Ruofan Wu

Tel: (734)747-1652 | Email: ruofanw@umich.edu | Homepage: https://ruofanwu.github.io/

I am a second year Ph.D. student in Computer Science and Engineering at the Univeristy of Michigan, advised by Prof. Mosharaf Chowdhury. My research interests lie in machine learning compilers and scalable machine learning systems, with recent and upcoming work aiming to build energy-efficient execution stacks for large model training, particularly for generative AI workloads.

Education

University of Michigan — Ann Arbor, MI, USA
Ph.D. student in Computer Science and Engineering — *Sep 2024 - Present*

Renmin University of China — Beijing, China
M.E. in Computer Application Technology — *Sep 2021 - Jun 2024*

Renmin University of China — Beijing, China
B.E. in Data Science and Big Data Technology — *Sep 2017 - Jun 2021*

Experience

Microsoft Software Technology Center Asia — Beijing, China
Research Intern at Bing, mentored by Dr. Zhen Zheng — *Oct 2023 - Jun 2024*

Alibaba Cloud — Hangzhou, China
Alibaba Innovative Research (AIR) Intern at Platform of Artificial Intelligence — *Jan 2022 - Sep 2023*

Microsoft Research Asia — Beijing, China
Research Intern at Systems Research Group, mentored by Dr. Jilong Xue and Dr. Fan Yang — *Apr 2021 - Dec 2021*

North Carolina State University — Remote
Collaborator with PICTure Research Group, mentored by Prof. Xipeng Shen — *Nov 2019 - Mar 2021*

Remin University of China — Beijing, China
Student at Database and Intelligent Information Retrieval Group, advised by Prof. Feng Zhang — *Oct 2019 - Jun 2024*

Selected Publications

- **Where Do the Joules Go? Diagnosing Inference Energy Consumption**, Jae-Won Chung, **Ruofan Wu**, Jeff J. Ma, Mosharaf Chowdhury, Preprint (under submission).
- **Kareus: Joint Reduction of Dynamic and Static Energy in Large Model Training**, **Ruofan Wu**, Jae-Won Chung, Mosharaf Chowdhury, Preprint (under submission).
- **The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization**, Jae-Won Chung, Jeff J. Ma, **Ruofan Wu**, Jiachen Liu, Oh Jun Kweon, Yuxuan Xia, Zhiyu Wu, Mosharaf Chowdhury, in **NeurIPS Datasets and Benchmarks (Spotlight)**, 2025.
- **TetriServe: Efficient DiT Serving for Heterogeneous Image Generation**, Runyu Lu, Shiqi He, Wenxuan Tan, Shenggui Li, **Ruofan Wu**, Jeff J. Ma, Ang Chen, Mosharaf Chowdhury, in **ASPLOS**, 2026.
- **PluS: Highly Efficient and Expandable ML Compiler with Pluggable Graph Schedules**, **Ruofan Wu**, Zhen Zheng, Feng Zhang, Chuanjie Liu, Zaifeng Pan, Jidong Zhai, Xiaoyong Du, in **USENIX ATC**, 2025.
- **ROLLER: Fast and Efficient Tensor Compilation for Deep Learning**, Zhu Hongyu, **Ruofan Wu**, Yijia Diao, Shanbin Ke, Haoyu Li, Chen Zhang, Jilong Xue, Lingxiao Ma, Yuqing Xia, Wei Cui, Fan Yang, Mao Yang, Lidong Zhou, Asaf Cidon, Gennady Pekhimenko, in **OSDI**, 2022.
- **DREW: Efficient Winograd CNN Inference with Deep Reuse**, **Ruofan Wu**, Feng Zhang, Jia Wei Guan, Zhen Zheng, Xiaoyong Du, Xipeng Shen, in **TheWebConf/WWW**, 2022.

Additional Information

- **Programming Languages:** Python, C/C++/CUDA
- **Tools and Frameworks:** NeMo, Megatron-LM, TransformerEngine, PyTorch (Dynamo), Nsight, Zeus, NVML
- **Teaching Assistant:** Parallel Architecture and Programming (Renmin University of China, Fall 2022)