

Research Interests

My research lies at the intersection of computer systems and AI. I study how emerging AI workloads challenge existing system abstractions by introducing new computational behaviors and application requirements. Based on these insights, I design efficient runtimes and resource-management mechanisms that adapt AI execution to heterogeneous computing platforms, from datacenter infrastructure to resource-constrained edge devices.

Education

Korea Advanced Institute of Science and Technology (KAIST)**Sep. 2025 – Expected Aug. 2027***Master of Science in Computer Science, School of Computing**Daejeon, Republic of Korea*

- Overall GPA 4.2/4.3; Recipient of the National Full Graduate Scholarship (Fall 2025 – Spring 2027)
- Advisor: Prof. Jaehyuk Huh, Computer Architecture and Systems Lab (CASYS)

Sungkyunkwan University**Mar. 2020 – Aug. 2025***Bachelor of Science in Computer Science and Engineering**Suwon, Republic of Korea*

- Overall GPA 4.2/4.5; Recipient of the Sungkyun Software and Research Promotion Scholarships (full tuition, 8 semesters)

Publications

Beyond LLM Serving: Characterizing Vision-Language-Action Workloads for Embodied AI System Design

- **Seonghun Jung**, Sieun Moon, Jiyoung Jeong, Jimin Lee, and Jaehyuk Huh
- ASPLOS 2027 (To Appear)

Experience

Samsung Electronics | AI Center**Jan. 2025 – Feb. 2025***Software Engineer Intern**Hwaseong, Republic of Korea*

- Traced slow response times on a semiconductor equipment data-analysis platform using Elastic APM/RUM and the ELK stack, reducing page-load latency by 26.45%.
- Implemented caching and optimized multi-table join queries in personalized dashboards, reducing server processing time by 6.49%.

Samsung Electronics | AI Center**Sep. 2024 – Dec. 2024***MLOps Engineer Intern**Hwaseong, Republic of Korea*

- Contributed to a distributed MLOps system for hyperparameter optimization built on Ray Cluster and Kubeflow, used by 100+ Samsung engineers.
- Developed a multi-tenant FastAPI backend that authenticated tenants with Keycloak JWT, tracked job state in Redis, and scheduled shared GPU resources through Volcano.
- Automated deployment with a CI/CD pipeline on ArgoCD, Helm, and Istio, and centralized logging with FluentBit and OpenSearch.

Computer System Lab | Prof. Euseong Seo**Sep. 2023 – Jun. 2024***Undergraduate Researcher**Suwon, Republic of Korea*

- Designed an elastic training platform that reallocates GPUs across distributed deep-learning jobs at runtime, using a scheduling algorithm based on job throughput gradients to improve cluster utilization.
- Built the underlying 4-node Kubernetes cluster with Containerd, the Kubeflow MPI Operator, and InfiniBand to run Elastic Horovod training at scale.

Projects

ProfilingSense**Sep. 2024 – May 2025***Non-Intrusive Kubernetes Scheduler for Distributed DL Based on Resource Efficiency Profiling**B.S. Graduation Project*

- Designed and implemented a non-intrusive Kubernetes scheduler that profiles jobs in an isolated pool using standard training logs and ranks them by measured throughput per resource and waiting time. It gang-schedules jobs by changing only their execution order, without modifying training code or reallocating GPUs.
- Built and validated a simulator using profiles of five workloads measured on a 6-GPU testbed, then replayed 2,312 Microsoft Philly jobs at 864-GPU scale. ProfilingSense reduced average JCT by 18.1% in physical experiments and by 22.1% over the default scheduler and 4.5% over FIFO-Gang at scale.

Honors & Skills

Honors: Outstanding Teaching Assistant Award, KAIST (2025, 2026); Gold Prize (2nd Place), SKKU Hackathon (2023); 15th Place, SKKU Programming Competition (2024, 2025)

Certifications: Certified Kubernetes Administrator, The Linux Foundation (2025); S/W Certification Professional, Samsung Electronics DX (2024)

Technical: Python, C/C++, CUDA, PyTorch, Kubernetes, Docker, Linux, Nsight Compute/Systems (NCU/NSYS)