JIANGFEI DUAN

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

BIOGRAPHY

I am a third-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. — 2021 - 2025(Expected)
  - Advisor: Prof. Dahua Lin.
- University of Chinese Academy of Sciences — Beijing, China
  - B.E. in Computer Science and Technology. — 2016 - 2020
  - GPA: 3.93/4.00 (Rank: 1/69)
  - Advisor: Prof. Shiguang Shan.

EXPERIENCE

- *Research Intern* at **Catalyst, CMU** — Apr. 2022 - May 2023, Remote
  - Advisors: Prof. Zhihao Jia, Prof. Minjia Zhang (UIUC), Dr. 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: Prof. Dahua Lin, Prof. Shengen Yan (PKU), Prof. Xiuhong Li (PKU)
  - 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: Prof. 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

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

[2] 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.

[3] 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

[4] 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.

[5] Proteus: Simulating the Performance of Distributed DNN Training. (**Under Review**)
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

- AEC Member: MLSys 2023

AWARDS AND HONORS

- First-class Academic Scholarship, University of Chinese Academy of Sciences (top 5%) — 2017, 2018
- Tang Lixin Scholarship — 2019
- Outstanding Graduate of University of Chinese Academy of Sciences — 2020
- Outstanding Graduate of Beijing — 2020