

Jihwan Oh

Stanford University, Stanford, CA, USA

jihwanoh@stanford.edu | [LinkedIn](#) | jihwan01.github.io (Website) | [GitHub](#)

RESEARCH INTERESTS

- Cross-layer optimization and hardware-software co-design for energy-efficient, high-performance computing
- Computer Architecture, Specialized Accelerators, and Systems for ML

EDUCATION

- **Stanford University** Sep. 2026 –
Incoming Ph.D. Student in Electrical Engineering
- **Korea Advanced Institute of Science and Technology (KAIST)** Feb. 2019 – Feb. 2026
B.S. in School of Electrical Engineering (Primary Major), School of Computing (Double Major)
GPA: 4.07/4.3, *Summa Cum Laude*; Dean's List: Spring 2019, Spring 2024
- **Georgia Institute of Technology** Jan. 2025 – Jul. 2025
Exchange Program, School of Electrical and Computer Engineering; GPA: 4.0/4.0

PUBLICATIONS

Regular-Compute-Communication Overlap Is Not Free: A Cross-Layer Sep. 2026

Characterization in GPU LLM Workloads

Jihwan Oh, Seokjin Go, Junkyum Kim, Jongse Park, Divya Mahajan

Accepted to IEEE International Symposium on Workload Characterization (IISWC) 2026

Characterizing Compute-Communication Overlap in GPU-Accelerated May. 2025

Distributed Deep Learning: Performance and Power Implications

Seonho Lee, Jihwan Oh, Junkyum Kim, Seokjin Go, Jongse Park, Divya Mahajan

Poster presented at ISPASS 2025, [arXiv:2507.03114](https://arxiv.org/abs/2507.03114)

RESEARCH EXPERIENCE

Systems Infrastructure and Architecture Research Lab, Georgia Tech Jan. 2025 – Aug. 2026

Researcher (Advisor: Prof. Divya Mahajan)

- Characterized compute-communication overlap in large-scale distributed LLM training and showed that the widely held assumption “maximizing overlap is always beneficial” breaks down, with aggressive overlap degrading both execution time and power efficiency across parallelism configurations. Published and presented as a poster at ISPASS 2025.
- Led a project to identify the hardware-level causes of compute-communication overlap overheads that remained hidden at the software layer through profiling and analysis.
- Designed a new NCCL cross-layer communication protocol that repurposes unused shared memory as a TMA-driven staging buffer to enable internal pipelining, achieving comparable or higher bandwidth with roughly 50% fewer SMs in microbenchmarks.

Computer Architecture and Systems Lab, KAIST Sep. 2024 – Dec. 2024

Undergraduate Researcher (Advisor: Prof. Jongse Park)

- Analyzed LLM inference on *NeuPIMs* (ASPLOS 2024) and found that FFN compute grows quadratically with hidden size, outpacing other operations and causing severe NPU-PIM load imbalance that limits model-size scalability. Designed a redistribution strategy that balances FFN execution across NPU and PIM, improving utilization and end-to-end performance.
- Supported research on applying quantization techniques to PIM architectures by surveying state-of-the-art methods and benchmarking accuracy across multiple models.

Urban Robotics Lab, KAIST

Feb. 2024 – Jun. 2024

Undergraduate Researcher (Advisor: Prof. Hyun Myung)

- Implemented SLAM algorithms and ROS-based motion planning systems for mobile robotics.

TEACHING & INDUSTRY EXPERIENCE

- CS.10001: Introduction to Programming, KAIST**, Teaching Assistant Sep. 2025 – Dec. 2025
Yulmae Academy, Programming Instructor Dec. 2023 – Dec. 2024
- Taught C, Python, and introductory machine learning to science high school students.
- CS.93000: Immersion Camp (Coding Camp), KAIST**, Teaching Assistant Oct. 2023 – Feb. 2024
Republic of Korea Air Force, Software Developer Aug. 2021 – May. 2023
- Developed and refined a VR flight-training system using Unreal Engine 4 and C++, integrating pilot feedback and Git-based collaboration to enhance realism and training effectiveness.
 - Selected to present the simulator as the Air Force representative at the 2022 ROKAF Information & Communication Development Conference.
 - Selected to serve as a counseling officer supporting soldier adaptation, dormitory operations, and mental-health guidance.

ACADEMIC ACTIVITIES

- Paper Presenter**, IISWC 2026 – Boulder, CO, USA Sep. 2026
- Presenting the first-author paper “Regular-Compute-Communication Overlap Is Not Free: A Cross-Layer Characterization in GPU LLM Workloads”.
- Poster Presentation**, ISPASS 2025 – Ghent, Belgium May. 2025
- Presented poster on “Characterizing Compute-Communication Overlap in GPU-Accelerated Distributed Deep Learning”.
- Attendee**, ISCA 2025 and uArch Workshop – Tokyo, Japan Jun. 2025
- Selected as a full travel grant recipient to attend ISCA 2025 and participate in the uArch Workshop.

AWARDS & SCHOLARSHIPS

- **Next-Generation Engineer Award: Highest Distinction** 2025
Institute for Promotion of Engineering and Science of Korea, IPESK
Awarded the top distinction as the single highest-ranked recipient in the department.
- **uArch Mentoring Workshop Full Travel Grant**, uArch @ ISCA 2025 2025
Awarded a fully funded travel grant for both the uArch workshop and ISCA 2025.
- **Student Travel Grant**, IEEE ISPASS 2025 2025
- **Dean’s List**, School of Electrical Engineering, KAIST 2024
- **Korea-U.S. Student Exchange Program Scholarship** 2024
Korea Institute for Advancement of Technology, KIAT
- **National Science and Technology Scholarship**, Korea Government 2021 – 2024
- **Dean’s List**, Freshman Division, KAIST 2019
- **Minister of Education Award**, Ministry of Education (STEAM R&E Program) 2018

TECHNICAL SKILLS

- **Programming Languages:** C/C++, Python, Rust, SystemVerilog, Verilog
- **Libraries and Tools:** CUDA, PyTorch, NCCL, CUTLASS, cuBLAS; Megatron-LM, Megatron-DeepSpeed; Nsight Compute, Nsight Systems; ROS, Unreal Engine 4; Git