Last updated in July 2025

Muhwan Kim

Seoul, South Korea | me@muhwan.dev | github.com/minty99

Summary

Versatile systems engineer focused on making AI models—especially LLMs—run efficiently on custom hardware by building compilers, schedulers, and performance tooling. Ramp up fast in new domains and apply new ideas rapidly to ship production systems. Optimize for real speedups, clear design, and reliability.

Experience

Senior Software Engineer, FuriosaAI May 2023 – present

Compiler Team; L5 since 2025-07 (company leveling introduced); previously Software Engineer.

- Designed and implemented a cycle-level performance estimator for the RNGD NPU in Rust
- Led the design and implementation of the instruction scheduler, enabling state-of-the-art LLMs to run on the RNGD NPU.
- Scaled the compiler test pipeline to keep iteration fast as model complexity grew, maintaining hundreds of LLM model-compilation tests per PR and the merge queue.
- *Tech: Rust (compiler), Python/TypeScript/React/Kubernetes/Tekton (tooling)*

Machine Learning Engineer, Kakao July 2020 – May 2023

Developed real-time DNN inference systems and personalized ad-targeting models for high-traffic services serving billions of ad requests daily.

- Engineered a high-performance C++ inference engine on ONNX Runtime, replacing TensorFlow Serving via cgo to meet production-latency requirements.
- Developed a contrastive-learning keyword-targeting model in PyTorch to improve ad relevance and performance.
- Built online training and feature-generation pipelines using Kubernetes and Go.
- Reviewed coding tests for intern and new-grad candidates and authored the official solution guide for the 2021 recruitment cycle, published on Kakao Tech Blog.
- *Tech: Python (ML), C++/Go (inference), Kubernetes/Argo (CI/CD), React/Flask (tooling), PySpark (data)*

Software Engineer (Part-time), MINDsLAB June 2019 – May 2020

CUDA-level profiling and optimization for PyTorch inference; ported Tacotron2 to Flashlight (C++).

Software Engineer (Intern), RideFlux Jan 2019 – Feb 2019

Built a real-time sensor visualization simulator for autonomous vehicle testing (React/Node.js).

Education

Seoul National University, B.S. in Computer Science and Engineering Mar 2017 – Feb 2023

- Major GPA: 3.7 / 4.3
- Completed all systems-related core courses with an A0 or higher

Awards

- **Google Hash Code 2019**: 9th Place (Final Round, Team: TwingTwingTwing Platinum)
- **Codeforces**: Max rating 1710 (Expert)

Skills

- **Programming Languages**: Rust, Python, C/C++, Go, TypeScript, JavaScript
- **ML & Computing**: PyTorch, TensorFlow, ONNX Runtime, CUDA, OpenCL
- **Infrastructure & DevOps**: Kubernetes, Docker, Argo CD, Argo Workflows, Tekton
- **Tools & Frameworks**: React, Node.js, Flask, PySpark
- **Languages**: Korean (Native); English: Professional working proficiency