Sangjin Choi

Postdoctoral Researcher
CASYS Lab, Information & Electronics Research Institute, KAIST
sjchoi@casys.kaist.ac.kr

RESEARCH STATEMENT

I build cost-driven AI infrastructure with a deep understanding of hardware, systems software, and AI workloads.

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Ph.D. in School of Computing — 2021-09 – 2026-08
Advisor: Youngjin Kwon, CASYS Lab

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, CASYS Lab

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Bachelor of Science in School of Computing — 2015-03 – 2019-08

WORK EXPERIENCE

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, South Korea
Postdoctoral Researcher, Information & Electronics Research Institute — 2026-09 – Present

Microsoft Research Asia — Vancouver, Canada
Research Intern, Vancouver Lab — 2026-03 – 2026-05

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

PUBLICATIONS

[1] **[arXiv 2026] Sangjin Choi**, Sukmin Cho, Yifan Xiong, Ziyue Yang, Youngjin Kwon, and Peng Cheng. "ELDR: Expert-Locality-Aware Decode Routing for PD-Disaggregated MoE Serving."

[2] **[MLSys 2026]** Hyunjae Lee, **Sangjin Choi**, Seungjae Lim, and Youngjin Kwon. "BEAM: Joint Resource–Power Optimization for Energy-Efficient LLM Inference under SLO Constraints."

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

[4] **[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."

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

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

[7] **[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." ([*]Equal contribution.)

SELECTED RESEARCH PROJECTS

- **LLM Inference Systems**

  ELDR *arXiv 2026* [Paper]

  –Led the end-to-end design, implementation, and evaluation of ELDR, a vLLM-based expert-locality-aware decode router that uses prefill-derived expert signatures to reduce median time per output token by 5.9–13.9% on PD-disaggregated MoE deployments of up to 40 GPUs without changing model outputs.

- **Energy-Efficient AI Systems**

  BEAM *MLSys 2026* [Paper]

  –Contributed technical feedback and manuscript development for a controller that jointly tunes GPU frequency, chunk size, and microbatch count to meet per-request SLOs, reducing GPU energy by up to 51% over vLLM.

  EnvPipe *USENIX ATC 2023* [Paper]

  –Designed and implemented a DeepSpeed-based pipeline-parallel training framework that reduces energy consumption by lowering GPU SM frequency during pipeline bubbles while limiting performance degradation.

- **GPU Memory Systems**

  HUVM *USENIX ATC 2022* [Paper]

  –Designed and implemented Hierarchical Unified Virtual Memory and its memHarvester memory manager, modifying NVIDIA's driver to harvest and access idle memory on neighboring GPUs over NVLink and expand the memory available to each GPU.

- **Systems for Emerging AI Hardware**

  PRIMO *IEEE/ACM ICCAD 2023* [Paper]

  –Implemented PRIMO's PIM driver stack as the software interface connecting the PIM compiler to the FPGA-based DRAM-PIM emulator, encoding PIM operations into hardware-ISA binaries and transferring them through virtual-memory scatter-gather DMA.