Sungmin Yun

sungmin.yun11@gmail.com
sungmin.yun@snu.ac.kr
sungmin.yun@scale.snu.ac.kr

Cell phone: +82-10-2745-7538
Google Scholar(link)

EDUCATION

Sep. 2025 ~ Current	**Seoul National University** Postdoctoral Researcher at Computer Laboratory Advisor: Jung Ho Ahn	Seoul, Korea
Sep. 2020 ~ Aug. 2025	**Seoul National University** *Doctor of Philosophy* in Artificial Intelligence College of Engineering, Interdisciplinary Program in Artificial Intelligence Advisor: Jung Ho Ahn	Seoul, Korea
Mar. 2017 ~ Feb. 2020	**Yonsei University** *Bachelor of* Science in Engineering College of Engineering, Department of Integrated Information Technology	Seoul, Korea
Mar. 2014 ~ Feb. 2017	**Seoul Science High School** Specialized high school for students talented in math and science	Seoul, Korea

INTERNATIONAL CONFERENCES AND JOURNALS

[1] K. Kyung, S. Yun, J. H. Ahn, "SSD Offloading for LLM Mixture-of-Experts Weights Considered Harmful in Energy Efficiency", in *IEEE Computer Architecture Letters, Early Access*

[2] S. Ko, H. Shim, W. Doh, **S. Yun**, J. So, Y. Kwon, S. Park, S. Roh, M. Yoon, T. Song, J. H. Ahn, "COSMOS: A CXL-Based Full In-Memory System for Approximate Nearest Neighbor Search", in *IEEE Computer Architecture Letters,* vol. 24, Issue. 1, 2025.

[3] J. Kim, **S. Yun**, H. Ji, W. Choi, S. Kim, J.H.Ahn, "Anaheim: Architecture and Algorithms for Processing Fully Homomorphic Encryption in Memory", in 2025 *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025.

[4] **S. Yun,** K. Kyung, J. Cho, J. Choi, J. Kim, B. Kim, S. Lee, K. Sohn, J. H. Ahn, "Duplex: A Device for Large Language Models with Mixture of Experts, Grouped Query Attention, and Continuous Batching", in *57th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2024.

[5] **S. Yun,** H. Nam, K. Kyung, J. Park, B. Kim, Y. Kwon, E. Lee and J.H. Ahn, "CLAY: CXL-based Scalable NDP Architecture Accelerating Embedding Layers," in *Proceedings of the 38th ACM International Conference on Supercomputing(ICS), 2024*.

[6] **S. Yun**, H. Nam, J. Park, B. Kim, J. H. Ahn and E. Lee, "GraNDe: Efficient Near-Data Processing Architecture for Graph Neural Networks," in *IEEE Transactions on Computers, 2023*.

[7] **S. Yun**, B. Kim, J. Park, H. Nam, J.H. Ahn and E. Lee, "GraNDe: Near-Data Processing Architecture With Adaptive Matrix Mapping for Graph Convolutional Networks," in *IEEE Computer Architecture Letters*, vol. 21, no. 2. (Best of CAL award at *29th* IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2023)

[8] B. Kim, J. Park, **S. Yun**, E. Lee, M. Rhu, and J. H. Ahn, "TRiM: Enhancing Processor-Memory Interfaces with Scalable Tensor Reduction in Memory," in *54th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2021.

AWARDS AND HONORS

Feb. 2023 Best of CAL award at *29th* IEEE International Symposium on High-Performance Computer Architecture

Mar. 2017 ~ Feb. 2019 Undergraduate Fellowship – ICT Consilience Creative Program (full tuition, monthly scholarship) Ministry of Science and ICT, Korea

INVITED PRESENTATIONS

Nov. 2023 Poster presentation at Samsung AI Forum

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