

Shiyang Li

✉ li004074@umn.edu 📍 Minneapolis, MN, USA 🌐 github.com/NKU-Yang

🌐 [NKU-Yang.github.io](https://github.com/NKU-Yang) 🎓 [Google Scholar](https://scholar.google.com/citations?user=0000-0003-3046-0414) 🆔 0000-0003-3046-0414

Interests: GPU Programming, Computer Architecture, Agentic MLSys, EDA, LLM Optimization

Skills: C/C++, CUDA, Python, Rust, Linux, Git

🎓 Education

- **University of Minnesota-Twin Cities** SEPT.2025 – Present
Ph.D. in Computer Science
Advisors: Prof. Caiwen Ding & Prof. Pen-Chung Yew
- **Nankai University Master of Engineering** SEPT.2022 – JUNE.2025
Major in Computer Technology
Advisor: Prof. Gong Xiaoli
- **Nankai University Bachelor of Engineering** SEPT.2018 – JUNE.2022
Major in Internet of Things Engineering

📖 Publications

- Yuebo Luo, **Li Shiyang**, Yifei Feng, Vishal Kancharla, Shaoyi Huang, Caiwen Ding,
“**GSR-GNN: Training Acceleration and Memory-Saving Framework of Deep GNNs on Circuit Graph**”
In Proceedings of the 63rd ACM/IEEE Design Automation Conference (DAC '26), June 2026, San Francisco, USA.
- **Li Shiyang**, Zijian Zhang, Winson Chen, Yuebo Luo, Mingyi Hong, Caiwen Ding,
“**StitchCUDA: An Automated Multi-Agents End-to-End GPU Programming Framework with Rubric-based Agentic Reinforcement Learning**”
International Conference on Machine Learning (ICML 2026), accepted. arXiv:2603.02637.
- **Li Shiyang**, Guangyan Sun, Jinwei Tang, Yanzhi Wang, Mingyi Hong, Caiwen Ding,
“**SparseDitto: Customizing GPU Kernels for Different Sparsity Patterns with LLM-Based Agentic System**”
arXiv preprint arXiv:2608.05033, Aug. 2026.
- **Li Shiyang**, Haoyang Chen, Mattia Fazzini, Caiwen Ding,
“**CUDABeaver: Benchmarking LLM-Based Automated CUDA Debugging**”
arXiv preprint arXiv:2605.08455, May 2026.
- **Li Shiyang**, Zijian Zhang, Guangyan Sun, Yuebo Luo, Winson Chen, Yanzhi Wang, Mingyi Hong, Caiwen Ding,
“**CUDAHercules: Benchmarking Hardware-Aware Expert-level CUDA Optimization for LLMs**”
arXiv preprint arXiv:2605.08467, May 2026.
- Ben Mechels, Ryan Billmeyer, Alexander Chen, **Li Shiyang**, Caiwen Ding,
“**FPTC: A Fast Parallel Transform-based Codec for Efficient Asymmetric Signal Compression**”
arXiv preprint arXiv:2605.01086, May 2026.
- Zijian Zhang, Rong Wang, **Li Shiyang**, Yuebo Luo, Mingyi Hong, Caiwen Ding,
“**CudaForge: An Agent Framework with Hardware Feedback for CUDA Kernel Optimization**”
arXiv preprint arXiv:2511.01884, Oct. 2025.
- Yuebo Luo, **Li Shiyang**, Junran Tao, Kiran Gautam Thorat, Xi Xie, Hongwu Peng, Nuo Xu, Caiwen Ding, Shaoyi Huang,
“**DR-CircuitGNN: Training Acceleration of Heterogeneous Circuit Graph Neural Network on GPUs**”
In Proceedings of the 39th ACM International Conference on Supercomputing (ICS '25), June 2025, Salt Lake City, USA.
- **Li Shiyang**, Tang Ruiqi, Zhu Jingyu, Zhao Ziyi, Gong Xiaoli, Wang Wenwen, Zhang Jin, and Yew Pen-Chung,
“**Liberator: A Data Reuse Framework for Out-of-Memory Graph Computing on GPUs**”
IEEE Transactions on Parallel and Distributed Systems, June 2023, doi:10.1109/TPDS.2023.3268662.
- **Li Shiyang**, Zhu J., Han Jiaxun, Peng Yuting, Wang Zhuoran, Gong Xiaoli, Wang Gang, Zhang Jin, Wang X.,
“**OneGraph: a cross-architecture framework for large-scale graph computing on GPUs based on oneAPI**”
CCF Transactions on High Performance Computing, Nov. 2023, doi:10.1007/s42514-023-00172-w.

🔬 Research Experience

- **CUDA Agentic Systems for GPU Programming and Optimization** Advisor: Prof. Caiwen Ding
Related publications: StitchCUDA, SparseDitto, CUDAHercules, CUDABeaver, CudaForge MAY.2025

– Present

Agentic MLSys

CUDA Programming

CUDA Optimization

GPU Profiling

- Led multi-agent workflows for end-to-end CUDA program generation, optimization, verification, and profiling.
- Incorporated rubric-based agentic reinforcement learning and hardware feedback to improve CUDA optimization quality.
- Built benchmarks for LLM-based CUDA debugging and expert-level hardware-aware CUDA optimization.

› **GPU Graph Computing and EDA-Oriented GNN Acceleration Advisors: Prof. Caiwen Ding, Prof. Gong Xiaoli, Prof. Pen-Chung Yew**

Related publications: GSR-GNN, DR-CircuitGNN, Liberator, OneGraph

2021 – Present

Graph Computing

EDA

GNN Acceleration

GPU Memory Management

- Developed GPU-accelerated graph and GNN systems for circuit graph learning and heterogeneous circuit graph workloads.
- Designed data reuse and memory-management techniques for out-of-memory graph computing on GPUs.
- Explored cross-architecture graph computing with oneAPI and large-scale dynamic temporal graph processing on distributed multi-GPUs.

› **GPU-Accelerated Signal Compression**

Advisor: Prof. Caiwen Ding

Related publication: FPTC

2026

Parallel Compression

GPU Decoder

Signal Processing

- Contributed to a high-throughput asymmetric signal codec with a lightweight encoder and massively parallel GPU decoder.
- Optimized transform-based decompression throughput while preserving reconstruction quality across multiple signal domains.

Teaching Experience

› **Teaching Assistant of Computer Architecture (CSCI 4203)**

Fall 2025

University of Minnesota-Twin Cities. Responsible for leading discussion sections and grading assignments on CPU pipelining, caching, and memory hierarchy.

› **Teaching Assistant of Operating System**

Autumn 2020, 2021, 2022

Responsible for teaching experimental courses and guiding junior undergraduate students in mastering the essential concepts of the Linux system and virtual machines. Supported students in developing a toy OS, with emphasis on managing threads, processes, memory, and the file system.

› **Teaching Assistant of Principles of Computer Organization**

Spring 2023

Responsible for teaching experimental courses; assisted students in gradually learning how to use Verilog to implement a 5-level pipeline CPU.

› **Teaching Assistant of Introduction to Parallel Programming**

Spring 2022

Responsible for teaching programming courses; taught students how to code with CUDA on NVIDIA GPUs and implemented the Gauss Elimination algorithm with CUDA.

Work Experience

Alibaba Cloud, Shanghai, China

System Engineer Intern

JUL.2021 – SEPT.2021

Container and OS kernel development with Rust

- › Developed Linux kernel with Rust. Mapped device memory between the Host and Hypervisor with MMIO.
- › Implemented passthrough and hotplug of the NVIDIA GPU in the container with VirtIO.

Awards

China National Scholarship (only one in college per year)

DEC.2023

Outstanding Student Leader of Nankai University

SEPT.2023

First-class Scholarship for Graduate Students of Nankai University (top 10%)

OCT.2023

Nankai University Special Scholarship (3 times)

2022–2024