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 — April 2024 - Present
Department of Computer Science, Stanford University — *Stanford, CA*
- Advisor: Prof. Kunle Olukotun
- Working on using LLM agents to automate high-performance library development.
- Designed a race-free protocol for reclaiming buffers with shared references for reconfigurable dataflow architecture. Presented at 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. Accepted by COLM 2024

Research Assistant — March 2022 - December 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 PLDI 2024.

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 *segment group*, a new abstraction for sparse compilation theory based on atomic parallelism. Accepted by CCF Transactions on High Performance Computing.

Selected Publications

- **Compilation of Modular and General Sparse Workspaces**
  Genghan Zhang, Olivia Hsu, Fredrik Kjolstad.
  Programming Language Design and Implementation (PLDI), 2024

- **CATS: Context-Aware Thresholding for Sparsity in Large Language Models**
  Donghyun Lee, Jaeyong Lee, **Genghan Zhang**, Mo Tiwari, Azalia Mirhoseini.
  Conference on Language Modeling (COLM), 2024

- **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.
  Machine Learning and Systems (MLSys), 2023

Service

- Reviewer: ICML 2025, ICLR 2025, NeurIPS 2024, ICLR 2025 DL4C Workshop, NeurIPS 2024 Sys2-Reasoning Workshop, NeurIPS 2022 GLFrontiers Workshop
- Artifact Evaluation Committee: ASPLOS 2025 summer, PLDI 2025
- Program Committee: LATTE 2025

Work Experience

Software Engineer — June 2024 - September 2024
NVIDIA — *Santa Clara, USA*
- Mentor: Andrew Kerr
- Compiler for deep learning libraries

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.

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.

Honors and Awards

Awards in Tsinghua University
- Academic Excellence Award 2020, 2021, Excellence Award 2022