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 Electrical Computer Engineering — *Aug. 2019 – Nov. 2024 (expected)*
- **University of Michigan** — Ann Arbor, MI
  Master of Engineering in Electrical Computer Engineering, GPA: 4.0/4.0 — *Aug. 2019 – Apr. 2022*
- **Tsinghua University** — Beijing, China
  Bachelor of Engineering in Electronics Engineering, GPA: 3.7/4.0 — *Aug. 2015 – July. 2019*

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.
- 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.
- P. Abillama, **Z. Fan**, et al. SONA: An Accelerator for Transform-Domain Neural Networks with Sparse Orthogonal Weights. *Submitted to International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 2023
- S. Shoouri, M. Yang, **Z. Fan**, et al. Efficient Computation Sharing for Multi-Task Visual Scene Understanding. *Submitted to International Conference on Computer Vision (ICCV)*. 2023

Research Experience

- **Michigan Integrated Circuit Lab (MICL),** University of Michigan — Ann Arbor, MI
  Co-advisors: Professor Dennis Sylvester and Professor David Blaauw — *Aug 2019 - Present*
  - [**Architecture**] **Multi-tasking NLP Accelerator**: Designed a heterogeneous architecture with dedicated dense and sparse cores for accelerating NLP multi-tasking operations. Proposed an energy-efficient task scheduling scheme to reduce off-chip memory access.

- [**Algorithm**] **Efficient Transfer Learning**: Proposed an efficient transfer learning algorithm for NLP and image/video applications. 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.
- [**Circuit**] **Multi-modal Signal Processing SoC**: Designed a 12mm^2 ultra-low-power multi-modal signal processor SoC in 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.

- **Nanoscale Integrated Circuits and Systems Lab (NICS),** Tsinghua University — Beijing, China
  Advisor: Professor Fei Qiao — *Aug 2017 - Jun 2019*
  - [**Circuit**] **Current-based PIM**: Worked on the switched-current based low-power processing-in-memory (PIM) 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.
  - [**Circuit**] **Analog Signal Processing SIFT Accelerator**: 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.

- **Computational Evolutionary Intelligence Lab (CEI),** Duke University — Durham, NC
  Co-advisor: Professor Yiran Chen and Professor Hai Li — *Jul 2018 - Sept 2018*
  - [**Architecture**] **ReRAM-based Deconvolution accelerator**: Worked on ReRAM-based 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× $\sim$ 31.15×.

Skills

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

Awards

- Qualcomm Innovation Fellowship (QIF) Finalist — 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