Lane (Lei) Huang

(+1) (646) 543-8495 leih5@illinois.edu lei-huang-winlere https://winlere.github.io

Highlights

- **HPC at scale and speed:** built both scalable distributed systems (ML infra and Storage Infra) and extremely fast program with low-level optimizations; ISC/SC award-winning results.
- **Modern C++ & Python & CUDA expert:** modern C++ and Python/PyTorch; applied in system and parallel programming; ICPC medalist.
- **Proven in top environments:** engineer roles at Sixie Capital (top quant trading firm in China), a startup at MIT, NUS and UIUC.

Education

University of Illinois Urbana-Champaign — Illinois, USA

Computer Science, *Master of Science* — Aug. 2025 - May 2027 (Expected)

ShanghaiTech University — Shanghai, China

Computer Science, *Bachelor of Engineering* | GPA 3.86/4.0 (Major) 3.70/4.0 (Overall) — Sep. 2021 - Jul. 2025

Bachelor's Concentration: ***Computational Mathematics***

Shanghai Municipal Outstanding Graduate (5% of the City)

Work Experience

Rsearch Assistant — Illinois USA

University of Illinois Urbana-Champaign — Aug. 2025 - Present

Parallel and distributed progarmming in Python. Scaling $\alpha\beta$-CROWN towards parallel, high-performant, distributed and scaled formal verifier, applying high perfomance computing techniques to the program in Python PyTorch.

ML System Engineer — Remote

Research Startup at MIT — Jul. 2025 - Aug. 2025

Built an end-to-end large language model pretraining system in HPC GPU and networking, including a data production cluster (10M tok/s output), distributed training system (NVIDIA Megatron) and evaluation tools. Produced 1 Trillion (10^{12}) tokens and finished the training of a 890M model on 16 H100 in 14 days.

System Engineer — Shanghai, China

Sixie Capital — Dec. 2024 - Jul. 2025

- **GPU Support**: Provided GPU/CUDA/PyTorch support for the quantitative research team. Discovered a PyTorch bug that was later triaged by the cutlass team.
- **Distributed ML Training**: Implemented multi-node, multi-GPU distributed model training on a GPU cluster, enabling one-click start/stop and saving significant time for the research team.
- **Distributed File System**: Automated the deployment and configuration of BeeGFS over RDMA Infiniband, serving as the storage infrastructure of the HPC cluster.
- **Auto Evolving Data Production Check**: Delivered a meta-rule based market data check framework, automatically covering all existing fields and incoming fields. Prevented five data production incidents.

Research Assistant — Singapore and Shanghai

National University of Singapore (NUS) — Dec. 2023 - Dec. 2024

- Developed CUDA implementations of the algorithm, achieving a 1000x speedup.

Projects

1. System Programming

PintOS: Operating System Kernel — profetia/pintos

C, CMake, make, GDB

To understand operating systems, we implemented an OS kernel featuring thread scheduling, system calls, virtual memory, and a file system. It can successfully boot in QEMU.

Rathernet: A Full 7-Layer OSI Computer Network Carried by Sound Waves — profetia/rather-net

Rust

To learn computer networking from the ground up, we started with sound wave modulation/demodulation to build a full 7-layer OSI network. We implemented physical, data link, network, and transport layers, installed our driver (presentation layer) on Windows 11, and successfully connected to the internet (application layer).

ChocoPy: A Compiler of Python Targeting RISC-V and LLVM-IR

cs131-chocopy/chocopy

C++, Flex, Bison, RISC-V, LLVM-IR

To study compiler principles, we built a compiler that processes a front-end language into an abstract syntax tree and generates LLVM IR and RISC-V stack machine code.

---

2. Parallel & High-Performance Programming

LBM: Fluid Simulation Optimized for Microarchitectural Features

Winlere/lbm

C, OpenMP, SSE2, AVX2

To practice efficient microarchitectural use, I optimized a fluid simulation program by employing cache-friendly memory access patterns and SIMD instructions, achieving a 200x speedup using 4 CPU cores.

BFAVerifier: CUDA-Accelerated Formal Verifier for Bit-Flip Attacks

zhangyedi/bfaverifier

CUDA/C++, Gurobi

Implemented the SymPoly algorithm from our published paper using CUDA, achieving a 1000x speedup compared to the CPU version.

Cuckoo Hash: CUDA Parallel Hash Table

Winlere/CuckooHash

CUDA/C++, CUDA Stream, NVIDIA Nsight

To study GPU programming, I implemented a GPU-based parallel hash table that achieves up to 1 × 10^9 insertions or 2 × 10^9 queries per second on an RTX 3090.

Competitions

Student Cluster Competition (SCC)

Hamburg, Germany; Denver, USA

ShanghaiTech University GeekPie_HPC Team — Jan. 2023 - Nov. 2023

- ISC'23 Student Cluster Competition: Third Place. Compiled, ran, analyzed, and optimized a fluid simulation program on FAU and Bridges-2 supercomputers.
- SC'23 Student Cluster Competition: Seventh Place, with an Outstanding Reproducibility Report. Over 48 continuous hours, compiled, ran, and analyzed large-scale matrix decomposition algorithms, successfully reproducing key results.

International Collegiate Programming Contest (ICPC)

Shanghai, Nanjing, Hefei (China)

- ICPC Asia Regional: Solved 7 complex algorithmic problems within five consecutive hours. As team captain and a core member, led the team to win 3 silver medals.

Publications

Verification of Bit-Flip Attacks against Quantized Neural Networks

OOPSLA 2025 (CCF-A)

Yedi Zhang, **Lei Huang**, Pengfei Gao, Fu Song, Jun Sun, Jin Song Dong — Feb. 2025

Skills

Programming	**Modern C/C++**, **Python**, **CUDA/C++**, **PyTorch**, Not limited to any single language
Speaking	Chinese (Native), English (Fluent)
Other Skills	**Linux System-Level Programming & Administration**, Pandas, Gurobi Solver