

Yinwei Dai

yinweid@princeton.edu <https://yinwei-dai.com>

EDUCATION

Princeton University

Doctor of Philosophy in Computer Science

- Advisor: Ravi Netravali

Princeton, NJ

Aug. 2022 - Present

University of Michigan

Master of Science in Engineering in Computer Science

- Advisor: Mosharaf Chowdhury & Harsha Madhyastha

Ann Arbor, MI

Sep. 2020 - May. 2022

Bachelor of Science in Engineering in Computer Science

- Summa Cum Laude

Sep. 2018 - May. 2020

Shanghai Jiao Tong University

Bachelor of Science in Electrical and Computer Engineering

Shanghai, China

Aug. 2016 - Aug. 2020

PUBLICATIONS

- [SOSP 24] **Apparate: Rethinking Early Exits to Tame Latency-Throughput Tensions in ML Serving** [Yinwei Dai*](#), Rui Pan*, Anand Iyer, Kai Li, Ravi Netravali
- [SOSP 24] **Improving DNN Inference Throughput Using Practical, Per-Input Compute Adaptation** Anand Iyer, Mingyu Guan, [Yinwei Dai](#), Rui Pan, Swapnil Gandhi, Ravi Netravali
- [SoCC 23] **Auxo: Efficient Federated Learning via Scalable Client Clustering** Jiachen Liu, Fan Lai, [Yinwei Dai](#), Aditya Akella, Harsha Madhyastha, Mosharaf Chowdhury
- [NSDI 23] **ModelKeeper: Accelerating DNN Training via Automated Model Transformation** Fan Lai, [Yinwei Dai](#), Harsha Madhyastha, Mosharaf Chowdhury
- [ICML 22] **FedScale: Benchmarking Model and System Performance of Federated Learning** Fan Lai, [Yinwei Dai](#), Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, Mosharaf Chowdhury
- [ResilientFL 21] **FedScale: Benchmarking Model and System Performance of Federated Learning** Fan Lai, [Yinwei Dai](#), Xiangfeng Zhu, Harsha Madhyastha, Mosharaf Chowdhury **Best Paper Award**

RESEARCH EXPERIENCE

Princeton University

Advised by Prof. Ravi Netravali, Prof. Kai Li and Professor Anand Iyer

Rethinking Early Exits to Tame Latency-Throughput Tensions in ML Serving

- Designed and built the first system that automatically injects and manages (at runtime) Early Exiting (EE) for ML model inference.
- Developed efficient online adaptation algorithms to adjust EE's configuration to meet accuracy & throughput constraints and maximize latency savings.
- Improved medium latency performance by 5X on traffic videos and 1.2X on NLP workloads while always adhering to accuracy and throughput constraints, compared to vanilla models.

Princeton, NJ

Aug. 2022 - Now

University of Michigan, EECS, SymbioticLab

Advised by Prof. Mosharaf Chowdhury and Prof. Harsha V. Madhyastha

Ann Arbor, MI

July. 2020 - July. 2022

FedScale: Benchmarking Model and System Performance of Federated Learning

- Codeveloped an open-sourced benchmark for FL that incorporates real-world client datasets for diverse tasks and supports the simulation of practical FL across millions of clients.
- Collected datasets across tasks and partitioned the raw data with unique client identification.
- Implemented baselines for vision tasks under federated learning settings and performed in-depth benchmark experiments for recent FL efforts.

ModelKeeper: Accelerating DNN Training via Automated Model Transformation

- Codeveloped a model service framework to accelerate DNN training by reducing the computation needed via automated model transformation.
- Developed a graph-matching algorithm to measure the transferability between models.
- Improved model training time performance by 1.5X on Imgelsmob and 3.2X on NASBench201 compared to random initialization.

PROJECTS

2-way superscalar P6 processor, Computer Architecture *Jan. 2022 - May. 2022*

- Implemented P6-structure pipeline to handle RISC-V instructions with 2-way super-scalar, associative no-blocking cache, prefetching and so on.

Paxos-based Key/Value Service, Distributed Systems *Sep. 2021 - Dec. 2021*

- Implemented Paxos and designed a key/value service that was fault-tolerant based on Paxos.

Test Input Generator, Programming Language *Jan. 2021 - May. 2021*

- Developed an input generator that can maximize branch coverage for a given C file.

Decaf Compiler, Compiler Construction *Jan. 2019 - May. 2019*

- Developed a compiler for Decaf from lexical analysis to assembly code generation.

TEACHING EXPERIENCE

Princeton University, COS *Princeton, NJ*

COS418 Distributed Systems *Dec. 2024 - May. 2024*

COS316 Principles of Computer System Design *Sep. 2023 - Dec. 2023*

University of Michigan, EECS *Ann Arbor, MI*

EECS442 Computer Vision *Jan. 2022 - May. 2022*

EECS489 Computer Networks *Sep. 2021 - Dec. 2021*

HONORS & AWARDS

- Participation Grant *ICML, 2022*
- Best Paper Award *SOSP ResilientFL, 2021*
- JI John Wu and Jane Sun Talent Scholarships (5 among 315) *SJTU, 2017*

SERVICE

- Conference Reviewer: NeurIPS (Datasets and Benchmarks) 2022, 2023, 2024
- Journal Reviewer: Transactions on Mobile Computing 2022
- Artifact Evaluation Committee: SIGCOMM 2022, MLSys 2023