

Sanghyeon Lee

Jul. 22, 2026

leesh6796@casys.kaist.ac.kr

Education

KAIST , Ph.D. in Computer Science Advisor: Prof. Jaehyuk Huh Dissertation: <i>SLO-aware Inference Serving Systems for Multi-Modal Generative-AI Workloads</i>	Mar. 2022 – Aug. 2026
KAIST , M.S. in Computer Science Advisor: Prof. Jaehyuk Huh Thesis: <i>Architectural Support for Exploiting Memory Locality of Spatial Multi-tenant Neural Processing Units</i>	Mar. 2020 – Feb. 2022
KAIST , B.S. in Computer Science	Mar. 2015 – Feb. 2020

Experience

Postdoctoral Researcher Computer Architecture and Systems Laboratory (CASys) School of Computing, KAIST	Aug. 2026 –
--	-------------

Research Interests

-
- | | |
|--|---|
| ◦ Multi-modal LLM Inference Serving | ◦ GPU and NPU Architecture |
| ◦ Processing-in-Memory for LLM Inference | ◦ Hierarchical KV Cache Offloading System |

Publications

A Hybrid Processing-in-Memory Architecture for Long Sequence LLM Inference with KV Cache Filtering
Soojin Hwang, **Sanghyeon Lee**, Juhyun Lee, and Jaehyuk Huh
The International Conference on Parallel Architectures and Compilation Techniques (PACT), 2026

An Operand Divergence-Aware All-Bank PIM Architecture for Sparse Attention
Soojin Hwang, **Sanghyeon Lee**, Juhyun Lee, and Jaehyuk Huh
IEEE Computer Architecture Letters (CAL), 2026

Throughput-Oriented LLM Inference via KV-Activation Hybrid Caching with a Single GPU
Sanghyeon Lee, Hongbeen Kim, Soojin Hwang, Guseul Heo, Minwoo Noh, and Jaehyuk Huh
The 43rd IEEE International Conference on Computer Design (ICCD), Nov. 2025

Improving Multi-tenant NPU Efficiency via Decoupled Tiling and Adaptive Memory Allocation
Sanghyeon Lee and Jaehyuk Huh
International Conference on ICT Convergence (ICTC), 2025

NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing
Guseul Heo, Sangyeop Lee, Jaehong Cho, Hyunmin Choi, **Sanghyeon Lee**, Hyungkyu Ham, Gwangsun Kim, Divya Mahajan, and Jongse Park
The ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Apr. 2024

Supporting Trusted Virtual Machines with Hardware-based Secure Remote Memory
Taekyung Heo, Seunghyo Kang, **Sanghyeon Lee**, Soojin Hwang, Joongun Park, and Jaehyuk Huh
The International Symposium on Memory Management (ISMM), Jun. 2024

Hardware Hardened Sandbox Enclaves for Trusted Serverless Computing
Joongun Park, Seunghyo Kang, **Sanghyeon Lee**, Taehoon Kim, Jongse Park, Youngjin Kwon, and Jaehyuk Huh
IEEE Transactions on Architecture and Code Optimization (TACO), 2024

Improving Data Reuse in NPU On-chip Memory with Interleaved Gradient Order for DNN Training
Jungwoo Kim, Seonjin Na, **Sanghyeon Lee**, Sunho Lee, and Jaehyuk Huh
The 56th IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 2023

Teaching Experience

CS510 Computer Architecture Teaching Assistant, KAIST School of Computing	2024
CS311 Computer Architecture Teaching Assistant, KAIST School of Computing	2022, 2023
CS230 System Programming Teaching Assistant, KAIST School of Computing	2020, 2022, 2023
CS206 Data Structures Teaching Assistant, KAIST School of Computing	2020, 2021

Service

Ad hoc Reviewer , <i>IEEE Transactions on Computers (TC)</i>	2026
---	------

Honors & Awards

Outstanding Teaching Assistant Award , CS311 Computer Architecture, KAIST	Spring 2022
Outstanding Teaching Assistant Award , CS206 Data Structures, KAIST	Spring 2021