Jungi Lee

jungi.lee@snu.ac.kr Seoul, Republic of Korea jungi-lee.github.io

RESEARCH INTERESTS

Computer Architecture, Microarchitecture for Emerging Workloads, Efficient AI Systems

EDUCATION

Seoul National University, Seoul, Republic of Korea *03/2023 - Present*
M.S./Ph.D. in Electrical and Computer Engineering
Computer Architecture and Systems Lab, advised by **Prof. Jaewoong Sim**.

Seoul National University, Seoul, Republic of Korea *03/2017 - 02/2023*
B.S. in Electrical and Computer Engineering
GPA: 3.87/4.30, Major GPA: 3.98/4.30

PROJECTS

- **Accelerator system for Large Language Model inference through algorithm-hardware co-design.**
  - Under low-bit inference, it achieves up to 2.63× speedup on average over other accelerators with higher accuracy.
- **Dynamic key-value cache management solution for efficient generative inference in large language model.**
  - Novel KV cache management framework that provides scalability under long-text generation, while achieving up to 3.00× speedup over other management methods.

PUBLICATIONS

[OSDI '24] InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management
Wonbeom Lee*, Jungi Lee*, Junghwan Seo, and Jaewoong Sim
Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (**OSDI**), 2024

[ISCA '24] Tender: Accelerating Large Language Models via Tensor Decomposition and Runtime Requantization
Jungi Lee*, Wonbeom Lee*, and Jaewoong Sim
Proceedings of the 51st ACM/IEEE International Symposium on Computer Architecture (**ISCA**), 2024

[ASPLOS '24] GSCore: Efficient Radiance Field Rendering via Architectural Support for 3D Gaussian Splatting
Junseo Lee, Seokwon Lee, Jungi Lee, Junyong Park, and Jaewoong Sim
Proceedings of the 2024 International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**), 2024

[ISCA '23] NeuRex: A Case for Neural Rendering Acceleration
Junseo Lee, Kwanseok Choi, Jungi Lee, Seokwon Lee, Joonho Whangbo, and Jaewoong Sim
Proceedings of the 50th ACM/IEEE International Symposium on Computer Architecture (**ISCA**), 2023

PATENTS

Accelerator and operating method using the same (1020240036408)
with Jaewoong Sim, Wonbeom Lee

SKILLS

- **Languages:** C/C++, CUDA, Python
- **Applications/Frameworks:** PyTorch, TVM, Intel Pin, LaTeX