Wookje Han

347-971-1195 | wookjeh@gmail.com | LinkedIn | Google Scholar

Summary

AI performance engineer working on LLMs, video models, and recommender systems across the stack, from inference engines to GPU kernels. Focused on GPU kernel optimization for low-latency inference and high-performance training on recent hardware architectures.

Technical Skills

Languages: Python, C/C++, CUDA C++, PTX
Frameworks, Libraries: CUTLASS, CuTeDSL, NCCL, vLLM, SGLang, FlashInfer, PyTorch, OpenAI Triton
Developer Tools: Docker, AWS, GCP, Nsight Systems, Nsight Compute

Education

Columbia University — New York, NY
MS in Computer Science — *Sep 2023 - Dec 2024*
Seoul National University (SNU) — Seoul, Korea
BS in Computer Science and Engineering, Graduated Summa cum laude — *Mar 2017 - Aug 2023*

Experience

Senior AI Technology Developer — Mar 2026 - Present
NVIDIA — *Santa Clara, CA*

- Collaborated with a partner company to optimize end-to-end inference for its LLM on vLLM and SGLang.
- Implemented custom attention kernels for Hopper, Blackwell, and Rubin across diverse attention masks and data types, achieving state-of-the-art training and inference performance.
- Developed high-performance GEMM on recent architectures for low-latency and high-throughput workloads.
- Optimized fused multi-GPU communication and computation kernels using NVLS for end-to-end inference.
- Contributed a custom GB300 kernel to NVIDIA's MLPerf Inference v6.0 submission.

AI Technology Developer — Feb 2025 - Feb 2026

- Optimized custom kernels and end-to-end deep learning workloads (e.g., recommender systems, LLMs, and DiTs) on recent architectures using CUDA, CUTLASS, CuTeDSL, and OpenAI Triton.
- Led a partner company's adoption of Blackwell for training by implementing custom attention forward and backward kernels in CUTLASS, achieving 4X over the OpenAI Triton kernel.
- Optimized high-performance Group GEMM forward and backward kernels on Hopper, 3.8X faster than OpenAI Triton across both low-latency and high-throughput regimes.
- Delivered technical lectures to partner companies on (i) CUTLASS and CuTeDSL, (ii) writing high-performance GEMM kernels, and (iii) recent architecture-specific optimization.

AI Technology Developer Intern — May 2024 - Aug 2024

- Implemented a CUDA kernel for the encoder layer of a Generative Recommendation model, utilizing Hopper Architecture features (TMA and WGMMA), resulting in a 2X performance improvement over the Triton kernel.
- Built MMA and epilogue fusion using CUTLASS, accelerating the Generative Recommendation Model by 1.16X.
- Analyzed the characteristics and detailed statistics of bottleneck kernels in the Generative Recommendation Model using Nsight Systems and Nsight Compute.

Software Systems Research Intern — Dec 2021 - Feb 2023
Software Platform Lab, SNU — *Seoul, Korea*

- Constructed a tool to distribute language model's memory load and computation to multiple GPUs with PyTorch.
- Established multi-node environment using docker on GCP for efficient memory management of DL training.

Projects

LLVM Compiler Optimization | *C++, LLVM Compiler, Git, Docker, Linux* — *Mar 2022 – Jul 2022*

- Led a team of 4, devising optimizations for LLVM compiler with C++ and reduced 42% of cost for a given machine.
- Built automated unit tests utilizing CMake and GitHub Actions for continuous integration.

ArchPresser | *Python, NumPy, pybind, Git, Linux* — *Sep 2021 – Dec 2021*

- Constructed a program to generate a panoramic dental x-ray image from 3D CT scans with Python and C++.
- Analyzed bottleneck and used pybind to integrate C++ code with Python, leading to 20x faster application.