Ahmet Inci

Email: inciaf@gmail.com • Website: https://inciaf.github.io • GitHub: https://github.com/inciaf • Mobile: 412-494-8068

RESEARCH INTERESTS

- Computer Architecture, Machine Learning, Hardware/Software Co-Design

EDUCATION

Carnegie Mellon University, Pittsburgh, PA — Aug 2017 – Jul 2022

- Ph.D. in Electrical and Computer Engineering
  - Advisors: Prof. Diana Marculescu & Prof. Gauri Joshi
  - Dissertation: Scalable and Efficient Systems for Deep Learning
- M.Sc. in Electrical and Computer Engineering

Sabanci University, Istanbul, Turkey — Sep 2012 – Jul 2017

- Bachelor of Science (B.Sc.) in Electronics Engineering
  - **GPA:** 3.84 / 4.00, *Salutatorian, Summa Cum Laude*

WORK EXPERIENCE

NVIDIA

- Senior Deep Learning Performance Architect — Jan 2024 – Present
  - Developed architectures to extend the state of the art in DL performance and energy efficiency

Apple

- Machine Learning Engineer, Neural Engine Compiler Team — Aug 2022 – Jan 2024
  - Research and development on neural engine compiler for ultra low-power devices
  - Implemented new functional features in the compiler stack using Apple Neural Engine simulation environment, with an emphasis on performance and power
  - Brought up new hardware silicon and added support for new hardware features in the compiler
  - Developed features for future products, bringing those features from initial concept, through development, leading to hardware bringup and validation

NVIDIA

- Research Intern, Architecture Research Group (ARG) — May 2021 – Aug 2021
  - Project: Optimizing Power Management of Deep Learning Systems with Reinforcement Learning
  - Optimized power management for state-of-the-art deep learning systems and achieved performance/watt improvements on various MLPerf inference workloads
  - Developed an automatic fine-grained RL-based power management framework, which automated the process of determining the optimal power management approach
  - Collaborated closely with product and AI Research teams to drive RL-based power management of the next generation deep learning platforms
- Research Intern, Architecture Research Group (ARG) — May 2020 – Aug 2020
  - Project: Towards Scalable and Efficient Reinforcement Learning on CPU-GPU Systems
  - Investigated, analyzed, profiled, and optimized distributed reinforcement learning training workloads on state-of-the-art hardware and software platforms
  - Examined the interplay between HW and SW in DL workloads to identify areas for improvement
  - Analyzed performance, cost, and performance/watt trade-offs by performing system-level performance bottleneck analysis and introducing a new system design metric
  - Collaborated with product and AI Research teams to guide HW/SW co-design of future generations of CPU-GPU based deep learning platforms

ARM

- Research Intern, ML Technology Group — May 2019 – Aug 2019
  - Implemented HW-aware neural architecture search (NAS) algorithms for heterogeneous systems
  - Leveraged state-of-the-art NAS techniques and tools to design neural networks that are customized for mobile devices, resulting in more efficient and effective models

Cadence Design Systems

- Research Intern, Virtuoso ML Team — May 2018 – Aug 2018
  - Created a machine learning based recommendation system for EDA tools, particularly for Virtuoso in order to alleviate the designer's workload, reduce design time, and improve productivity

RESEARCH EXPERIENCE

Energy-Aware Computing Lab, Carnegie Mellon University

- Advisor: Prof. Diana Marculescu — Aug 2017 – Jul 2022
  - Designing scalable and efficient systems and ML models using HW/ML model co-design techniques to achieve the best of both worlds. I worked on quantization-aware DNN accelerator and model co-exploration through architecture-level modeling and efficient design space exploration. Before that, I worked on scalable and efficient reinforcement learning training on CPU-GPU systems. Additionally, my previous work has explored how to utilize emerging non-volatile memories in GPU architectures for DL workloads.

Performance and Energy-Aware Computing Lab, Boston University

- Advisor: Prof. Ayse Coskun — Jun 2016 – Sep 2016
  - Project: Temperature Dependent DRAM Power and Performance Model
  - Modeling 3D-stacked DRAM power consumption under various temperatures and embedding this temperature dependent power model into already existing DRAM simulators to optimize overall power and performance of 3D-stacked systems

Signal Processing and Information Systems Lab, Sabanci University

- Advisor: Prof. Mujdat Cetin — Jan 2015 – Jul 2017
  - I had multiple projects within the common theme of signal processing and machine learning. In my junior year, I worked on error-related potentials (ErrP) in brain-computer interfaces applications to better understand the relation between ErrP and error severity.

Neuroelectronics Lab, University of California, San Diego

- Advisor: Prof. Duygu Kuzum — Jun 2015 – Sep 2015
  - Calculating local field potentials (LFP) by using a network and performing simulations on NEURON simulator. Understanding the contributions of spikes and synaptic potentials to sharp wave-ripple complexes.

PUBLICATIONS

CONFERENCES

[1] **Inci, A.**, Isgenc, M., Marculescu, D., "**DeepNVM: A Framework for Modeling and Analysis of Non-Volatile Memory Technologies for Deep Learning Applications**" *DATE'20*

WORKSHOPS

[1] **Inci, A.**, Virupaksha, S., Jain, A., Thallam, V., Ding, R., Marculescu, D., "**QADAM: Quantization-Aware DNN Accelerator Modeling for Pareto-Optimality**" *ML for Computer Architecture and Systems Workshop (ISCA'21)*

[2] **Inci, A.**, Virupaksha, S., Jain, A., Thallam, V., Ding, R., Marculescu, D., "**QAPPA: Quantization-Aware Power, Performance, and Area Modeling of DNN Accelerators**" *2nd On-Device Intelligence Workshop (MLSys'21)*

[3] **Inci, A.**, Isgenc, M., Marculescu, D., "**Cross-Layer Design Space Exploration of NVM-based Caches for Deep Learning**" *12th Non-Volatile Memories Workshop (NVMW'21)*

[4] **Inci, A.**, Bolotin, E., Fu, Y., Dalal, G., Mannor, S., Nellans, D., Marculescu, D., "**The Architectural Implications of Distributed Reinforcement Learning on CPU-GPU Systems**" *6th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2'20)*

[5] **Inci, A.**, Marculescu, D., "**Solving the Non-Volatile Memory Conundrum for Deep Learning Workloads**" *8th Workshop on Architectures and Systems for Big Data, (ISCA'18)*

JOURNALS

[1] **Inci, A.**, Virupaksha, S., Jain, A., Chin, R., Thallam, V., Ding, R., Marculescu, D., "**QUIDAM: A Framework for Quantization-Aware DNN Accelerator and Model Co-Exploration**" *ACM Transactions on Embedded Computing Systems, September 2022*

[2] **Inci, A.**, Isgenc, M., Marculescu, D., "**DeepNVM++: Cross-Layer Modeling and Optimization Framework of Non-Volatile Memories for Deep Learning**" *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, November 2021*

[3] Canakci, S., Toy, M. F., **Inci, A.**, Liu X., and Kuzum, D., "**Computational Analysis of Network Activity and Spatial Reach of Sharp Wave-Ripples**" *PLoS One, September 2017*

BOOK CHAPTERS

[1] **Inci, A.**, Isgenc, M., Marculescu, D., “**Efficient Deep Learning Using Non-Volatile Memory Technology**” *Embedded Machine Learning for Cyber Physical, IoT, and Edge Computing , June 2023*

PATENTS

[1] **Inci, A.**, Loh, D., Meng, L., Suda, N., Kunze, E. “**Specializing Neural Networks for Heterogeneous Systems**” *US Patent Application 16/724,849, Filed: December 2019*

HONORS AND AWARDS

- Finalist for Qualcomm Innovation Fellowship 2020
  - Hardware-Aware Multimodal 3D Object Detection for On-Device Augmented Reality Applications
- Bob Lee Gregory Fellowship, Carnegie Mellon University 2019
- Best Project Award for *Hardware Architectures for Machine Learning* 2018
  - MAGNETO: Evaluation of Non-Volatile Memory Technologies for Deep Learning Workloads
- CMU ECE Finalist for Google PhD Fellowship 2018
- Best Project Runner-Up Award for *Energy-Aware Computing* 2017
  - Power/Performance Analysis and Optimization for Deep Learning on a CPU-GPU Platform
- Best Project Award for *Networks in the Real World* 2017
  - Who Speaks to Whom? Spatiotemporal Analysis of Phone Call Networks
- Carnegie Institute of Technology Dean’s Fellow 2017
- Graduated as *Salutatorian* (2nd highest ranking) student in Electronics Engineering Department 2017
- Dean’s High Honor List for all semesters 2013 – 2017
- Massachusetts Institute of Technology - Sabanci University Freshman Scholars Program 2015

  Chosen for MIT - Sabanci University Freshman Scholars Program for outstanding success in freshman courses
- Dilek Sabanci Scholarship, Sabanci University 2015

  Full-tuition scholarship with stipend for undergraduate studies. It is only given to 5 students each year.
- Sakip Sabanci Encouragement Scholarship, Sabanci University 2014

  Full-tuition scholarship with stipend for undergraduate studies
- Merit Scholarship, Sabanci University 2012 – 2017

  Awarded for ranking in top 0.15 percent among 1.8 Million participants in the Nationwide University Entrance Exam

SKILLS

- **Programming Languages:** Python, C / C++, C#, Verilog, Assembly, MATLAB, Java, SKILL
- **Tools:** TensorFlow, Caffe, PyTorch, gem5, GPGPU-Sim, HotSpot, DRAMSim2, McPAT, Sniper
- **CAD Tools:** Xilinx ISE, Cadence Virtuoso, Mentor Graphics ModelSim, Synopsys Design Compiler, Cadence SoC Encounter, Agilent ADS

COURSEWORK

Carnegie Mellon University, Pittsburgh, PA

- Hardware Architectures for Machine Learning, Energy-Aware Computing, Machine Learning, Computer Architecture and Systems, System-on-Chip Design, Networks in the Real World

Sabanci University, Istanbul, Turkey

- Computer Architectures, VLSI Systems Design, Data Structures, Operating Systems, Digital IC, Microcomputer Based System Design

TEACHING EXPERIENCE

Carnegie Mellon University, Pittsburgh, PA

- TA for Energy-Aware Computing (18-743) Fall 2018
  - Instructor: Prof. Diana Marculescu
  - Designed and evaluated research projects, graded reports, presentations, and homeworks, and held weekly office hours.
- TA for ULSI Technology Status and Roadmap for SoC and SiP (18-664) Fall 2020
  - Instructor: Prof. Andrzej Strojwas
  - Gave tutorials on several architectural tools, evaluated research projects and presentations.

Sabanci University, Istanbul, Turkey

- TA for Introduction to Computing (CS-201) Spring 2015
  - Instructor: Gulsen Demiroz
  - Held weekly office hours and helped students to overcome their problems on programming concepts.
- TA for Logic and Digital System Design (CS-303) Fall 2016
  - Instructor: Prof. Ilker Hamzaoglu
  - Held weekly office hours, supervised students in laboratory sessions, and evaluated their performances.

[*CV compiled on 2024-01-20*]