Zichen Fan

https://zichenfan.github.io/

Email : zcfan@umich.edu

Mobile : +1-210-724-9052(US) +86-18813101800(CN)

Education

- **University of Michigan** Ann Arbor, MI
  PhD Candidate in ECE, Advisor: Prof. Dennis Sylvester and Prof. David Blaauw *Aug. 2019 – Nov. 2024 (exp.)*
- **University of Michigan** Ann Arbor, MI
  Master of Engineering in ECE, GPA: 4.0/4.0 *Aug. 2019 – Apr. 2022*
- **Tsinghua University** Beijing, China
  Bachelor of EE, GPA: 3.7/4.0 *Aug. 2015 – Jul. 2019*
- **Duke University** Durham, NC
  Summer Research Intern in ECE, Advisor: Prof. Yiran Chen *Jul. 2018 – Sept. 2018*

Industry

- **Qualcomm** San Diego, CA
  Engineering Intern, Qualcomm Research Center *May 2023 – Aug. 2023*

Selected Publication

- **Z. Fan**, et al. TaskFusion: An Efficient Transfer Learning Architecture with Dual Delta Sparsity for Multi-Task Natural Language Processing. *International Symposium on Computer Architecture (ISCA)*. 2023.
- S. Shoouri, M. Yang, **Z. Fan**, et al. Efficient Computation Sharing for Multi-Task Visual Scene Understanding. *International Conference on Computer Vision (ICCV)*. 2023
- P. Abillama, **Z. Fan**, et al. SONA: An Accelerator for Transform-Domain Neural Networks with Sparse Orthogonal Weights. *International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 2023 (Best Paper)
- C-W. Tseng, Z. Feng, **Z. Fan**, et al. Reconfigurable Analog FIR Filter Achieving -70dB Rejection with Sharp Transition for Narrowband Receivers *IEEE Symposium on VLSI Circuits (VLSI-Symp)*. IEEE, 2023.
- H. An, Y. Chen, **Z. Fan**, et al. A 8.09TOPS/W Neural Engine Leveraging Bit-Sparsified Sign-Magnitude Multiplications and Dual Adder Trees. *International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023.
- **Z. Fan**, et al. Audio and Image Cross-Modal Intelligence via a 10TOPS/W 22nm SoC with Back-Propagation and Dynamic Power Gating. *IEEE Symposium on VLSI Circuits (VLSI-Symp)*. IEEE, 2022.
- Q. Zhang, H. An, **Z. Fan**, et al. A 22nm 3.5TOPS/W Flexible Micro-Robotic Vision SoC with 2MB eMRAM for Fully-on-Chip Intelligence. *IEEE Symposium on VLSI Circuits (VLSI-Symp)*. IEEE, 2022.
- **Z. Fan**, et al. ASP-SIFT: Using Analog Signal Processing Architecture to Accelerate Keypoint Detection of SIFT Algorithm. *IEEE Transactions on Very Large Scale Integration (T-VLSI) Systems*. 2019.
- Z. Liu, **Z. Fan**, et al. Design of Switched-Current Based Low-Power PIM Vision System for IoT Applications. *2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 2019.
- **Z. Fan**, et al. RED: A ReRAM-based Deconvolution Accelerator. *Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019.

Working Paper

- **Z. Fan**, et al. AIMMI: Audio and Image Multi-Modal Intelligence via a 10TOPS/W 22nm Low Power SoC with 2MByte On-chip MRAM for IoT Devices. *Submitted to IEEE Journal of Solid-State Circuits (JSSC)*. IEEE, 2023.

Research

- **Multi-task Processing Transformer Accelerator** University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim *Mar. 2022 - Present*
  - **Publications**: 1^{st} author paper on multi-task natural language processing accepted by ISCA 2023, extended idea to multi-task image/video processing accepted by ICCV 2023
  - [**Architecture**] **Efficient Multi-task Transformer Architecture**: Designed a heterogeneous architecture with dedicated dense, sparse, and attention cores for accelerating Transformer-based multi-tasking scenarios. Proposed an energy-efficient task scheduling scheme to reduce off-chip memory access and increase computation utilization.
  - [**Algorithm**] **Efficient Multi-task Transformer Inference**: Proposed an efficient Transformer inference algorithm by sharing both weight and activation of sub-task execution from base task. Transformed dense computation to sparse computation for inference acceleration.
  - [**Algorithm**] **Data-sharing Boosted Transfer Learning**: Proposed an efficient transfer learning algorithm for NLP and image/video applications by adding regularization on both delta weight and delta activation during training. Taking advantage of the pre-trained and fine-tuning scheme, the method achieved more than 5x the number of FLOPs reduction for each sub-task with negligible accuracy loss.
- **Multi-modal Signal Processing Accelerator** University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim *Aug. 2019 - Feb. 2022*
  - **Publications**: 1^{st} author paper accepted by VLSI-symp 2022, extended journal paper submitted to JSSC. MRAM-related part is used in a low-power vision system accepted by VLSI-symp 2022. Audio interface part is used in a low-power KWS system accepted by ISSCC 2023, extended journal paper submitted to JSSC
  - [**Circuit**] **Multi-modal Signal Processing SoC**: Designed a 12mm^2 ultra-low-power multi-modal signal processor SoC in TSMC 22nm technology that integrated a versatile deep neural network engine with audio and image signal processing accelerators for cross-modal IoT intelligence. The SoC achieves up to 3-10 TOPS/W peak energy efficiency and consumes only 0.25-3.84 mW. Being the first to demonstrate CNN, GAN, and back-propagation (BP) on a single accelerator SoC for cross-modal fusion, it outperforms state-of-the-art DNN processors by 1.4 - 4.5× in energy efficiency. **The first real silicon I taped out and tested successfully.**
  - [**Circuit**] **MRAM**: Used TSMC 22nm 2MB MRAM macro on two different real-silicon tape-outs and tested correctly. Proposed MRAM dynamic power gating technology and MRAM-weight cache architecture. Demonstrated MRAM's advantages in IoT applications.
  - [**EDA**] **CPF Flow**: Used Cadence Innovus CPF Flow to handle multiple power domains (wrote the first example script for the group). The chip was tested correctly and the scripts were used in another 2 real silicon tape-outs.
  - [**Algorithm**] **Back-propagation GAN for Audio Compression**: Worked on the GAN-based audio compression algorithms, which result in low data rate (less than 1kbps) for audio. Used BPGAN to compress audio and reduced the complexity of the network to be compatible with the computation ability of the hardware.
- **Transform Domain Neural Network Accelerator** University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim *Aug. 2021 - Mar. 2022*
  - **Publications**: 2^{nd} author paper accepted by ASAP 2023, Best Paper Award
  - [**Architecture**] **Transform Domain NN Accelerator**: Collaboratively designed a novel transform-domain neural network accelerator in which convolution operations are replaced by element-wise multiplications with sparse-orthogonal weights. Architecture simulation showed up to 5.2× performance gain from traditional convolution accelerators.
  - [**Circuit**] **MRAM-logic Asynchronous Interface**: Helped designed a MRAM-logic asynchronous interface. Collaboratively designed a data pre-loading scheme. Helped tape out a TSMC 22nm testing chip.
- **Sign-Magnitude Neural Network Accelerator** University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim *Oct. 2021 - Mar. 2022*
  - **Publications**: 3^{rd} author paper accepted by ISSCC 2023
  - [**Circuit**] **Bit-sparsified Sign-Magnitude NN Accelerator**: Collaboratively designed a neural network accelerator using bit-sparsified sign-magnitude multiplications and dual adder trees. Helped tape out a TSMC 28nm testing chip with both two's complement NN accelerator and sign-magnitude NN accelerator. Results showed 50% energy efficiency gain comparing to traditional two's complement NN accelerator.
- **GNN-based Cost Prediction on Compiled Graph** Qualcomm Research Center
  Mentors: Mahesh Balasubramanian and Apoorva Gokhale Manager: Rishi Chaturvedi *May 2023 - Present*

  - **[Compiler] DNN Compiler**: Used DNN compiler for compiling more than 400 DNN models and analyzed both high-level IR and low-level IR. Run compiled graphs on Qualcomm Cloud AI 100 Chip and extracted instruction features and node-level execution durations as custom dataset.
  - **[Algorithm] GNN-based Cost Prediction**: Proposed a GNN-based cost prediction algorithm that used GNN pre-trained model and finetuning scheme.

- **Analog Signal Processing SIFT Accelerator** — Tsinghua University
  Advisor: Prof. Fei Qiao — *Jul. 2017 - Jul. 2018*
  - **Publications**: 1st author paper accepted by T-VLSI in 2019.
  - **[Circuit] Analog Computing**: Worked on the SIFT acceleration system using analog signal processing architecture. Proposed an analog circuit network that realized the keypoint detection part of SIFT. Results showed that the system can process 2.3k VGA frames per second, which is at least 3.26 times faster than the state-of-art digital hardware accelerators.
- **Near-sensor Current-based CIM Accelerator** — Tsinghua University
  Advisor: Prof. Fei Qiao — *Oct. 2018 - Jun. 2019*
  - **Publications**: 1st co-author paper accepted by ISVLSI 2019.
  - **[Circuit] Current-based CIM**: Worked on the switched-current based low-power computer-in-memory (CIM) vision system. Designed current mode APS and current mode PIM architecture. The designed system outperformed the state-of-the-art designs in terms of power consumption (1.45mW) and achieves energy efficiency up to 28.25TOPS/W.
- **RRAM-based CIM Deconvolution Accelerator** — Duke University
  Co-advisors: Prof. Yiran Chen and Prof. Hai Li — *Jul. 2018 - Sept. 2018*
  - **Publications**: 1st author paper accepted by DATE 2019, extended journal paper accepted by TCAD in 2020
  - **[Architecture] RRAM-based CIM Architecture for Deconvolution**: Worked on ReRAM-based compute-in-memory deconvolution accelerator design, which aims to accelerate the deconvolution operation in Generative Adversarial Networks and Fully Convolutional Networks using ReRAM. Proposed the pixel-wise mapping scheme for reducing redundancy and zero-skipping data flow for increasing the computation parallelism which could accelerate the deconvolution operation by 3.69× ∼ 31.15×.

Academic Service

- **Reviewer**: JSSC, TCAS-II, MICRO

Skills

- **Languages**: C, C++, Python, MATLAB, Verilog
- **Tools**: Synopsys & Cadence Design Tools, Pytorch, Vivado, Spice

Awards

- ASAP 2023 Best Paper Award — 2023
- Qualcomm Innovation Fellowship (QIF) Finalist — 2023
- International Symposium on Computer Architecture (ISCA) Student Travel Grant — 2023
- Rackham Conference Travel Grant — 2022, 2023
- 3rd Place in Low-Power Image Recognition Challenge (LPIRC 2018), Track2 — 2018
- 2nd Prize in Tsinghua Excellent Student Research Training (SRT) Project **(top 10%)** — 2018
- 3rd Prize in 35th Challenge cup of Tsinghua University **(top 15%)** — 2017
- Meritorious Winner in Mathematical Contest in Modeling, COMAP — 2017
- 1st Prize for the 32rd National Undergraduate Physics Olympic — 2016
- Scholarship for Scientific and Technological Innovation — 2016-18
- Scholarship for Academic Excellence — 2015-16