

# Wonbeom Lee

☎ (+82) 02-880-1779 📍 Seoul, South Korea ✉ wonbeom@snu.ac.kr 🏠 leewonbeom.github.io

## RESEARCH INTERESTS

---

Systems for AI, Computer Architecture, Hardware-Software Co-design

## EDUCATION

---

**M.S./Ph.D. in Electrical and Computer Engineering**

03/2023-Present

Seoul National University

Computer Architecture and Systems Lab (SNU-CompArch)

**B.S. in Electrical and Computer Engineering**

03/2019-08/2022

Seoul National University

Early Graduation, GPA: 3.84/4.30, major GPA: 3.94/4.30

## SELECTED PUBLICATIONS

---

[OSDI '24] **InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management**

Wonbeom Lee\*, Jungi Lee\*, Junghwan Seo, Jaewoong Sim

Acceptance Rate: 49/282  $\approx$  17.4%

[ISCA '24] **Tender: Accelerating Large Language Models via Tensor Decomposition and Runtime Requantization**

Jungi Lee\*, Wonbeom Lee\*, Jaewoong Sim

## PATENTS

---

**Accelerator and operating method using the same (1020240036408)**

with Jaewoong Sim, Jungi Lee

## RESEARCH EXPERIENCES

---

**Research Assistant**

03/2023-Present

Seoul National University (Advisor: Prof. Jaewoong Sim)

- **Tender: Accelerating Large Language Models via Tensor Decomposition and Runtime Requantization**

- Algorithm-hardware co-design solution that offers high performance and accuracy without the need of mixed-precision compute units or custom data types even for low-bit quantization.
- Decomposed quantization technique in which the scale factors of the decomposed matrices have multiples of integer two relationships for implicit requantization with negligible rescaling overhead and minimal hardware extension.
- Up to  $2.63\times$  speedup on average over other outlier-aware accelerators. Less than a 0.1 increase in perplexity for INT8 quantization and a lower perplexity than any other outlier-aware quantization techniques for INT4 quantization.

- **InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management**

- Novel KV cache management framework tailored for long-text generation, which synergistically works with modern offloading-based inference systems.
- Minimal rehearsal with the input of the current layer can speculate a few important tokens that are essential for computing the subsequent attention layer which minimizes the data transfer overhead in offloading-based LLM serving systems.
- Up to  $2.98\times$  speedup over the existing KV cache management methods while providing better model accuracy.

## SKILLS

---

- **Languages:** C/C++, Python

- **Applications/Frameworks:** PyTorch, Intel Pin, LaTeX