

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

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%

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

Jungi Lee*, Wonbeom Lee*, Jaewoong Sim

Acceptance Rate: 83/423 \approx 19%

RESEARCH EXPERIENCES

Research Assistant

03/2023-Present

Seoul National University (Advisor: Prof. Jaewoong Sim)

- **InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management**
 - A novel KV cache management framework tailored for long-text generation, which synergistically works with modern offloading-based inference systems.
 - A minimal rehearsal with the inputs of the current layer enables prefetching a few important tokens that are essential for computing the subsequent layer, thereby mitigating the data transfer overhead in offloading-based LLM serving systems.
 - Up to $3.00\times$ speedup over the existing KV cache management methods while offering substantially better model accuracy.
- **Tender: Accelerating Large Language Models via Tensor Decomposition and Runtime Requantization**
 - An 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.
 - A decomposed quantization technique in which the scale factors of the decomposed matrices are powers of two apart, enables implicit requantization with negligible rescaling overhead and minimal hardware extension.
 - Up to $2.63\times$ speedup on average over the outlier-aware accelerators while achieving better model accuracy.

TEACHING EXPERIENCES

Seoul National University

Spring 2023

Graduate Teaching Assistant

• **430.322: Computer Organization**

Designed part of course projects, led recitation sessions, and graded projects/exams

SKILLS

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