JIHEON CHOI

HPC Scheduling, LLM Systems, and GPU Computing Researcher

unidev@ajou.ac.kr | jiheonchoi.com | Suwon, Republic of Korea

SUMMARY

I am a **Graduate Research Assistant** at Ajou University developing reliable scheduling systems for HPC and AI infrastructure. My research studies how runtime uncertainty, calibration error, and workload drift affect scheduling decisions in large-scale computing systems. I design uncertainty-aware prediction, buffering, and calibration methods that reduce scheduler-level failures, including premature job termination, backfill failure, and utilization loss. My broader interests include AI infrastructure systems, LLM inference serving, and AI-driven scientific computing systems, particularly scalable acceleration of molecular GNN-based interatomic potentials. My prior industry experience includes security automation and production SaaS delivery.

Research Interests: HPC Systems · AI Infrastructure Systems · AI-Driven Scientific Computing Systems · Reliable Scheduling under Runtime Uncertainty · Runtime Prediction and Calibration · Workload Drift Adaptation · LLM Inference Serving · GPU Cluster Scheduling · Molecular GNN-IP Acceleration · Power-aware and Carbon-aware AI/HPC Scheduling

EDUCATION

Ph.D. in Artificial Intelligence — Sep. 2023 – Present
Ajou University, Suwon, Republic of Korea — Advisor: Sangyoon Oh

B.Sc. in Software and Computer Engineering — Aug. 2023
Ajou University, Suwon, Republic of Korea

PUBLICATIONS

*: 1st co-authors †: corresponding authors C: conferences J: journals W: workshops P: preprints

[C3] **Jiheon Choi** and Sangyoon Oh†, **S-CQR: Stratified Calibration for Runtime Prediction in HPC Backfill Scheduling**, *32nd International European Conference on Parallel and Distributed Computing (Euro-Par), 2026.*

[C2] **Jiheon Choi** and Sangyoon Oh†, **Reducing Backfill Failures with Lightweight Uncertainty Buffers under Workload Drift in HPC Job Scheduling**, *26th IEEE International Symposium on Cluster, Cloud, and Internet Computing (CCGrid), 2026.*

[J3] **Jiheon Choi** and Sangyoon Oh†, **UARP: Uncertainty-Aware Runtime Prediction for Preventing Scheduler Termination under Wallclock Constraints in HPC**, *Journal of Supercomputing, 2026. IF 2.7; JCR Q2.*

[J2] **Jiheon Choi** and Sangyoon Oh†, **Lightweight Multi-Layered De-Identification Architecture: Secure Client Selection in Federated Learning**, *Journal of Systems Architecture, 2025. IF 4.1; JCR Q1.*

[C1] **Jiheon Choi**, Jaehyun Lee, Minsol Choo, Taeyoung Yoon, Oh-Kyoung Kwon, and Sangyoon Oh†, **When HPC Scheduling Meets Active Learning: Maximizing the Performance with Minimal Data**, *High Performance Computing in Asia-Pacific Region (HPC Asia), 2025.*

[J1] Miri Yu, **Jiheon Choi**, Jaehyun Lee, and Sangyoon Oh†, **Staleness Aware Semi-asynchronous Federated Learning**, *Journal of Parallel and Distributed Computing, 2024. IF 4.0; JCR Q1.*

RESEARCH EXPERIENCE

Graduate Research Assistant, Distributed and Parallel Lab — Sep. 2023 – Present
Ajou University, Suwon, Republic of Korea

- **LLM Inference Acceleration:** Profiled vector-indexing bottlenecks in high-concurrency RAG workloads and redesigned storage access around a multi-tier hierarchy to support real-time LLM serving.
- **Graph Processing on GPU:** Developed memory-efficient graph partitioning and fine-grained GPU kernels for billion-scale graph workloads, including GNN inference acceleration for molecular dynamics simulations.
- **HPC Scheduling Optimization** [C1, J3, C2, C3]: Designed runtime prediction, uncertainty buffering, and stratified calibration methods to absorb wall-clock estimation errors and reduce premature job termination in HPC schedulers.
- **Privacy-Preserving Machine Learning** [J1, J2]: Developed a lightweight multi-layered de-identification architecture for secure client selection and studied staleness-aware aggregation in semi-asynchronous federated learning.
- **Lab Management:** Coordinated project timelines across concurrent R&D grants. Recruited and onboarded six summer research interns in 2025; three joined the lab as M.S. students and five contributed to active research projects.

RESEARCH PROJECTS

National Research Foundation of Korea (NRF): *Efficient hyperscale LLM inference based on scale-out context memory* — Jul. 2026 – Jun. 2029

HL Mando: *Cloud-updatable SDV chassis software platform for heterogeneous zonal control units (ZCUs)* — Apr. 2026 – Dec. 2029

KISTI / National Research Council of Science and Technology (NRCST): *GPU-based graph parallel processing algorithms* — Jun. 2024 – Present

Ministry of Science and ICT: *Multi-tiered, multi-purpose autotuning framework for exa-scale supercomputers* — Nov. 2023 – Present

Samsung Display Inc.: *High-efficiency HPC job scheduling; raised cluster utilization from 72% to 93%* Jan. 2024 – Feb. 2025
Korea Automotive Technology Institute: *Safety of the Intended Functionality (SOTIF) for perception and decision-making insufficiency* Sep. 2022 – Present
IITP: *MR-IoT convergence AI for disaster countermeasures* Jan. 2023 – Dec. 2023
Ajou University: *User engagement in social/mobile platforms and welfare value* Jan. 2022 – Oct. 2022
Ministry of SMEs and Startups: *Web application vulnerability assessment software* Apr. 2022 – Dec. 2022

PATENTS

[P3] **Task Scheduling Device, Task Scheduling Method Thereof and Computing System**, Korean Patent Application No. 10-2025-0078984, filed.

[P2] **Apparatus and Method for Hotspot Prevention Based on Bloom Filter in Distributed Database Management System**, Korean Patent Application No. 10-2024-0162896, filed.

[P1] **Method and Apparatus for Classifying Workload Using Artificial Intelligence Model**, Korean Patent Application No. 10-2023-0029398, filed.

OTHER EXPERIENCE

BSKB's Summer Patent Seminar Jun. 2023
Birch, Stewart, Kolasch & Birch, LLP
Completed focused training on U.S. patent law, including patent prosecution, litigation strategy, recent developments, and case studies.

INDUSTRY EXPERIENCE

Lead Software and Security Research Engineer Aug. 2020 – Dec. 2023
MONTHLY HACKING Inc., Seoul, Republic of Korea

- Built backend services for an automated penetration testing SaaS platform that turned recurring security assessments into production delivery workflows.
- Led development of an automated web vulnerability scanner that passed a government-certified performance evaluation for an SME R&D grant.
- Developed an automatic Log4J vulnerability detection and reporting module and integrated the scanning engine as a Python Flask API into a Spring Boot production backend.
- Implemented static-analysis logic for detecting vulnerabilities in mobile application binaries and shipped it as a B2B SaaS product using Nest.js and TypeScript.

Software Engineer Oct. 2019 – Aug. 2020
WAPAS SYSTEMS Inc., Incheon, Republic of Korea

- Built a mobile web ordering and delivery platform for Philippine market operations using Node.js, React, and AWS.
- Designed a cash-based settlement system tailored to local payment infrastructure, enabling order-to-delivery workflows for household and enterprise customers in a market with limited credit card adoption.

PROFESSIONAL SERVICE AND TEACHING

Reviewer, SCI/SCIE Journals
The Journal of Supercomputing (2026); Cluster Computing (2026); Discover Artificial Intelligence (2026)

Teaching Assistant Jan. 2022 – Jun. 2026
Ajou University

- Distributed and Parallel Computing: Fall 2025
- Object-oriented Programming: Spring 2024, Spring 2025, Spring 2026
- Web System Design: Fall 2023, Fall 2024
- Artificial Intelligence Convergence Capstone Design: Fall 2022, Spring 2023
- Data Analytics–Machine Learning: Spring 2022

Junior Mentoring Nov. 2021 – May 2023
Team Sparta Inc.; Elice Inc.

- Mentored learners in modern web frameworks including React and Node.js.

TECHNICAL SKILLS

Programming Languages: Python, C, C++, Java, JavaScript, TypeScript
HPC and Cluster Management: SLURM, IBM LSF, MPI
Machine Learning and LLM Systems: PyTorch, vLLM, DeepSpeed, Hugging Face, LangChain
Web and Backend Systems: Node.js, React, Next.js, Nest.js, Spring Boot, Flask, AWS
Security Tools: Burp Suite, Nmap, Wireshark, Frida
Data Analysis and Visualization: Pandas, Matplotlib