Sungmin Yun

Computer Architectures and Systems for AI

Email: sungmin.yun11@gmail.com
GitHub: github.com/sungmin-yun
Web: sungminyun.me
Phone: +82-10-2745-7538

About Me

Computer architecture and systems researcher at Seoul National University, focusing on **AI acceleration and large-scale LLM serving system**. I have extensive hands-on experience in **accelerator design, hardware profiling, kernel programming, and performance modeling**, along with strong expertise in **modern AI model architectures** (e.g., MoE, MLA, RAG). I am currently leading research on accelerating Multi-LoRA inference and maximizing hardware utilization in large-scale LLM serving. I also developed an **open-source end-to-end LLM performance simulator**, LLMSimulator 🔗, enabling rapid evaluation of emerging hardware and AI model architectures. My research aims to **reduce the computational cost of AI models and make them broadly accessible**.

Selected Publications

[MICRO 2024] Duplex: A Device for Large Language Models with Mixture of Experts, Grouped Query Attention, and Continuous Batching, **S. Yun**, K. Kyung, J. Cho, J. Choi, J. Kim, B. Kim, S. Lee, K. Sohn, J. H. Ahn, (Acceptance Rate: **22**% (113 among 497))

[ICS 2024] CLAY: CXL-based Scalable NDP Architecture Accelerating Embedding Layers, **S. Yun**, H. Nam, K. Kyung, J. Park, B. Kim, Y. Kwon, E. Lee, J. H. Ahn, (Acceptance Rate: **36**% (45 among 125))

[IEEE TC 2023] GraNDe: Efficient Near-Data Processing Architecture for Graph Neural Networks, **S. Yun**, H. Nam, J. Park, B. Kim, J. H. Ahn, E. Lee, (**Best of CAL Award**)

Experience

Seoul National University Sep. 2025 – Present
Postdoctoral Researcher
- Leading research projects to accelerate Multi-LoRA inference at scale
- Developing system-level guidelines for efficient LLM serving with modern architectures (e.g., MLA, MoE)
- Advising on research projects to accelerate RAG pipelines

Samsung Electronics Mar 2023 – Aug 2023
Internship
- Implemented an controller in RTL to perform general matrix–vector multiplication using HBM-PIM.
- By analyzing the limitations of HBM-PIM, I designed an LLM accelerator (Duplex) using custom HBM integrated with xPUs.

Education

Seoul National University, Postdoctoral Researcher — Sep 2025 – Present

Seoul National University, Ph.D. in Artificial Intelligence — GPA: 4.1/4.3 — Aug 2020 – Aug 2025

Yonsei University, B.S. in Integrated Information Technology — GPA: 3.6/4.3 — Mar 2017 – Feb 2020

Technical Skills

Languages	C/C++, CUDA, Python, System Verilog
Frameworks & Tools	Pytorch, vLLM, SGLang, Nsight Compute, Nsight Systems, Ramulator

Honors & Awards

Best Ph.D. Dissertation Award Aug 2025
Seoul National University

Best of CAL Award Feb 2023
29th IEEE International Symposium on High-Performance Computer Architecture

Invited Presentations

Poster presentation at Samsung AI Forum — Nov 2023

Oral presentation on Best of CAL session at IEEE International Symposium on High-Performance Computer Architecture — Feb 2023

Program Committee and Paper Review

Lightweight Program Committee Member, ISCA 2026

IEEE Transactions on Computers

IEEE Transactions on Circuits and Systems for Artificial Intelligence

IEEE Computer Architecture Letters

Future Generation Computer Systems

Full Publications

[arXiv 2025] The New LLM Bottleneck: A Systems Perspective on Latent Attention and Mixture-of-Experts, **S. Yun**, S. Park, H. Nam, Y. Lee, G. Lee, K. Kyung, S. Kim, N. S. Kim, J. Kim, H. Kim, J. Cho, S. Baek, J. H. Ahn

[IEEE CAL 2025] SSD Offloading for LLM Mixture-of-Experts Weights Considered Harmful in Energy Efficiency, K. Kyung, **S. Yun**, J. H. Ahn

[IEEE CAL 2025] COSMOS: A CXL-Based Full In-Memory System for Approximate Nearest Neighbor Search, S. Ko, H. Shim, W. Doh, **S. Yun**, J. So, Y. Kwon, S. Park, S. Roh, M. Yoon, T. Song, J. H. Ahn

[HPCA 2025] Anaheim: Architecture and Algorithms for Processing Fully Homomorphic Encryption in Memory, J. Kim, **S. Yun**, H. Ji, W. Choi, S. Kim, J. H. Ahn

[MICRO 2024] Duplex: A Device for Large Language Models with Mixture of Experts, Grouped Query Attention, and Continuous Batching, **S. Yun**, K. Kyung, J. Cho, J. Choi, J. Kim, B. Kim, S. Lee, K. Sohn, J. H. Ahn

[ICS 2024] CLAY: CXL-based Scalable NDP Architecture Accelerating Embedding Layers, **S. Yun**, H. Nam, K. Kyung, J. Park, B. Kim, Y. Kwon, E. Lee, J. H. Ahn

[IEEE TC 2023] GraNDe: Efficient Near-Data Processing Architecture for Graph Neural Networks, **S. Yun**, H. Nam, J. Park, B. Kim, J. H. Ahn, E. Lee

[IEEE CAL 2022] GraNDe: Near-Data Processing Architecture With Adaptive Matrix Mapping for Graph Convolutional Networks, **S. Yun**, B. Kim, J. Park, H. Nam, J. H. Ahn, E. Lee

[MICRO 2021] TRiM: Enhancing Processor-Memory Interfaces with Scalable Tensor Reduction in Memory, B. Kim, J. Park, **S. Yun**, E. Lee, M. Rhu, J. H. Ahn