Sangjin Choi

Ph.D. Student
School of Computing, KAIST
sjchoi@casys.kaist.ac.kr

RESEARCH STATEMENT

My research aims to realize sustainable and efficient AI systems. I am particularly interested on developing next-generation AI systems that maximize resource efficiency and boost performance.

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Ph.D. Student in School of Computing — 2021-09 – Present
Advisor: Youngjin Kwon

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Master of Science in School of Computing — 2019-09 – 2021-08
Advisor: Youngjin Kwon

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Bachelor of Science in School of Computing — 2025-03 – 2019-08

WORK EXPERIENCE

Microsoft Research Asia — Vancouver, Canada
Research Intern, Systems and Networking Research Group — 2026-03 – 2026-06

Microsoft Research Asia — Beijing, China
Research Intern, Systems and Networking Research Group — 2025-07 – 2026-01

PUBLICATIONS

[1] [**ACM Eurosys 2026**] Changjun Lee, **Sangjin Choi**, and Youngjin Kwon. "MTTM: Dynamic Fast Memory Partitioning with Bandwidth Optimization for Multi-tenant Cloud".

[2] [**Findings of NAACL 2025**] Sukmin Cho, **Sangjin Choi**, Taeho Hwang, Jeongyeon Seo, Soyeong Jeong, Huije Lee, Hoyun Song, Jong C. Park, and Youngjin Kwon. "Lossless Acceleration of Large Language Models with Hierarchical Drafting based on Temporal Locality in Speculative Decoding".

[3] [**IEEE/ACM ICCAD 2023**] Jehoon Heo, Yongwon Shin, **Sangjin Choi**, Sungwoong Yune, Junghoon Kim, Youngjin Kwon, Hyojin Sung, and Joo-Young Kim. "PRIMO: A Full-Stack Processing-in-DRAM Emulation Framework for Machine Learning Workloads".

[4] [**USENIX ATC 2023**] **Sangjin Choi**, Inhoe Koo, Jeongseob Ahn, Myeongjae Jeon, and Youngjin Kwon. "EnvPipe: Performance-preserving DNN Training Framework for Saving Energy".

[5] [**USENIX ATC 2022**] **Sangjin Choi***, Taeksoo Kim*, Jinwoo Jeong, Rachata Ausavarungnirun, Myeongjae Jeon, Youngjin Kwon, and Jeongseob Ahn. "Memory Harvesting in Multi-GPU Systems with Hierarchical Unified Virtual Memory". (*Co-first author).

RESEARCH EXPERIENCE

- **ML System**

  CoSpec (Sep 2023 – Oct 2024)
  – Efficient LLM serving system overcoming speculative decoding limitations in real-world scenarios (high request rates, temperature sensitivity). It introduces Dynamic Colocation for smart GPU resource management across batch sizes and Selective Validation to minimize wasted computation and maintain accuracy.

  Hierarchical Drafting (Aug 2024 – Oct 2024)
  – A novel lossless speculative decoding drafting approach that organizes various token sources into multiple databases in a hierarchical framework based on temporal locality. Ensures consistent, robust inference acceleration across diverse tasks and model sizes.

  EnvPipe (May 2022 – Jan 2023)
  – Designed and implemented a framework for training large language models (LLMs) that significantly reduces energy consumption while maintaining minimal performance degradation.

  - Optimized pipeline parallelism by strategically scheduling bubbles and selectively lowering SM frequency.
  - Developed the system on top of DeepSpeed (https://github.com/casys-kaist/EnvPipe).

- **Memory System**

  HUVM (Mar 2021 – Jan 2022)
  - Developed Hierarchical Unified Virtual Memory (HUVM), a system that creates the illusion of a unified virtual memory space for GPUs by utilizing temporarily idle memory of neighboring GPUs.
  - Designed and implemented a novel memory manager, memHarvester, to efficiently harvest and utilize neighbor GPUs' memory via NVLink.
  - Built a prototype system based on NVIDIA's driver (https://github.com/casys-kaist/HUVM).

  TribuOS (Sep 2019 – Jun 2020)
  - Contributed to the development of a disaggregated memory system designed for high availability.
  - Implemented a memory server with support for 1GB huge pages and a RDMA RPC recovery mechanism to enable faster recovery times.

- **AI with FPGA**

  PRIMO (Oct 2022 – May 2023)
  - Participated in developing a full-stack processing-in-DRAM emulation framework named PRIMO, which is the first emulation framework that can model and analyze DRAM-PIM for end-to-end ML inference.
  - Collaborated with a compiler team and an FPGA team to implement a PIM driver that enables efficient execution of PIM operations by providing seamless integration between the SW and HW components.

  Accelerating DNN with FPGA (Dec 2020 – Dec 2021)
  - Participated in developing an acceleration framework for DNNs using FPGA.
  - Implemented functionalities on top of Xilinx driver to support layer encoding to FPGA instructions and quantized models to 8-bit fixed-point format to accelerate model inference.