Wei Cheng

michael623@iis.sinica.edu.tw | weicheng14159.github.io | github.com/WeiCheng14159 | linkedin.com/in/WeiCheng14159 | Wei Cheng (Michael)

Research Interests

Computer Architecture, Non-Volatile Memory, Graph Processing, Compute-In-Memory Architecture. Asides from research interests, I contribute to open source RISC-V projects.

Education

National Cheng Kung University (NCKU) — *Tainan*

M.S. in Computer Science and Information Engineering, GPA: 4.3/4.3 — *Sep 2020 - Jun 2022*

- Supervisor: *Prof. Ing-Chao Lin*
- Paper [1] published in **ICCAD'22**, Paper [2] published in **TCAS-I'22**
- Courses Taken: Computer Architecture, Deep Learning Integrated Circuit Design and Acceleration, Digital IC Design, AI-on-chip for Machine Learning and Inference, VLSI System Design, Computer Vision and Deep Learning

The University of Hong Kong (HKU) — *Hong Kong*

B.S. in Computer Engineering, GPA: 3.08/4.3 (2nd honor division one) — *Sep 2014 - Jun 2018*

- Supervisor: *Prof. Cho-Li Wang*
- Unpublished work: The performance optimization on TensorFlow framework on Mobile GPU devices using OpenCL

Research Experiences

Research Assistant, Institute of Information Science, Academia Sinica — *Taipei*

Graph processing on dual-addressing memory [**ICCAD'22**], *Prof. Yuan-Hao Chang* — *Feb 2021 - Now*

- Design a graph processing accelerator for dual-addressing memory (RCNVM).
- Propose two methods: Vertex-Merging (VM), and Aggressive-Vertex Merging (AVM). Both methods try to maximize cache block utilization and increase graph processing speed.
- VM acquires speedup by merging vertices in a graph while AVM merge vertices more aggressively to achieve more speedup at the expense of tolerable resulting accuracy.

Research Assistant, Computer Architecture & IC Design Lab, NCKU — *Tainan*

CNN accelerator with CLIP-Q network quantization on FPGA [**TCAS-I'22**], *Prof. Ing-Chao Lin* — *Sep 2020 - Jun 2022*

- Design a hardware-software codesigned CNN accelerator based on the CLIP-Q network quantization algorithm.
- Implement the CLIP-Q algorithm from scratch and modify it such that it fits in resource-limited computing platform.
- Propose a hardward architecture that consists of 5 ×5 reconfigurable convolutional arrays.

Summer Intern, TCL Corporate Research (HK) Co., Ltd — *Hong Kong*

Assist researchers on Structure from Motion, SLAM algorithms — *Summer 2016*

Publications

[1] **Wei**. **Cheng**, C.-F. Wu, Y.-H. Chang, and I.-C. Lin, "GraphRC: Accelerating Graph Processing on Dual-addressing Memory with Vertex Merging," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD'22)*, 2022.

[2] **Wei**. **Cheng**, I.-C. Lin, and Y.-Y. Shih, "An Efficient Implementation of Convolutional Neural Network With CLIP-Q Quantization on FPGA," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 10, pp. 4093–4102, 2022.

Teachings

2021 **Teaching Assist**, Computer Organization Course (Undergraduate level) with 100+ students — *Tainan*

2021 **Teaching Assist**, Deep Learning IC Design Course (Graduate level) — *Tainan*

Honors & Awards

2014-2018 **HKU Foundation Scholarship for International Students**, 240k HKD in total — *Hong Kong*

2016-2017 **Reaching Out Award scholarship from HKSAR gov.**, Sponsor my summer school study in UC Berkeley — *Hong Kong*

2018 **IEEE (Hong Kong) Final Year Project Merit Award**, Project Title: The performance optimization on TensorFlow framework on Mobile GPU devices using OpenCL — *Hong Kong*

Projects

A complete SW/HW co-design system for mask detection SoC *Tainan*

National Cheng Kung University *Sep 2021 - Dec 2021*

- SoC consists of: pipelined RV32I core, I-cache/D-cache, AXI bus, DMA, DRAM/ROM controller, Interrupt manager, and CNN acceleration unit.
- Apply network compression and quantization on a mask detection NN model.
- Inference the compressed NN model on our SoC with HW acceleration.
- SoC handles the booting sequence, data movements, the control of acceleration unit, and system interrupts.
- Github: `https://github.com/WeiCheng14159/VSD_CNN_accelerator`

Contribute to ria-jit (an open source RISC-V to x86 binary translator) *Tainan*

National Cheng Kung University *Sep 2020 - Dec 2020*

- Expose and fix a divide by zero bug with RISC-V compliance tests.
- Details: `https://hackmd.io/x6qVjoU1RE-yjjgviKEVBQ`

Contribute to srv32 (an open source 3-stage pipeline RV32IM core) *Tainan*

National Cheng Kung University *Sep 2020 - Dec 2020*

- Verify and contribute RV32C instructions to the existing implementation.
- Details: `https://hackmd.io/2WC89qPcSimKqCY4LxYoRA`

Other Experiences

Taiwan Semiconductor Research Institute *Tainan*

Cell-based Digitial IC Tapeout *2021*

- Participate in the design, tapeout, and verification of an UMC 0.18 um process digital IC

University of Hong Kong *Hong Kong*

Class representative for CE major students *2015 - 2016, 2017 - 2018*

Skills

Programming	Python, C++/C, Verilog/SystemVerilog
AI Frameworks	PyTorch, TensorFlow
EDA tools	NC Verilog, Design Compiler, IC Compiler/Innovus
Miscellaneous	Linux, Shell, LaTeX, Markdown, Git
Other Skills	Qi-gong (A Chinese system of physical exercises and breathing control)

Languages

Mandarian, Taiwanese	Native
English, Cantonese	Fluent