Fan Lai

Department of Computer Science, 4828 BBB, 2260 Hayward Street, Ann Arbor, MI, 48105, USA
Email: fanlai@umich.edu Homepage: https://web.eecs.umich.edu/~fanlai/

INTERESTS

Machine learning systems, cloud computing, and networked systems.

During my Ph.D. research, I developed systems support across different layers of the software stack for efficient datacenter-scale and wide-area computing, including deep learning systems for the cloud, and distributed systems for fast federated computation over the Internet.

EDUCATION

University of Michigan, Ann Arbor 2023 (Expected)
Ph.D. Candidate, Computer Science
Advisor: Mosharaf Chowdhury

Shanghai Jiao Tong University 2017
Bachelor of Computer Science, IEEE Honor Class

AWARDS & HONORS

- Finalist, Meta Fellowship *(Top 4% among 2300+ worldwide PhD applicants)* *2022*
- Best Paper Award (ACM SOSP ResilientFL) *2021*
- Distinguished Artifact Award (USENIX OSDI) *2021*
- Outstanding Graduates of SJTU *2017*
- National Scholarship (*Awarded by Ministry of Education of China)* *2016*
- Scholarship for Creative Research (Top 2% SJTU Students) *2016*
- Academic Excellence Scholarship of SJTU *2014 – 2016*

PUBLICATIONS

- F. Lai, Y. Dai, H. Madhyastha, M. Chowdhury
  ModelKeeper: Accelerating Model Training via Automated Model Transformation
  NSDI 2023
- F. Lai, Y. Dai, S. Singapuram, J. Liu, X. Zhu, H. Madhyastha, M. Chowdhury
  FedScale: Benchmarking Model and System Performance of Federated Learning at Scale
  ICML 2022 (Best Paper Award@SOSP ResilientFL)
- F. Lai, X. Zhu, H. Madhyastha, M. Chowdhury
  Oort: Efficient Federated Learning via Guided Participant Selection
  OSDI 2021 (Distinguished Artifact Award)
- F. Lai, J. You, X. Zhu, H. Madhyastha, M. Chowdhury
  Sol: Fast Distributed Computation Over Slow Networks
  NSDI 2020
- F. Lai, M. Chowdhury, H. Madhyastha
  To Relay or Not to Relay for Inter-Cloud Transfers?
  HotCloud 2018
- Y. Cui, M. Médard, E. Yeh, D. Leith, F. Lai, K. R. Duffy
  A Linear Network Code Construction for General Integer Connections Based on CSPs
  IEEE/ACM TON 2017

- F. Lai, F. Qiu, W. Bian, Y. Cui, E. Yeh
  Scaled VIP Algorithms for Joint Dynamic Forwarding and Caching in NDN
  ACM ICN 2016
- Y. Cui, F. Lai, E. Yeh, R. Liu
  Enhanced VIP Algorithms for Forwarding, Caching, and Congestion Control in NDN
  IEEE GLOBECOM 2016
- Y. Cui, F. Lai, S. Hanly, P. Whiting
  Optimal Caching and Scheduling in Cache-enabled Heterogeneous Wireless Networks
  IEEE GLOBECOM 2016

In Submission

- H. Zheng, R. Liu, F. Lai, A. Prakash
  Coverage-centric Coreset Selection for High Pruning Rates
- S. Sanjay, F. Lai, C. Hu, M. Chowdhury
  Swan: A Neural Engine for Efficient DNN Training on Smartphone SoCs
- N. Shi, F. Lai, R. Kontar, M. Chowdhury
  Fed-ensemble: Improving Generalization via Model Ensembling in Federated Learning
- Y. Wang, D. Sun, K. Chen, F. Lai, M. Chowdhury
  Egeria: An Efficient DNN Training System with Knowledge-Guided Layer Freezing
- J. You, S. Yang, F. Lai, M. Chowdhury, S. Khuller
  System H: A Framework for Optimizing Hybrid Multi-Cloud Analytics

INDUSTRY EXPERIENCE

Visiting Researcher, Meta AI, Menlo Park — Feb. 2022 - Aug. 2022
Implemented, evaluated, and merged research prototypes in the production platform to accelerate recommendation models

Research Intern, Meta AI — Summer 2021
Analyzed and implemented recommendation models to accelerate model training

RESEARCH

Machine Learning Systems:

- **ModelKeeper** is a cluster model manager that accelerates the training of deep neural networks by repurposing the weights of previously-trained models in the cluster.
- **Oort** is a client orchestration system that improves the performance of federated machine learning across millions of edge devices via guided participant selection.

Federated Computation Systems:

- **FedScale** is a scalable and extensible open-source system for both practical deployment of federated computation at scale and benchmarking.
- **Sol** is a federated execution engine that enables sub-second execution latency for federated data analytics over the high-latency and low-bandwidth Internet.

Application-Aware Networking:

- **System H** considers the performance of parallel communication patterns (e.g., shuffle in Spark) and network resource to schedule traffics over the wide-area network (WAN).
- **Relay** relies on the heterogeneity of WAN bandwidth and latency to determine the relay of network traffics across datacenters.

SERVICE

- **Technical Program Committee**
  2022: DistributedML
- **(External) Reviewer**
  2022: SIGCOMM, CVPR, ICML, NeurIPS, IEEE VTC, IEEE Network Magazine

STUDENT MENTORED

I am fortunate to mentor and co-advise the research of four graduate students and seven undergraduate students.

- **Ph.D. Students**: Jiachen Liu (Umich), Sanjay S. Singapuram (Umich)
- **Masters**: Yinwei Dai (now Ph.D. student at Princeton), Yile Gu
- **Undergraduates**: Xiangfeng Zhu (now Ph.D. student at UW), Yuxuan Zhu, Chengsong Zhang, Junwen Tan, Xuye He, Yunzhen Liu, Xiang Shen

PRESENTATIONS

Slides from all talks are available at https://web.eecs.umich.edu/~fanlai/

Systems Support for Federated Computation at Scale

- Meta AI, Menlo Park, CA, May 2022
- LinkedIn, remote, May 2022

FedScale: Benchmarking Model and System Performance of Federated Learning

- Conference talk at ICML, Baltimore, MD, July 2022
- Workshop talk at SOSP ResilientFL, remote, October 2021

Oort: Efficient Federated Learning via Guided Participant Selection

- Conference talk at OSDI, remote, July 2021

Sol: Fast Distributed Computation Over Slow Networks

- Qubole Marketing and Infrastructure Group, remote, June 2020
- Conference talk at NSDI, Santa Clara, CA, February 2020

To Relay or Not To Relay for Inter-Clouds Transfers?

- Workshop talk at HotCloud, Boston, MA, July 2018