Chien-Yu Lin

PhD Candidate
Computer Science and Engineering
University of Washington

Email: cyulin@cs.washington.edu
Website: https://cylinbao.github.io

Research Interests

I believe **efficiency** is key to the **continued growth** and **sustainability** of future machine learning. My research spans a **variety of ML workloads**, including CNNs, GNNs, NeRF, and LLMs, and covers **multiple domains** such as accelerators, GPU kernels, and efficient training and inference via algorithm and system co-design. Moving forward, I aim to continue cross-stack research to build highly efficient multi-modal models, explore model architectures beyond transformers, and investigate the robustness of efficient methods.

Education

Sep 2018 - (Jun 2025) Ph.D., Computer Science and Engineering
University of Washington, USA
Advisor: Prof. Luis Ceze
Thesis Topic: Efficient Machine Learning Systems

Sep 2015 - Jun 2017 M.Sc., Electronics Engineering
National Yang Ming Chiao Tung University, Taiwan
Advisor: Prof. Bo-Cheng Lai
Thesis Topic: Accelerator for Sparse Convolutional Neural Networks

Jan 2015 - Jun 2015 Exchanged student
Koc University, Istanbul, Turkey

Sep 2011 - Jan 2015 B.Sc., Electronics Engineering
Minor in Computer Science
National Yang Ming Chiao Tung University, Taiwan
GPA: 3.82 / 4.0

Experience

Sep 2018 - Present **Research Assistant**
SAMPL Lab, University of Washington, Seattle, USA
- Algorithm and software co-design for efficient machine learning systems.

Mar 2023 - Jun 2023 **Machine Learning Research Intern**
Oct 2021 - Sep 2022 AI/ML org., Apple Inc, Seattle, USA
- First time hosts: Anish Prabhu and Carlo Del Mundo.
- Second time hosts: Thomas Merth and Anurag Rajan.
- Research on model compression and efficient 3D rendering algorithms.
- Published one ECCV and one WAVC paper.

Jan 2018 - Aug 2018 **Algorithm Engineer Intern**
Ambarella Inc, Santa Clara, USA
- Developed efficient lane and object detection algorithms for self-driving cars.

Sep 2015 - Jun 2017 **Research Assistant**
PCS Lab, NYCU, Hsinchu, Taiwan
- Designed an efficient accelerator for sparse CNNs.

July 2014 - Aug 2014 **Compiler Engineer Intern**
Marvell, Hsinchu, Taiwan
- Built a verification tool for an advanced in-house C++ compiler.

Publications

(* indicates equal contribution)

[C4] Atom: Low-bit Quantization for Efficient and Accurate LLM Serving [pdf].
Yilong Zhao, **Chien-Yu Lin**, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, Baris Kasikci.
In Conference on Machine Learning Systems (MLSys), 2024 (accept rate 22%).
Cited by 81 times in one year; over 280 stars on Github..

[C3] FastSR-NeRF: Improving NeRF Efficiency on Consumer Devices with A Simple Super-Resolution Pipeline [pdf].
Chien-Yu Lin, Qichen Fu, Thomas Merth, Karren Yang, Anurag Ranjan.
In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, **Oral (Top 2.6%)**.

[C2] SPIN: An Empirical Evaluation on Sharing Parameters of Isotropic Networks [pdf].
Chien-Yu Lin*, Anish Prabhu*, Thomas Merth, Sachin Mehta, Anurag Ranjan, Maxwell Horton, and Mohammad Rastegari.
In European Conference on Computer Vision (ECCV), 2022.

[C1] Supporting Compressed-Sparse Activations and Weights on SIMD-like Accelerator for Sparse Convolutional Neural Networks [pdf].
Chien-Yu Lin and Bo-Cheng Lai.
In the 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), 2018.

[J1] Enhancing Utilization of SIMD-Like Accelerator for Sparse Convolutional Neural Networks [pdf].
Bo-Cheng Lai, Jyun-Wei Pan, and **Chien-Yu Lin**.
In IEEE Transactions on Very Large Scale Integration Systems (TVLSI), Feb. 2019.

Preprints

[A5] TeleRAG: Efficient Retrieval-Augmented Generation Inference with Lookahead Retrieval.
Chien-Yu Lin*, Keisuke Kamahori*, Yiyu Liu, Xiaoxiang Shi, Madhav Kashyap, Rulin Shao, Yile Gu, Zihao Ye, Kan Zhu, Arvind Krishnamurthy, Stephanie Wang, Rohan Kadekodi, Luis Ceze, Baris Kasikci.
In submission to OSDI 2025.

[A4] Palu: Compressing KV Cache via Low-Rank Projection [pdf].
Chi-Chih Chang*, Wei-Cheng Lin*, **Chien-Yu Lin***, Yu-Fang Hu, Pei-Shuo Wang, Chong-Yan Chen, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, Kai-Chiang Wu.
In submission to ICLR 2025 (average review score: 5.75).

[A3] NanoFlow: Towards Optimal Large Language Model Serving Throughput [pdf].
Kan Zhu, Yilong Zhao, Liangyu Zhao, Gefei Zuo, Yile Gu, Dedong Xie, Yufei Gao, Qinyu Xu, Tian Tang, Zihao Ye, Keisuke Kamahori, **Chien-Yu Lin**, Stephanie Wang, Arvind Krishnamurthy, Baris Kasikci.
In submission to OSDI 2025. **Over 680 stars on Github.**.

[A2] Efficient Encoder-Decoder Transformer Decoding for Decomposable Tasks [pdf].
Bo-Ru Lu, Nikita Haduong, **Chien-Yu Lin**, Hao Cheng, Noah A. Smith, Mari Ostendorf.
ArXiv:2403.13112, May 2024.

[A1] Accelerating SpMM Kernel with Cache-First Edge Sampling for Graph Neural Networks [pdf].
Chien-Yu Lin, Liang Luo, and Luis Ceze.
ArXiv:2104.10716, April 2021.

Teaching Experience

Fall 2024 **Guest Instructor and Teaching Assistant**
Systems for Machine Learning, CSE 599K, UW

- With Prof. Arvind Krishnamurthy
- Taught three lectures on LLM performance optimizations and ML hardware
- Designed an assignment on attention performance analysis.
- Link: `https://courses.cs.washington.edu/courses/cse599k/24au/`

Spring 2024 **Teaching Assistant**
High-Performance Scientific Computing, Amath 483/583 A, UW
- With Prof. Kenneth Roche.
- Parallel computing class in UW.
- Topics cover pthreads, multi-process, MPI, and CUDA.

Spring 2022 **Teaching Assistant**
Computer Architecture II, CSE 470, UW
- With Prof. Luis Ceze.

Fall 2016 **Teaching Assistant**
Fall 2015 Computer Architecture (Grad Level), EE, NYCU
Spring 2015 Computer Organization (Undergrad Level), EE, NYCU
- With Prof. Bo-Cheng Lai.
- Designed several new course projects. Topics included acceleration of image processing and dense/sparse neural networks.
- Tools involved RISC-V toolchain, Multi2Sim and CUDA programming.

Service

2023 - 2025 Lab seminar organizer, SAMPL Lab, UW
- Events link: https://sampl.cs.washington.edu/talks.html

2025 PhD admission committee area chair, UW CSE
2024 Reviewer, 3DV
2021 PhD admission committee, UW CSE
2020 Artifact evaluation committee, ASPLOS
2020 Prospective student committee chairs, CSE, UW
2013 - 2014 Student system administrator, EE, NYCU

Awards

2024 MLSys student travel grant.
2014 Outstanding student, System and architecture talent incubation program, Taiwan.

Mentoring

I find great joy in helping junior students develop skills and achieve their goals. I am fortunate to have mentored the following students.

Spring 2025 - present Ruby Lee (UW), ECE undergrad at UW.
Fall 2024 - present Yiyu Liu (SJTU), now applying CS PhD program in US.
Spring 2024 - present Chi-Chih Chang (NYCU), now an ECE PhD student in Cornell.
Summer - Fall 2023 Yilong Zhao (SJTU), now an EECS PhD student in UC Berkeley.
Spring 2017 Jyun-Wei Pan (NYCU), now an engineer at MediaTek.

Invited Talks

Nov. 2024 KV-Cache compression with low-rank projection, at UW CSE research day.
May. 2024 Low-bit quantization for LLMs, at MLSys.
Jan. 2024 Low-bit quantization for LLMs, at NCKU.
Jan. 2024 Fast NeRF with super resolution, at WACV.
Jan. 2018 Accelerator for sparse CNN, at ASP-DAC.
Jun. 2016 A Survey of CNN Accelerators, at MediaTek

Patents

[P1] Apparatus and Method of Using Dual Indexing in Input Neurons and Corresponding Weights of Sparse Neural Network [pdf].
Chien-Yu Lin, and Bo-Cheng Lai.
US Patent Application 15/594,667, 2018.

Mountain Leadership

In addition to my research, I have a strong passion in exploring nature, particularly through mountaineering and backcountry skiing. I frequently lead groups on mountain expeditions and summit attempts of challenging peaks. These experiences have taught me invaluable lessons in team leadership, risk management, and resilience - skills that I apply in my professional work.