Jiwon Kim

Seoul, South Korea | jeewonbob@gmail.com | jj1kim.github.io(currently deprecated) | github.com/jj1kim

Interests

The pursuit of **building computer systems that maximize efficiency in terms of time, cost, and space** is the driving force behind my study and research in computer engineering. My interests span a wide spectrum—from designing hardware architectures tailored to the unique demands of specific domains, to enhancing efficiency at highly abstracted layers of software. Previously, my research focused on optimizing operating systems and other system components to handle large-scale user traffic. Recently, I am deeply engaged in exploring system designs and hardware architectures that enable various deep learning models to train and run with both speed and resource efficiency.

Beyond technical inquiry, I am passionate about harnessing the power of computer engineering to **tackle real-world challenges and pioneer untapped markets**. Through a spirit of entrepreneurship, I seek to uncover problems ripe for technological innovation and work alongside outstanding minds to introduce transformative experiences to society.

Education

Seoul National University, BS in Computer Science & Engineering, double major in Entrepreneurship, double major in Civil & Environmental Engineering — Mar 2022 – Present

- GPA: 3.83/4.3
- **Core Coursework:** Hardware System Design(A), Computer Programming (A), Operating System (A), Scalable High Performance Computing(A), Digital Computer Concept and Practice(A), Logic Design(A-), Data Structure(A-)

Sejong Science High School, normal science — Mar 2020 – Feb 2022

- Early Graduate in 2 years
- **Top 10** of all graduates

Research Experience

SNU Scalable Computer Architecture Laboratory, Research Intern — June 2025 – Nov 2025

- Advisor : Jungho Ahn
- Main research topic : Efficient LLM serving, Accelerating Generative AI, Computer architecture
- Currently researching on scheduling policy of LLM serving system

Publications

From Tokens to Layers: Redefining Stall-Free Scheduling for LLM Serving with Layered Prefill — Oct 2025

Gunjun Lee, ***Jiwon Kim***, Jaiyoung Park, Younjoo Lee, Jung Ho Ahn

arXiv preprint(Under review)

Projects

SGS

- Developed a Kubernetes-based multi-node GPU auto-allocation system with user-level isolation and robust multi-cluster storage integration
- Tools Used: Golang, Kubernetes, Grafana, Prometheus, Harbor, Ceph, Cilium
- github.com/bacchus-snu/sgs

Layered Prefill

- Developed a LLM serving framework which changes the scheduling axis from tokens to layers and removes redundant MoE weight reloads while keeping decode stall free

- github.com/scale-snu/layered-prefill

Hodu

- Built a secure, resource-efficient, and high-speed coding test grading server using custom lightweight isolation techniques
- Tools Used: Rust, Kubernetes, GRPC
- github.com/wafflestudio/hodu

CareWise

- Fine-tuned a large language model to deliver expert-level clothing care advice and developed a system that analyzes care labels to generate personalized maintenance tips
- Designed and deployed a low-latency web infrastructure, enabling the service to scale effectively.
- Tools Used: Langchain, Pytorch, AWS, React.js, Django
- github.com/jj1kim/Carewise (currently deprecated)

Honors & Awards

MICRO 2025 AI Model Benchmarking Competition, 3rd place Oct 2025
Organized by IEEE/ACM International Symposium on Microarchitecture

- Participated in MICRO 2025 held in Seoul and achieved 3rd place in the AI Benchmarking Competition
- Developed an AI model that optimized both inference latency and accuracy for image classification tasks on the given NPU
- Strategically leveraged the hardware characteristics of the NPU and advanced model training techniques to maximize performance

ESG Smartcity Startup Hackerthon, 2nd place Jun 2023
Organized by Korean Standards Association

- Collaborated with an IoT startup to explore ESG-oriented strategies for enhancing indoor temperature management productivity
- Discovered untapped spatial temperature distribution data within the company's database and proposed a data-sharing partnership with the interior design industry
- Contributed to designing and implementing the data refinement and analysis pipeline, which led to measurable improvements in the partner company's business model and revenue

Academic Highlights

CUDA Neuralnet Accelerating Competition

- As part of Scalable High Performance Computing lecture
- Project on model inference throughput optimization using CUDA C++
- 1st place(of 120) by performance

Custom Accelerator Architecture Design Competition

- As part of Hardware System Design lecture
- Desiged CNN accelerator with Amaranth and Implemented efficient memory controller
- 1st place(of 50) by performance(reducing cycle)

Club Experience

Bacchus Mar 2024 – Present

- Served as the president of Bacchus, a student system administration club, and oversaw the full operation of all GPU and CPU servers within the university's Computer Science department
- Clustered all servers using Proxmox and enforced virtualization standards by deploying all services as VMs, which formed the base for a Kubernetes cluster with integrated dashboards and logging systems
- Developed a Kubernetes-native GPU resource scheduler to automate allocation and ensure fair, efficient usage for all students in the department

- Maintained essential student services such as account management and community platforms, incorporating ongoing feedback from administrative staff and users
- Standardized and managed all hardware lab computers to ensure consistent environments and implemented secure systems for managing user-specific configurations across sessions

WaffleStudio Sep 2023 – Jul 2025

- Provided Kotlin Springboot training to fellow students and collaboratively developed an internal community service platform
- Participated in the development of a software solution to provide resource-efficient isolation environments within computer systems
- Managed system components and resources in Kubernetes infrastructure

SNU LikeLion Mar 2024 – Present

- Educated students about infrastructure and backend development as part of a tech entrepreneurship club, SNU LiekLion
- Held several ideathons to find real problems that can be solved by technological solutions, and based on the results, built services quickly to see how the market responds
- Managed overall operations of the club as Administator

TA

Server System Management TA, SNU Semiconductor Specialized University Mar 2024 – Sep 2025

- Participated as a member of the Semiconductor Specialized University initiative and oversaw end-to-end management of GPU servers
- Led the deployment of GPU clusters and implemented usability-enhancing systems such as automated resource allocation and activity logging
- Provided technical consultation for key infrastructure decisions and educated students on effective and responsible use of GPU resources

Scholarship

Full Merit Scholarship, Hyungae Scholarship Foundation Feb 2024 – Present

- Received the scholarship upon being selected as **one of the top 10 undergraduate students** across the entire university.
- The scholarship guarantees full tuition coverage until graduation, amounting to **over $17,000** in total.

Full Merit Scholarship, Seoul National University Alumni Association Aug 2023

- Received the scholarship as the **highest-ranked student** in the CS undergraduate cohort

Technologies & Skills

Languages: C, C++, Java, Kotlin, Python, Go, Rust, Scala, TypeScript, SQL

Frameworks: CUDA, vLLM, PyTorch, Django, Spring, React, Next.js, JPA

Proficiencies: Kubernetes, Terraform, Ansible, ArgoCD, Prometheus, Grafana, Kafka, Over 6 years of GNU/Linux experience, Over 4 years of AWS experience.

Personal Details

Languages: native in Korean, fluent in English, elementary in Japanese