Seonho Lee

seonho.lee@gatech.edu | linkedin.com/in/seonho-lee | +1 470 349 0845

Professional Summary

Master of Science in Electrical and Computer Engineering specializing in **computer architecture**, **hardware-software co-design**, **performance and power analysis**, and **machine learning systems**. Hands-on experience optimizing AI training GPU systems at **AMD**. First-author publications in the top-tier computer architecture conference **ASPLOS** and the reputable journal **CAL**.

Education

Georgia Institute of Technology — Atlanta, GA, United States
Master of Science in Electrical and Computer Engineering (GPA: 4.0/4.0) — *Aug. 2023 – May 2025 (Expected)*

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Bachelor of Science in Electrical Engineering (GPA: 4.01/4.3) — *Feb. 2023*

Experience

AMD — Bellevue, WA, United States
Research Intern – Manager: Dr. Zicheng Liu — *May 2024 – Aug. 2024*

- Conducted GPU power consumption analysis for LLM training, focusing on the influence of hardware architecture, clock frequency, and workload characteristics to enhance power efficiency and performance.

Georgia Institute of Technology — Atlanta, GA, United States
Graduate Research Assistant – Advisor: Dr. Divya Mahajan — *Aug. 2023 – Dec. 2024*

- Developed NeuSight, a performance modeling framework combining machine learning methods with tile-based decomposition and performance laws to predict GPU performance for unseen hardware and workloads
- NeuSight, trained solely on older GPU data, reduced latency prediction errors from 121.4% and 30.8% in prior works to 2.3% on the NVIDIA H100 GPU for GPT-3 training and inference, not in the training set
- Conducted kernel-level profiling using PyTorch and CUDA toolkits
- First-author publication in *ASPLOS*

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Research Intern – Advisor: Dr. Minsoo Rhu, Dr. Jongse Park — *Mar. 2021 – Jun. 2022*

- Designed HAMMER, a low-power hardware accelerator for self-attention in large language models, leveraging efficient approximation algorithms
- Achieved up to 1.6× performance and 1.5× energy efficiency improvements over state-of-the-art accelerators
- Developed C++ cycle-level simulator and Verilog RTL model
- Performed area and power estimation on Samsung 65nm node using Synopsys Design Compiler
- First-author publication in *IEEE CAL*

Publications

Forecasting GPU Performance for Deep Learning Training and Inference
Seonho Lee, Amar Phanishayee, Divya Mahajan
The 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Mar. 2025.
Paper: arxiv.org/abs/2407.13853 Code: github.com/sitar-lab/NeuSight

HAMMER: Hardware-friendly Approximate Computing for Self-Attention with Mean-Redistribution and Linearization
Seonho Lee, Ranggi Hwang, Jongse Park, Minsoo Rhu
IEEE Computer Architecture Letters (CAL), Jan. 2023.
Paper: ieeexplore.ieee.org/document/10005793

Technical Skills

Languages: C, C++, Python, CUDA, SystemVerilog, Bash, LaTeX
Frameworks and Tools: PyTorch, Pandas, NumPy, OpenGL, CUDA Toolkit, ModelSim, Git, Linux, Docker

Relevant Coursework

Computer Engineering: Computer Architecture, Parallel Computer Architecture, Digital System Design, Data Structures and Algorithms, Circuit Theory, Operating Systems, Computer Networks, Embedded Systems, Hardware Acceleration for Machine Learning, Hardware Software Co-Design for Machine Learning
Data Analysis: Machine Learning, Statistics, Linear Algebra, Discrete Mathematics, Big Data Analysis