Genghan Zhang

zhang677.github.io | ghzhang19@gmail.com

Education

Stanford University September 2023 - Present
PhD Student in Computer Science *Stanford, USA*

- Research Interests: Domain-specific compiler and computer architecture

Tsinghua University August 2019 - June 2023
Bachelor of Engineer in Electronic Information Science and Technology *Beijing, China*

- GPA: 3.94/4.00 (Top 3%)

Research Experience

Research Assistant September 2023 - Present
Department of Computer Science, Stanford University *Stanford, CA*

- Advisor: Prof. Kunle Olukotun
- Designed a race-free protocol for reclaiming buffers with shared references for reconfigurable dataflow architecture.
- Accepted by 6th Young Architect Workshop (in conjunction with ASPLOS 2024).

Research Assistant January 2024 - March 2024
Department of Computer Science, Stanford University *Stanford, CA*

- Advisor: Prof. Azalia Mirhoseini
- Proposed GPU kernel fusion techniques to accelerate FFN layers for LLM inference by utilizing the sparsity of activation.
- Submitted to Conference on Language Modeling (COLM) 2024

Research Assistant March 2022 - August 2023
Department of Computer Science, Stanford University *Remote*

- Advisor: Prof. Fredrik Kjølstad
- Designed an algorithm template and code generation algorithm for *sparse workspace* to solve the sparse scattering problem with a sparse tensor algebra compiler called TACO.
- Accepted by Programming Language Design and Implementation (PLDI) 2024.

Undergraduate Research Assistant January 2022 - October 2022
Innovative Data-centric Efficient Architecture Lab (IDEAL), Tsinghua University *Beijing, China*

- Advisor: Prof. Mingyu Gao
- Proposed a novel paradigm, *Kernel Architecture Search (KAS)* that automatically designs efficient neural network layers with system budgets as first-priority constraints. Implemented a system called Canvas to examine the idea.
- Achieved on average 1.5× speedups than previous state-of-the-arts with acceptable accuracy loss.

Undergraduate Research Assistant August 2021 - July 2022
Nanoscale Integrated Circuits and Systems Lab (NICS), Tsinghua University *Beijing, China*

- Advisors: Prof. Yu Wang, Prof. Guohao Dai (SJTU) and Prof. Sitao Huang (UC Irvine)
- Proposed *atomic parallelism*, a new optimization space for sparse-dense hybrid algebra and implemented it to a high performance sparse kernel CUDA library called dgSPARSE, achieving on average 1.6× ∼ 2.3× speedup.
- Proposed *segment group*, a new abstraction for sparse compilation theory based on atomic parallelism and implemented it to TACO compiler, achieving up to 1.2× speedup.
- Accepted by CCF Transactions on High Performance Computing.

Work Experience

Part-time Research Assistant October 2022 - November 2022
HPC-AI Tech *Beijing, China*

- Mentor: Prof. Yang You (NUS)
- Developed novel automatic parallelization techniques for gaint deep learning models.

Part-time Research Assistant May 2023 - July 2023
Infinigence Tech *Beijing, China*

- Mentor: Prof. Xiuhong Li (PKU)
- Assembled an in-house GPU kernel library for LLM inference which demos the company's first-generation product.
- Optimized fused attention on GPU with collaborators.

Publications

- **Sgap: Towards Efficient Sparse Tensor Algebra Compilation for GPU**
  Genghan Zhang, Yuetong Zhao, Yanting Tao, Zhongming Yu, Guohao Dai, Sitao Huang, Yuan Wen, Pavlos Petoumenos, Yu Wang.
  CCF Transactions on High Performance Computing, 2023
- **HyperGef: A Framework Enabling Efficient Fusion for Hypergraph Neural Network on GPUs**
  Zhongming Yu, Guohao Dai, Shang Yang, **Genghan Zhang**, Hengrui Zhang, Feiwen Zhu, June Yang, Jishen Zhao, Yu Wang.
  Proceedings of Machine Learning and Systems, 2023
- **Canvas: End-to-End Kernel Architecture Search in Neural Networks**
  Chenggang Zhao, **Genghan Zhang**, Mingyu Gao.
  arXiv preprint, 2023

Honors and Awards

Awards in Tsinghua University

- Tsinghua University Academic Excellence Award 2020, 2021
- Tsinghua University Comprehensive Excellence Award 2022

Technical Skills

Programming Languages & Software Tools

- Most experienced: CUDA, Python, Matlab, PyTorch
- Some experience: Verilog HDL, Rust, LtSpice, Cadence