MINSU KIM

Mail: mskim@casys.kaist.ac.kr
Homepage: kms040411.github.io

EDUCATION

2017 - 2021 **B.S. in Computer Science, KAIST**

2021 - 2023 **M.S. in Computer Science, KAIST**
Advised by Prof. Jongse Park

2023 - Current **Ph.D. in Computer Science, KAIST**
Advised by Prof. Jongse Park

EXPERIENCE

July 2023 - August 2024 **Intern, Software Developer**
HyperAccel
Developed software stack (Python) and compiler (C++) for large language model inference accelerator.

PUBLICATIONS

- M. Kim*, S. Hong*, R. Ko, S. Choi, H. Lee, J. Kim, J. Kim, J. Park, "Oaken: Fast and Efficient LLM Serving with Online-Offline Hybrid KV Cache Quantization", in International Symposium on Computer Architecture (ISCA), June 2025.
**: Equal contribution*
- S. Moon, J. Kim, J. Kim, S. Hong, J. Cha, M. Kim, S. Lim, G. Choi, D. Seo, J. Kim, H. Lee, H. Park, R. Ko, S. Choi, J. Park, J. Lee, J. Kim, "LPU: A Latency-optimized and Highly Scalable Processor for Large Language Model Inference" in IEEE Micro, special issue on Contemporary Industry Products, 2024.
- J. Cho, M. Kim, H. Choi, G. Heo, J. Park, "LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving Systems at Scale" in IEEE International Symposium on Workload Characterization (IISWC), September 2024.
- M. Kim, J. Hwang, G. Heo, S. Cho, D. Mahajan, J. Park, "Accelerating String-key Learned Index Structures via Memoization-based Incremental Training" in International Conference on Very Large Databases (VLDB), August 2024.
- J. Hwang, M. Kim, D. Kim, S. Nam, Y. Kim, D. Kim, H. Sharma, J. Park, "CoVA: Exploiting Compressed-Domain Analysis to Accelerate Video Analytics" in USENIX Annual Technical Conference (ATC), July 2022.

AWARDS

- Best Paper Award & Distinguished Artifact Award
At IISWC 2024, "LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving Systems at Scale"
- Best Paper Award
At KAIST, "Accelerating String-key Learned Index Structures via Memoization-based Incremental Training"

SKILLS

Programming Languages: Python, C, C++, Chisel, SystemVerilog

Technologies: Linux, PyTorch, Triton, CUDA

TEACHING ASSISTANT

KAIST	CS230: System Programming	2021 Fall
KAIST	CS311: Computer Organization	2022 Spring
KAIST	CS510: Computer Architecture	2023 Spring
KAIST	CS311: Computer Organization	2024 Spring
KAIST	CS230: System Programming	2024 Fall

REFERENCE

Jongse Park, Associate Professor, KAIST *jspark@casys.kaist.ac.kr*