Yanyue Xie

xie.yany@northeastern.edu ◇ https://yanyuexie.com

RESEARCH INTERESTS

- Efficient deep learning (quantization, pruning)
- Algorithm/hardware co-design
- FPGA acceleration for deep learning
- Electronic design automation

EDUCATION

Northeastern University, Boston, MA, USA *Sept. 2020 – Present*
Ph.D. student, Department of Electrical and Computer Engineering
Advisors: Prof. Xue (Shelley) Lin and Prof. Yanzhi Wang
GPA 3.97/4.0

Fudan University, Shanghai, China *Sept. 2016 – June 2020*
B.E. in Microelectronic Science and Engineering
GPA 3.60/4.0

Nanyang Technological University, Singapore *Jan. 2019 – May 2019*
Exchange student, School of Electrical and Electronic Engineering
GPA 4.83/5.0

PROFESSIONAL EXPERIENCE

Futurewei Technologies Inc., CA, USA *May. 2023 – Sept. 2023*
Research Intern
Efficient learned image compression model

PUBLICATIONS

[C12] **Yanyue Xie***, Peiyan Dong*, Geng Yuan, Zhengang Li, Masoud Zabihi, Chao Wu, Sung-En Chang, Xufeng Zhang, Xue Lin, Caiwen Ding, Nobuyuki Yoshikawa, Olivia Chen, and Yanzhi Wang, "SuperFlow: A Fully-Customized RTL-to-GDS Design Automation Flow for Adiabatic Quantum-Flux-Parametron Superconducting Circuits", *2024 Design, Automation & Test in Europe Conference (DATE 2024) (*Equal contributions).*

[C11] Zhengang Li, Geng Yuan, Tomoharu Yamauchi, Zabihi Masoud, **Yanyue Xie**, Peiyan Dong, Xulong Tang, Nobuyuki Yoshikawa, Devesh Tiwari, Yanzhi Wang, and Olivia Chen, "SupeRBNN: Randomized Binary Neural Network Using Adiabatic Superconductor Josephson Devices", *56th IEEE/ACM International Symposium on Microarchitecture (MICRO 2023).*

[C10] Dana Diaconu, **Yanyue Xie**, Mehmet Gungor, Suranga Handagala, Xue Lin, and Miriam Leeser, "Machine Learning Across Network-Connected FPGAs", *2023 IEEE High Performance Extreme Computing Conference (HPEC 2023).*

[C9] Zhengang Li*, **Yanyue Xie***, Peiyan Dong*, Olivia Chen, and Yanzhi Wang, "Algorithm-Software-Hardware Co-Design for Deep Learning Acceleration", *60th ACM/IEEE Design Automation Conference (DAC 2023) (*Equal contributions).*

[C8] Sung-En Chang, Geng Yuan, Alec Lu, Mengshu Sun, Yanyu Li, Xiaolong Ma, Zhengang Li, **Yanyue Xie**, Minghai Qin, Xue Lin, Zhenman Fang, and Yanzhi Wang, "ESRU: Extremely Low-Bit and Hardware-Efficient Stochastic Rounding Unit Design for 8-Bit DNN Training", *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE 2023).*

[C7] Zhenglun Kong, Haoyu Ma, Geng Yuan, Mengshu Sun, **Yanyue Xie**, Peiyan Dong, Xin Meng, Xuan Shen, Hao Tang, Minghai Qin, Tianlong Chen, Xiaolong Ma, Xiaohui Xie, Zhangyang Wang, and Yanzhi Wang, "Peeling the Onion: Hierarchical Reduction of Data Redundancy for Efficient Vision Transformer Training", *37th AAAI Conference on Artificial Intelligence (AAAI 2023).*

[C6] Peiyan Dong, Mengshu Sun, Alec Lu, **Yanyue Xie**, Kenneth Liu, Zhenglun Kong, Xin Meng, Zhengang Li, Xue Lin, Zhenman Fang, and Yanzhi Wang, "HeatViT: Hardware-efficient adaptive token pruning for vision transformers", *29th IEEE International Symposium on High-Performance Computer Architecture (HPCA 2023)*.

[C5] Geng Yuan, Sung-En Chang, Qing Jin, Alec Lu, Yanyu Li, Yushu Wu, Zhenglun Kong, **Yanyue Xie**, Peiyan Dong, Minghai Qin, Xiaolong Ma, Xulong Tang, Zhenman Fang, and Yanzhi Wang, "You Already Have It: A Generator-Free Low-Precision DNN Training Framework using Stochastic Rounding", *European Conference on Computer Vision (ECCV 2022)*.

[C4] Zhengang Li, Mengshu Sun, Alec Lu, Haoyu Ma, Geng Yuan, **Yanyue Xie**, Hao Tang, Yanyu Li, Miriam Leeser, Zhangyang Wang, Xue Lin, and Zhenman Fang, "Auto-ViT-Acc: An FPGA-Aware Automatic Acceleration Framework for Vision Transformer with Mixed-Scheme Quantization", *32nd International Conference on Field-Programmable Logic and Applications (FPL 2022)*, 2022.

[C3] Peiyan Dong*, **Yanyue Xie***, Hongjia Li*, Mengshu Sun, Olivia Chen, Nobuyuki Yoshikawa, and Yanzhi Wang, "TAAS: A Timing-Aware Analytical Strategy for AQFP-Capable Placement Automation", *59th ACM/IEEE Design Automation Conference (DAC 2022)*, San Francisco, Jul 10-14, 2022 (*Equal contributions).

[C2] Zhifeng Lin, **Yanyue Xie**, Gang Qian, Jianli Chen, Sifei Wang, Jun Yu, and Yao-Wen Chang, "Timing-Driven Placement for FPGAs with Heterogeneous Architectures and Clock Constraints", *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE 2021)*, pp. 1564-1569, 2021.

[C1] Zhifeng Lin, **Yanyue Xie**, Gang Qian, Sifei Wang, Jun Yu, and Jianli Chen, "Late Breaking Results: An Analytical Timing-Driven Placer for Heterogeneous FPGAs", *57th ACM/IEEE Design Automation Conference (DAC 2020)*, pp. 1-2, San Francisco, Jul 20-24, 2020.

[J2] Jianli Chen, Zhifeng Lin, **Yanyue Xie**, Wenxing Zhu, and Yao-Wen Chang, "Mixed-Cell-Height Placement with Complex Minimum-Implant-Area Constraints", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 11, pp. 4639-4652, Nov. 2022, doi: 10.1109/TCAD.2021.3133855.

[J1] Zhifeng Lin, **Yanyue Xie**, Peng Zou, Sifei Wang, Jun Yu, and Jianli Chen, "An Incremental Placement Flow for Advanced FPGAs with Timing Awareness", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 9, pp. 3092-3103, Sept. 2022, doi: 10.1109/TCAD.2021.3120070.

[P1] Masoud Zabihi, **Yanyue Xie**, Zhengang Li, Peiyan Dong, Geng Yuan, Olivia Chen, Massoud Pedram, Yanzhi Wang, "A Life-Cycle Energy and Inventory Analysis of Adiabatic Quantum-Flux-Parametron Circuits", *arXiv preprint arXiv:2307.12216.*

RESEARCH EXPERIENCES

Accelerating Sparse Neural Networks on FPGA Using Data-Flow Architecture *Jan. 2021 – Present*
Research Assistant
Advisor: Prof. Xue (Shelley) Lin and Prof. Miriam Leeser, Northeastern University

- Pruned convolutional neural networks by applying the ADMM-based pruning algorithm and trained neural networks using quantization-aware training,
- Modified a data-flow architecture to make the compressed neural networks suitable for inference on FPGAs [C10],
- Efficient implementation of vision transformer models on FPGAs including mixed-scheme quantization [C4] and token pruning [C6], [C7], and stable diffusion UNet model on FPGAs.

Timing-Driven Placement for AQFP Superconducting Circuits *Feb. 2021 – Present*
Research Assistant
Advisor: Prof. Yanzhi Wang, Northeastern University

- Proposed a timing model for the deep-pipelined, four-phase AQFP superconducting circuits,
- Integrated timing cost into the objective function of an analytical placer and improved the maximum operating frequency by 19.17% on average compared with previous methods [C3], [C12],
- Efficient implementation of binarized neural networks on AQFP-based crossbar synapse array [C9], [C11].

Timing-Driven Placement Algorithm for FPGAs *Sept. 2019 – May 2021*
Research Assistant
Advisor: Prof. Jianli Chen, School of Microelectronics, Fudan University

- Proposed a timing model for xc7k325t device, validated using Vivado, and integrated the model into a timing-driven placer [C1], [C2], [J1].

AWARDS

DAC Young Fellow	Design Automation Conference	2021
First Prize Scholarship	Fudan University	2020

SKILLS

Programming Languages
C/C++, Python, Verilog, MATLAB, LaTeX

EDA Tools: Xilinx Vitis/Vivado Design Suite, Synopsys Design Compiler, Modelsim
Deep Learning Frameworks: PyTorch, TensorFlow, ONNX