Chung-En Ho

404-203-5834 | cho322@gatech.edu | linkedin.com/in/chung-en-ryan-ho/ | github.com/chungen04 | Personal Page

Education

Georgia Institute of Technology Aug. 2025 - May 2027 (Expected)
Masters of Science in Computer Science (On Campus) *Atlanta, GA*
- Coursework: GPU Hardware and Software, Systems for Machine Learning

National Taiwan University Sep. 2020 - Jun. 2024
B.S. Electrical Engineering, Minor Computer Science, GPA: 4.20/4.30 *Taipei, Taiwan*
- Leadership: Co-director of NTUEE+, an alumni networking organization

Professional Experience

Skymizer Inc. Oct. 2024 - Jul. 2025
Software Engineer | [Company Website] *Taipei, Taiwan*
- Developed a customized backend in HuggingFace TGI LLM serving framework, demonstrating the first successful product integration with open-source frameworks, boosting serving throughput by 12x with continuous batching
- Implemented LLM/VLMs including LLaVA, LLaMA-3, Mamba-2 in C, ensuring bit-true operator validation on Skymizer's Language Processing Unit
- Designed a token streaming strategy to compute encoder-based Transformer attention, achieving minimal on-chip SRAM usage (1MB) for vision-language models

IBM Research Almaden Jun. 2023 - Sep. 2023
Research Intern, Analog AI team *San Jose, CA*
- Researched compute-in-memory (CIM) architecture to enable high-throughput LLM inference, analyzing design trade-off between volatile-memory and non-volatile memory-based system on a data-transfer scale
- Implemented a highly pipelined architecture for Transformer encoder-based LLM inference, achieving 2x area efficiency on volatile-memory-based CIM systems than the previous works [IEDM'23]
- Contributed to a C++-based simulator to evaluate CIM systems performance, improving its ability to analyze CIM system architectures under different settings

Selected Projects

Efficient VLM Serving via Traffic-aware Token Reduction Sep. 2025 - Dec. 2025
Skills: vLLM, LLM serving, LLM systems benchmarking *Atlanta, GA*
- Implemented a VLM serving system with encoder, prefill and decode disaggregation based on vLLM, deployed on 2xH100 GPUs, and support up to 10 query per second within a 2 second TTFT budget for Qwen2.5-VL-3B VLM
- Proposed and implemented a traffic-aware token reduction algorithm to reduce the encoder workload via image resizing, reducing time-to-first-token (TTFT) by up to 37% with negligible performance drop on Microsoft Azure production trace
- Developed an evaluation pipeline to benchmark trade-offs between VLM performance and TTFT service-level objectives (SLOs)

Research experiences

Efficient and Intelligent Computing (EIC) Lab Aug. 2025 - Present
Graduate Researcher, Supervised by Prof. Yingyan (Celine) Lin *Atlanta, GA*
- Proposing efficient algorithms for diffusion LLM inference, boosting token throughput via efficient sampling

Cyber Physical Systems Lab Oct. 2023 - Oct. 2024
Undergraduate Researcher, Advised by Prof. Chung-Wei Lin *Taipei, Taiwan*
- Proposed a quantization strategy for communication-efficient deep neural network inference in vehicular edge computing to reduce the inference latency and communication traffic
- Employed deep reinforcement learning as the quantization decision process, reducing inference latency and communication time by 14-21% and 31-44% while complying with accuracy constraints
- Led the research and authored a journal paper, currently under peer review at *IEEE ESL*

Publications

[1] **Chung-En Ho**, Chung-Ting Tsai, I-Ching Tseng, Chung-Wei Lin. A Quantization Strategy for Communication-Efficient DNN Inference in Vehicular Edge Computing. *Under Review at IEEE ESL.*

[2] G. W. Burr, H. Tsai, W. Simon, I. Boybat, S. Ambrogio, **C.-E. Ho**, Z.-W. Liou et al. Design of Analog-AI Hardware Accelerators for Transformer-based Language Models. *International Electron Devices Meeting (IEDM)*, 2023.

Technical Skills

Programming Languages: Python, C, C++, Rust, Bash, JavaScript, Verilog, MATLAB
ML infrastructure: CUDA, Triton, Nsight, NCCL, vLLM, HuggingFace TGI
Software: UNIX systems, Conda, uv, LaTex, React, Flask, GraphQL
Developer Tools: Git, Docker, Google Cloud Platform, VS Code
Frameworks: NumPy, Matplotlib, PyTorch, gRPC