Seonho Lee

seonho.lee08@gmail.com | linkedin.com/in/seonho-lee | +1 470 349 0845

Professional Summary

M.S. in ECE focused on computer architecture, GPU performance, and ML acceleration. Proficient in C/C++, Python, GPU Programming (CUDA), ML frameworks (PyTorch), and LLM optimizations.

Education

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

Korea Advanced Institute of Science and Technology (KAIST) Daejeon, South Korea
Bachelor of Science in Electrical Engineering (GPA: 4.0/4.3) *Feb. 2023*

Experience

Apple Greater Boston Area
Software Engineer *Jun. 2025 – Present*

- Machine learning framework engineer; GPU Graphics and Machine Learning team.
- Optimize LLM inference on large-scale GPU clusters.

AMD Greater Seattle Area
Research Intern *May 2024 – Aug. 2024*

- Analyzed GPU power consumption and performance trade-offs in LLM training. Focused this analysis on workload characteristics, clock frequency, and compute-communication overlap to improve power efficiency.
- Manager: Dr. Zicheng Liu

Georgia Institute of Technology Atlanta, GA, United States
Graduate Research Assistant *Aug. 2023 – May 2024*

- Developed NeuSight, a framework to forecast GPU performance for DL training/inference on unseen hardware and workloads.
- Characterized GPU performance and power implications of compute-communication overlap in distributed deep learning training.
- Advisor: Dr. Divya Mahajan

Korea Advanced Institute of Science and Technology (KAIST) Daejeon, South Korea
Research Intern *Mar. 2021 – Jun. 2022*

- Co-designed HAMMER, self-attention hardware accelerator. Developed its C++ cycle-level simulator, designed Verilog RTL, and performed synthesis for power/area estimation.
- Advisor: Dr. Minsoo Rhu, Dr. Jongse Park

Technical Skills

Languages: C, C++, Python, CUDA, SystemVerilog, Java, Bash

Frameworks & Tools: PyTorch, CUDA Toolkit, LLVM Compiler, NumPy, Pandas, Git, Linux, Docker

Publications

Characterizing Compute-Communication Overlap in GPU-Accelerated Distributed Deep Learning: Performance and Power Implications
Seonho Lee, Jihwan Oh, Seokjin Go, Divya Mahajan
IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), May 2025 (Poster)
Paper: arxiv.org/pdf/2507.03114

Forecasting GPU Performance for Deep Learning Training and Inference
Seonho Lee, Amar Phanishayee, Divya Mahajan
ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Mar. 2025
Paper: dl.acm.org/doi/10.1145/3669940.3707265 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