Yu-Ju Huang

yh885@cornell.edu
https://www.cs.cornell.edu/~yjhuang
https://www.linkedin.com/in/yu-ju-huang/

Education

Cornell University, Ithaca, NY
Ph.D. in Computer Science *Aug. 2019 – Present*
Advisor: Prof. Robbert van Renesse

National Chiao Tung University, Hsinchu, Taiwan
M.S. in Computer Science, GPA: 4.0 *Sep. 2013 - Jun. 2015*
Thesis: A KVM-based Hypervisor for Heterogeneous System Architecture
Advisor: Prof. Wei-Chung Hsu

National Chiao Tung University, Hsinchu, Taiwan
B.S. in Computer Science, GPA: 3.91 *Sep. 2009 - Jun. 2013*

Professional Summary

Computer Systems Researcher & System Software Engineer
* Research Expertise: distributed systems, databases, ML systems, operating systems, compilers.
- PhD research on designing high-performance, strongly consistent consensus protocols to enhance the performance of **transactional databases and streaming data processing**.
- As a research intern at AWS, worked on **vector databases for Generative AI, streaming LLM** inference pipelines, and **data lake** optimization.
- Conducted research on operating systems, focusing on enabling **virtualization** for emerging hardware features, such as virtualizing unified address spaces for CPU and GPU.

* Software Development: 3+ years of experience in industry-quality software development.
- Developed **compilers and runtimes for in-house deep learning accelerators** (DLA) at a leading IC design company.

* Programming Languages: Proficient in C, C++, Rust, Java, Python, and Go.

Work Experience

Applied Scientist Intern - Amazon Web Services (AWS) *Cambridge, UK | May-Aug 2024*
* **Vector Database for Streaming Data and GenAI/RAG**
- Designed an end-to-end Flink pipeline for vector search, enabling real-time data integration for Generative AI (GenAI) and Retrieval-Augmented Generation (RAG).

* **Vector Database for Large-Scale Data**
- Developed an ANN-based VectorDB supporting vector mutation and hosting > 1TB data.

*This internship resulted in **two patent** filings.*

Applied Scientist Intern - Amazon Web Services (AWS) *Cambridge, UK | May-Aug 2023*
* **Data Lake Optimization**
- Applied statistical analysis to optimize Parquet tables, reducing file sizes and improving query performance.

* **Streaming LLM Inference**
- Implemented LLM inference on Flink using either in-memory ML model or external ML agent.

Applied Scientist Intern - Amazon Web Services (AWS) *Seattle, US | May-Aug 2022*
* **Transactional Key-Value Store Verification**
- Built an infrastructure in Rust to verify invariants of a transaction KVS library.

System Software Engineer - MediaTek, Office of CTO Hsinchu, Taiwan | Dec 2015-Jun 2019

* **Compiler & Runtime for ML Inference Frameworks**
  - Developed a compiler for an in-house deep learning accelerator (DLA).
  - Built frameworks for running AI models (TensorFlow, Android NN) on CPUs, GPUs, and DLAs.
  - Led a taskforce to optimize DLA performance.

* **Android Runtime & Compiler Optimization**
  - Implemented a staged compiler using LLVM to optimize Android applications.

* **QoS-Based Framework**
  - Developed a quality-of-service (QoS) framework to optimize Android runtime by dynamically adjusting system resources based on QoS hints.

Awards

* **Cornell University Fellowship**, 2019-2020
* **Columbia University Presidential Fellowship** (declined), 2019-2023
* **Best Paper Award**, 12th International Conference on Virtual Execution Environments (*VEE'16*)

Professional Service

Shadow Program Committee Member
18th European Conference on Computer Systems (EuroSys'23)

Program Committee Member and Conference Session Chair
13th International Conference on Virtual Execution Environments (VEE'17)

Publications

Ziplog: A Totally Ordered Log combining Low Latency with Scalable Throughput
***Yu-Ju Huang**, Shubham Chaudhary, Lorenzo Alvisi,* Robbert van Renesse
Under submission

Fast Replica Coordination with ZIP
***Yu-Ju Huang**, Shubham Chaudhary, Rafael Soares, Shir Cohen, Lorenzo Alvisi, Luis Rodrigues,* Robbert van Renesse
Under submission

Disaggregated Applications Using Nanoservices
Xinwen Wang, **Yu-Ju Huang**, Tiancheng Yuan, Robbert van Renesse
Workshop On Resource Disaggregation and Serverless (*WORD'21*), *April 2021*

Building a KVM-based Hypervisor for a Heterogeneous System Architecture Compliant System
Yu-Ju Huang, Hsuan-Heng Wu, Yeh-Ching Chung, Wei-Chung Hsu.
12th International Conference on Virtual Execution Environments (*VEE'16*), *April 2016*
Best Paper Award

Teaching

Cornell University
Head TA for CS 4411: Practicum in Operating Systems
Head TA for CS 4410: Operating Systems
Grad TA for CS 6410: Advanced Systems
Grad TA for CS 3410: Computer System Organization and Programming