

HAN-TING LIANG

☎ 217-200-5283

✉ htliang2@illinois.edu

🌐 [linkedin.com/in/han-ting-liang](https://www.linkedin.com/in/han-ting-liang)

🐙 github.com/lhan0123

Education

University of Illinois Urbana-Champaign

Aug 2023 – Present

Ph.D. in Computer Science (Advised by Professors Francis Yan and Indranil Gupta)

National Tsing Hua University

Sep 2021 – Jun 2023

Master of Science in Computer Science

Chung Yuan Christian University

Sep 2017 – Jun 2021

Bachelor of Science in Information and Computer Engineering

Technical Skills

Programming Languages: C/C++, Python, Go, JavaScript, Java, SQL

Systems: Distributed Systems, Kubernetes, Clouds, Serverless, AI Inference and Automation

Libraries/Models: PyTorch, TensorFlow, vllm, CUDA, scikit-learn, OpenMPI

Databases & Data Systems: MongoDB, Kafka, Spark, Hadoop, Redis

Observability & Monitoring: Prometheus, Grafana, Jaeger, OpenTelemetry

System Optimization: Performance modeling and debugging, Resource management, Scheduling, Scaling, Reliability

Professional Experience

Graduate Research Assistant, University of Illinois Urbana-Champaign

USA, Aug 2023 – present

- Designing a safety evaluation framework to assess the safety and performance of AI-generated changes in production systems.
- Constructing feedback-driven learning loops that use evaluator signals to improve AI agents for incident response and planning.
- Implemented adaptive polling for IoT systems, reducing RPC overhead by up to 86% and improving routine automation security.

Software Engineering Intern (Data-Technical Infrastructures-IaaS), ByteDance

USA, May 2025 – Aug 2025

- Prototyped an LLM-based natural language shell to enable safe and reversible automation of Linux and cloud API operations.
- Redesigned the overlay file system to support transaction-like rollback and fine-grained undo for LLM-generated commands.
- Integrated a sandboxing mechanism to eliminate overhead from deep file system stacking.

Graduate Research Assistant, National Tsing Hua University

Taiwan, Sep 2021 – Jun 2023

- Developed a communication-efficient algorithm that improved runtime performance of microservice applications by 44.66%.
- Designed a caching policy to speed up the function launch process on Kubernetes with 42.4% latency reduction.

AI Algorithm Engineer Intern, Industrial Technology Research Institute

Taiwan, Jul 2022 – Aug 2022

- Optimized machine learning model deployment with Triton Inference Server and Kubernetes, reducing startup time by 67.4%.
- Developed an adaptive start-time scheduling algorithm to enable concurrent model execution and improve GPU utilization.
- Implemented distributed ML applications, including healthcare and payment applications, using PyTorch and Flask.

Cloud Software Assistant, Chung Yuan Christian University

Taiwan, Jul 2018 – Aug 2021

- Contributed to the successful deployment of computationally efficient telecom systems for 5G Network Function Virtualization.
- Designed an optimized auto-scaling algorithm into a practical cloud system, increasing success call rate by 8.77%.
- Reduced virtual machine performance degradation by 5.2% using SDN, OVS, and CPU/memory affinity.

Projects

Distributed Systems Design and Implementation

Aug 2023 – Dec 2023

- Processed a distributed system from scratch using Go, including the functions of Distributed Log Querier, Group Membership Protocol, Gossip Protocol, Leader Election, and Distributed File System Replication across 10 VMs. Emulated MapReduce framework and SQL commands on top of the systems and compared its performance against Apache Hadoop.

Fast Stable Diffusion Inferencing System by Patch Reusing

Aug 2024 – Dec 2024

- Collaborated with a team of three to optimize serving for open-source text-to-image diffusion models. Focused on implementing a k-round caching policy using PyTorch and vector database and enabling faster inference by creating a shortcut using similarity to avoid regenerating new patches at every step.

User-Aware Scheduling for Coordinated IoT Routines

Jan 2024 – May 2024

- Designed a user-centric IoT scheduler that shifts the scheduling optimization focus from system performance to user experience, introducing a co-completion-based strategy to coordinate routine completion across smart devices.

Interpreter/Assembler Design and Implementation

Aug 2019 – Jan 2020

- Designed a 9000+ line program interpreter/assembler, including the functions of lexical analysis, syntax analysis, conditional expressions, calculation, function calls, and loops in C++.

Publications

- Anna Karanika, Kai Wang, **Han-Ting Liang**, Indranil Gupta. “RASC: Enhancing Observability and Programmability in Smart Spaces.” *USENIX Symposium on Networked Systems Design and Implementation (NSDI '26)*, 2026.
- **Han-Ting Liang**, Jerry Chou. “HPA: Hierarchical Placement Algorithm for Multi-Cloud Microservices Applications.” *IEEE International Conference on Cloud Computing Technology and Science (CloudCom '22)*, 2022.
- **Han-Ting Liang**, Yun-He Wang, Wu-Chun Chung. “Performance Impacts of Scaling Policies for virtual IP Multimedia Subsystem on the Cloud.” *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN '22)*, 2022.