YINMIN ZHONG

Yanyuan Building 816, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China
Homepage: yinminzhong.com ⋄ Email: zhongyinmin@pku.edu.cn ⋄ Github: PKUFlyingPig

EDUCATION

Peking University *Sep 2022 - Present*
Ph.D. in Computer Science.
Advisor: Xin Jin

Peking University *Sep 2018 - June 2022*
B.S. in Computer Science. **GPA: 3.79/4.0 (top 5%)**

RESEARCH INTEREST

My research interest lies in the intersection of Deep Learning and Distributed Systems. I use the insight from a system view to improve different aspects (efficiency, performance, scalability) of deep learning training and serving.

PUBLICATIONS

DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving
Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, Hao Zhang
USENIX Symposium on Operating Systems Design and Implementation (OSDI 2024)

Scaling Large Language Model Training to More Than 10,000 GPUs
Ziheng Jiang*, Haibin Lin*, **Yinmin Zhong***, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng, Xiang Li, Cong Xie, Shibiao Nong, Yulu Jia, Sun He, Hongmin Chen, Zhihao Bai, Qi Hou, Shipeng Yan, Ding Zhou, Yiyao Sheng, Zhuo Jiang, Haohan Xu, Haoran Wei, Zhang Zhang, Pengfei Nie, Leqi Zou, Sida Zhao,Liang Xiang, Zherui Liu, Zhe Li, Xiaoying Jia, Jianxi Ye, Xin Jin, Xin Liu
(*Equal Contribution)
USENIX Symposium on Networked Systems Design and Implementation (NSDI 2024)

DistMind: Efficient Resource Disaggregation for Deep Learning Workloads
Xin Jin, Zhihao Bai, Zhen Zhang, Yibo Zhu, **Yinmin Zhong**, Xuanzhe Liu
IEEE/ACM Transactions on Networking (TON 2024)

AlpaServe: Statistical Multiplexing with Model Parallelism for Deep Learning Serving
Zhuohan Li*, Lianmin Zheng*, **Yinmin Zhong***, Vincent Liu, Ying Sheng, Xin Jin, Yanping Huang, Zhifeng Chen, Hao Zhang, Joseph E. Gonzalez, Ion Stoica (*Equal Contribution)
USENIX Symposium on Operating Systems Design and Implementation (OSDI 2023)

ElasticFlow: An Elastic Serverless Training Platform for Distributed Deep Learning
Diandian Gu, Yihao Zhao, **Yinmin Zhong**, Yifan Xiong, Zhenhua Han, Peng Cheng, Fan Yang, Gang Huang, Xin Jin, Xuanzhe Liu
ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2023)

Fast Distributed Inference Serving for Large Language Models
Bingyang Wu*, **Yinmin Zhong*,** Zili Zhang*, Gang Huang, Xuanzhe Liu, Xin Jin
(*Equal Contribution)
In preprint

EXPERIENCE

ByteDance AML Team **Aug 2023 - Present**
Research Intern

- **Mentor: Xin Liu**

Peking University **Sep 2022 - Present**
Research Assistant

- **Mentor: Xin Jin**

Alibaba DAMO Academy **Sep 2021 - Sep 2022**
Research Intern

- **Mentor: Pengyu Zhang**

AI Innovation Center, Peking University **Sep 2020 - Mar 2021**
Software Engineer Intern

- **Mentor: Ming Lei**

PROJECTS

Large Language Model Serving

DistServe **Jan 2024**
DistServe improves the performance of large language models (LLMs) serving by disaggregating the prefill and decoding computation. Given the application latency requirements, DistServe co-optimizes the resource allocation and parallelism strategy tailored for each phase.

FastServe **April 2023**
Existing LLM serving systems use run-to-completion processing for inference jobs, which suffers from head-of-line blocking and long JCT. FastServe exploits the autoregressive pattern of LLM inference to enable preemption at the granularity of each output token and uses preemptive scheduling to minimize JCT with a novel skip-join Multi-Level Feedback Queue scheduler.

AlpaServe **Oct 2022**
AlpaServe targets the multi-tenancy setting for LLM serving and determines an efficient strategy for placing and parallelizing collections of large deep learning models across a distributed cluster. It will trade-off between the overhead introduced by model parallelism and the opportunity to exploit statistical multiplexing to reduce serving latency in the presence of bursty workloads.

Deep Learning Training

Optimus (deployed in ByteDance) **Sep 2023**
Optimus is a production system for training large language models (LLMs) at the scale of more than 10,000 GPUs. Optimus takes a full-stack approach that co-designs the algorithmic and system components across operator optimization, model block, optimizer design, computation, communication overlapping, data pipeline, and network performance tuning. We develop a set of diagnosis tools to monitor system components and events deep in the stack, identify root causes, and derive effective techniques to achieve fault tolerance and mitigate stragglers.

Alpa **Mar 2022**
Alpa automates model-parallel training of large deep learning (DL) models by generating execution plans that unify data, operator, and pipeline parallelism. Alpa makes it simple to train and serve large models like GPT-3 by adding a few lines of code.

ElasticFlow **Sep 2021**

ElasticFlow is an elastic serverless training platform for distributed deep learning. ElasticFlow provides a serverless interface with two distinct features: (i) users specify only the deep neural network (DNN) model and hyperparameters for a job, but not the number of GPUs; (ii) users specify the deadline for a job, but not the amount of time to occupy GPUs. In contrast to existing server-centric platforms, ElasticFlow provides performance guarantees in terms of meeting deadlines while alleviating tedious, low-level, and manual resource management for deep learning developers.

Side Projects

TacOS (developed as course project at Peking University) **Feb 2024**

TacOS is an educational Operating System implemented in Rust. It borrows many design philosophies from PintOS, which is a C project adopted by many Operating System courses in the top universities. We provide detailed documentation, comprehensive test cases, and debugging tools for students to accomplish a series of labs based on the TacOS skeleton code.

csdiy.wiki **Nov 2021**

csdiy.wiki collects various learning materials and online courses in computer science (CS) from top universities in the world. It covers almost all the areas in CS and provides a comprehensive guidance for anyone who would like to self-learn CS. This project has earned over 40k stars on Github and repeatedly ranked the top popular repository in Github Trending.

TEACHING

Head TA, Operating Systems (Honor Track)	**2023 Spring**
TA, Introduction to Computing	**2022 Fall**
Head TA, Operating Systems (Honor Track)	**2022 Spring**
TA, Introduction to Computer System	**2020 Fall**

AWARDS & HONORS

Outstanding Graduate of Peking University	**2022**
Third Prize Scholarship of Peking University	**2020**
Merit Student of Peking University	**2019**
Tian Chuang Scholarship	**2019**
Zhongying Tang Scholarship	**2019**
Merit Student of Peking University	**2018**