Suwon, Republic of Korea
+82-10-2054-2036

Channoh Kim

channoh@gmail.com
linkedin.com/channoh-kim

An experienced NPU hardware researcher for 2+ years at Samsung and a software engineer intern at Naver (was offered a full-time job after the internship). Specialty in hardware-software co-design for emerging applications and optimizations across systems and architectures. **Area of Interests:** Computer Architecture, Parallel Computing, Performance Analysis and Optimization

Experience

Google Korea — **Seoul, Korea**
Silicon Architect, Edge TPU — December 2023 — Present

- Researched efficient neural network accelerators for Google Mobile SoC

MangoBoost — **Seoul, Korea**
System Architect, Network offload project — August 2022 — November 2023

- Implemented a full-stack TCP offloading engine on a Xilinx FPGA that achieves line-rate throughput.
- Designed network security offloading engines such as TLS and IPsec.

Samsung Advanced Institute of Technology — **Suwon, Korea**
Staff Researcher, Distributed Deep Learning Training project — March 2022 — August 2022

- Exploiting the parallelism of distributed training for NLP models on extreme scales

Staff Researcher, Neural Network Accelerators on Flagship Mobile project — March 2019 — February 2022

- Researched efficient and flexible neural network accelerators for Samsung Flagship Mobile SoC
- Modeled performance simulators to demonstrate the feasibility of NPU accelerators using SystemC
- Analyzed functional operations and data statistics of popular neural network models
- Studied memory access pattern of neural networks and designed optimized hardware for it
- Published a top-conference paper [**ISCA21**]

Architecture & Code Optimization (ARC) Lab, Seoul National University (SNU) — **Seoul, Korea**
Student Lead, Typed Architecture project — March 2015 — February 2019

- Serving as student lead for a $1M research project (funded by Samsung) and developed a novel processor architecture to accelerate dynamically typed scripting languages such as JavaScript, Lua, and Python
- Built a performance simulator using Gem5 and prototyped it on Xilinx FPGA using RISC-V Rocket Core
- Ported Firefox SpiderMonkey JavaScript engine and Lua to RISC-V and added support for ISA extension to GCC
- Published two top-conference papers [**ASPLOS17**] [**ISCA16**] and filled two patent applications in Korea and US

Research Assistant, Deep Learning-Optimized DRAM project — December 2017 — September 2018

- Analyzed memory system performance of popular DNNs by porting Caffe to GPGPU-sim
- Investing novel DRAM device architecture optimized for DNNs (sponsored by SK Hynix)

Research Assistant, CPU/GPU Parallelization project — January 2012 — December 2014

- Devised an efficient CPU-GPU work-sharing scheduler for data-parallel JavaScript kernels on WebKit
- Investigated speculative parallelization for variable-length decompression algorithms such as gzip and bzip2
- Co-authored four conference papers (PPoPP, ISLPED, WWW, LCTES) and filed two domestic patents

Naver Clova AI — **Seongnam, Korea**
Summer Internship — July — August 2018

- Implemented a fast homerun scene detection which is started from pitching in the video stream
- Offered a full-time position for outstanding performance immediately upon completion of the internship

Intensive English Program — **Auburn, AL, USA**
Visiting Student — October 2008 — August 2009

- Participated in an intensive American language and culture program at Auburn University

Military Service (Compulsory) — **June 2006 — August 2008**

Education

Sungkyunkwan University (SKKU) — **Suwon, Korea**
M.S./Ph.D. Student, Department of Electrical and Computer Engineering — September 2012 — February 2019

- Dissertation: Architectural Techniques to Accelerate Scripting Languages
- Thesis supervisor: Prof. Jae W. Lee at Seoul National University (SNU)

B.S. Student, Department of Electronic and Electrical Engineering — March 2005 — August 2012

Technical Skills

Language Skills	Korean (native), English (working proficiency)
Programming Skills	C/C++ (incl. OpenCL, CUDA, SystemC), Python (incl. PyTorch, Tensorflow), Scala (Chisel), JavaScript (incl. AJAX, WebCL), Lua
Tools	Simulator (Gem5, GPGPU-sim, ZSim, DRAMsim2/3), Compiler (LLVM), Version control (Git, SVN), LaTex

Publications

[ISCA'21] J.-W. Jang, S. Lee, D. Kim, H. Park, A. S. Ardestani, Y. Choi, **C. Kim**, Y. Kim, H. Yu, H. Abdel-Aziz, J.-S. Park, H. Lee, D. Lee, M. W. Kim, H. Jung, H. Nam, D. Lim, S. Lee, J.-H. Song, S. Kwon, J. Hassoun, S. Lim, and C. Choi, "Sparsity-Aware and Re-configurable NPU Architecture for Samsung Flagship Mobile SoC," in *48th ACM/IEEE International Symposium on Computer Architecture (**ISCA**)*, 2021.

[ASPLOS'17] **C. Kim**, J. Kim, S. Kim, D. Kim, N. Kim, G. Na, Y. H. Oh, H. G. Cho, and J. W. Lee, "Typed Architectures: Architectural Support for Lightweight Scripting," in *22nd ACM Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, 2017.

[ISCA'16] **C. Kim**, S. Kim, H. G. Cho, D. Kim, J. Kim, Y. H. Oh, H. Jang, and J. W. Lee, "Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors," in *43rd IEEE/ACM International Symposium on Computer Architecture (**ISCA**)*, 2016.

[PPoPP'15] X. Piao, **C. Kim**, Y. Oh, H. Li, J. Kim, H. Kim, and J. W. Lee, "JAWS: A JavaScript Framework for Adaptive CPU-GPU Work Sharing," in *20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (**PPoPP**)*, 2015, (poster).

[ISLPED'14] W. Lee, **C. Kim**, H. Song, and J. W. Lee, "QPR.js: A runtime framework for QoS-aware power optimization for parallel JavaScript programs," in *IEEE/ACM International Symposium on Low Power Electronics and Design (**ISLPED**)*, 2014.

[PRISM'14] Y. Oh, X. Piao, **C. Kim**, and J. W. Lee, "Automatic Runtime Selection of Best Hardware for Data Parallel JavaScript Kernels via Lifelong Profiling," in *2nd International Workshop on Parallelism in Mobile Platform (**PRISM** in conjunction with the ISCA-41)*, 2014.

[WWW'14] X. Piao, **C. Kim**, Y. Oh, H. Kim, and J. W. Lee, "Efficient CPU-GPU Work Sharing for Data Parallel JavaScript," in *23rd International World Wide Web Conference (**WWW**)*, 2014, (poster).

[LCTES '13] H. Jang, **C. Kim**, and J. W. Lee, "Practical Speculative Parallelization of Variable-Length Decompression Algorithms," in *14th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems (**LCTES**)*, 2013.