Liangyu Zhao

Email: liangyu@cs.washington.edu **Website**: https://liangyurain.github.io/

Research Interests

Machine learning systems, distributed systems, collective communications; broadly speaking, I am interested in formulating and solving mathematical problems in computer systems and networking.

Education

University of Washington Seattle, WA
Ph.D. in Computer Science 2021 – Present
Direction: Systems & Networking
Advisor: Prof. Arvind Krishnamurthy

University of Washington Seattle, WA
M.S. in Computer Science (unfinished) 2020 – 2021

University of Washington Seattle, WA
B.S. in Computer Science,
B.S. in Applied & Computational Mathematical Sciences
(Discrete Math and Algorithms track) 2015 – 2020

Industry Experience

Microsoft Research, Research in Software Engineering (RiSE) Redmond, WA
Part-Time Researcher Summer 2024 – Present

Microsoft Research, Research in Software Engineering (RiSE) Redmond, WA
Research Intern Summer 2023
Mentor: Saeed Maleki
Optimizing collective communications on machine learning GPUs (e.g., NVIDIA DGX A100, AMD MI250).

ByteDance, AI-Lab Bellevue, WA
Research Intern, ML System Summer 2020
Mentor: Yibo Zhu
Working on automatic learning-rate schedule.

Microsoft, Azure Compute Core Redmond, WA
Software Engineer Intern Autumn 2019

Google, Ads Infra Mountain View, CA
Software Engineer Intern Summer 2019

Microsoft, Azure Compute Core Redmond, WA
Software Engineer Intern Summer 2018

Zap Surgical Systems San Carlos, CA
Software Engineer Intern Summer 2017

Publications

ForestColl: Efficient Collective Communications on Heterogeneous Network Fabrics
Liangyu Zhao, Saeed Maleki, Aashaka Shah, Ziyue Yang, Hossein Pourreza, Arvind Krishnamurthy
arXiv preprint, in submission

NanoFlow: Towards Optimal Large Language Model Serving Throughput
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
arXiv preprint, in submission

Efficient Direct-Connect Topologies for Collective Communications
Liangyu Zhao, Siddharth Pal, Tapan Chugh, Weiyang Wang, Jason Fantl, Prithwish Basu, Joud Khoury, Arvind Krishnamurthy
USENIX Symposium on Networked Systems Design and Implementation (NSDI ’25)

Rethinking Machine Learning Collective Communication as a Multi-Commodity Flow Problem
Xuting Liu, Behnaz Arzani, Siva Kesava Reddy Kakarla, **Liangyu Zhao**, Vincent Liu, Miguel Castro, Srikanth Kandula, Luke Marshall
ACM Special Interest Group on Data Communication (SIGCOMM ’24)

Efficient all-to-all Collective Communication Schedules for Direct-connect Topologies
Prithwish Basu, **Liangyu Zhao**, Jason Fantl, Siddharth Pal, Arvind Krishnamurthy, Joud Khoury
International Symposium on High-Performance Parallel and Distributed Computing (HPDC ’24)

AutoLRS: Automatic Learning-Rate Schedule by Bayesian Optimization on the Fly
Yuchen Jin, Tianyi Zhou, **Liangyu Zhao**, Yibo Zhu, Chuanxiong Guo, Marco Canini, Arvind Krishnamurthy
International Conference on Learning Representations (ICLR ’21)

Nexus: A GPU Cluster Engine for Accelerating DNN-Based Video Analysis
Haichen Shen, Lequn Chen, Yuchen Jin, **Liangyu Zhao**, Bingyu Kong, Matthai Philipose, Arvind Krishnamurthy, Ravi Sundaram
ACM Symposium on Operating Systems Principles (SOSP ’19)

Invited Talks

ForestColl: Efficient Collective Communications on Heterogeneous Network Fabrics
Microsoft Research August, 2024

Automatic Generation of Collective Communication Algorithms: The Principles of ForestColl
ByteDance August, 2024

ForestColl: Efficient Collective Communications on Heterogeneous Network Fabrics
AMD Research July, 2024

Efficient Direct-Connect Topologies for Collective Communications
FOCI Annual Symposium, University of Washington October, 2023

Efficient Direct-Connect Topologies for Collective Communications
Harvard Cloud Networking and Systems Group July, 2023