

Guseul Heo

Jan 6, 2026

Education

KAIST, BS in Computer Science	Feb 2022
KAIST, MS in Computer Science	Feb 2024
KAIST, PhD in Computer Science	Mar 2024 –

Research Interest

Computer Architecture, System for Machine Learning, Processing-in-Memory, Large Language Models

Publications

Throughput-Oriented LLM Inference via KV-Activation Hybrid Caching with a Single GPU

Sanghyeon Lee, Hongbeen Kim, Soojin Hwang, *Guseul Heo*, Minwoo Noh, Jaehyuk Huh
ICCD, 2025

Déjà Vu: Efficient Video-Language Query Engine with Learning-based Inter-Frame Computation Reuse

Jinwoo Hwang, Daeun Kim, Sangyeop Lee, Yoonsung Kim, *Guseul Heo*, Hojoon Kim, Yunseok Jeong, Tadiwos Meaza, Eunhyeok Park, Jeongseob Ahn, Jongse Park
VLDB, 2025

ONNXim: A Fast, Cycle-level Multi-core NPU Simulator

Hyungkyu Ham*, Wonhyuk Yang*, Yunseon Shin, Okkyun Woo, *Guseul Heo*, Sangyeop Lee, Jongse Park, Gwangsun Kim (*: co-first authors)
IEEE Computer Architecture Letters (CAL), 2024

LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving at Scale

Jaehong Cho, Minsu Kim, Hyunmin Choi, *Guseul Heo*, Jongse Park
IISWC, 2024, Best Paper Award & Distinguished Artifact Award

Accelerating String-key Learned Index Structures via Memoization-based Incremental Training

Minsu Kim, Jinwoo Hwang, *Guseul Heo*, Seiyoon Cho, Divya Mahajan, Jongse Park
VLDB, 2024

NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing

Guseul Heo, Sangyeop Lee, Jaehong Cho, Hyunmin Choi, Sanghyeon Lee, Hyungkyu Ham, Gwangsun Kim, Divya Mahajan, Jongse Park
ASPLOS, 2024

Projects

NeuPIMs Simulator: Cycle-accurate NPU+PIM simulator

- github.com/casys-kaist/NeuPIMs
- Tools Used: C++

Technologies

Languages: C, C++, CUDA, Python, Javascript, Typescript, Java, Kotlin, Scala, Verilog, C#

Tools & Libraries: Pytorch, OpenCV, Node.js, React, Vue

Experience

Front-end Engineer , Data Monsters – Sejong, South Korea	Aug 2020 – Feb 2022
• Visualization of data analysis results	

- Frontend Development, React, Vue, Typescript

Machine Learning Research Intern, Park Systems Corp. – Suwon, South Korea

Sep 2020 – Feb 2021

- Image quality assessment model for Atomic Force Microscopy
- Deep Learning Model Engineering, Pytorch, Numpy, OpenCV, Pandas

Front-end Engineer Intern, kakao Corp. – Seongnam, South Korea

Dec 2019 – Feb 2020

- Implemented edit mode for the Daum Plusmap
- Frontend Development, Vanilla JS, HTML