Han-Ting Liang

217-200-5283 htliang2@illinois.edu www.linkedin.com/in/han-ting-liang github.com/lhan0123

Education

University of Illinois Urbana-Champaign Ph.D. in Computer Science **Expected May 2028**

National Tsing Hua University M.S. in Computer Science **June 2023**

Chung Yuan Christian University B.S. in Computer Science **June 2021**

Technical Skills

Programming Languages: C/C++, Python, Go, JavaScript, Java, Shell/Base Script, HTML, SQL
Systems: Docker, Kubernetes, Hadoop, Spark, Microservices, HPC clusters, AWS Cloud, Azure, OpenStack, Linux
Libraries/Models: Pytroch, Cuda, OpenMPI, Tensorflow, OpenCV, Flask, CNN for image processing, Text-conditional Diffusion
Databases: DynamoDB, Redis, Cassandra, MySQL, Google Firebase
Networking: Restful API, grpc, SDN, OVS, Ansible
Optimization: auto-scaling design, scheduling design, resource management orchestration, fault-tolerance design

Professional Experience

Graduate Research Assistant, University of Illinois Urbana-Champaign USA, Aug 2023 – present

- Collaborated with a team of three to bridge device-state inconsistency on IoT systems, reducing RPC overhead by 44%-86%.
- Accelerated scheduling parallelism by 10-50% for multiple routines in real Home Assistant with 30+ smart devices using Python.
- Managed a Microservices project to propose a wise way of writing and deploying distributed applications.

Graduate Research Assistant, National Tsing Hua University Taiwan, Sep 2021 - June 2023

- Conducted research on Microservices and Serverless, including optimizing resource management and service runtime.
- Developed a novel communication-efficient algorithm that plays the runtime of microservice applications 44.66% better.
- Designed a caching policy to speed up the function launching process on Kubernetes with 42.4% latency reduction.
- Simulated and built a real cross-cloud platform using C++ and AWS services, including Lambda, EC2, EKS, S3, API Gateway.

AI Algorithm Engineer Intern, Industrial Technology Research Institute Taiwan, July 2022 - Aug 2022

- Worked with Triton Inference and Kubernetes to improve the performance of ITRI's existing AI service systems.
- Enhanced the deployment process and execution time of machine learning models by 67.4%.
- Implemented AI health services with PyTorch and Flask.

Cloud Software Assistant, Chung Yuan Christian University Taiwan, July 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 virtualization performance degradation by 5.2% using SDN, OVS, and CPU/memory affinity.
- Wrote customized SIPP scenarios with XML for testing various vIMS features.

Projects

Fast Stable Diffusion Inferencing System by Patch Reusing Aug 2024 – present

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

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.

Full-Stack Customer Feedback Portal Jan 2023 – May 2023

- Implemented an iOS mobile application for collecting customer feedback using the Flutter framework, with HTML/CSS for user input forms. On the backend, built a data analysis and authorization system using AWS Lambda and integrated Google Firebase for data storage. Utilized Node.js to ensure seamless communication between the frontend and backend.

Hadoop-like Framework Design and Implementation Aug 2021 – Dec 2021

- Developed multi-threaded MapReduce tasks with data partitioning and shuffling to parallelize computation across multiple GPU nodes. Built robust coordination mechanisms to synchronize tasks effectively across machines using C++, Cuda, and OpenMPI.

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

- **Han-Ting Liang**, Jerry Chou. HPA: Hierarchical Placement Algorithm for Multi-Cloud Microservices Applications.(https://www.doi.org/10.1109/CloudCom55334.2022.00013)
- **Han-Ting Liang**, Yun-He Wang, Wu-Chun Chung. Performance Impacts of Scaling Policies for virtual IP Multimedia Subsystem on the Cloud.(https://www.doi.org/10.1109/NFV-SDN56302.2022.9974953)