Bingyi Zhang

Email: bingyizh@usc.edu, bingyizh233@gmail.com
(571)-443-9187
Los Angeles, CA 90007
Person website: https://zjjzby.github.io/

EDUCATION

University of Southern California (USC) — Los Angeles, USA
Ph.D. of Computer Engineering, Ming Hsieh Dept of ECE — *Sept. 2019 - Sept. 2024 (expected)*
GPA: 4.00/4.00, Ph.D. Advisor: Professor Viktor Prasanna
Award: Annenberg Fellowship

Fudan University — Shanghai, China
Master of Engineering, Integrated circuit engineering — *Sept. 2017 - Jun. 2019*
GPA: 3.70/4.00
Award: 2018 China National Scholarship for Graduate Student (Top 1 of Dept)

Fudan University — Shanghai, China
B.Eng., Microelectronic science and engineering — *Sept. 2013 - Jun. 2017*
GPA: 3.52/4.00

RESEARCH FIELD

High performance computing (HPC), Computer architecture, Reconfigurable computing, Graph-based machine learning, Very large-scale integrated circuit (VLSI)

SELECTED PUBLICATION (Google Scholar)

- **Bingyi Zhang**, Hanqing Zeng, Viktor Prasanna, GraphAGILE: An FPGA-based Overlay Accelerator for Low-latency GNN Inference, Transactions on Parallel and Distributed Systems (TPDS).
- **Bingyi Zhang**, Jun Han, Zhize Huang, Jianwei Yang, Xiaoyang Zeng, "A Realtime and Hardware-efficient Processor for Skeleton-based Action Recognition with Lightweight Convolutional Neural Network", IEEE Transaction on Circuits and Systems II: Express Briefs
- **Bingyi Zhang**, Viktor Prasanna, Dynasparse: Accelerating GNN Inference through Dynamic Sparsity Exploitation, 37th IEEE International parallel and distributed processing symposium, (IPDPS 2023).
- **Bingyi Zhang**, Hanqing Zeng, Viktor Prasanna, Low-latency Mini-batch GNN Inference on CPU-FPGA Heterogeneous Platform, 29th IEEE international conference on high performance computing, data, & analytics (HiPC 2022)
- **Bingyi Zhang**, Rajgopal Kannan, Viktor Prasanna, Carl Busart, Accurate, Low-latency, Efficient SAR Automatic Target Recognition on FPGA, International Conference on Field Programmable Logic and Applications (FPL), 2022.
- Yi Chien Lin, **Bingyi Zhang (Co-first author)**, Viktor Prasanna, "HP-GNN: Generating High Throughput GNN Training Implementation on CPU-FPGA Heterogeneous Platform", The 30th ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2022).
- Hongkuan Zhou, **Bingyi Zhang (Co-first author)**, Rajgopal Kannan, Viktor Prasanna, Carl Busart, Model-Architecture Co-Design for High Performance Temporal GNN Inference on FPGA, The 36th IEEE Internatioal Parallel and Distributed Processing Symposium. (IPDPS 2022).
- **Bingyi, Zhang**; Kannan, Rajgopal ; Prasanna, Viktor. BoostGCN: A Framework for Optimizing GCN Inference on FPGA. FCCM, 2021.

PROFESSIONAL SKILLS

Expertise: Parallel and distributed computing; FPGA design; Machine learning; Algorithm design and Analysis
Programming language: C, C++, Python, Matlab, Perl
Parallel programming library: Pthread, OpenMP, Nvidia CUDA, OpenCL
Hardware design language: AMD Xilinx High-level Synthesis (HLS), Intel oneAPI, Verilog.
Machine learning library: TensorFlow, PyTorch, PyTorch Geometric.

SELECTED PROJECTS

University of Southern California, FPGA/Parallel Computing Lab *Sept. 2021 - Sept. 2023*

Topic: Model-architecture codesign for high-performance and energy-efficient SAR ATR on FPGA [paper]

- Designed a high-accuracy and low-complexity GNN model for Synthetic Aperture Radar (SAR) Automatic Target Recognition (ATR). Achieved an impressive accuracy rate of 99.38% on the MSTAR dataset..
- Developed parameterized hardware templates. Achieved an impressive throughput of up to 27,014 images per second for the proposed GNN architecture

University of Southern California, FPGA/Parallel Computing Lab *Sept. 2019 - Sept. 2022*

Topic: GraphAGILE: An FPGA-based Overlay Accelerator for Low-latency GNN Inference [paper]

- Designed an instruction set architecture to support various GNN models, including GCN, GIN, GraphSAGE, etc.
- Developed a compiler to translate GNN models into the designed instruction set architecture
- Achieves up to 47.1× and 3.9× speedup compared with state-of-the-art CPU and GPU platforms.

State Key Laboratory of ASIC and System, Fudan University *Aug. 2017 - May. 2018*

Topic: Model-architecture codesign for real-time skeleton-based human action recognition (HAR) [paper]

- Designed an lightweight 1D-CNN to perform real-time HAR, achieving 20x acceleration compared with the state-of-the-art. Quantize and prune the 1D-CNN for hardware implementation.
- Designed a processor to perform the proposed algorithm, achieving low-power and hardware-efficient inference.

State Key Laboratory of ASIC and System, Fudan University *Sep. 2016 - June. 2017*

Topic: Visual object tracking on embedded FPGA platform [paper]

- Designed a siamese neural network for object tracking, which is computation-saving and storage-saving.
- Designed a hardware architecture on FPGA and achieve real-time (18.6 frame/s) and robust tracking with power consumption 1.284W, outperforming the state-of-the-art.

ACADEMIC COMPETITION AND AWARD

1st place at 36th IEEE International Parallel Distributed Processing Symposium, Ph.D. Forum [Link] *May. 2023*

Topic: Software and Hardware codesign for High Performance Graph Neural Network Inference

Outstanding Student Paper Award at 2021 IEEE High Performance Extreme Computing Virtual Conference [Link] *Sept. 2021*

Topic: Efficient Neighbor-Sampling-based GNN Training on CPU-FPGA Heterogeneous Platform

Third National Undergraduate Integrated Circuit Innovation and Entrepreneurship Competition of China [Link] *Feb. 2019*

Topic: Hardware acceleration for convolutional neural networks. Result: Second price out of 500+ teams

First National Undergraduate Integrated Circuit Innovation and Entrepreneurship Competition of China [Link] *Feb. 2017*

Topic: Design neural network accelerator. Result: Rank 3rd out of 150+ teams