

Chao Jin

PH.D. CANDIDATE · PEKING UNIVERSITY

Yanyuan Building 818, No.5 Yiheyuan Road, Haidian District, Beijing, Republic of China

✉ chaojin@pku.edu.cn | 🏠 chaojin0310.github.io | 📧 chaojin0310 | 🎓 Chao Jin

Education

Peking University

PH.D. IN COMPUTER SOFTWARE AND THEORY

- Advisor: Prof. Xin Jin

Beijing, China

Sep. 2023 - Present

Peking University

B.SC. IN COMPUTER SCIENCE AND TECHNOLOGY

- GPA: 3.82/4.0, Ranking: 5/205

Beijing, China

Sep. 2018 - Jun. 2023

Experience

StepFun System Team

RESEARCH INTERN

- Mentor: Ranchen Ming

Beijing, China

Apr. 2025 - Aug. 2025

ByteDance Seed

RESEARCH INTERN

- Mentor: Ziheng Jiang, Haibin Lin

Beijing, China

Jul. 2023 - Mar. 2025

Software Engineering Institute, PKU

RESEARCH ASSISTANT

- Advisor: Prof. Xin Jin

Beijing, China

Sep. 2021 - Jun. 2023

Alibaba Cloud

RESEARCH INTERN

- Mentor: Rui Miao

Beijing, China

Mar. 2022 - Aug. 2022

Publications

MegaScale-MoE: Large-Scale Communication-Efficient Training of Mixture-of-Experts Models in Production

Edinburgh, UK

EUROPEAN CONFERENCE ON COMPUTER SYSTEMS (EUROSYS 2026)

Apr. 2026

- **Chao Jin***, Ziheng Jiang*, Zhihao Bai, Zheng Zhong, Juncai Liu, Xiang Li, Ningxin Zheng, Xi Wang, Cong Xie, Qi Huang, Wen Heng, Yiyuan Ma, Wenlei Bao, Size Zheng, Yanghua Peng, Haibin Lin, Xuanzhe Liu, Xin Jin, Xin Liu
- (* Equal Contribution)

Efficient RL for LLMs with Dynamic and Online Speculative Decoding

Hong Kong, China

THE 1ST FRONTIER AI SYSTEMS WORKSHOP (FAISys 2025)

Nov. 2025

- **Chao Jin**, Yinmin Zhong, Zili Zhang, Yimin Jiang, Yibo Zhu

MegaScale-Infer: Serving Mixture-of-Experts at Scale with Disaggregated Expert Parallelism

Coimbra, Portugal

ACM SPECIAL INTEREST GROUP ON DATA COMMUNICATION (SIGCOMM 2025)

Sep. 2025

- Ruidong Zhu*, Ziheng Jiang*, **Chao Jin***, Peng Wu, Cesar A. Stuardo, Dongyang Wang, Xinlei Zhang, Huaping Zhou, Haoran Wei, Yang Cheng, Jianzhe Xiao, Xinyi Zhang, Lingjun Liu, Haibin Lin, Li-Wen Chang, Jianxi Ye, Xiao Yu, Xuanzhe Liu, Xin Jin, Xin Liu
- (* Equal Contribution)

RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation

ACM TRANSACTIONS ON COMPUTER SYSTEMS (TOCS 2025)

Sep. 2025

- **Chao Jin**, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Shufan Liu, Xuanzhe Liu, Xin Jin

Ditto: Efficient Serverless Analytics with Elastic Parallelism

New York City, NY, U.S.

ACM SPECIAL INTEREST GROUP ON DATA COMMUNICATION (SIGCOMM 2023)

Sep. 2023

- **Chao Jin**, Zili Zhang, Xingyu Xiang, Songyun Zou, Gang Huang, Xuanzhe Liu, Xin Jin

HydraServe: Minimizing Cold Start Latency for Serverless LLM Serving in Public Clouds

Renton, WA, U.S.

USENIX SYMPOSIUM ON NETWORKED SYSTEMS DESIGN AND IMPLEMENTATION (NSDI 2026)

May 2026

- Chiheng Lou, Sheng Qi, **Chao Jin**, Dapeng Nie, Haoran Yang, Xuanzhe Liu, Xin Jin

StreamRL: Scalable, Heterogeneous, and Elastic RL for LLMs with Disaggregated Stream Generation

IN PREPRINT.

Apr. 2025

- Yinmin Zhong, Zili Zhang, Xiaoniu Song, Hanpeng Hu, **Chao Jin**, Bingyang Wu, Nuo Chen, Yukun Chen, Yu Zhou, Changyi Wan, Hongyu Zhou, Yimin Jiang, YiBo Zhu, Daxin Jiang

FaaSPR: Latency-oriented Placement and Routing Optimization for Serverless Workflow Processing

IEEE/ACM TRANSACTIONS ON NETWORKING (TON 2025)

Mar. 2025

- Yunshan Jia, **Chao Jin**, Qing Li, Xuanzhe Liu, Xin Jin

Pyxix: Scheduling Mixed Tasks in Disaggregated Datacenters

IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS (TPDS 2024)

Jun. 2024

- Sheng Qi, **Chao Jin**, Mosharaf Chowdhury, Zhenming Liu, Gang Huang, Xuanzhe Liu, Xin Jin

Jolteon: Unleashing the Promise of Serverless for Serverless Workflows

Santa Clara, CA, U.S.

USENIX SYMPOSIUM ON NETWORKED SYSTEMS DESIGN AND IMPLEMENTATION (NSDI 2024)

Apr. 2024

- Zili Zhang, **Chao Jin**, Xin Jin

Fast, Approximate Vector Queries on Very Large Unstructured Datasets

USENIX SYMPOSIUM ON NETWORKED SYSTEMS DESIGN AND IMPLEMENTATION (NSDI 2023)

Boston, MA, U.S.

Apr. 2023

- Zili Zhang, **Chao Jin**, Linpeng Tang, Xuanzhe Liu, Xin Jin

Melon: Breaking the Memory Wall for Resource-Efficient On-Device Machine Learning

Portland, OR, U.S.

ACM INTERNATIONAL CONFERENCE ON MOBILE SYSTEMS, APPLICATIONS, AND SERVICES (MOBISYS 2022)

June 2022

- Qipeng Wang, Mengwei Xu, **Chao Jin**, Xinran Dong, Jinliang Yuan, Gang Huang, Xin Jin, Yunxin Liu, Xuanzhe Liu

Teaching

Operating Systems (Honor Track)

TEACHING ASSISTANT

Peking University

Feb. 2024 - Jun. 2024

Introduction to Computer System

TEACHING ASSISTANT

Peking University

Sep. 2021 - Jan. 2022

Services

IEEE Transactions on Computers (TC)

REVIEWER

Jan. 2026

IEEE Transactions on Services Computing (TSC)

REVIEWER

Sep. 2025

Honors & Awards

UNIVERSITY AWARDS

- 2024 **Merit Student**, Peking University Award
- 2024 **Huawei Scholarship**, Peking University Award
- 2023 **Presidential Scholarship**, Peking University Award
- 2021 **Learning Excellence Award**, Peking University Award
- 2020 **Merit Student**, Peking University Award
- 2020 **Xiaomi Scholarship**, Peking University Award

Peking University

Peking University

Peking University

Peking University

Peking University

Peking University