Yuzong Chen

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

Education

- **Cornell University** Aug. 2022 – Dec. 2026 (Expected)
  Ph.D. in Electrical and Computer Engineering
  Advisor: Prof. Mohamed S. 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 novel hardware architectures for accelerating deep learning applications. This includes efficient inference algorithms and accelerator co-design.

Research Experience

- **Stanford University** Jan. 2026 – Apr. 2026
  Visiting Scholar *Advisor: Prof. Thierry Tambe*
  Work on large language model acceleration via compute-optimal test-time scaling.
- **Cornell University** Aug. 2022 – Present
  Graduate Research Assistant *Advisor: Prof. Mohamed S. Abdelfattah*
  Work on algorithm-hardware co-design for machine learning acceleration, with a special focus on quantization numerics, FPGA architectures, and processing in-memory.
- **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, Singapore** Feb. 2020 – Aug. 2021
  Project Officer *Advisor: Prof. Tony Tae-Hyoung Kim*
  Conduct and lead projects about processing in-memory circuit design based on static random access memory (SRAM) and resistive random access memory (ReRAM). Tape-out several chips in 65nm technology.

Industry Experience

- **Meta**, **AI & Systems Co-Design Team** May. 2026 – Aug. 2026
  Incoming Research Scientist Intern *Menlo Park, CA*
  Explore efficient quantization numerics for LLM training/inference acceleration to guide the next-generation Meta Training Inference Accelerator (MTIA) chip.
- **Samsung Semiconductor**, **AGI Computing Lab** Jun. 2025 – Aug. 2025
  Architecture Research Intern *San Jose, CA*
  Propose novel mixed-precision LLM quantization algorithms by jointly optimizing all matrix multiplication modules, including both linear and attention layers.
- **Qualcomm AI Research** Sept. 2024 – Dec. 2024
  Research Intern *San Diego, CA*
  Implement performance modelling framework for LLM inference on heterogeneous NPU-PIM systems.

Publications

- **Conference Proceedings**

[1] **Yuzong Chen**, Chao Fang, Xilai Dai, Yuheng Wu, Thierry Tambe, Marian Verhelst, Mohamed S. Abdelfattah, "Unlocking Efficient Processing-In-Memory for Edge LLM Inference with Hybrid Numerical Formats", submitted to ACM/*IEEE International Symposium on Computer Architecture (ISCA),* 2026. **[Under Review]**

[2] **Yuzong Chen**, Ahmed AbouElhamayed, Xilai Dai, Yang Wang, Marta Andronic, George A. Constantinides,

Mohamed S. Abdelfattah, “BitMoD: Bit-serial Mixture-of-Datatype LLM Acceleration”, *IEEE International Symposium on High-Performance Computer Architecture (HPCA),* 2025.

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

[4] Xilai Dai, **Yuzong Chen**, Mohamed S. Abdelfattah, “Kratos: An FPGA Benchmark for Unrolled Deep Neural Networks with Fine-Grained Sparsity and Mixed Precision”, *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, Zhiru Zhang, "Learning from Students: Applying t-Distributions to Explore Accurate and Efficient Formats for LLMs”, *International Conference on Machine Learning (ICML)*, 2024.

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

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

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

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

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

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

[12] Vishal Sharma, Ju Eon Kim, Yong-Jun Jo, **Yuzong Chen**, Tony Tae-Hyoung Kim, “AND8T SRAM Macro with Improved Linearity for Multi-bit In-Memory Computing”, *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”, *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020.

- **Journal Articles**

[1] **Yuzong Chen**, Chi-Chih Chang, Xilai Dai, Ahmed AbouElhamayed, Marta Andronic, George A. Constantinides, Mohamed S. Abdelfattah “Bit-serial Acceleration of LLM Inference with Mixture-of-Datatype Quantization”, *IEEE Transactions on Computers (TC),* 2025.

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

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

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

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

Teaching Experience

- Modern Computer Systems and Architecture (ECE5755 / CS5754, Cornell Tech) Fall 2025
  Head Teaching Assistant *Instructor: Prof. Udit Gupta*
- Applied Machine Learning (CS5785, Cornell Tech) Fall 2023
  Teaching Assistant *Instructor: Prof. Kyra Gan*

HONOURS AND AWARDS

- Student Travel Grant: HPCA 2025
- Qualcomm Innovation Fellowship (QIF), Finalist, 2024
- Cornell Graduate Fellowship, 2022
- NTU Singapore Undergraduate Dean's List (top 5% of the cohort), 2015 – 2017
- Singapore Science and Engineering Undergraduate Scholarship, 2015 – 2019

INVITED TALKS

- **Efficient Computing In-memory Architectures for FPGA-based Deep Learning Acceleration**
  - FCCM'23, Los Angeles, CA, May 2023
  - FPT'23, Japan, Dec. 2023
  - Centre for Spatial Computational Learning (SpatialML), Online, Mar. 2024
- **Leveraging Bit-serial Computation for Deep Learning Acceleration**
  - Samsung AI Research Cambridge, Online, Jul. 2024
  - MICRO'24, Austin, TX, Nov. 2024
  - Southeast University, China, Dec. 2024
- **BitMoD: Bit-serial Mixture-of-Datatype LLM Acceleration**
  - Qualcomm AI Research, San Diego, CA, Nov. 2024
  - Intel Research Review, Online, Nov. 2024
  - HPCA'25, Las Vegas, NV, Mar. 2025
  - Computer Architecture Day @ Columbia University, New York, May 2025

EDITORIAL SERVICE

- **Journal Reviewer**
  IEEE Transactions on Computers (TC)
  IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I)
  IEEE Transactions on Very Large Scale Integration Systems (TVLSI)
  IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)
- **Conference Secondary Reviewer**
  International Symposium on Field-Programmable Gate Arrays (FPGA), 2024
  Design Automation Conference (DAC), 2023
- **Artifact Evaluation Committee**
  International Symposium on High-Performance Computer Architecture (HPCA), 2025