairy** “Efficient Utilization of GPGPU Cache Hierarchy,” May 2015, M.Sc. dissertation, Cairo University, Egypt

Grants

I assisted and contributed with my PhD adviser to write an accepted NSF grant #1910924 “Addressing Challenges for the Next Decade of Massively Parallel NUMA Accelerators”.

Open-Source Tools

- **Accel-Sim**: Principle contributor and technical lead
  Project website: https://accel-sim.github.io/
- **Accel-Wattch**: Co-author
  Project website: https://accel-sim.github.io/accelwattch.html
- **Principle Kernel Analysis**: Co-author
  Project github: https://github.com/cesar-avalos3/micro-2021-artifact
- **SIMTec**: Co-author
  Project github: https://github.com/LAhmos/x86_tracing_pintool
- **SST-GPU**: Co-author
  Project website: http://sst-simulator.org/SSTPages/SSTElementBalar/

Awards & Honors

- ISCA-50 Retrospective for the Accel-Sim paper: One of 98 papers selected from 1,077 submissions between 1996 and 2020.
- Purdue Graduate Fellowship, Purdue University, 2017-2018
- Fully-funded internship from SAFARI group, Carnegie Mellon University, 2015
- M.Sc. Scholarship from Cairo University, September 2012 - May 2015
- B.Sc. with Honors from Cairo University, May 2011
- Fully-funded undergraduate internship from Technical University of Munich, Germany, 2010
- Undergraduate Academic Outstanding Award from the Egyptian Government, 2006-2010

Press & Media Report

- *HPCWire*
  Purdue Researchers Peer into the ‘Fog of the Machine Learning Accelerator War’
  https://www.hpcwire.com/2021/09/27/purdue-researchers-peer-into-the-fog-of-the-machine-learning-accelerator-war/
- *ZDnet*
  AI computer maker Cerebras nabs TotalEnergies SE as first energy sector customer
  https://www.zdnet.com/article/ai-computer-maker-cerebras-nabs-totalenergies-se-as-first-energy-sector-customer/
- *Analytics India Magazine*
  Thinking Beyond Generative AI, One Token At A Time
  https://analyticsindiamag.com/thinking-beyond-generative-ai-one-token-at-a-time/

Professional Experience

Member of Research Staff, Advanced Micro Devices (AMD) Aug 2022 to Present

High Performance Computing Researcher, AMD Research, Santa Clara, CA

- Conducting applied research to reduce data movement in Multi-GPU Multi-Chiplet architecture for high-performance computing (HPC) and cutting-edge deep learning workloads—such as large language models (LLM). Collaborating regularly with the product team to identify current and future challenges and deliver applicable and fast solutions.
- Profiling and identifying key bottlenecks in LLM inference workloads on the MI300X GPU, and proposing software and hardware optimizations to minimize data movement between GPU chiplets and reduce power consumption.
- Publishing scientific papers in top-tier conferences and submitting impactful and novel patents, with four patents filed.

Research Assistant, Purdue University Jan 2017 to July 2022

Computer Engineering Department, Purdue University, IN, US

- Principal contributor to a new open-source GPU simulation infrastructure, Accel-Sim, that models next-generation GPU architectures quickly and accurately with tensor cores, deep learning support and multi-GPU systems. This work appeared at the **ISCA 2020** venue. Accel-Sim website: https://accel-sim.github.io/

- Investigating a transparent compiler-assisted solution to overcome the non-uniform-memory-access (NUMA) overhead for multi-GPU DGX server and next-generation multi-chiplet-modules GPUs. This work appeared at the **MICRO 2020** venue.
- Performance profiling of data center microservices-based workloads and identifying performance bottlenecks and hardware acceleration opportunities. This work appeared at the **MICRO 2022** venue.
- Collaborating with Sandia Lab to integrate GPGPU-Sim simulator into SST infrastructure to enable simulating large-scale GPU-based supercomputer.
  Project website: http://sst-simulator.org/SSTPages/SSTElementBalar/
- Deep learning system evaluation and comparing between the leading ML training hardware solutions, focusing on efficiency metrics and identifying the ML training bottlenecks. See my Medium and SigArch articles for further details about this work.
- Mentoring a group of undergraduate, master's and fresh PhD students to put their first hands on academic research and assisting them through their own research ideas. This has resulted in several 2nd author publications.

Software Engineer, Microsoft Dec 2015 to Dec 2016

Data Mining and Tooling team, Bing Ads, Microsoft, Redmond, WA, US

- Developing and improving a microservices-based data analytics platform that collects data from distributed data sources in order to monitor Bing Ads performance metrics in real-time and analyze the marketplace trends. In particular, I was fully responsible for managing anomaly detection microservice that dynamically checks data on a regular basis and notifies our users if any anomaly exists.

Research Assistant Intern, CMU June 2015 to September 2015

Carnegie Mellon University, Pittsburgh, Penn, US.

- Research: Studying the implications of High Bandwidth Memory (HBM) on next-generation GPGPU workloads and architectures.

Research and Teaching Assistant, Cairo University Sept 2012 to May 2015

Computer Engineering Department, Cairo University, Egypt.

- Research: Investigating high-performance cache management techniques to mitigate GPGPU cache contention. This work results in [GPGPU 2015] and [TPDS 2016] publications.
- Teaching: Computer Architecture, GPGPU Programming, Modeling & Simulation courses.

Software Engineer, Mentor Graphics Nov 2011 to July 2012

Mentor Graphics, Cairo, Egypt

I worked on R&D project which aims to build a new electronic design automation (EDA) tool for automated VLSI layout migration. My responsibilities included software performance profiling and improving the tool's execution time through algorithm optimization and parallel programming. My contributions resulted in 2.5x overall speedup.

Teaching Experience

Teaching Assistant at Computer Engineering Department, Cairo University

- Computer Architecture: Spring 2012, Spring 2013, Spring 2014
- Modeling & Simulation: Fall 2012, Fall 2013, Fall 2014
- GPGPU Programming: Summer 2013

Invited Talks

- "Scalable and Energy-Efficient SIMT Systems for Deep Learning and Data Center Microservices"
  - Intel AI, April 2022
  - University of Rochester, March 2022
  - AMD Research, February 2022
  - University of Central Florida, February 2022
- MICRO 2022 conference, "SIMR: Single Instruction Multiple Request Processing for Energy-Efficient Data Center Microservices", October 2022
- MICRO 2020 conference, "Locality-Centric Data and Threadblock Management for Massive GPUs", September 2020
- ISCA 2020 conference, "Accel-Sim: An Extensible Simulation Framework for Validated GPU Modeling", June 2020

Community Services

- ISPASS 2024 Travel Grant Chair
- Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD) Journal - External Reviewer - 2023
- ISPASS 2023 Program Committee Member and Poster Chair
- IEEE Transactions on Computer (TC) Journal - External Reviewer - 2022
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD) Journal - External Reviewer - 2022

Research Mentorships

- Vadim Nikiforov, Undergraduate Research. Currently a PhD student at Berkeley
  Research: Efficient CUDA kernel scheduling and data placement for multi-GPU systems
- Jason Shen, M.Sc. Research. Currently a hardware engineering at Intel
  Research: Demystifying modern GPU architecture through microbenchmarking
- Weili An, Undergraduate Research
  Research: Supporting and studying AMD ISA traces and architecture in Accel-Sim framework
- Cesar Avalos, 2nd year PhD student
  Research: Efficient GPU simulation of large scale workloads, like MLPerf benchmarks
- Ahmad Alawneh, 3rd year PhD student
  Research: Building a new simulation tool that identifies the SIMT GPU acceleration opportunities for MIMD CPU-based workloads

Technical Skills

Programming languages: C/C++, C#, Python, Assembly
Libraries and SDKs: CUDA, HIP, MPI, OpenMP/Pthread, PIN, LLVM
ML SDKs: PyTorch, VLLM, NVIDIA Nsight, NvProf, AMD ROCm SDK
Developments tools: Visual Studio, Linux development tools, Git Source Control
Simulators: GPGPU-Sim, Accel-Sim, SST, Gem5

Work Eligibility

Green card holder

References

References available upon request.