Huangxu Chen

Tel: (86) 15817666127 | Email: hchen499@connect.hkust-gz.edu.cn

EDUCATION

The Hong Kong University of Science and Technology (Guangzhou) **Sep 2024 - Present**

- Master of Philosophy | Microelectronics Thrust | Grade-Point Average: 3.79/4
- Key courses: Parallel Computer Architecture (A+), Customized Computing with FPGAs (A-), Accelerated Computing with GPU (A-), etc.

South China University of Technology **Sep 2020 - Jul 2024**

- Bachelor of Engineering | Microelectronics Science and Technology | Grade-Point Average: 3.68/4 (27/112)
- Key courses: Digital Electronic Technology (98), Signals and Systems (89), Digital Signal Processing (92), Course Design of Digital Integrated Circuits (A), Verilog Design and FPGA (96), etc.

RESEARCH INTERESTS

- Large Language Model Acceleration, Hardware-Software Co-design, Computer Architecture, Neural network Quantization and Compression.

PUBLICATIONS

[1] **Huangxu Chen***, Yingbo Hao*, Yi Zou, and Xinyu Chen, “OA-LAMA: An Outlier-Adaptive LLM Inference Accelerator with Memory-Aligned Mixed-Precision Group Quantization”, in International Conference on Computer-Aided Design (ICCAD), 2025. (Accepted)

[2] Yingbo Hao*, **Huangxu Chen***, Yi Zou, and Yanfeng Yang, “An Algorithm-Hardware Co-design Based on Revised Microscaling Format Quantization for Accelerating Large Language Models”, in Design Automation Conference (DAC), 2025.

[3] Yingbo Hao*, **Huangxu Chen***, Yi Zou, and Yanfeng Yang, “A Dynamic Logic Based Reconfigurable Digital CIM Macro for Edge Computing”, in International Conference on Circuits and Systems (ICCS), 2024.

[4] **Huangxu Chen***, and Mingjian Zhao, "Low Resource-Cost Depthwise Separable Convolutional Co-Processor Architecture", International Conference on Electronic Information Engineering and Computer Science (EIECS), 2023.

RESEARCH/PROJECT EXPRIENCE

Singular Value-Aware Mixed-precision quantization with Dynamic Floating-Point Format **Apr 2025 - Present**

- Designing a singular value-aware rotation mechanism that enables dynamic feature reorganization for mixed-precision quantization.
- Proposing a bias-enhanced dynamic floating-point format to augment data expressiveness, effectively addressing precision loss in ultra-low-bit quantization scenarios.
- **Outcome:** Paper in preparation.

Outlier-Adaptive LLM Inference Accelerator with Mixed-Precision Quantization **Nov 2024 - Apr 2025**

- Created OA-LAMA, a co-designed framework featuring memory-aligned mixed-precision group quantization (OAMAG) and novel outlier reordering technique.
- Proposed distribution-aware group allocation strategy addressing inter-layer outlier variance and designed hardware with three-level accumulation architecture.
- **Outcome:** One paper accepted to the 2025 IEEE/ACM ICCAD.

Revised Microscaling Format Quantization for LLM Acceleration **Jul 2024 - Nov 2024**

- Proposed an algorithm-hardware co-design featuring a two-level Revised MX Format Quantization (RMFQ) and a dedicated Revised MX Format Accelerator (RMFA) architecture.
- Developed the revised MX (RMX) format with innovative group-wise quantization.
- **Outcome:** One paper accepted to the 2025 IEEE/ACM DAC.

Design of Dynamic Reconfigurable Digital In-memory Computing Macro **Sep 2023 - May 2024**

- Adopt dynamic logic instead of static logic to reduce the number of transistors and transmission delay.
- Design a reconfigurable Manchester carry chain that supports flexible configuration of multiple data bit widths.
- Integrate seven reusable computing operators, optimize circuit area and energy efficiency, and achieve a multiplicative energy efficiency of 0.76TOPS/W@0.8V under TSMC 40nm process.

- **Outcome:** Documented in undergraduate dissertation, one paper accepted to the 2024 IEEE ICCS, and won the best oral presentation.

RISC-V Micro-architecture Optimization and Instruction Set Expansion **Jun 2023 - Aug 2023**

- Optimized the memory access path and prediction unit of the hummingbird E203 processor, reducing the probability of memory access conflicts and prediction failures.
- Expanded the floating-point instruction set and integrated integer and floating-point instruction hardware.
- **Outcome:** Won the first prize in the 2023 South China Integrated Circuit Innovation Competition and the second prize in the 2023 National Integrated Circuit Innovation Competition.

Hardware Accelerator for Depthwise Separable Convolution **Dec 2022 - Aug 2023**

- Design the multiplier multiplexing channel and AXI4 control module for depthwise separable convolution computation based on the DDR3 memory interface of the AXI4 protocol.
- **Outcome:** One paper accepted to the 2023 IEEE EIECS.

SKILLS

- **Language:** Mandarin (Native), English (IELTS 6.5).
- **Tool skills:**
  1. Program Language: Python, Verilog, CUDA, C.
  2. RTL simulation software: Modelsim, Vcs+Verdi, Verilator+GTKWave.
  3. IC Back-end Tools: Synosys Design Compiler, Cadence Encounter.
  4. FPGA software: Vivado, Quartus, Pango, Gowin.