ZHICHEN ZENG

Tel: +1 (607)-327-2454 E-mail: zczeng@uw.edu Home: zhichenzzz Publication: Google Scholar

Education

University of Washington **Sept. 2024 -**

- *Doctor of Philosophy in Electrical and Computer Engineering*
- Advisor: Ang Li and Banghua Zhu
- Research interest: ML systems, efficient systems for LLMs

University of Science and Technology of China (USTC) **Sept. 2020 – June 2024**

- *Bachelor of Science in Physics*
- GPA: 4.05/4.3 (Top 1%)
- **Guo Moruo Scholarship**, Highest honor at USTC

Past Research Experience

Microsoft Research Asia | ***Research Intern advised by Dr. Shijie Cao and Dr. Ting Cao***

Project: Learning Intrinsic Sparse Attention for Long-Context LLMs Mar. 2024 – Aug. 2024

- Proposed a learning-based method for intrinsic attention sparsity in LLMs, enabling efficient post-training and fine-tuning. Our learned attention sparsity outperformed over predefined patterns or heuristic approximations.
- Achieved 90% sparsity at 32k context with only 5% perplexity loss, delivering a 5.67× speedup over FlashAttention.

Cornell University | ***Research Intern advised by Prof. Zhiru Zhang***

Project: A Programming Model for Composable Accelerator Design Jul. 2023 – Nov. 2023

- Developed an MLIR-based compiler and DSL for modular, high-performance hardware accelerators. Achieved 1.7× faster inference latency and 5.4× higher energy efficiency than A100 GPU on GPT-2.

USTC | ***Undergrad Researcher advised by Prof. Xi Jin***

Project: Bit-weight dimension optimizations for TensorCore Jul. 2023 – Nov. 2023

- Focusing on the bit-weight dimension of the multiply-accumulator (MAC) to optimize tensor processing engines (TPEs) in GPUs and domain-specific architectures. Achieved up to better area and energy efficiency improvements over existing architectures.

Rewards & Honors

Meta-UW AI Mentorship Program *Apr. 2025*

Guo Moruo Scholarship*, highest honor at USTC (35/1958)* *Jun. 2024*

Microsoft Research Asia "Stars of Tomorrow" *Aug. 2024*

Outstanding Student Scholarship, Gold*, USTC* *Nov. 2021, 2022, 2023*

Scholarship for Talent Program in Basic Disciplines, Class A*, USTC* *Oct. 2021, 2022, 2023*

Service

Artifact Evaluation Committee: MLSys 2025, ASPLOS 2025, HPCA 2025, MICRO 2024

Conference Reviewer: ICLR 2025, ACL 2025, NeurIPS 2024

Teaching Assistant: CSE 469: Computer Architecture, Spring 2025, UW

Skills

Languages: English, Chinese

Program Languages: Python, C/C++, CUDA, System Verilog, MATLAB, Mathematica

Frameworks: PyTorch, Triton, CUTLASS