

Chien-Yu Lin

Doctor of Philosophy (Ph.D)
Computer Science and Engineering
University of Washington

Email: cyulin@cs.washington.edu
Website: <https://cylinbao.github.io>
Phone: (+1) 415-818-6916

Research Interests

I'm passionate about **making machine learning efficient**. My research spans a **wide range of ML workloads**, including LLMs, NeRFs, GNNs and CNNs, and covers **multiple domains**, such as efficient training and inference algorithms, GPU kernels and accelerator designs. Moving forward, I aim to expand my cross-stack research to develop highly efficient systems for multi-modal models, explore novel model architectures beyond transformers, and investigate the robustness for efficient techniques.

Education

Sep 2020 - Jun 2025	Ph.D., Computer Science and Engineering University of Washington, USA Advisor: Prof. Luis Ceze Thesis: Toward Efficient Machine Learning Systems with Sampling and Compression
Sep 2018 - Jun 2020	M.Sc., Computer Science and Engineering University of Washington, USA
Sep 2015 - Jun 2017	M.Sc., Electronics Engineering National Yang Ming Chiao Tung University, Taiwan Advisor: Prof. Bo-Cheng Lai Thesis: A Dual-Sparsity 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

Work Experience

Sep 2018 - Jun 2025	Research Assistant SAMPL Lab, University of Washington, Seattle, USA <ul style="list-style-type: none">Algorithm and system co-design for efficient machine learning.
Mar 2023 - Jun 2023 Oct 2021 - Sep 2022	Machine Learning Research Intern AI/ML org., Apple Inc, Seattle, USA <ul style="list-style-type: none">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 <ul style="list-style-type: none">Developed efficient CNN models for self-driving cars.
Sep 2015 - Jun 2017	Research Assistant Parallel Computing System Lab, NYCU, Hsinchu, Taiwan <ul style="list-style-type: none">Accelerator design for sparse CNNs.

July 2014 - Aug 2014 **Compiler Engineer Intern**

Marvell, Hsinchu, Taiwan

- Built a verification tool for an advanced in-house C++ compiler.

Teaching Experience

- Fall 2024 **Guest Lecturer 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 course 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.

Publications

(* indicating equal contribution)

- [1] TeleRAG: Efficient Retrieval-Augmented Generation Inference with Lookahead Retrieval [pdf].
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 SOSP 2025. ArXiv:2502.20969, 2025.
- [2] 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 The 19th USENIX Symposium on Operating Systems Design and Implementation (**OSDI**), 2025.
- [3] 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 The Thirteenth International Conference on Learning Representations (**ICLR**), 2025
First low-rank based KV-Cache compression (Developed concurrently to DeepSeek's MLA).
- [4] 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 The Seventh Conference on Machine Learning and Systems (**MLSys**), 2024 (Accept rate 22%)
First W4A4 quantization for LLMs; More than 100 citations within one year.

- [5] 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 The IEEE/CVF Winter Conference on Applications of Computer Vision (**WACV**), 2024.
Oral (Top 2.6%).
- [6] 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, 2024.
- [7] 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 The European Conference on Computer Vision (**ECCV**), 2022.
- [8] Accelerating SpMM Kernel with Cache-First Edge Sampling for Graph Neural Networks [pdf].
Chien-Yu Lin, Liang Luo, and Luis Ceze.
ArXiv:2104.10716, 2021.
- [9] 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.
- [10] 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.

Service

2023 - 2025	Lab seminar organizer, SAMPL Lab, UW <ul style="list-style-type: none"> Events link: https://sampl.cs.washington.edu/talks.html
2024 - 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.
2024	Oral, top 2.6%, WACV 2024.
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 mentor the following students.

Fall 2024 - Spring 2025	Yiyu Liu (SJTU), incoming PhD student in Harvard CS program.
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.

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.

Invited Talks

April. 2025	Toward efficient machine learning systems with sampling and compression, Amazon AWS AI.
Mar. 2025	Toward efficient machine learning systems with sampling and compression, Meta FAIR.
Mar. 2025	Efficient 3D neural rendering and LLM inference, Apple MLR.
Jan. 2025	Efficient RAG inference with lookahead retrieval, at NYCU, Taiwan.
Jan. 2025	LLM quantization and KV-Cache compression, at NTU, Taiwan.
Nov. 2024	KV-Cache compression with low-rank projection, at UW CSE research day, Seattle.
May. 2024	Low-bit quantization for LLMs, at MLSys, Santa Clara.
Jan. 2024	Low-bit quantization for LLMs, at NCKU, Taiwan.
Jan. 2024	Fast NeRF with super resolution, at WACV.
Nov. 2023	Low-bit quantization for large language models, at UW CSE research day, Seattle.
Jan. 2018	Accelerator design for sparse CNNs, at ASP-DAC, South Korea.
Jun. 2016	A Survey of CNN Accelerators, at MediaTek, Hsinchu.

Mountain Experience

In addition to my research, I have a strong passion for exploring nature, particularly through **mountaineering** and **backcountry skiing**. I frequently lead groups on mountain expeditions and have had many successful summits and ski descents on challenging peaks in the Pacific Northwest (PNW), including Mt. Rainier (14,410 ft), Mt. Shasta (14,179 ft), and Mt. Hood (11,249 ft). These experiences have taught me invaluable lessons in team leadership, risk management, and resilience - skills that I apply in my professional work.