Mengyue Xi

Email: ximy@mail2.sysu.edu.cn | Tel: +8617358300383 | WeChat: mengyue414141 | Personal Website

Basic Information

Mengyue Xi, Master's student at Sun Yat-sen University, supervised by Prof. Xianwei Zhang. Interested in **GPU hardware-software co-design**, with a focus on **cache management** and **compiler optimization** techniques.
- Languages: C++/C, CUDA/HIP, Python, PyTorch, Verilog
- Technologies: CUDA/ROCm ecosystem, LLVM framework, LLM, basic machine learning models, etc.

Education

Sun Yat-sen University, Master of Computer Science and Technology — Sep. 2023 – June 2026
- GPA: 94.42/100.00, Rank: 2/78

Chongqing University, Bachelor of Computer Science and Technology — Sep. 2019 – June 2023
- GPA: 3.81/4.00, Rank: 16/219

Projects

GPU Software Layer Resource Management, Three works — Dec. 2023 - Nov. 2024
General Program of the National Natural Science Foundation of China

- Multi-level Cache Bypassing on GPUs: Developed a system using **LLVM** and **C++** to dynamically select optimal cache levels for bypassing based on load instructions, achieving 1.15× performance improvement over the default cache policy.
- LLM-assisted GPU Cache Management for Kernel Concurrency: Developed CLLM using **LLVM, C++**, and **LLMs** to optimize cache utilization for concurrent GPU kernels, reducing cache contention and achieving 1.337× performance improvement.
- Fine-grained GPU Kernel Fusion: Developed GoPTX, a kernel fusion method that enhances instruction-level parallelism (ILP) by weaving **PTX-level** instructions. Reduced pipeline stalls and optimized scheduling, achieving 11.2% average speedup.

Compiler Construction Course Reform - Teaching Assistant — Feb. 2024 - Aug. 2024
CCF Excellence in Computer Education Teaching Award – First Prize

- Contributed to the reform of "Compiler Theory" and "Compiler Construction" courses by developing an LLVM-based teaching framework with CMake, Docker, and VSCode. Led the creation of a syntax analysis framework using Flex and Bison (or ANTLR) for syntax analysis, type checking, and abstract syntax tree construction.

Customs Declaration AI Platform Development - Lab Collaboration — Apr. 2024 - Aug. 2024

- Built a prototype system integrating OCR, RAG, and open-source LLMs.
- Deployed on A100, RTX 3090, and RTX 4090, achieving scalable performance and robust accuracy.

Reconfigurable Architecture Optimization with MLIR - Bachelor's Thesis — Jan. 2023 - June 2023

- Implemented loop interchange optimization using polyhedral models for CGRA architectures in LLVM.
- Achieved 1.14-1.16x performance improvement for identifying optimal loop structures.

Publications

Mpache: Interaction Aware Multi-level Cache Bypassing on GPUs - **First Author** *Accepted by ASP-DAC 2025*
CLLM: Leveraging LLMs for Cache Management to Enhance GPU Kernel Concurrency - **First Author** *Under review at DAC 2025*
GoPTX: Fine-grained GPU Kernel Fusion by PTX-level Instruction Flow Weaving - **Third Author** *Under review at DAC 2025*

Experience

Recommendation System Engineer Intern, Cider – Beijing — Sep. 2022 – June 2023
Implemented and tested backend systems using C++, Python, Flask, Milvus, Airflow, and Kubernetes, with a focus

on efficient data processing.

Miscellaneous

National Scholarship, Sun Yat-sen University Samsung Scholarship
Reviewer for NAS 2024, Class Leader, Member of the English Debate Team, CET-6: 517