

Yifan Yang

Tel: +1 617-909-9768, Email: yifany@csail.mit.edu, <https://yang-yifan.github.io/>

RESEARCH INTERESTS

Computer Architecture, Accelerator for Sparse and Irregular Applications (transformer, CNN, tensor algebra, graph analytics), HW/SW Co-design, Domain-specific Accelerator, FPGA

EDUCATION

Massachusetts Institute of Technology, EECS

Cambridge, MA

Ph.D. in Electrical Engineering and Computer Science

September 2019 - May 2024 (expected)

S.M. in Electrical Engineering and Computer Science

September 2019 - June 2021

- GPA: 5.0/5.0
- Thesis advisor: Prof. Daniel Sanchez and Prof. Joel Emer
- Thesis title: Architectural Support for Efficient Processing of Sparse and Irregular Applications
- Coursework: Computer System Architecture, Hardware Architecture for Deep Learning, TinyML and Efficient Deep Learning Computing, Advances in Computer Vision, Theory of Computation

Tsinghua University, Department of Physics

Beijing, China

B.S. in Mathematics and Physics

August 2015 - July 2019

- GPA: 3.90/4.00
- Thesis advisor: Prof. Leibo Liu
- Thesis title: A Message Passing-based Graph Processing Framework on CPU-FPGA Heterogeneous System
- Selected coursework: Fundamental of Digital Logic and Processor, Digital Integrated Circuit Analysis and Design (English), Computer Organization and Architecture, Operating System, Principles and Practice of Compiler Construction, Distributed Computing, Principles of Analog Circuits, Data Structure, Introduction to Artificial Intelligence, Autonomous Driving

RESEARCH EXPERIENCE

Graduate Research Assistant

September 2019 - present

MIT Computer Science and Artificial Intelligence Laboratory

Cambridge, MA

- Advisor: Prof. Daniel Sanchez and Prof. Joel Emer
- (Ongoing) A unified accelerator to accelerate matrix multiplication at all sparsity levels
 - Conduct performance analysis (C/PyTorch) of matrix multiplication on a broad set of application domains and sparsity levels (transformer, CNN, tensor algebra, graph analytics scientific computing, etc.)
 - Codesign dataflow and hardware architecture to efficiently support dense, mildly sparse, and highly sparse matrix multiplication with modest area overhead (evaluated using Verilog)
- Worked on leveraging inter-layer pipelining to accelerate sparse CNNs
 - Proposed input-stationary output-stationary (IS-OS) dataflow that minimizes inter-layer storage requirement
 - Designed the **ISOSceles** accelerator that implements IS-OS dataflow to efficiently pipeline multiple sparse layers in CNNs
 - Conducted end-to-end C++ simulation evaluation of ISOSceles on several sparse CNNs and achieved large speedup and off-chip traffic reduction over state-of-the-art accelerators
- Worked on offering architectural support to accelerate irregular applications using data compression

Designed **SpZip**, specialized hardware support for traversing and generating compressed data structures in irregular applications

Proposed the Dataflow Configuration Language (DCL) for the SpZip programmable hardware

Evaluated SpZip using zsim simulation on a broad set of irregular applications, results show large performance gains and data movement reductions over prior systems

Undergraduate Research Assistant

Research Center for Mobile Computing, Tsinghua University

March 2017 - August 2019

Beijing, China

- Advisor: Prof. Leibo Liu
- Worked on designing graph processing framework across the hardware software interface

Proposed the asynchronous Block Coordinate Descent (BCD) view of iterative graph algorithms to reveal the algorithmic and system trade-offs on achieving fast algorithm convergence

Designed **GraphABCD**, an asynchronous graph processing framework on heterogeneous system

Prototyped GraphABCD whole system in Verilog on Intel HARPv2 CPU-FPGA heterogeneous platform

Undergraduate Research Assistant

Berkeley Artificial Intelligence Research Lab (BAIR), UC Berkeley

July 2018 - September 2018

Berkeley, CA

- Advisor: Prof. Kurt Keutzer, collaboration with Kees Vissers and Michaela Blott from Xilinx Research Labs
- Worked on co-designing Convolutional Neural Network and its hardware accelerator

Designed **DiracDeltaNet** using PyTorch, a hardware efficient neural network targeting image classification task

Quantized DiracDeltaNet down to 4bit weight and 4bit activations with minor accuracy loss

Designed and implemented **Synetgy**: the hardware accelerator for DiracDeltaNet on embedded FPGA using HLS

PUBLICATIONS

1. **Yifan Yang**, Joel S. Emer, and Daniel Sanchez, "Trapezoid: A Versatile Accelerator for Dense and Sparse Matrix Multiplications", submitted to ISCA 2024.
2. **Yifan Yang**, Joel S. Emer, and Daniel Sanchez, "[ISOSceles: Accelerating Sparse CNNs through Inter-Layer Pipelining](#)", in Proceedings of the 29th international symposium on High Performance Computer Architecture (HPCA-29), 2023.
3. **Yifan Yang**, Joel S. Emer, and Daniel Sanchez, "[SpZip: Architectural Support for Effective Data Compression In Irregular Applications](#)", in Proceedings of the 48th annual International Symposium on Computer Architecture (ISCA-48), 2021.
4. **Yifan Yang**, Zhaoshi Li, Yangdong Deng, Zhiwei Liu, Shouyi Yin, Shaojun Wei, and Leibo Liu, "[GraphABCD: Scaling Out Graph Analytics with Asynchronous Block Coordinate Descent](#)", in Proceedings of the 47th annual International Symposium on Computer Architecture (ISCA-47), 2020.
5. **Yifan Yang**, Qijing Huang, Bichen Wu, Tianjun Zhang, Liang Ma, Giulio Gambardella, Michaela Blott, Luciano Lavagno, Kees Vissers, John Wawrzynek, and Kurt Keutzer, "[Synetgy: Algorithm-hardware Co-design for ConvNet Accelerators on Embedded FPGAs](#)", in Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2019.

INDUSTRY & TEACHING EXPERIENCE

Platform Architecture Intern

Platform Architecture, Apple

- CPU cache subsystem performance research

May 2022 - August 2022
Cupertino, CA

Graduate Teaching Assistant

6.812/6.825 *Hardware Architecture for Deep Learning*, MIT

- Held recitations and office hours
- Developed and graded lab assignments and paper review sessions
- Mentored and graded final design projects

January 2022 - May 2022
Cambridge, MA

HONORS & AWARDS

- MIT Jacobs Presidential Fellowship Cambridge, 2019
- Chi-sun Yeh Nominee Prize (8%) Beijing, 2019
- CNPC Scholarship (5%) Beijing, 2018
- National Scholarship (1/91) Beijing, 2017
- Outstanding Award in the 35th Tsinghua 'Challenge Cup' (Top 6/381)
(Student Extracurricular Academic Science and Technology Works Competition) Beijing, 2017
- Zhang Mingwei Scholarship (10%) Beijing, 2016
- Bronze Medal in the 31st China Physics Olympiad Hangzhou, 2014

TECHNICAL SKILLS

Programming: C/C++, Python, Verilog, System Verilog, Vivado HLS, Unix/Linux, Golang, R, MATLAB, Assembly

Tools: Pytorch, Intel Pin, zsim, Intel(Altera) toolchain (Quartus, Qsys), Xilinx toolchain (Vivado, Vivado HLS, PYNQ), Modelsim, VCS

Updated on December 5, 2023