Yuzong Chen

Mobile: +1 (607) 262-1616, **Email:** yc2367@cornell.edu, **Website:** https://yc2367.github.io

EDUCATION

- **Cornell University** *Aug. 2022 – Present*
  Ph.D. in Electrical and Computer Engineering
  Advisor: Prof. Mohamed Abdelfattah
- **Nanyang Technological University, Singapore** *Aug. 2015 – Jun. 2019*
  B.Eng. in Electrical & Electronic Engineering
  GPA: 4.74 / 5.00, Honours (Highest Distinction)

RESEARCH INTERESTS

Efficient Hardware for Deep Learning: I am interested in developing new hardware architectures for accelerating deep neural networks, with a special focus on sparsity and mixed precision. This includes deep neural networks compression and quantization and FPGA/ASIC design.

RESEARCH EXPERIENCE

- **Cornell University** *Aug. 2022 – Present*
  Graduate Research Assistant *Advisor: Prof. Mohamed Abdelfattah*
  Currently working on efficient algorithm-hardware co-design for sparse and mixed-precision deep neural networks and large language models.
- **National University of Singapore** *Sept. 2021 – Jul. 2022*
  Research Engineer *Advisor: Prof. Heng Chun-Huat*
  Work on a joint project with NXP Semiconductors to design an RF switched-capacitor power amplifier for high-speed communication. Help tape-out the chip in 22nm FDSOI technology.
- **Nanyang Technological University** *Feb. 2020 – Aug. 2021*
  Project Officer *Advisor: Prof. Tony Tae-Hyoung Kim*
  Conduct and lead projects about computing in-memory circuit design based on static random access memory (SRAM) and resistive random access memory (ReRAM). Help tape-out several chips in 65nm technology.

INDUSTRY EXPERIENCE

- **Qualcomm AI Research**, San Diego, USA *Sept. 2024 – Present*
  Research Intern
  Implement performance modelling framework for deep learning on heterogeneous hardware platforms.

PUBLICATIONS

- **Conference Proceedings**

[1] Yuzong Chen*, Xilai Dai*, Chi-Chih Chang*, Yash Akhauri and Mohamed S. Abdelfattah, "The Power of Negative Zero: Datatype Customization for Quantized Large Language Models," in *Conference on Machine Learning and Systems (MLSys),* 2025. (* Equal contribution) [Submitted]

[2] Yuzong Chen, Ahmed AbouElhamayed, Xilai Dai, Yang Wang, Marta Andronic, George A. Constantinides and Mohamed S. Abdelfattah, "BitMoD: Bit-serial Mixture-of-Datatype LLM Acceleration," in *IEEE International Symposium on High-Performance Computer Architecture (HPCA),* 2025.

[3] Yuzong Chen, Jian Meng, Jae-sun Seo and Mohamed S. Abdelfattah, "BBS: Bi-directional Bit-level Sparsity for Deep Learning Acceleration," in *IEEE/ACM International Symposium on Microarchitecture (MICRO),* 2024.

[4] Xilai Dai, Yuzong Chen and Mohamed S. Abdelfattah, "Kratos: An FPGA Benchmark for Unrolled Deep Neural Networks with Fine-Grained Sparsity and Mixed Precision," in *IEEE International Conference on Field-Programmable Logic and Applications (FPL)*, 2024.

[5] Jordan Dotzel, Yuzong Chen, Bahaa Kotb, Sushma Prasad, Gang Wu, Sheng Li, Mohamed S Abdelfattah and Zhiru Zhang, "Learning from Students: Applying t-Distributions to Explore Accurate and Efficient Formats for LLMs," in *International Conference on Machine Learning (ICML)*, 2024.

[6] Yuzong Chen, Jordan Dotzel and Mohamed S. Abdelfattah, "M4BRAM: Mixed-Precision Matrix-Matrix Multiplication in FPGA Block RAMs," in *IEEE International Conference on Field Programmable Technology (FPT)*, 2023.

[7] Yuzong Chen and Mohamed S. Abdelfattah, "BRAMAC: Compute-in-BRAM Architectures for Multiply-Accumulate on FPGAs," in *IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2023.

[8] Yuzong Chen, Junjie Mu, Hyunjoon Kim, Lu Lu, and Tony Tae-Hyoung Kim, "A Reconfigurable 8T SRAM Macro for Bit-Parallel Searching and Computing In-Memory," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022.

[9] Yuncheng Lu, Zehao Li, Yuzong Chen, and Tony Tae-Hyoung Kim, "A 181μW Real-Time 3-D Hand Gesture Recognition System based on Bi-directional Convolution and Memoryless Clustering," in *IEEE Custom Integrated Circuits Conference (CICC)*, 2022.

[10] Yuzong Chen, Lu Lu, Yuncheng Lu, and Tony Tae-Hyoung Kim, "A Multi-Functional 4T2R ReRAM Macro Enabling 2-Dimensional Access and Computing In-Memory," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[11] Lu Lu, Yuzong Chen, and Tony Tae-Hyoung Kim, "A Configurable Randomness Enhanced RRAM PUF with Biased Current Sensing Scheme," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[12] Vishal Sharma, Ju Eon Kim, Yong-Jun Jo, Yuzong Chen, and Tony Tae-Hyoung Kim, "AND8T SRAM Macro with Improved Linearity for Multi-bit In-Memory Computing," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[13] Yuzong Chen, Lu Lu, Bongjin Kim, and Tony Tae-Hyoung Kim, "Reconfigurable 2T2R ReRAM with Split Word-lines for TCAM Operation and In-Memory Computing," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020.

- **Journal Articles**

[1] Yuzong Chen, Junjie Mu, Hyunjoon Kim, Lu Lu, and Tony Tae-Hyoung Kim, " BP-SCIM: A Reconfigurable 8T SRAM Macro for Bit-Parallel Searching and Computing In-Memory," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2023.

[2] Donghyuk Kim, Chengshuo Yu, Shanshan Xie, Yuzong Chen, Joo-Young Kim, Bongjin Kim, Jaydeep Kulkarni, and Tony Tae-Hyoung Kim, "An Overview of Processing-in-Memory Circuits for Artificial Intelligence and Machine Learning," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, 2022. [Featured as one of the most popular papers in IEEE JETCAS]

[3] Yuzong Chen, Lu Lu, Bongjin Kim, and Tony Tae-Hyoung Kim, "A Reconfigurable 4T2R ReRAM Computing In-Memory Macro for Efficient Edge Applications," in *IEEE Open Journal of Circuits and Systems*, 2021.

[4] Yuzong Chen, Lu Lu, Bongjin Kim, and Tony Tae-Hyoung Kim, "Reconfigurable 2T2R ReRAM Architecture for Versatile Data Storage and Computing In-Memory," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2020.

- **Book Chapters**

[1] Tony Tae-Hyoung Kim, Yuzong Chen, and Lu Lu, "ReRAM-based Processing-in-Memory (PIM)," in *Processing-in-Memory for AI from Circuits to Systems*, Springer, 2022, pp. 93-120.

HONOURS AND AWARDS

- Qualcomm Innovation Fellowship (QIF), Finalist — 2024
- Cornell Graduate Fellowship — 2022
- Undergraduate Dean's List (top 5% of the cohort) — 2015, 2016
- Singapore Science and Engineering Undergraduate Scholarship — 2015 – 2019

INVITED TALKS

- "Leveraging Bit-serial Computation for Deep Learning Acceleration," at Samsung AI Research Cambridge, [Online], July 11, 2024.

TEACHING EXPERIENCE

- CS 5785, Applied Machine Learning Fall 2023

EDITORIAL SERVICE

- **Journal Reviewer:** IEEE Transactions on Very Large Scale Integration Systems (TVLSI); IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I); IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS). IEEE Open Journal of Circuits and Systems (OJCAS).
- **Conference Secondary Reviewer:** International Symposium on Field-Programmable Gate Arrays (FPGA), Design Automation Conference (DAC).

TECHNICAL SKILLS

- Python, C++, SystemVerilog, Pytorch, Git, Cadence Virtuoso, Vivado HLS, Synopsys Design Compiler.