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, Co-advisors: Prof. Dennis Sylvester and Prof. David Blaauw — *Aug. 2019 – Fall 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

- **Nvidia** — Santa Clara, CA
  Research Intern, Nvidia Research — *May 2024 – Aug. 2024*
- **Qualcomm** — San Diego, CA
  Research Intern, Qualcomm Research Center — *May 2023 – Aug. 2023*

Selected Publication (google scholar)

- **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. *IEEE Journal of Solid-State Circuits (JSSC)*. IEEE, 2024.
- P. Abillama, Q. Zhang, **Z. Fan**, et al. A 22nm 9.51 TOPS/W Neural Engine with 2MB MRAM Leveraging Sparse-Orthogonal Walsh-Hadamard Transform Computations and Dynamic Power Gatings. *IEEE European Solid-State Circuits Conference (ESSCIRC), 2024.*
- Q. Zhang, **Z. Fan**, et al. RoboVisio: A Micro-Robot Vision Domain-Specific SoC for Autonomous Navigation Enabling Fully-on-Chip Intelligence via 2-MB eMRAM. *IEEE Journal of Solid-State Circuits (JSSC)*. IEEE, 2024.
- **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)
- 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.
- J-H. Seol, H. Yang, R. Rothe, **Z. Fan**, et al. A 1.5μW End-to-End Keyword Spotting SoC with Content-Adaptive Frame Sub-Sampling and Fast-Settling Analog Frontend. *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. Li, B. Li, **Z. Fan**, et al. RED: A ReRAM-based Efficient Accelerator for Deconvolutional Computation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. 2020.
- **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***, Z. Li*, B. Li*, et al. RED: A ReRAM-based Deconvolution Accelerator. *Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019.

Research

- **Diffusion Model Acceleration** | Algorithm, Architecture — Nvidia Research
  Mentor: Steve Dai Manager: Brucek Khailany — *May 2024 - Aug. 2024*
  - Quantized diffusion model to 4bit weight and 4bit activation with negligible image generation quality loss.
  - Explored time-step per-tensor activation sparsity in ReLU-based diffusion models.
  - Designed a sparsity aware diffusion model accelerator to accelerate diffusion model process.
  - Algorithm submitted to Nvidia internel conference NTECH 2024. Hardware proposal submitted to provisional patent.

- **GNN-based Cost Prediction on Compiled Graph** | Algorithm, Compiler — Qualcomm Research Center
  Mentors: Mahesh Balasubramanian and Apoorva Gokhale Manager: Rishi Chaturvedi — *May 2023 - Aug. 2023*
  - Used DNN compiler for compiling different DNN models and analyzed both high-level IR and low-level IR.
  - Ran compiled graphs and extracted instruction features and node-level cost as custom dataset.
  - Proposed a GNN-based cost prediction algorithm that used GNN pre-trained model and finetuning scheme.
  - Tested node attribute masking as pre-trained task and cost prediction as downstream task.
  - Proposed algorithm achieved $> 95\%$ cost prediction accuracy and more than $3\times$ speed-up than previous modeling.

- **Multi-task Processing Transformer Accelerator** | Algorithm, Architecture — University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim — *Mar. 2022 - Present*
  - Publications: ISCA 2023 (first author), ICCV 2023, ISSCC 2025 (first author, submitted)
  - Proposed an efficient Transformer inference algorithm by sharing both weight and activation from base task.
  - Proposed an efficient transfer learning algorithm for NLP and image/video applications
  - Designed a heterogeneous architecture for accelerating Transformer-based multi-task scenario.
  - Designed a task scheduling scheme to reduce off-chip memory access and increase computation utilization.
  - Algorithm achieved averagely 73% number of FLOPs reduction with negligible accuracy loss.
  - Built cycle-accurate simulator and energy simulator using Python.
  - Architecture achieved 1.48×-2.43× speed-up and 1.62×-3.77× higher efficiency than SOTA Transformer accelerators

- **Multi-modal Signal Processing Accelerator** | Architecture, Circuit — University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim — *Aug. 2019 - Feb. 2022*
  - Publications: VLSI-symp 2022 (first author), JSSC (first author), ISSCC 2023
  - Worked on the GAN-based audio compression algorithm, which achieved more than 64× compression rate for audio.
  - Worked on the audio-image/video cross modal verification algorithm.
  - Designed a $12mm^2$ ultra-low-power multi-modal signal processor SoC in TSMC 22nm technology and taped out.
  - Designed an architecture that integrates a versatile neural engine with audio and image processing accelerators.
  - Proposed MRAM dynamic power gating technology and MRAM-weight cache architecture.
  - Demonstrated MRAM's advantages in IoT applications regarding power consumption and energy efficiency.
  - Proposed a dedicated system power domain design, used Cadence Innovus CPF Flow to handle power domains.
  - The 0.25-3.84 mW SoC achieved 3-10 TOPS/W energy efficiency: 1.4 - 4.5× higher than SOTA DNN processors.

- **Millimeter-Scale Ultra-low-power Audio System** | Circuit, System — University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim — *Mar. 2022 - Present*
  - Helped build the test environment for millimeter-scale ultra-low-power audio system.
  - Tested the voice activity detection behavior in audio system.
  - Joint tested the audio system with aforementioned multi-modal signal processing SoC.

- **Transform Domain Neural Network Accelerator** | Architecture, Circuit — University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim — *Aug. 2021 - Mar. 2022*
  - Publications: ASAP 2023 (Best Paper Award)

- 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.
- Helped design a MRAM-logic asynchronous interface. Helped design a MRAM-based data pre-loading scheme.

- **Sign-Magnitude Neural Network Accelerator** | Architecture, Circuit — University of Michigan
  Co-advisors: Prof. Dennis Sylvester, Prof. David Blaauw and Prof. Hun-Seok Kim — *Oct. 2021 - Mar. 2022*
  - Publications: ISSCC 2023
  - Collaboratively designed a NN accelerator using bit-sparsified sign-magnitude multiplications and dual adder trees.
  - Helped tape out a 28nm test chip with both 2's complement NN accelerator and sign-magnitude NN accelerator.
  - System achieved 50% energy efficiency gain comparing to traditional 2's complement NN accelerator.

- **Analog Signal Processing SIFT Accelerator** | Circuit — Tsinghua University
  Advisor: Prof. Fei Qiao — *Jul. 2017 - Jul. 2018*
  - Publications: T-VLSI 2019 (first author)
  - Worked on the SIFT acceleration system using analog signal processing architecture.
  - Proposed an analog circuit network that realized the keypoint detection part of SIFT.
  - System simulation achieved 2.3k VGA frames/s, which is 3.26× faster than SOTA digital accelerators.

- **Near-sensor Current-based CIM Accelerator** | Circuit — Tsinghua University
  Advisor: Prof. Fei Qiao — *Oct. 2018 - Jun. 2019*
  - Publications: ISVLSI 2019 (first author)
  - Proposed a switched-current based low-power computer-in-memory vision system.
  - Designed current mode active pixel sensor and current mode compute-in-memory architecture.
  - System simulation showed 1.45mW power consumption and up to 28.25TOPS/W energy efficiency.

- **RRAM-based CIM Deconvolution Accelerator** | Architecture — Duke University
  Co-advisors: Prof. Yiran Chen and Prof. Hai Li — *Jul. 2018 - Sept. 2018*
  - Publications: DATE 2019 (first author), T-CAD 2020
  - Proposed an architecture for accelerating deconvolution using ReRAM-based compute-in-memory.
  - Proposed the RRAM pixel-wise mapping scheme and deconvolution zero-skipping data flow.
  - The proposed architecture achieved 3.69× $\sim$ 31.15× speed-up and up to 88.36% energy reduction on deconvolution.

Course Projects

- EECS 427 (VLSI Design I [A+]): A 16bit RISC Processor with a 4-bit Time-Domain Mixed-Signal MAC Processor
- EECS 627 (VSLI Design II [A+]): A 130nm Configurable Neural Engine Design for Deep Learning Applications
- EECS 470 (Computer Architecture [A]): A P6-style 2-way Superscalar Out-of-Order Processor
- EECS 413 (Monolithic Amplifier Circuits [A+]): A 5Gbps Wide-bandwidth Transimpedance Amplifier
- EECS 545 (Machine Learning [A]): Music Style Transfer based on Autoencoder
- EECS 504 (Foundations of Computer Vision [A]): Segmentation Guidance for 3D Object Detection

Academic Service

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

Skills

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

Recent Awards

- Best Paper Award in International Conference on Application-specific Systems, Architectures and Processors 2023 — 2023
- Qualcomm Innovation Fellowship (QIF) Finalist — 2023
- International Symposium on Computer Architecture (ISCA) Student Travel Grant — 2023
- Rackham Conference Travel Grant — 2022, 2023