

Bin-Lun Li

 [mike911209](#) |  [my site](#) |  mike911209@gmail.com |  +886 982-880-498

EDUCATION

National Tsing Hua University (NTHU)

Sep 2021 - Jun 2025

B.S. in Computer Science | Advisor: [Jerry Chou](#)

Average GPA 4.2/4.3 on last 4 semesters.

Teaching Assistant of "Operating Systems" and "Hardware Design and Lab".

SKILLS

Programming Languages C/C++, Python, Verilog, Go

Tools & Libraries Git, Unix-like shells, Docker, Kubernetes, CUDA, MPI, OpenMP, Triton

Languages Chinese: Native, English: Fluent (TOEIC 915)

EXPERIENCE

Machine Learning Engineer Intern - Lasertec Taiwan

Dec 2024 - Present

PyTorch/Triton/CUDA/Nsight Systems/Linux

- Accelerated morphological operations from CPU to GPU, achieving a **297x speedup** by leveraging PyTorch for initial parallelization and Triton for custom kernel optimization.
- Profiled performance bottlenecks using Nsight Systems, identifying and resolving inefficiencies in GPU computation.
- Integrated optimized pipelines into a **production-grade GPU software** system, enabling deployment for industrial applications.

President of Computer Science Student Association - NTHU

Sep 2022 - Aug 2023

Leadership/Communication

- **Led a team of 10 people**, collaborated with faculty, industry professionals, and student groups.
- Organized technical workshops, introducing 50+ students to tools like Git, Docker, and LaTeX, fostering early adoption of industry-standard technologies.

AWARDS

Second Prize - 2024 Meichu Hackathon

[Link to Project](#)

Python/Flask/Backend/LLM

- Achieved **2nd place out of 230 contestants**, first prize in Logitech group.
- Multi-agents chatroom, where each agent embodies a distinct personality, enabling dynamic interactions and valuable insights through conversations.
- **Containerized** the backend, ensuring consistent deployment across different environments.

PROJECTS

Scalable LLM Inference Serving System

Go/Kubernetes/Prometheus/Grafana/GPU sharing/System Design

- Leveraged **GPU sharing technique** to allocate resources, resolved GPU underutilization issue.
- Implemented **dynamic resource allocation** strategies based on LLM inference workload analysis.
- Delivered a **59% reduction** in mean token processing time by optimizing inference throughput and resource utilization.

All-Pairs Shortest Path

C/C++/CUDA/Parallel Programming

- Optimized both SRAM and host memory access using parallel programming techniques.
- Achieved **10th place** out of 110 contestants in the course competition.