Jiamin Li

Pronouns: she/her/hers ⋄ Email: jiaminli.icy@gmail.com ⋄ Tel: +1 737-342-9176 (Preferred) / +852 5484 7038
Ph.D. Candidate (5th year) ⋄ Computer Science ⋄ City University of Hong Kong

RESEARCH INTERESTS

I am broadly interested in Systems for Machine Learning with a specific focus on the following topics:

- Systems tailored for large language models: Optimizing the LLM training and inference
- Resource scheduling in GPU clusters: Designing efficient scheduling algorithms for DNN workloads
- Distributed training & inference: Accelerating large-scale distributed DNN tasks
- Adaptive & sparse computation: Exploring new computing paradigms to effectively scale DNN models

EDUCATION

The University of Texas at Austin, Austin, TX, United States Apr. 2023 – Dec. 2023
Visiting Researcher at UTNS Group, Department of Computer Science.
Supervisor: Prof. Aditya Akella
Project: LLEGO: Finer-Grained Large Language Model Serving in Multi-tenant Clouds

City University of Hong Kong, Kowloon, Hong Kong Sep. 2019 – Jan. 2024 (Expected)
Ph.D. Candidate, Department of Computer Science.
Supervisors: Prof. Cong Wang, Prof. Hong Xu (The Chinese University of Hong Kong)
Thesis: "Efficient Scheduling of Distributed Deep Neural Network Workloads"

City University of Hong Kong, Kowloon, Hong Kong Aug. 2015 – Jul. 2019
B.S., Department of Computer Science (First Class Honours).
Advisor: Prof. Shiqi Wang
Final Year Project: "Mobile Face Anti-spoofing with Deep Learning"

University of Missouri, Columbia, MO, United States Jul. 2017 – Aug. 2017
Big Data Analysis Summer Program, Department of Computer Science.

PUBLICATIONS

Preprints

[P3] **Jiamin Li**, Le Xu, Cong Wang, Hong Xu, Aditya Akella, "LLEGO: Finer-Grained Large Language Model Serving in Multi-tenant Clouds", in submission.

[P2] Xin Tan, **Jiamin Li**, Yitao Yang, Jingzong Li, Hong Xu, "Arlo: Serving Transformer-based Language Models with Dynamic Input Lengths", under revision.

[P1] **Jiamin Li**, Cheng Luo, Ziyue Yang, Lei Qu, Peng Cheng, Cong Wang, Hong Xu, "Merak: An Analytical Performance Simulator for Large Scale Distributed Training", under revision.

Conference Proceedings

[C4] **Jiamin Li**, Qiang Su, Yitao Yang, Yimin Jiang, Cong Wang, Hong Xu, "Adaptive Gating in Mixture-of-Experts based Language Models", Empirical Methods in Natural Language Processing (**EMNLP** Main conference), 2023. (Acceptance Rate = 21.3%)

[C3] **Jiamin Li**, Yimin Jiang, Yibo Zhu, Cong Wang, Hong Xu, "Accelerating Distributed MoE Training and Inference with Lina", USENIX Annual Technical Conference (**ATC**), 2023. (Acceptance Rate = 18.4%) [Link]

[C2] **Jiamin Li**, Hong Xu, Yibo Zhu, Zherui Liu, Chuanxiong Guo, Cong Wang, "Lyra: Elastic Cluster Scheduling for Deep Learning", ACM European Conference on Computer Systems (**EuroSys**), 2023. (Acceptance Rate = 16.1%) [Link] (*Deployed partially at ByteDance*)

[C1] Kaiwei Mo, Chen Chen, **Jiamin Li**, Hong Xu, Chun Jason Xue, “Two-Dimensional Learning Rate Decay: Towards Accurate Federated Learning with Non-IID Data”, IEEE International Joint Conference on Neural Networks (**IJCNN**), 2021.

Journals

[J2] Libin Liu, Hong Xu, Zhixiong Niu, Jingzong Li, Wei Zhang, Peng Wang, **Jiamin Li**, Jason Xue Chun, Cong Wang, “ScaleFlux: Efficient Stateful Scaling in NFV”, IEEE Transactions on Parallel and Distributed Systems, 2022 (**TPDS**).

[J1] Libin Liu, Chengxi Gao, Peng Wang, Hongming Huang, **Jiamin Li**, Hong Xu, Wei Zhang, “Bottleneck-Aware Non-Clairvoyant Coflow Scheduling with Fai”, IEEE Transactions on Cloud Computing, 2021 (**TCC**).

WORK & RESEARCH EXPERIENCE

Research Intern, MLSys - AI Lab, ByteDance May. 2019 – May. 2021
Supervisor: Dr. Yibo Zhu
We design an elastic scheduler for GPU training clusters. The key idea is to exploit cluster-level elasticity by loaning idle inferences servers for training jobs and job-level elasticity by scaling training jobs to better utilize the dynamic resource pool.

Part-time Research Assistant, City University of Hong Kong May. 2018 – May. 2019
Supervisor: Prof. Hong Xu
We propose an adaptive gradient aggregation approach to accelerate distributed DNN training. The key idea is to selectively transmit the gradients with non-negligible large updates, effectively reducing data size.

Backend Developer Intern, Jardine Matheson & Co. Limited May. 2017 – May. 2018
Design and develop web services to facilitate employee recruitment for the Group Human Resources department.

SELECT PROJECTS

Finer-Grained Serving for Large Language Models Ongoing
Collaborate with Dr. Le Xu, Supervised by Prof. Aditya Akella
We propose a finer-grained serving system for large language models in multi-tenant clouds. The key idea is to partition LLMs into smaller blocks to enable the reuse of model components and independent provisioning to improve the computation efficiency.

Performance Simulator for Large-scale Distributed DNN Training Ongoing
Collaborate with Networking Research Group at Microsoft Research Asia
We design a simulator to predict the step time of large-scale distributed DNN training tasks. Merak consists of an analytical formulation to compute the all-reduce kernel running time and an ML model to predict the running time slowdown caused by concurrent execution of kernels.

SELECT AWARDS

ATC Student Travel Grant	USENIX ATC	2023
ML and Systems Rising Stars	ML Commons	2023
Research Activity Funds	City University of Hong Kong	2023
EuroSys Student Travel Grant	ACM EuroSys	2023
Full Postgraduate Studentship	City University of Hong Kong	2019 – 2023
Dean’s List (College of Engineering)	City University of Hong Kong	2015 – 2019

TEACHING ASSISTANT

2022 Fall, 2019 Fall	CS2311, Computer Programming
2022 Spring, 2020 Spring	CS4296 & CS5296, Cloud Computing
2021 Fall	CS4394 & CS5294, Information Security and Management
2021 Spring	CS4293 & CS6290, Topics on Computer Security

2020 Fall	CS5222, Computer Networks and Internets

PROFESSIONAL SERVICES

Artifact Evaluation Committee	ACM CoNEXT 2022, MLSys 2023, ACM EuroSys 2023, USENIX OSDI 2023, USENIX ATC 2023, ACM SOSP 2023, ACM SIGCOMM 2023, ACM EuroSys 2024, USENIX FAST 2024
Technical Program Committee	IEEE IJCNN 2023
Reviewer	IEEE/ACM Transactions on Networking

TECHNICAL SKILLS

Programming Languages	C++, Python, Bash, Go, LaTeX
Operating System	Linux/UNIX
Machine Learning Systems	PyTorch, MXNet, DeepSpeed, HuggingFace, NCCL, CUDA