Jae Hyung Ju

jhju@gatech.edu

Interest

My research focuses on computer architecture and systems for AI workloads. Specifically, my recent works include data movement optimization in CPU-GPU systems for LLM inference, GPU memory system design for emerging technology, and efficient AI algorithms.

Education

Georgia Institute of Technology — Atlanta, GA, United States
Ph.D. in Electrical and Computer Engineering — 2024 - 2029 (Expected)
Advisor: Prof. Moinuddin Qureshi

Seoul National University (SNU) — Seoul, South Korea
B.S. in Electrical and Computer Engineering — 2018 - 2024
Summa Cum Laude, GPA: 4.24/4.30

Research Experience

FAST Lab, Georgia Tech (Advisor: Prof. Moinuddin Qureshi) — Jan 2025 - Present

- Proposed GPU software and microarchitecture changes to incorporate High-Bandwidth Flash (HBF) on GPUs, mitigating read amplification caused by its large granularity while retaining the capacity benefit [P1].
- Developed a memory allocation system for concurrent GPU and CPU memory access in LLM inference, increasing goodput by 55% without any kernel code changes [S1].
- Developed a policy that activates fewer experts in mixture-of-experts (MoE) LLMs, reducing time between tokens by 23% with no accuracy loss [S3].

Scalable Computer Architecture Lab, SNU (Advisor: Prof. Jung Ho Ahn) — Jan - Dec 2023

- Proposed algorithms that accelerate CNN inference under fully homomorphic encryption by 1.5–2.7× [P4].

Work Experience

Microsoft — May-Aug 2026
Research Intern, Azure Systems Research/Microsoft Systems Research — Redmond, WA

- Analyzed scale-out network bottlenecks in shared KV cache storage using GitHub Copilot agentic traces.
- Developed a multi-rack simulator to design GB200/VR200 NVL72 servers for agentic workloads.

Qualcomm — May-Aug 2025
Engineering Intern, On-target software team — San Diego, CA

- Improved traceability in the ExecuTorch AI compiler for edge devices.
- Developed an automated tracing tool of torch.fx graph transformations with negligible overhead.

CryptoLab — Apr-Jul 2024
Research Intern, Core Development Team — Seoul, South Korea

- Developed multi-GPU acceleration for Llama 2 and ResNet-18 under fully homomorphic encryption (FHE).
- Designed the initial architecture and abstractions for a new FHE library.

Honors

Overseas PhD Scholarship, Korea Foundation for Advanced Studies (KFAS) — 2024 - 2029

Outstanding Graduate Award, Seoul National University (SNU) — 2024

Presidential Science Scholarship, Korea Student Aid Foundation (KOSAF) — 2018 - 2023

Publications

*Equal contribution

[P1] **J. H. Ju***, E. Chung*, H. Taneja, A. Saxena, S. Jeong, H. Kim, and M. Qureshi, "TileLens: Efficiently Using Large-Granularity Memory Systems with Transparent Two-Dimensional Memory Layout", *IEEE International Symposium on High-Performance Computer Architecture*, 2027.

[P2] E. Chung, **J. H. Ju**, and H. Kim, "Making Locality-Aware GEMM Compatible With Page-Granularity Placement on Chiplet GPUs", *IEEE Computer Architecture Letters*, vol. 25, no. 02, pp. 387–390, Jul. 2026.

[P3] C. Shin, J. Song, H. Jang, D. Kim, J. Sung, T. Kwon, **J. H. Ju**, F. Liu, Y. Choi, and J. Lee, "Piccolo: Large-Scale Graph Processing with Fine-Grained In-Memory Scatter-Gather", *IEEE International Symposium on High-Performance Computer Architecture*, 2025.

[P4] **J. H. Ju***, J. Park*, J. Kim, M. Kang, D. Kim, J. H. Cheon, and J. H. Ahn, "NeuJeans: Private Neural Network Inference with Joint Optimization of Convolution and FHE Bootstrapping", *ACM Conference on Computer and Communications Security*, 2024.

[P5] C. Shin, T. Kwon, J. Song, **J. H. Ju**, F. Liu, Y. Choi, and J. Lee, "A case for in-memory random scatter-gather for fast graph processing", *IEEE Computer Architecture Letters*, vol. 23, no. 1, pp. 73–77, 2024.

Preprints

[S1] A. Saxena*, **J. H. Ju***, H. Taneja, P.-A. Tsai, A. Jaleel, C. Kozyrakis, and M. Qureshi, "BOOST: Concurrent Access to Host Memory and HBM to Accelerate LLM Inference", *arXiv preprint arXiv:2609.13592*, 2026.

[S2] H. Taneja, A. Saxena, A. Revinipati, **J. H. Ju**, N. C. Crago, and M. Qureshi, "SiFAR: Synchronization-Free All-Reduce for Low-Latency LLM Inference", *arXiv preprint arXiv:2607.08973*, 2026.

[S3] V. Gupta, **J. H. Ju**, K. Sinha, A. Gavrilovska, and A. Iyer, "Lynx: Enabling Efficient MoE Inference through Dynamic Batch-Aware Expert Selection", *arXiv preprint arXiv:2411.08982*, 2026.

[S4] H. Kim, **J. H. Ju**, H. S. Choi, H. Roh, and W.-S. Choi, "NeuralEQ: Neural-Network-Based Equalizer for High-Speed Wireline Communication", *arXiv preprint 2308.02133*, 2023.

Patents

[T1] **J. H. Ju**, J. Park, J. H. Cheon, and J. H. Ahn, "Electronic device, server device for handling homomorphic encrypted data, and methods thereof", U.S. Patent App. Pub. 2026/0012327 A1, Jan. 2026.

[T2] **J. H. Ju**, J. Park, J. H. Cheon, and J. H. Ahn, "Server device for handling homomorphic encrypted data, and methods thereof", U.S. Patent App. Pub. 2026/0012326 A1, Jan. 2026.

Teaching

- Teaching Assistant for ECE 4100/6100 - Advanced Computer Architecture — Fall 2025
- Teaching Assistant for ECE 4100/6100 - Advanced Computer Architecture — Spring 2025

Projects

- Linux kernel hacking: custom scheduler and lock — Spring 2023
- CNN accelerator for VGGNet in Verilog on Arty A7 FPGA — Fall 2022
- Training and evaluation of a neural-network-based wireline equalizer — Fall 2022
- Simulator for a 6T2R-based spiking restricted Boltzmann machine chip — Summer 2022
- 16-bit pipelined MIPS CPU with cache and cycle-stealing DMA in Verilog — Spring 2022

Relevant Coursework

- **Computer Architecture**: Memory Systems Architecture, Hardware Software Co-Design for Machine Learning Systems, Advanced Computer Architecture, Computer Organization and Design
- **Systems**: Systems for AI, Operating Systems, Systems Programming, Data Communication Networks
- **Machine Learning**: Conversational AI, Generative Deep Learning, Machine Learning Fundamentals
- **Digital Design**: Digital Systems Design and Experiments, Digital Integrated Circuits

Skills

C++, CUDA, Python, SystemVerilog, vLLM, Pytorch, gem5, ChampSim, Accel-Sim, MacSim, Scale-Sim, ASTRA-Sim