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

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

Summary

My current work focuses on **designing efficient computer architectures and serving systems for large-scale AI models** through algorithm–hardware co-design and system-level optimization. To this end, I continuously track the latest trends in both AI models and hardware developments.

My expertise lies in computer architectures for running AI models and in understanding their computational characteristics. I have developed accelerators for various AI models, including Large Language Models (LLMs), Graph Neural Networks (GNNs), and Recommender Systems (RecSys). These projects have given me a thorough knowledge of the computational behavior of commonly used layers in AI models. By closely analyzing recent LLM trends, I have also gained strong expertise in modern architectures such as Mixture-of-Experts (MoE) and Multi-Head Latent Attention (MLA).

In addition, I specialize in serving systems for deploying AI models at scale. I understand how model parallelism affects inference performance—both latency and throughput—in multi-device serving environments. I also possess extensive knowledge of system-level techniques such as disaggregated systems, chunked prefill, and prefix caching, and I have practical experience with widely used inference frameworks, including vLLM and TensorRT.

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

Experience

Postdoctoral Researcher, Seoul National University — Sep 2025 – Present

Samsung Electronics Internship — Mar 2023 – Aug 2023

Mandatory Military Service as Professional Research Personnel — Sep 2023 – Present (expected Aug 2026)

Awards

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

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

Paper Review

IEEE Transactions on Computers

Skills

C/C++, Python, PyTorch, CUDA, Verilog