JIANGFEI DUAN

EMAIL: dj021@ie.cuhk.edu.hk GITHUB: https://github.com/JF-D

BIOGRAPHY

I am a final year Ph.D. student at MMLab, CUHK, advised by Prof. Dahua Lin. My research interests lie in broad area of MLSys, especially efficient LLM training and inference. Before joining CUHK, I received my Bachelor's degree in Computer Science from University of Chinese Academy of Sciences, advised by Prof. Shiguang Shan.

EDUCATION

- The Chinese University of Hong Kong — Hong Kong
  - Ph.D. Candidate in MMLab, Department of Information Engineering. — Aug. 2021 - July 2025
  - Advisor: Prof. Dahua Lin.
- University of Chinese Academy of Sciences — Beijing, China
  - B.E. in Computer Science and Technology. — Aug. 2016 - June 2020
  - GPA: 3.93/4.00 (Rank: 1/69)
  - Advisor: Prof. Shiguang Shan.

EXPERIENCE

- *Applied Scientist Intern* at **AWS** — Aug. 2024 - Jan. 2025, Santa Clara, CA
  - Advisors: Liangfu Chen, Zhen Jia, Zhihao Jia
  - Optimizing LLM inference performance on Amazon homegrown chip. We propose a sequence sharding strategy to shard LLMs more efficiently and an attention mechanism designed to minimize padding while adhering to static shape constraints.
- *Research Intern* at **Hao AI Lab, UCSD** — Aug. 2023 - Aug. 2024, Remote
  - Advisors: Hao Zhang
  - We proposed and built MuxServe to enable efficient multiple LLM serving via spatial-temporal multiplexing.
- *Research Intern* at **Catalyst, CMU** — Apr. 2022 - May 2023, Remote
  - Advisors: Zhihao Jia, Minjia Zhang, Xupeng Miao
  - We proposed and built Parcae to enable cheap, fast, and scalable DNN training on preemptible instances by proactively adjusting the parallelization stratey.
  - We proposed and built SpotServe, the first distributed LLM serving system on preemptible instances.
- *Research Assistant* at **MMLab, CUHK** — Sep. 2020 - Apr. 2022, Hong Kong
  - Advisors: Dahua Lin, Shengen Yan, Xiuhong Li
  - We explored to automatically parallelize DNN training on a given cluster.
  - We proposed and built Proteus to accurately model the performance of various parallelization strategies.
- *Research Assistant* at **MMLab, CUHK** — July 2019 - July 2020, Hong Kong
  - Mentors: Dahua Lin, Xingcheng Zhang
  - We built a system to accelerate large scale data parallel training performance. With sparse communication and system optimization, **we trained AlexNet in 1 minute on a cluster of 1000 V100 GPUs with Parrots** (a framework similar to PyTorch).
  - We also explored large language model distributed training and acceleration techniques.

PUBLICATIONS

* indicates equal contribution.

Conference

[1] Libra: Toward Efficient Balance of Variable-length Data Parallel Large Model Training (**Under Review**)
Chang Chen, Tiancheng Chen, **Jiangfei Duan**, Qianchao Zhu, Zerui Wang, Qinghao Hu, Peng Sun, Xiuhong Li, Chao Yang, and Torsten Hoefler

[2] MxMoE: Mixed-precision Quantization for MoE with Accuracy and Performance Co-Design (**ICML '25**)
Haojie Duanmu, Xiuhong Li, Zhihang Yuan, Size Zheng, **Jiangfei Duan**, Xingcheng Zhang, and Dahua Lin

[3] Tropical: Enhancing SLO Attainment in Disaggregated LLM Serving via SLO-Aware Multiplexing (**DAC '25**)
Jinming Ma, Jiefei Chen, Xiuhong Li, **Jiangfei Duan**, Haojie Duanmu, Xingcheng Zhang, Chao Yang and Dahua Lin

[4] SampleAttention: Near-Lossless Acceleration of Long Context LLM Inference with Adaptive Structured Sparse Attention (**MLSys '25**)
Qianchao Zhu, **Jiangfei Duan**, Chang Chen, Siran Liu, Xiuhong Li, Guanyu Feng, Xin Lv, Huanqi Cao, Chuanfu Xiao, Xingcheng Zhang, Dahua Lin, and Chao Yang

[5] SKVQ: Sliding-window Key and Value Cache Quantization for Large Language Models (**COLM '24**)
Haojie Duanmu, Zhihang Yuan, Xiuhong Li, **Jiangfei Duan**, Xingcheng Zhang, and Dahua Lin

[6] MuxServe: Flexible Multiplexing for Efficient Multiple LLM Serving (**ICML '24**)
Jiangfei Duan, Runyu Lu, Haojie Duanmu, Xiuhong Li, Dahua Lin, Ion Stoica, Hao Zhang.

[7] Centauri: Enabling Efficient Scheduling for Communication-Computation Overlap in Large Model Training via Communication Partitioning (**ASPLOS '24**)
Chang Chen, Xiuhong Li, Qianchao Zhu, **Jiangfei Duan**, Peng Sun, Xingcheng Zhang and Chao Yang.
Best Paper Award

[8] SpotServe: Serving Generative Large Language Models on Preemptible Instances. (**ASPLOS '24**)
Xupeng Miao*, Chunan Shi*, **Jiangfei Duan**, Xiaoli Xi, Dahua Lin, Bin Cui, Zhihao Jia.
Distinguished Artifact Award
IEEE Micro Top Picks Honorable Mention

[9] Parcae: Proactive, Liveput-Optimized DNN Training on Preemptible Instances. (**NSDI '24**)
Jiangfei Duan*, Ziang Song*, Xupeng Miao*, Xiaoli Xi, Dahua Lin, Harry Xu, Minjia Zhang, and Zhihao Jia.

Journal

[1] Efficient Training of Large Language Models on Distributed Infrastructures: A Survey (**Under Review**)
Jiangfei Duan*, Shuo Zhang*, Zerui Wang*, Lijuan Jiang, Wenwen Qu, Qinghao Hu, Guoteng Wang, Qizhen Weng, Hang Yan, Xingcheng Zhang, Xipeng Qiu, Dahua Lin, Yonggang Wen, Xin Jin, Tianwei Zhang and Peng Sun.

[2] Proteus: Simulating the Performance of Distributed DNN Training. (TPDS '24)
Jiangfei Duan, Xiuhong Li, Ping Xu, Xingcheng Zhang, Shengen Yan, Yun Liang, and Dahua Lin.

TEACHING

- TA, IERG3050: Simulation and Statistical Analysis — Fall 2021, CUHK
- TA, CSCI2100: Data Structure — Spring 2022, CUHK

SERVICES

- PC Member: IJCAI 2025
- Reviewer: TPDS 2024, ACM TIST 2024, IJCNN 2025, ICML 2025

- AEC Member: MLSys 2023, OSDI 2024, ATC 2024, ASPLOS 2025

TALKS

- Optimizing Cost and Efficiency in Training and Serving Large Language Models
  Huawei Cloud InnovWave, 2024
- Parcae: Proactive, Liveput-Optimized DNN Training on Preemptible Instances
  CMU Catalyst Seminar, 2024

AWARDS AND HONORS

- Best Paper Award, ASPLOS 2024
- Distinguished Artifact Award, ASPLOS 2024
- Outstanding Graduate of Beijing 2020
- Outstanding Graduate of University of Chinese Academy of Sciences 2020
- Tang Lixin Scholarship 2019
- First-class Academic Scholarship, University of Chinese Academy of Sciences (top 5%) 2017, 2018