Jaehong Cho

Master Student

(82)10-2963-9902 | Daejeon, South Korea
jhcho@casys.kaist.ac.kr | linkedin.com/in/jaehongcs20 | github.com/jaehongcs20

Summary

Master's student at the School of Computing, KAIST, in the Computer Architecture and Systems Lab under the supervision of Prof. Jongse Park. Specializing in Processing-in-Memory(PIM) and LLM inference systems. Experienced in working with and modifying various hardware, system, and serving open-source simulators. Published papers on LLM inference acceleration with PIM and developing a new large-scale LLM inference simulation infrastructure, earning recognition for innovative contributions. Interested in exploring advanced scheduling techniques and heterogeneous systems to enhance next-generation AI serving frameworks.

Education

KAIST — Daejeon, South Korea

Master of Science in School of Computing — Sep 2024 – Aug 2026 (Expected)

- Computer Architecture and Systems Lab under the supervision of Prof. Jongse Park

KAIST — Daejeon, South Korea

Bachelor of Science in School of Computing, Minor in Electrical Engineering (Summa Cum Laude) — Mar 2020 – Aug 2024

- GPA of 4.14/4.3

Internship

Undergraduate Student Intern — KAIST – Daejeon, South Korea

Computer Architecture and Systems Lab (Prof. Jongse Park) — Dec 2021 – Aug 2024

- Accelerating LLM inference using NPU-PIM heterogeneous system

Project Experience

Building Systolic Array System in FPGA to Accelerate CNN — KAIST – Daejeon, South Korea

Team Member — Fall 2023

- Course Project for Systems for Artificial Intelligence, School of Computing, KAIST
- Developed a systolic array on a PYNQ-Z2 FPGA board to accelerate the GEMM operation in YOLO-v2-tiny

Reducing Energy Consumption in On-device Image and Sound Classification — KAIST – Daejeon, South Korea

Team Leader — Spring 2023

- Course Project for Operating Systems, School of Computing, KAIST
- Used MobileNet for image classification and YamNet for audio classification with TensorFlow Lite (TFLite)
- Using conditional sensing and adaptive inference rate, achieved a 2.44x reduction in energy consumption

Enhancing AudioCaps with a Transformer-based Architecture — KAIST – Daejeon, South Korea

Team Leader, Presenter — Fall 2022

- Course Project for Introduction to Artificial Intelligence, School of Computing, KAIST
- AudioCaps generates captions for audio using a bi-directional LSTM, which struggles to capture long-range dependencies
- Replaced the bi-directional LSTM encoder layers with transformer encoder layers, achieved a 14.3% improvement in accuracy

Publications

LLMServingSim2.0: A Unified Simulator for Heterogeneous Hardware and Serving Techniques in LLM Infrastructure

1st Author — **IEEE Computer Architecture Letters (CAL) 2025**

LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving at Scale **IISWC 2024**
1st Author

- Best Paper Award & Distinguished Artifact Award

LLMServingSim: A Simulation Infrastructure for LLM Inference Serving Systems **MLArchSys 2024**
1st Author

NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing **ASPLOS 2024**
3rd Author

Awards & Achievements

Sep 2024	**Best Paper Award & Distinguished Artifact Award,** LLMServingSim paper	**IISWC 2024**
Sep 2024	**Encouragement Award,** Undergraduate Research Participation (URP) Program	**KAIST**
Spring 2023	**Dean's List,** for Academic Excellence	**College of Engineering, KAIST**
Fall 2022	**Dean's List,** for Academic Excellence	**College of Engineering, KAIST**
Spring 2022	**Dean's List,** for Academic Excellence	**College of Engineering, KAIST**
Fall 2021	**Dean's List,** for Academic Excellence	**College of Engineering, KAIST**
Spring 2021	**Semester High,** for top-of-the-class throughout the semester	**School of Computing, KAIST**

Teaching Experience

Teaching Assistant **KAIST – Daejeon, South Korea**

2024 Fall CS230 System Programming TA (Prof. Jongse Park) Sep 2024 – Dec 2024

Skills

Programming Languages

- **Proficient:** Python, C/C++
- **Familiar:** Verilog, Chisel, Scala, CUDA

Software Proficiency

- **Development Tools:** Git, Docker, Conda
- **Machine Learning & Data Science Tools:** PyTorch, TensorFlow, ONNX, Pandas, Matplotlib, NumPy, Scikit-learn
- **Open-Source Hardware Simulators:** Ramulator, ZSim, DRAMSim3, Accel-Sim, Genesys, NeuPIMs, mNPUSim, ONNXim
- **Open-Source System Simulators:** ASTRA-Sim, LLMServingSim
- **Open-Source LLM Serving Frameworks:** NVIDIA Triton, NVIDIA Dynamo, vLLM, SGLang, LMCache
- **Open-Source LLM Optimization Libraries:** NVIDIA FasterTransformer, FlashAttention
- **Profiling Tools:** NVIDIA Nsight, Intel VTune

English Proficiency

- **TOEIC Score: 955 (Jul 2021)**

Interests

LLM Inference System

- **Request scheduling and memory management for efficient LLM inference**

Processing-in-Memory (PIM)

- **Accelerating LLM inference using heterogeneous systems with NPU and PIM**

Large Distributed System

- **Task scheduling and parallelization schemes for large-scale LLM inference**