Yifan Yang

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

INTERESTS & EXPERTISE

GPU, Deep Learning Inference, Low Latency Inference, Low latency kernels, LLM, Computer Architecture, Accelerator for Sparse and Irregular Applications, 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*

- GPA: 5.0/5.0
- Thesis advisor: Prof. Daniel Sanchez and Prof. Joel Emer
- Thesis title: Effective and Flexible Acceleration of Sparse Computations

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

INDUSTRY & TEACHING EXPERIENCE

Senior Deep Learning Architect *June 2024 - present*
Deep Learning Inference Architecture, Nvidia *Santa Clara, CA*

- Compute and memory locality optimization and programming model exploration of Blackwell/Rubin and future GPU architecture (e.g. MPS Memory Locality Optimized Partitions)
- Low latency inference optimized Blackwell kernel authoring (CuTe/Triton)
  - GEMM and batched GEMM (open sourced in FlashInfer)
  - Flash attention/decoding (will be open sourced soon)
  - (WIP) MoE layer
- TRT-LLM Llama/DeepSeek R1/GPT-OSS low latency inference optimization
- Prototyping/Exploring speculative decoding techniques for popular LLM models (e.g. DeepSeek MTP)
- Memory layout optimization for improved kernel DRAM bandwidth efficiency (In production TRTLLM-gen kernels)
- Performance exploration of using prefetching for low latency LLM inference
- Hardware feature exploration for low latency LLM inference in future GPU architectures

- Developing tools for better silicon performance observability of low latency LLM workloads on GPUs

Platform Architecture Intern *May 2022 - August 2022*
Platform Architecture, Apple *Cupertino, CA*
- CPU cache subsystem performance research

Graduate Teaching Assistant *January 2022 - May 2022*
6.812/6.825 Hardware Architecture for Deep Learning, MIT *Cambridge, MA*

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

RESEARCH EXPERIENCE

Graduate Research Assistant *September 2019 - May 2024*
MIT Computer Science and Artificial Intelligence Laboratory *Cambridge, MA*
- Advisor: Prof. Daniel Sanchez and Prof. Joel Emer
- Worked on 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)

  Evaluated the accelerator on a wide range of sparse levels and application domains, results show large perf/area gains over state-of-the-art accelerators

- 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 *March 2017 - August 2019*
Research Center for Mobile Computing, Tsinghua University *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 *July 2018 - September 2018*
Berkeley Artificial Intelligence Research Lab (BAIR), UC Berkeley *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. Axel Feldmann, Courtney Golden, **Yifan Yang**, Joel S. Emer, and Daniel Sanchez, "Azul: An Accelerator for Sparse Iterative Solvers Leveraging Distributed On-Chip Memory", in Proceedings of the 57th annual international symposium on Microarchitecture (MICRO-57), 2024.
2. **Yifan Yang**, Joel S. Emer, and Daniel Sanchez, "Trapezoid: A Versatile Accelerator for Dense and Sparse Matrix Multiplications", in Proceedings of the 51th annual International Symposium on Computer Architecture (ISCA-51), 2024.
3. **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.
4. **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.
5. **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.
6. **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.

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 35^{th} Tsinghua 'Challenge Cup' (Top 6/381) *Beijing, 2017*
  (Student Extracurricular Academic Science and Technology Works Competition)
- Zhang Mingwei Scholarship (10%) *Beijing, 2016*
- Bronze Medal in the 31^{st} China Physics Olympiad *Hangzhou, 2014*

TECHNICAL SKILLS

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

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

Updated on December 19, 2025