Mark Zhao

myzhao@stanford.edu | 418 Gates Computer Science, Stanford, CA 94305 | (662)-801-1496
https://web.stanford.edu/~myzhao/

Research Interests

My research focuses on building **computer systems for machine learning (ML) pipelines** to train and deploy large-scale ML models. I take a holistic approach by co-designing the multiple interconnected systems that compose modern ML pipelines, including training data systems, secure and distributed accelerators, and application-centric inference serving frameworks. To improve the efficiency, security, and scalability of these diverse pipelines, I leverage tools across the computing stack, including computer systems, computer architecture, security, databases, and machine learning.

Education

Stanford University 2025 (expected)
Ph.D. in Electrical Engineering
Dissertation: *Performant and Scalable Systems Across the Machine Learning Pipeline*
Advisor: Christos Kozyrakis

Cornell University 2018
B.S. in Electrical and Computer Engineering, summa cum laude
Research Advisor: Edward Suh

Industry Research Experience

Meta Platforms 2020 – 2022
Visiting Researcher, FAIR SysML & Capacity Engineering and Analysis
Mentors: Carole-Jean Wu and Niket Agarwal
· Built, deployed, and optimized distributed systems to improve the performance and efficiency of Meta's production machine learning infrastructure. Projects included a disaggregated data preprocessing service (DPP), a flash storage tier for ML datasets (Tectonic-Shift), and deduplication optimizations for recommendation model training infrastructure (RecD).

Intel Corporation 2019
Graduate Cloud Engineering Intern, Data Center Group
Mentor: Arindam Saha
· Developed an inference serving framework that dynamically manages ML accelerator designs on Intel FPGAs to maximize serving performance across diverse inference requests.

Peer-Reviewed Publications

cedar: Optimized and Unified Machine Learning Input Data Pipelines 2025
Mark Zhao, Emanuel Adamiak, and Christos Kozyrakis
[**VLDB 2025**] *Proceedings of the VLDB Endowment, Volume 18*

ReCycle: Resilient Training of Large DNNs using Pipeline Adaptation 2024
Swapnil Gandhi, **Mark Zhao**, Athinagoras Skiadopoulos, and Christos Kozyrakis
[**SOSP 2024**] *30th ACM Symposium on Operating Systems Principles*

High-throughput and Flexible Host Networking for Accelerated Computing 2024
Athinagoras Skiadopoulos, Zhiqiang Xie, **Mark Zhao**, Qizhe Cai, Saksham Agarwal, Jacob Adelmann, David Ahern, Carlo Contavalli, Michael Goldflam, Vitaly Mayatskikh, Raghu Raja, Daniel Walton, Rachit Agarwal, Shrijeet Mukherjee, and Christos Kozyrakis
[**OSDI 2024**] *2024 USENIX Symposium on Operating Systems Design and Implementation*

Tectonic-Shift: A Composite Storage Fabric for Large-Scale ML Training 2023
Mark Zhao, Satadru Pan, Niket Agarwal, Zhaoduo Wen, David Xu, Anand Natarajan, Pavan Kumar, Shiva Shankar P, Ritesh Tijoriwala, Karan Asher, Hao Wu, Aarti Basant, Daniel Ford, Delia David, Nezih Yigitbasi, Pratap Singh, Carole-Jean Wu, and Christos Kozyrakis
[**ATC 2023**] *2023 USENIX Annual Technical Conference*
Invited fast-track submission to ACM Transactions on Storage

RecD: Deduplication for End-to-End Deep Learning Recommendation Model Training Infrastructure 2023
Mark Zhao, Dhruv Choudhary, Devashish Tyagi, Ajay Somani, Max Kaplan, Sung-Han Lin, Sarunya Pumma, Jongsoo Park, Aarti Basant, Niket Agarwal, Carole-Jean Wu, and Christos Kozyrakis
[**MLSys 2023**] *6th Conference on Machine Learning and Systems*

Understanding Data Storage and Ingestion for Large-Scale Deep Recommendation Model Training 2022
Mark Zhao, Niket Agarwal, Aarti Basant, Buğra Gedik, Satadru Pan, Mustafa Ozdal, Rakesh Komuravelli, Jerry Pan, Tianshu Bao, Haowei Lu, Sundaram Narayanan, Jack Langman, Kevin Wilfong, Harsha Rastogi, Carole-Jean Wu, Christos Kozyrakis, and Parik Pol
[**ISCA 2022**] *49th IEEE/ACM International Symposium on Computer Architecture*

ShEF: Shielded Enclaves for Cloud FPGAs 2022
Mark Zhao, Mingyu Gao, and Christos Kozyrakis
[**ASPLOS 2022**] *27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*

Llama: A Heterogeneous & Serverless Framework for Auto-tuning Video Analytics Pipelines 2021
Francisco Romero*, **Mark Zhao***, Neeraja J Yadwadkar, and Christos Kozyrakis
[**SoCC 2021**] *12th ACM Symposium on Cloud Computing*
(* denotes equal contribution)

HyperFlow: A High-Assurance Processor Architecture for Practical Timing-Safe Information Flow Security 2018
Andrew Ferraiuolo, **Mark Zhao**, Andrew C. Myers, and G. Edward Suh
[**CCS 2018**] *25th ACM Conference on Computer and Communications Security*

FPGA-Based Remote Power Side-Channel Attacks 2018
Mark Zhao and G. Edward Suh
[**S&P 2018**] *39th IEEE Symposium on Security and Privacy*
Distinguished Practical Paper Award
2022 Top Pick in Hardware and Embedded Security

Technical Articles, Preprints, and Working Manuscripts

Adaptive Semantic Prompt Caching with VectorQ 2025
Luis Gaspar Schroeder, Shu Liu, Alejandro Cuadron, **Mark Zhao**, Stephan Krusche, Alfons Kemper, Matei Zaharia, Joseph E. Gonzalez
Under Submission

ReCoOpt: A Framework for HW/SW Optimization for Recommendation Inference Pipelines from Retrieval to Ranking 2025
Zhanqiu Hu, **Mark Zhao**, Zhiru Zhang, and Udit Gupta
Under Submission

Dynamic Memory Management for Efficient Mixture-of-Experts Training 2025
Athinagoras Skiadopoulos, **Mark Zhao**, Swapnil Gandhi, Thomas Norrie, Rachit Agarwal, Shrijeet Mukherjee, and Christos Kozyrakis
In preparation

Remote Power Side-Channel Attacks on FPGAs 2024
Mark Zhao and G. Edward Suh
IEEE Design & Test, 2024

Counting Spree: Color Recognition and Segmentation in Real-time Video to Detect Manufacturing Defects 2018
Mark Zhao and Claire Chen
Circuit Cellar Magazine, Issue #333, April 2018

Awards and Honors

Meta Ph.D. Fellowship in AI System HW/SW Co-Design 2023
· ***Full funding and stipend for two academic years***

MLCommons Machine Learning and Systems Rising Star 2023

Top Pick in Hardware and Embedded Security 2022
· *For FPGA-based Remote Power Side-Channel Attacks*

Stanford Graduate Fellowship 2018
· ***Full funding and stipend for three academic years***

Distinguished Practical Paper Award, IEEE Symposium on Security and Privacy 2018
· *For FPGA-based Remote Power Side-Channel Attacks*

Sibley Prize, Cornell ECE 2018
· *Awarded to the top graduating senior in Electrical and Computer Engineering*

Meinig Family Cornell National Leadership Scholar, Cornell University 2014
· *University-wide scholarship for demonstrating "an outstanding degree of leadership"*

United States Presidential Scholar, U.S. Department of Education 2014
· *Program established in 1964, by executive order of the President, to "recognize and honor some of our nation's most distinguished graduating high school seniors"*

Invited Talks

End-to-End Optimization of Large-Scale ML Training Systems
· *AMD Research and Advanced Development* 2024
· *UCF ECE Computer Architecture Seminar Series* 2024
· *SRC JUMP 2.0 ACE Center for Evolvable Computing Annual Review* 2023

Understanding and Optimizing Data Storage and Ingestion Systems
· *SRC JUMP 2.0 ACE Center for Evolvable Computing Liason Meeting* 2023
· *Cornell Systems Lunch* 2023
· *ByteDance Infrastructure Research Group* 2023
· *Stanford SystemX Fall Conference* 2022

FPGA-Based Remote Power Side-Channel Attacks
· *Top Picks in Hardware and Embedded Security Workshop* 2022

Llama: A Heterogeneous & Serverless Framework for Auto-Tuning Video Analytics Pipelines
· *Stanford Systems Seminar* 2021
· *Stanford Platform Lab Retreat* 2020

ShEF: Shielded Enclaves for Cloud FPGAs
· *Stanford SystemX Fall Conference* 2019
· *Stanford Platform Lab Review* 2019

Teaching Experience

CS 349D: Cloud Computing Technology, *Course Assistant* Spring 2024
Stanford University

CS 349D: Cloud Computing Technology, *Course Assistant* Spring 2023
Stanford University

EE 180: Digital Systems Architecture, *Course Assistant* Winter 2023
Stanford University

ECE 5760: Advanced Microcontroller Design, *Teaching Assistant* Spring 2018
Cornell University

ECE 4760: Designing with Microcontrollers, *Teaching Assistant* Fall 2017
Cornell University

ECE 3140: Embedded Systems, *Teaching Assistant* Spring 2017
Cornell University

PHYS 2213: Physics II (Electromagnetism), *Undergraduate Teaching Assistant* Fall 2015
Cornell University

MATH 1920: Multivariable Calculus for Engineers, *Course Assistant* Fall 2015
Cornell University

Service

Workshop on Machine Learning and Systems (EuroMLSys at EuroSys'25), *Technical Program Committee* 2025

Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models (SCOPE at ICLR'25), *Technical Program Committee*	2025
IEEE International Symposium on Circuits and Systems (ISCAS'25), *External Reviewer*	2024
Stanford EE Faculty Search Committee, *Graduate Student Member*	2024
Workshop on ML for Computer Architecture and Systems (MLArchSys at ISCA'24), *Technical Program Committee*	2024
Workshop on Machine Learning and Systems (EuroMLSys at EuroSys'24), *Technical Program Committee*	2024
Workshop on ML for Computer Architecture and Systems / Architecture and System Support for Transformer Models Workshop (MLArchSys/ASSYST at ISCA'23), *Technical Program Committee*	2023
IEEE Transactions on Circuits and Systems II: Express Briefs, *External Reviewer*	2022
Design Automation Conference (DAC), *External Reviewer*	2019

Mentorship

Zhanqiu (Summer) Hu (Ph.D. @ Cornell Tech) · *End-to-End Optimization of Recommendation Systems*	2024–Present
Suze van Adrichem (B.S. @ Stanford) · *PandoRT: A Distributed Serving System for Compound LLM Applications*	2024–Present
Jenny Wei (B.S. @ Stanford) · *Building and Optimizing Systems for LLM Pipelines*	2024–Present
Laasya Konidala (B.S. @ Stanford) · *Building and Optimizing RAG for LLM Pipeline Serving Systems*	2024
Ethan Zhang (B.S. @ Stanford) · *A New Frontier for Model Routing*	2024
Emanuel Adamiak (B.S. @ Stanford) · *cedar: Optimized and Unified Machine Learning Input Data Pipelines*	2023 – 2024
Andrew Woen (B.S. @ Stanford) · *Optimizing Data Storage and Ingestion Pipelines for ML Training*	2023