

TAEHYUN KIM

Dept. of ECE, Seoul National University, Seoul, South Korea

★ E-mail: taehyunzzz@snu.ac.kr; kimth@capp.snu.ac.kr; ★ Homepage: <https://taehyunzzz.github.io/>

Research Interests

- **Efficient AI Algorithms:** Parameter-Efficient Finetuning (PEFT), Speculative Decoding, Mixture-of-Experts (MoE)
- **Hardware Architecture:** Processing-in-Memory (PIM), Near-Data Processing (NDP), CXL Memory
- **High-Performance System Design:** Accelerator Parallelism, Algorithm-System Co-design, Memory Optimizations

Work Experience

Computer Architecture and Parallel Processing Lab @ SNU <i>Ph.D. Researcher</i>	Mar. 2019 – Present <i>Seoul, South Korea</i>
Computer Architecture and Parallel Processing Lab @ SNU <i>Undergraduate Researcher</i>	Aug. 2019 – Dec. 2019 <i>Seoul, South Korea</i>
DISCO HI-TEC <i>Undergraduate Intern</i>	Aug. 2018 <i>Seongnam, South Korea</i>
LG Electronics <i>Undergraduate Intern</i>	Jan. 2018 – Feb. 2018 <i>Seoul, South Korea</i>

Education

Seoul National University (SNU) <i>Doctor of Philosophy in Electrical and Computer Engineering</i> <ul style="list-style-type: none">• Advisor : Prof. Hyuk-Jae Lee• GPA: 3.94 / 4.3	Mar. 2019 – Present <i>Seoul, South Korea</i>
Seoul National University (SNU) <i>Bachelor of Science in Electrical and Computer Engineering</i> <ul style="list-style-type: none">• GPA: 3.61 / 4.3	Mar. 2013 – Feb. 2019 <i>Seoul, South Korea</i>
Tokyo Korean School (TKS) <i>High School Education</i>	Apr. 2010 – Feb. 2013 <i>Tokyo, Japan</i>

Publications (Featured)

- MoNDE: Mixture of Near-Data Experts for Large-Scale Sparse Models**
Taehyun Kim, Kwansoek Choi, Youngmook Cho, Jaehoon Cho, Hyuk-Jae Lee, Jaewoong Sim
61th ACM/IEEE Design Automation Conference (DAC), Jun. 2024.
- An FPGA-based Evaluation Platform for Testing Memory Prototype Chips**
Youngmook Cho, Taehyun Kim, and Hyuk-Jae Lee
International Conference on Electronics, Information, and Communication (ICEIC), Jan. 2024.
- Analyzing the Scaling Characteristics of Transformer Feed-forward Networks for the Trillion-Parameter Era and Beyond**
Taehyun Kim and Hyuk-Jae Lee
International Conference on Electronics, Information, and Communication (ICEIC), Jan. 2024.
- Exploring the Inefficiencies of a Large-scale Deep Neural Network Training Framework**
Taehyun Kim, Youngmook Cho, and Hyukjae Lee
International Conference on Electronics, Information, and Communication (ICEIC), Feb. 2023.
- Virtual Keyboards with Real-time and Robust Deep Learning-based Gesture Recognition**
Tae-Ho Lee, Sunwoong Kim, Taehyun Kim, Jin-Sung Kim, Hyuk-Jae Lee
IEEE Transactions on Human-Machine Systems (THMS), Volume 52 Issue 4, Apr. 2022.
- Smart Refrigerator Inventory Management using Convolutional Neural Networks**
Tae-Ho Lee, Shin-Woo Kang, Taehyun Kim, Jin-Sung Kim, Hyuk-Jae Lee
3rd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Jun. 2021.
- GradPIM: A Practical Processing-in-DRAM Architecture for Gradient Descent**
Heesu Kim, Hanmin Park, Taehyun Kim, Kwanheum Cho, Eojin Lee, Soojung Ryu, Hyuk-Jae Lee, Kiyoun Choi, Jinho Lee
27th IEEE International Symposium on High-Performance Computer Architecture (HPCA), Jan. 2021.

Patents

METHOD FOR ROUTING TOKEN AND APPARATUS THEREFOR

Inventor: Taehyun Kim, Jaewoong Sim, Jaehoon Cho, Hyuk-Jae Lee

Submitted: Dec. 10, 2024. (*Under Review*)

NEAR-DATA PROCESSING SYSTEM FOR LARGE-SCALE MIXTURE-OF-EXPERTS AI MODEL

Inventor: Taehyun Kim, Hyuk-Jae Lee, Jaewoong Sim, Jaehun Cho, Kanseok Choi, Youngmock Cho

Submitted: Dec. 29, 2023. (*Under Review*)

Research Projects

Intelligent In-Memory Error-Correction Device for Highly-Reliable Memory

- Mar. 2021 - Dec. 2024
- Served as the lead manager of the project.
- Developed and verified a hardware architecture of next-generation error-correcting code (ECC) memory.
- Hands-on HDL experience including Verilog design, hardware synthesis and FPGA emulation.
- Conducted research on near-data-processing (NDP) in emerging CXL memory that could work aside ECCs.
- 1 major publication, 2 patents under review.

High-Efficiency Deep Learning based on Memory-Aware Optimizations

- Mar. 2020 - Oct. 2020
- Participated in exploring optimizations for mixed-precision DNN training in an NPU accelerator setting.
- Analyzed the effect of on-chip buffers on energy efficiency and processing speed.
- Designed an PIM accelerator that exploits bank-level parallelism in modern DRAMs to accelerate the memory-intensive parameter update operation in mixed-precision DNN training settings.
- 1 major publication.

Smart Refrigerator Stock Management System

- Jun. 2018 - Dec. 2019
- Developed an automatic stock management system based on YOLOv3.
- Tracks the movement of stored items inside the refrigerator for stock management.
- Exhibited at CES 2020.

Scholarships & Funding

- Samsung, Device Solutions System LSI Division, scholarship contract

Skills & Tools

- Languages (proficiency): Korean (native), English (high), Japanese (high)
- Programming Languages: Python, C++, CUDA, Verilog
- DL Frameworks: PyTorch, HuggingFace, Deepspeed, TensorRT-LLM (FasterTransformer), vLLM
- Simulators: DRAMsim3, Ramulator
- Other tools: SystemC, Synopsys Design Compiler, ModelSim, NVIDIA NSight Systems, Xilinx FPGA, vim

Services

External Reviewer

42nd IEEE International Conference on Computer Design (ICCD)

2024

Research Assistant

Graduation Project for Undergraduate Students

Spring 2019, Spring & Fall 2024

Teaching Assistant

400.018 Creative Engineering Design

Fall 2019-2022

Republic of Korea Army

Administrative staff @ AFMC

2015-2016