CAGRI ERYILMAZ

(they/them)

Results-driven computer architect and engineer with a proven track record in optimizing ML accelerators, SoC performance modeling, simulators & tools, high-level-to-HW mapping, and new architecture explorations. Specializes in benchmarking, PPA trade-offs, pre-silicon system exploration, and HW-SW co-design.

📧 email@cagri.email 🔗 github.com/66mhz 🌐 cagri.dev

WORK EXPERIENCE

SoC Architect
03/2025 - Current

Intel

- Developing comprehensive cycle-accurate simulators for next-generation datacenter and AI SoC architectures
- Designing simulation methodologies to support pre-silicon performance analysis and architectural decision-making

ML Hardware & Software Architect
03/2024 – 03/2025

Rain.AI

- Designed a new simulator framework on SystemC from scratch for compiler optimizations & architecture explorations for the next-gen product
- Ensured the accuracy is within 10% compared to cycle accurate simulator

ML Hardware Architect
03/2023 – 03/2024

Rain.AI

- Defined and designed performance architecture model for the first-generation accelerator IP, optimizing compute efficiency & utilization
- Conducted throughput and QoS analysis, identifying power-performance bottlenecks
- Modeled dataflows for the SoC and subsystems, optimizing interconnect bandwidth
- Conducted pre-silicon performance modeling, refining system architecture projections
- Designed our first SoC utilizing the developed SystemC simulator, significantly affecting the SoC architecture and microarchitecture
- Created the first HW model, achieving 10% accuracy in pre-silicon performance estimation, compared to RTL

SoC Architect
01/2022 – 03/2023

Intel

- Collaborated on implementing coherency protocol engine in C++ for multi-domain system simulation frameworks
- Contributed to developing inter-component communication protocols ensuring simulator coherence and data integrity

Deep Learning Engineer
04/2020 – 01/2022

Advanced Micro Devices

- Profiled deep learning frameworks for hardware-software co-optimization and inference tuning
- Collaborated with AMD ROCm team on application exploration and performance characterization for AI workloads
- Developed and optimized ML model training performance on ROCm using MI100 and MI200, increasing efficiency by ~20%
- Enabled key kernel implementations in MIGraphX deep learning inference library for AMD GPUs

SoC Performance Architect
05/2017 – 04/2020

Intel

- Performance tuned server chipsets and conformed PPA, developed special tools to scan parameters for best design
- Evaluated performance and validated RTL for a mobile SoC (Lakefield), identifying bottlenecks in execution pipelines
- Architected performance validation methodologies for Sapphire Rapids server platform chipset, including PCIe lane optimization and I/O subsystem analysis
- Evaluated and ensured performance for Tiger Lake & Alder Lake chipsets
- Developed automated performance analysis tools and methodologies used across multiple Intel SoC projects

Graduate Research Assistant
01/2015 – 01/2017

The University of Texas at Austin

- Admitted with PhD scholarship, focused on research throughout the time at the graduate school
- Thesis: Fine-Grain Acceleration of Graph Algorithms on Heterogeneous Chips
- Conducted research on compute performance for irregular workloads

EARLY CAREER & INTERNSHIPS

SoC Performance Architect Intern

01/2017 – 08/2016 (Multiple Terms)

Intel

- Performance tuned server chipsets and conformed PPA, developed special tools to scan parameters for best design
- Developed Python/Perl interfaces for RTL simulation environments, streamlining design validation processes

Research Engineer

06/2015 - 09/2015

AMD Research

- Developed an internal computer architecture simulator, enhancing design validation for next-gen architectures
- Implemented a memory module for virtual-to-physical address mapping, improving system performance analysis
- Conducted architectural studies, influencing future SoC designs

Research Intern

06/2013 - 09/2013

Barcelona Supercomputing Center

- Investigated cache reliability versus process variations using Cadence tools, leading to a published paper at DATE Conference
- Conducted simulations on memory architectures to assess performance impact of hardware-level variability

Teaching Assistant

08/2014 – 12/2016

The University of Texas at Austin

- Assisted in courses "Introduction to Computing" and "Circuit Theory", mentoring students in foundational computing concepts

EDUCATION

MSE - Thesis in Computer Architecture and Embedded Processors

The University of Texas at Austin - Electrical and Computer Engineering

08/2014 – 05/2017

Thesis: Fine – grain acceleration of graph algorithms in a heterogeneous chip (Written in OpenCL)

Courses: Computer Architecture, System-on-a-chip Design, Computer Architecture Parallelism and Locality, Microarchitecture, High-Speed Computer Arithmetic, Mobile Computing, Verification of Digital Systems, Intellectual Property

B.Sc. in Electrical and Electronics Engineering

Middle East Technical University, Ankara, Turkey

09/2009 – 06/2014

Minor in Computer Engineering

Middle East Technical University, Ankara, Turkey

09/2011 – 06/2014

Courses: Data Structures, Algorithms, Software Engineering, Parallel Computing, Data Communications and Networking, File Structures, Database Management Systems

TECHNICAL SKILLS

Programming

C/C++, Python, SystemC, TLM

Pre-Silicon Estimation

Power-performance trade-offs, cycle-accurate simulation, benchmarking

Performance Modeling & Architecture

SoC/CPU/GPU performance-power-area analysis, throughput & QoS modeling

CERTIFICATIONS

Structuring Machine Learning Projects
ML project approaches with practical industry decision making

Sequence Models
RNN, GRU, LSTM models for NLP, audio, speech recognition

Docker for DevOps
Container technologies for development operations

Improving Deep Neural Networks
Hyperparameter tuning, regularization, optimization algorithms

Introduction to Data Science in Python
Data manipulation, cleaning, statistical interpretation

Inspiring and Motivating Individuals
Team leadership, performance drivers, diversity

PROGRAMMING SKILLS

Daily Command

C/C++
Daily @Intel and @Rain.AI as SoC architect, inference library development at AMD

Python
Significant utilization at Intel and Rain.AI for LT and AT approaches

SystemC/TLM
Daily @AMD for deep learning/ML, various efforts at Intel

Good Command

CUDA
Graduate parallelization course, mainly CUDA with MPI

X86
6-stage pipelined x86 processor at UT-Austin

OpenCL
MS-Thesis work based on OpenCL, lasted two years

MPI
Graduate course for parallelization and computer architecture

Prior Experience

Java
Database development in undergraduate course

QEMU
ARM cores + accelerator implementation using Xilinx tools

Verilog
Six-stage pipelined x86 CPU design, FPGA game development

RTL
Kogge-stone parallel prefix adder power analysis

HONORS & AWARDS

★**2018 November** - Intel Corporation Division Recognition Award
★**2018 March** - Intel Corporation Award for Valuable Contributions
★**2010-2013** - High Honors at Middle East Technical University

PUBLICATIONS

C. Eryilmaz, A. Seyedi, O. Unsal & A. Cristal, **"Analysis of Random Dopant Fluctuations and Oxide Thickness on a 16nm L1 Cache Design"** *MEDIAN'14*

P. Reviriego, S. Can, C. Eryilmaz, J. Maestro & O. Ergin, **"Exploiting Processor Features to Implement Error Detection in Reduced Precision Matrix Multiplications"** *Microprocessors and Microsystems Journal*