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 Analytics Accelerators

I have been interested in the hardware/software co-design of computing systems since my undergraduate study. I was first fascinated by the power of parallel computing (especially GPUs) and software-level optimizations; however, I then noticed that it is necessary to understand the hardware system to come up with a thorough solution. Therefore, I have been polishing my hardware design skills and designing my own chips (FPGAs and ASICs) since my master's study. Moreover, I set foot in the field of emerging memory and found that it is a promising candidate in the post-Moore's law era, so my master's thesis is about the performance modeling of a graph analytics accelerator based on non-volatile memory systems. Aside from these research topics, I have experience in some open-source RISC-V emulators and RV-based CNN accelerators.

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**
- Thesis: A Dual-Addressing Graph Processing Accelerator with Vertex Coalescing
- 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

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

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

- Supervisor: *Prof. Cho-Li Wang*
- Thesis: 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**], *Prof. Ing-Chao Lin* — *Sep 2020 - Jun 2022*

- Design a hardware-software co-designed 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] W. Cheng, C.-F. Wu, Y.-H. Chang, and I.-C. Lin, "GraphRC: Accelerating Graph Processing on Dual-Addressing Memory with Vertex Merging," in *Proc. of the 41st IEEE/ACM Int. Conf. on Comput.-Aided Des.*, San Diego CA, Oct. 2022, pp. 1–9. [Online]. Available: https://dl.acm.org/doi/10.1145/3508352.3549408

[2] W. Cheng, I.-C. Lin, and Y.-Y. Shih, "An Efficient Implementation of Convolutional Neural Network With CLIP-Q Quantization on FPGA," *IEEE Trans. Circuits Syst. I*, vol. 69, no. 10, pp. 4093–4102. [Online]. Available: https://ieeexplore.ieee.org/document/9849674/

Teachings

2022	**Teaching Assist for a Short Course Taught by Prof. H. T. Kung**, AI Accelerator with Good Performance Course (reserved for teachers and graduate students only) with 100+ students	*Tainan*
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

2022	**IEEE Tainan Section Best Master Thesis Award**, Thesis title: A Dual-Addressing Graph Processing Accelerator with Vertex Coalescing	*Tainan*
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/@WeiCheng14159/BJuwQJy_s`

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/@WeiCheng14159/ryh1iJ1_o`

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	GRE: 320 (∼ 2027), TOEFL: 104 (∼ 2024)