Hojoon Leo Kim

hojoon.kim@snu.ac.kr | hojoonleokim.github.io

Research Interests

Machine Learning System, Efficiency, HW-SW Co-Design

As an **undergraduate researcher** passionate about Deep Learning and system-level engineering, I explore innovative solutions that enhance model efficiency through integrated hardware-software approaches. My work focuses on **co-design optimizations for emerging ML workloads**, spanning low-bit quantization, storage-assisted inference, and most recently, cache-driven planning for **embodied AI agents**, with publications at **OSDI'25**, **ICML'25 (Spotlight)**, and submissions to **MLSys'26**. I aim to develop system architectures that address the unique challenges of next-generation ML applications by bridging algorithmic innovations with hardware constraints for real-world deployment.

Education

- **Seoul National University** *Mar 2020 – Present*
  Undergraduate, Electrical and Computer Engineering Seoul, South Korea
  - GPA: 3.88/4.00 (overall: 4.01/4.30, major GPA: 4.10/4.30)
- **Stanford University** *Jun 2025 – Aug 2025*
  Visiting student researcher Stanford, CA, USA
  - GPA: 4.15/4.30; Courses: CS107 - Computer Architecture (A+), CS161 - Algorithm (A0)

Publications

C=Conference, S=In Submission

* Equal contribution

[S.2] Hojoon Kim, Yuheng Wu, and Thierry Tambe. **AgenticCache: Cache-Driven Asynchronous Planning for Embodied AI Agents.** *In submission to MLSys'26*

[S.1] Seong Hoon Seo, Donghyun Lee, Geonha Lee, Hojoon Kim, Yeonhong Park, and Jae W. Lee. **QUESO: Storage-Assisted Quantization Error Compensation for On-Device LLM Inference.** *In submission to MLSys'26*

[C.2] Seung Yul Lee, Hojoon Kim, Yutack Park, Dawoon Jeong, Seungwu Han, Yeonhong Park, and Jae W. Lee. **FlashTP: Fused, Sparsity-Aware Tensor Product for Machine Learning Interatomic Potentials.** *ICML'25* (Spotlight Poster, Acceptance Rate: 2.6%) [Paper PDF]

[C.1] Yeonhong Park*, Jake Hyun*, Hojoon Kim, and Jae W. Lee. **DecDEC: A Systems Approach to Advancing Low-Bit LLM Quantization.** *OSDI'25*. (Acceptance Rate: 15.9%) [Paper PDF]

Research and Work Experience

- **Stanford Tambe Lab [Tambe Lab]** *Jun 2025 – Present*
  Research Intern (P.I: Professor Thierry Tambe) Stanford, CA
  - Conducted research on cache-driven asynchronous planning for embodied AI agents by exploiting plan-level locality. (MLSys'26 submission)
- **SNU Architecture and Code Optimization Lab [ARC Lab]** *Sep 2024 – Present*
  Research Intern (P.I.: Professor Jae W. Lee) Seoul, South Korea
  - Researched error compensation at prefill stage with low-precision decoding to minimize GPU memory usage while improving model accuracy. (MLSys'26 submission)
  - Conducted figure design, experimental evaluation, and kernel performance analysis for Tensor-Product acceleration project. (ICML'25 Spotlight)
  - Built AWQ-quantized 3/3.5/4-bit LLMs using lutgemm-based weight packing (bitplane-wise packing + GEMV) for dynamic error compensation. (OSDI'25)
- **SNU Computer Architecture and Systems Lab [COMPARCH Lab]** *Dec 2023 – Sep 2024*
  Research Intern (P.I.: Professor Jaewoong Sim) Seoul, South Korea
  - Authored undergraduate thesis based on Tender (ISCA'24) quantization architecture; designed hardware module to shift offline computation to runtime, improving model accuracy.
  - Applied delayed-aggregation (Mesorasi) to point cloud models to analyze throughput and accuracy.
- **MODULABS [MODULABS]** *Jun 2021 – Sep 2021*
  Backend Software Engineer Seoul, South Korea
  - Developed backend for AI education platform: implemented payment processing system with time-synchronized functions and managed backend database infrastructure for AIFFEL, a new business initiative.

Academic Projects

- **A File System with Full-Path Indexing: Operating Systems** *Dec 2024*
  - Achieved 1st place (evaluated by throughput). Extended the xv6-riscv OS kernel to support full-path indexing and faster file access. Optimizations included (1) hash-based file lookup and (2) a dedicated sleep-wait stack to reduce scheduling overhead. Reduced file lookup latency from 24,000 to 993 cycles.
- **Pipelined Central Processing Unit with Perceptron Branch Predictor: Computer Organization** *Jun 2024*
- **L1, L2, and Memory Three-Level Memory Simulator with MSHR: Computer Organization** *Jun 2024*
- **A CNN Accelerator on FPGA(ARTIX-7): Digital Systems Design and Lab** *Dec 2023*
  - Ranked 1st based on throughput and task completion time. Optimized performance by addressing memory and compute bottlenecks: (1) implemented multi-bank BRAM architecture and DMA-driven dataflow to fetch data in parallel, resolving memory-bound issues; (2) used multiple output-stationary systolic arrays to boost compute throughput; and (3) pipelined the Load and Compute states to enable weight prefetch and overlap memory transfer with computation.

Honors and Awards

- **SNU Semiconductor Specialized Scholarship (USD 15,000)** *Jan 2024 – Present*
  SNU Semiconductor Specialization Project Group
- **Global Leadership Program (airfare and program stipend)** *Jun 2025 – Aug 2025*
  College of Engineering
- **Merit-Based Tuition Scholarship (70% of tuition)** *Aug 2023 – Jun 2025*
  Department of Electrical and Computer Engineering
- **1st Prize – Advanced CUDA Accelerator Programming School** *Feb 2025*
  SNU Thunder Research Group
- **4th Prize – Military AI Contest** *Dec 2022*
  Ministry of National Defense & Ministry of Science and ICT

Teaching Experience

- **Peer Tutor, Computer Organization; Programming Methodology** *Sep 2024 – Jun 2025*
  Department of Electrical and Computer Engineering

Extracurricular Activities

- **Member, SNU Business School Tennis Club** *Mar 2024 – Dec 2024*
- **Peer Mentor, SNU Center for Campus Life & Culture** *Mar 2024 – Jun 2024*
  - Received the Best Mentor Award and Best Mentoring Team Award.
- **SIGINT Specialist, Republic of Korea Army** *Jan 2022 – Jul 2023*
  - Two-time recipient of the Best SIGINT Analyst Award; recognized as a First-Class Soldier (fitness, mental readiness, and specialty proficiency).

Relevant Courses

- System Programming (Prof. Jae W. Lee)
- Operating Systems (Prof. JinSoo Kim)
- Computer Organization (Prof. Jaewoong Sim)
- Advanced Computer Architecture (Prof. Sungjoo Yoo)
- Digital Systems Design (Prof. Jaewoong Sim)
- Advanced Compiler (Prof. SooMook Moon)
- Computer Networks (Prof. KyoungSoo Park)
- Linear Algebra (Prof. Young-Hoon Kiem)
- Parallel Distributed Computing (Prof. Jinho Lee)
- Computer Organization & Design (Prof. Jangwoo Kim)

Technical Strengths

- **Computer Languages:** C/C++, Verilog/SystemVerilog, Python
- **Frameworks and Skills:** CUDA, PyTorch, Nsight System/Compute

Language Proficiency

Korean (Native), English (TOEFL 106 [R:30/L:28/S:22/W:26])