

# Liangyu Zhao

Email: liangyu@cs.washington.edu

Website: <https://liangyuzhao.me/>

**Research Interests** Machine learning systems, distributed systems, collective communications; broadly speaking, I am interested in applying mathematical techniques to design and build efficient, scalable computer systems.

**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 (incomplete) 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** **Meta**, Superintelligence Lab (MSL) Infra Menlo Park, CA  
Research Scientist Intern Jun – Oct 2025  
Mentors: Bingzhe Liu, Liang Luo, Amar Phanishayee  
Improved recommendation model QPS via novel embedding table sharding and pipeline parallelism.

**NVIDIA**, Applied Deep Learning Research (ADLR) Redmond, WA  
Research Intern Mar – Jun 2025  
Mentors: Vijay Anand Korthikanti & Deepak Narayanan  
Designed and optimized customized communication kernels for Megatron-LM.

**Microsoft Research**, Research in Software Engineering (RiSE) Redmond, WA  
Part-Time Researcher Jul – Nov 2024

**Microsoft Research**, Research in Software Engineering (RiSE) Redmond, WA  
Research Intern Jun – Sep 2023  
Mentor: Saeed Maleki  
Developed collective communication algorithms for multi-node GPU clusters.

**ByteDance**, AI-Lab Bellevue, WA  
Research Intern, ML System Jul – Oct 2020  
Mentor: Yibo Zhu  
Worked on automatic learning-rate schedule.

<b>Microsoft</b> , Azure Compute Core Software Engineer Intern	Redmond, WA Sep – Dec 2019
<b>Google</b> , Ads Infra Software Engineer Intern	Mountain View, CA Jun – Sep 2019
<b>Microsoft</b> , Azure Compute Core Software Engineer Intern	Redmond, WA Jun – Aug 2018
<b>Zap Surgical Systems</b> Software Engineer Intern	San Carlos, CA Jun – Sep 2017

## Publications

## *ForestColl: Throughput-Optimal Collective Communications on Heterogeneous Network Fabrics*

**Liangyu Zhao**, Saeed Maleki, Yuanhong Wang, Zezhou Wang, Ziyue Yang,  
Hossein Pourreza, Arvind Krishnamurthy  
*arXiv preprint, in submission*

*FLASH: Fast All-to-All Communication in GPU Clusters*  
Yiran Lei, Dongjoo Lee, **Liangyu Zhao**, Daniar Kurniawan, Chanmyeong Kim,  
Heetaek Jeong, Changsu Kim, Hyeonseong Choi, Liangcheng Yu, Arvind Krishnamurthy, Justine Sherry, Eriko Nurvitadhi  
*arXiv preprint, in submission*

*Tactic: Adaptive Sparse Attention with Clustering and Distribution Fitting for Long-Context LLMs*  
Kan Zhu\*, Tian Tang\*, Qinyu Xu\*, Yile Gu, Zhichen Zeng, Rohan Kadekodi,  
**Liangyu Zhao**, Ang Li, Arvind Krishnamurthy, Baris Kasikci  
*arXiv preprint, in submission*

*NanoFlow: Towards Optimal Large Language Model Serving Throughput*  
Kan Zhu, Yufei Gao, Yilong Zhao, **Liangyu Zhao**, Gefei Zuo, Yile Gu, Dedong Xie, Tian Tang, Qinyu Xu, Zihao Ye, Keisuke Kamahori, Chien-Yu Lin, Ziren Wang, Stephanie Wang, Arvind Krishnamurthy, Baris Kasikci  
*USENIX Symposium on Operating Systems Design and Implementation (OSDI '25)*

*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

*Efficient Direct-Connect Topologies for Collective Communications*

➤ USENIX NSDI '25 April, 2025

➤ ACE Liaison Meeting Theme 3

ACE Center for Evolvable Computing January, 2025

➤ Future of Cloud Infrastructure (FOCI) Annual Symposium

University of Washington October, 2023

➤ Harvard Cloud Networking and Systems Group

Harvard University July, 2023

*ForestColl: Throughput-Optimal Collective Communications on Heterogeneous Network Fabrics*

➤ Machine Learning System Seminar

Rice University October, 2025

➤ Network and Mobile System Group

Massachusetts Institute of Technology July, 2025

➤ Distributed Systems Laboratory (DSL) Seminar

University of Pennsylvania November, 2024

➤ NLP Reading Group

NVIDIA November, 2024

➤ Paul G. Allen School Annual Research Showcase

University of Washington October, 2024

➤ Research in Software Engineering (RiSE)

Microsoft Research August, 2024

➤ ByteDance  
➤ AMD Research

August, 2024  
July, 2024