Tianci Hou

tiancihou0102@outlook.com • https://tiancihou.me

PROFESSIONAL EXPERIENCE

NVIDIA Semiconductor Technology (Shanghai) Co., Ltd.
Deep Learning Performance Architect Intern Jul 2024 – Nov 2024

- Developed data collection scripts for cuDNN performance testing on Hopper GPUs, establishing a critical performance baseline for the next-generation Blackwell architecture.
- Contributed to the MLIR-based compiler by implementing kernel fusion passes and extending type test coverage, enhancing the compiler's optimization capabilities.

EDUCATION

University of California, San Diego Sep 2025 – May 2027 (Expected)
Master of Science, Computer Science
The Chinese University of Hong Kong, Shenzhen Sep 2021 – May 2025
Bachelor of Engineering, Computer Science and Engineering

- GPA: 3.83/4.0（Major: 4.0/4.0); Outstanding Graduate; Academic Performance Scholarship; Dean's List.
- Teaching Assistant in Data Structure, Computer Architecture, Operating System and Parallel Programming.

University of California, Berkeley Jan 2024 – May 2024
Visiting Student; GPA: 4.0/4.0

AWARDS

- **Gold Medal, Third Place** – Guangdong Collegiate Programming Contest Jun 2022
- **Gold Medal** – China Collegiate Programming Contest, Harbin Site Nov 2021
- **Gold Medal** – The ACM-ICPC Asia Xuzhou Regional Contest Mar 2019

PROJECTS

Parallel Programming Course | *GitHub Link* Sep 2023 – Dec 2023

- Accelerated algorithms for projects in Image Processing, Matrix Multiplication, Sorting and CNN/DNN by applying diverse parallelization strategies with technologies including SIMD, OpenMP, MPI, CUDA and Triton.

Database Messing System | *GitHub Link* Sep 2022 – Dec 2022

- Developed a simple Database Management System with a SQL interpreter using Java.
- Implemented support for basic syntax (e.g., load, store, select), advanced syntax (e.g., join, group by) and version control features (e.g., snapshot, rollback).

CUHKSZOJ | *GitHub Link* Sep 2021 – May 2025

- Engineered and maintained a university-wide Online Judge platform by customizing and extending the open-source DMOJ and HydroOJ project, serving over 1,000 students across 10+ computer science courses.
- Engineered a sandboxed auto-judging system on university clusters, using NixOS and Docker for process isolation to securely run students' code; created a RAG-based AI assistant for students learning.

RESEARCH

Tuning Block Size to Optimize MoE GEMM Performance in vLLM
CTHPC 2025 Workshop Oct 2024 – May 2025

- Optimized vLLM's Mixture-of-Experts (MoE) GEMM kernel, demonstrating up to 5x speedup over PyTorch's default by tuning Triton block sizes on an NVIDIA RTX 4080 Super GPU.
- Analyzed how block size selection impacts performance and memory utilization, emphasizing its criticality for efficient GPU-accelerated LLM inference.

Efficient Maximal Motif-Clique Enumeration over Large Heterogeneous Information Networks (HINs)
DOI: 10.14778/3681954.3681975 (VLDB 2024) Feb 2023 – Mar 2024

- Processed datasets such as Freebase and DBPedia, increasing the size of test HINs, and conducted experiments on five real-world large HINs to demonstrate the efficiency and scalability of our algorithm.

SKILLS

- **Languages:** C/C++ (Expert), Python (Expert), Java (Advanced), SQL (Advanced)
- **Developer Tools:** Git, Linux, Docker