Jae Hyung Ju

jhju@gatech.edu

Education

Georgia Institute of Technology — Atlanta, United States
Ph.D. in Electrical and Computer Engineering (GPA: 4.00/4.00) — 2024 - 2029 (Expected)

Seoul National University (SNU) — Seoul, South Korea
B.S. in Electrical and Computer Engineering (GPA: 4.24/4.30, Rank: 2/107) — 2018 - 2024
Summa Cum Laude

Research Experience

FAST Lab, Georgia Tech (Advisor: Moinuddin Qureshi) — Jan 2025 - Current

LLM (Large Language Model) Offloading Scheme

- Increases LLM inference throughput by offloading weights/KV cache to memory tiers outside GPU HBM.

Reducing MoE (Mixture of Experts) LLM Latency through Expert Reduction [1]

- Developed a policy to decrease MoE decode latency by reducing the total number of activated experts.
- Implemented inside vLLM, achieved up to 23% speedup with no accuracy loss.

Increasing DRAM Bandwidth Utilization by Minimizing Bus Turnaround Overhead [2]

- Analyzed the effect of different DRAM address mappings on delays between read/write commands.
- Evaluated by modifying ChampSim. Achieved up to 11% speedup, and 3.5% on real systems.

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

Optimization of CNN Inference Latency within FHE (Fully Homomorphic Encryption) [3]

- Proposed a state-of-the-art algorithm for evaluating CNN inference within FHE.
- Implemented using C++, CUDA, and the HEAAN library. Achieved x1.5–2.7 speedup for ResNet18/50.

Accelerated Intelligent Systems Lab, SNU (Advisor: Jinho Lee) — Mar-Jun 2023

DRAM PIM (Processing In Memory) Design and Evaluation [4][5]

- Accelerates random access workloads by internally gathering scattered data.
- Evaluated by modifying and integrating gem5 and Ramulator.

Publications

[1] V. Gupta, **J. H. Ju**, K. Sinha, A. Gavrilovska, and A. Iyer, "Prowl: Efficient moe inference through dynamic, workload-agnostic expert reduction", *in submission*,

[2] H. Taneja, **J. H. Ju**, A. Saxena, and M. Qureshi, "Splitstream: Maximizing system bandwidth utilization via read-write buffer management", *in submission*,

[3] **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.

[4] 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*, 2024.

[5] 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 (to appear).

[6] 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 arXiv:2308.02133*, 2023.

Scholarships and Awards

Overseas PhD Scholarship, Korea Foundation for Advanced Studies (KFAS) 2024 - 2029
Research grant $13k/year

Presidential Science Scholarship, Korea Student Aid Foundation (KOSAF) 2018 - 2023
Full tuition and stipend for eight semesters, total $44k

Work Experience

Qualcomm May-Aug 2025
Engineering Intern, On-target software team
- Improving traceability of the Executorch AI compiler for edge devices
- Designed and implemented the automatic tracing of torch.fx graph modifications with less than 3% overhead

CryptoLab Apr-Jul 2024
Research Intern, Core Development Team
- Multi-GPU acceleration of Llama2 and ResNet18 within FHE.
- Using C++ and CUDA, designed and implemented the initial structure of a new FHE library.

Teaching

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

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

Relevant Coursework

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

Academic Projects

Linux Kernel: Custom Scheduler and Read/Write Lock Spring 2023
- Implemented a WRR scheduler with periodic load balancing, and a range-based lock in the Linux kernel.

FPGA CNN Accelerator Fall 2022
- Using Verilog, implemented computation modules for a CNN accelerator on FPGA (Arty A7), and tested the inference its inference performance for a VGGNet variant.

16 bit Pipelined CPU with Cache and DMA Spring 2022
- Using Verilog, implemented a 16 bit pipelined CPU that supports a simplified MIPS ISA, with a write-back, write-allocate cache. Also implemented a simple DMA logic with cycle stealing.

Skills

- **Programming**: C++, CUDA, C, Python, Matlab
- **Machine Learning Library**: Executorch, vLLM, Pytorch
- **HDL and Architecture Simulators**: SystemVerilog, gem5, ChampSim, Ramulator, Scale-Sim, ASTRA-Sim
- **Tools**: Xilinx Vivado, Xschem, Ngspice, Magic VLSI

Extracurricular Activities

Nongnet Agricultural Commodity Price Prediction AI Competition Fall 2022
Achieved a top 13% ranking out of 69 participating teams.
Developed an AI model for price prediction utilizing a 10-year agricultural transaction database.

7th Airforce Communication Service Group, Republic of Korea Air Force Nov 2019 - Jun 2021
Crewman of the mobile TACAN (TACtical Air Navigation) system.