Ya Wang

ywangmu@connect.ust.hk | (+852)98165561

https://yawang.xyz

Education

The Hong Kong University of Science and Technology, Hong Kong SAR, China **Sept. 2022 – Sept.2026**
PhD. in Electronic & Comp Eng **Hong Kong PhD Fellowship Scheme (HKPFS)**

Nanjing University, Nanjing, China **Sept. 2018 - June 2022**
B.S. in Electronic Information Science and Technology with a Minor in Computer Science
GPA : **4.54/5.0**; **Ranking: 3/201** Member of **Outstanding Engineer Class** in Nanjing University

MITxPRO Program **Jan. 2021 -Mar. 2021**
Took the "Applying Machine Learning to Engineering and Science" course, and got **full marks.**

Employment

Alibaba DAMO Academy, Beijing,China **June 2022 – Present**
- Research Intern at Computational Technology Laboratory, Mentor: *Sicheng Li,* Leader: *Prof.Yuan Xie*

The Hong Kong University of Science and Technology, Hong Kong SAR, China **July 2021 – Dec. 2021**
- Research Intern at the Department of Electronic & Computer Engineering, *Advisor:Prof. Wei Zhang*

Publication

- H. Song*, Y. Wang*, M. Wang,and Z. Wang, "UCViT: Hardware-Friendly Vision Transformer via Unified Compression" 2022 IEEE International Symposium on Circuits and Systems (ISCAS), 2022

Research Experience

RISC-V CPU Analytical Performance Model Enable Design Space Exploration. **Sept.2022-Present**

- Developed a Performance Analytical Model for RISC-V CPUs, which incorporates a wider range of architectural parameters compared to previous models. Particularly, the model provides more detailed representations of two crucial components: branch predictors and memory subsystems. This improved model demonstrates a strong correlation with cycle-accurate simulation results, enabling extensive design space exploration for various hardware architectural parameters of out-of-order RISC-V processors.

AU3:Hardware and Software Co-Design for Vision Transformer via Orthogonal Multi-level Compression
Sept.2022- Apr. 2023
- Developed a hardware-software co-design compression framework to accelerate the computation of Vision Transformer (ViT). The framework orthogonally compresses the attention layer computation at multiple levels, including the decomposition and compression of embedding features, token pruning, and attention map pruning. Additionally, I restructured the data flow and designed a hardware accelerator to support these optimizations, achieving a significant improvement in the overall performance of ViT computations.

UCViT: Hardware-Friendly Vision Transformer via Unified Compression **Oct. 2020 – Oct. 2021**
Research Assistant, *Lab of Integrated Circuits and Intelligent System (ICAIS), Nanjing University*
- Developed a Unified Compression Framework for Vision Transformer (UCViT) aimed at compressing the original ViT model by integrating low bit-width quantization and dense matrix decomposition. By employing aggressive quantization, most matrix multiplications are transformed into hardware-friendly shift and addition operations. Additionally, we integrated a small module into the quantized model by exploiting the unique characteristic of multi-head attention during matrix decomposition. Thanks to the effective fusion of various compression techniques and hardware-friendly operations, the proposed model achieves up to 98% energy consumption reduction in inference compared to the original ViT model. Furthermore, it maintains a highly compact structure with a competitive compression ratio (up to 6.7 times) on CIFAR-10 and CIFAR-100 image classification tasks.

Automatic Generation Framework of Large Verification Samples for EDA Tools **July 2021-Nov. 2021**
Research Intern, Reconfigurable Computing System Laboratory (RCSL), Advisor:Prof. Wei Zhang, HKUST
- Proposed a Chipyard-based large-scale System on Chip (SoC) generation and verification framework to address the verification requirements of EDA tools. The proposed framework allows for easy generation of customized large-scale SoCs by parameterizing the number and type of CPU cores, CPU micro-architecture parameters, bus structure, and more while automatically deploying them in a multi-FPGA system. The generated SoC system is

verified at different levels, including software-level simulation, multi-core memory access and communication verification before Linux system startup, and OpenMP-based multi-threaded software testing after bootloading a Linux system.

High Speed Chip Interconnection System and Design of Optimized Coding **Sept. 2020 – Sept. 2021**
Research Assistant, Advisor: Prof. Yuan Du, Nanjing University

- Constructed FPGA based channel testing system using GTX transceivers for high-speed differential signals.
- Analyzed and verified high-speed signal integrity with different encoding methods and various channels made of PCB using IBERT tools.
- Explored the optimized encoding method and channel for high-speed signal transmission based on the premise of low BER and jitter.

Prize and Awards

- **Hong Kong PhD Fellowship Scheme (HKPFS)**
- RedBrid Scholarship
- **National Scholarship** (Highest scholarship for Chinese undergraduates, 3 out of 204), 2020
- Puxin Elite Scholarship (1 out of 204, Reward outstanding talents in the field of integrated circuits) ,2021
- CETC The 14th Research Institute Guorui Scholarship, 2021
- People's Scholarship in Nanjing University, 2019&2020&2021
- Outstanding Student Cadre Model (only one selected in whole college, 1 out of 800) , 2021
- **Competitions Expert of Kaggle Competition** (top 1.7% in the world), two silver medals (top 5%) ,2021
- **Best Performance Awards of 2021 Xilinx Women in Technology (WIT) Hackathon**
- Second Prize in Final Contest at National Undergraduate IOT Design Contest, 2020; Grand Prize in the eastern China division (top 5% nationally)
- Honorable Mention in 2020 MCM
- Second prize in TI cup Jiangsu college students electronic design competition (Quadrotor aircraft), 2020

Skills

Programming Languages: Python, C/C++, Verilog, Chisel, Matlab, Latex
Framwroks: PyTorch, Keras, Chipyard, OpenCV