Chengzhi Lu

CONTACT INFORMATION

Phone: 86 18565796826
Email: chengzhilu1994@gmail.com
NTU Singapore

EDUCATION

University of Chinese Academy of Sciences, China, (Sep. 2019 – Jan. 2025)

Ph.D., Computer Science
Advisor: Prof. Chengzhong Xu, Dean of FST, University of Macau

Zhejiang University, China, (Sep. 2016 - Sep. 2018)

M.S., Software Engineering

Wuhan University, China, (Sep. 2012 - Jun. 2016)

B.S., Computer Science

CONFERENCE PUBLICATIONS

[**EuroSys'26**] Yanying Lin, Shijie Peng, **Chengzhi Lu**, ChengZhong Xu, and Kejiang Ye. FlexPipe: Adapting Dynamic LLM Serving Through Inflight Pipeline Refactoring in Fragmented Serverless Clusters.

[**EuroSys'25**] Wenyan Chen, **Chengzhi Lu**, Huanle Xu, Kejiang Ye, and Cheng-Zhong Xu. Multiplexing Dynamic Deep Learning Workloads with SLO-awareness in GPU Clusters.

[**SC'24**] **Chengzhi Lu**, Huanle Xu, Yudan Li, Wenyan Chen, Kejiang Ye, and Chengzhong Xu. SMIless: Serving DAG-based Inference with Dynamic Invocations under Serverless Computing.

[**ICWS'24**] Yanying Lin, Shijie Peng, Shuaipeng Wu, Yanbo Li, **Chengzhi Lu**, Chengzhong Xu, Kejiang Ye. Planck: Optimizing LLM Inference Performance in Pipeline Parallelism with Fine-Grained SLO Constraint.

[**EuroSys'23**] **Chengzhi Lu***, Huanle Xu*, Kejiang Ye, Guoyao Xu, Liping Zhang, Guodong Yang, and Chengzhong Xu. Understanding and Optimizing Workloads for Unified Resource Management in Large Cloud Platforms.

[**SoCC'21**] Shutian Luo*, Huanle Xu*, **Chengzhi Lu**, Kejiang Ye, Guoyao Xu, Liping Zhang, Yu Ding, Jian He, Chengzhong Xu. Characterizing Microservice Dependency and Performance: Alibaba Trace Analysis (**Best Paper Award**).

[**CLUSTER'21**] Wenyan Chen, **Chengzhi Lu**, Kejiang Ye, Yang Wang, Chengzhong Xu. RPTCN: Resource Prediction for High-dynamic Workloads in Clouds based on Deep Learning.

[**ICPADS'19**] **Chengzhi Lu**, Kejiang Ye, Wenyan Chen, Chengzhong Xu. ADGS: Anomaly Detection and Localization based on Graph Similarity in Container-based Clouds.

[**ICPADS'18**] Kejiang Ye, Yanmin Kou, **Chengzhi Lu**, Yang Wang, Chengzhong Xu. Modeling Application Performance in Docker Containers Using Machine Learning Techniques.

[**Big Data'17**] **Chengzhi Lu**, Kejiang Ye, Guoyao Xu, Cheng-Zhong Xu, Tongxin Bai. Imbalance in the cloud: an analysis on Alibaba cluster trace.

JOURNAL PUBLICATIONS

[**TPDS,2022**] Shutian Luo*, Huanle Xu*, **Chengzhi Lu**, Kejiang Ye, Guoyao Xu, Liping Zhang, Jian He, Chengzhong Xu. An in-depth study of microservice call graph and runtime performance.

[**JCST,2020**] Wen-Yan Chen, Ke-Jiang Ye, **Chengzhi-Lu**, Dong-Dai Zhou, Cheng-Zhong Xu Interference analysis of co-located container workloads: a perspective from hardware performance

counters.

Recent Research Projects

System Optimization for Supporting Efficient AI Applications

- Improve performance of LLM inference service by co-optimizing hardware resources and application configurations.

Cloud Computing Resource Management, Performance Optimization

- Design scheduling algorithms with an optimistic fault tolerance mechanism to improve throughput and cluster resource utilization.

Resource Allocation Algorithms for Large-Scale Clusters with Performance Guarantees

- Design ML-based resource configuration strategy for the large-scale co-located cluster.

Academic Experience

NTU Singapore
Postdoc — Sep. 2025 - Now

- Resource management for LLM serving.

University of Macau
Postdoc — Jan. 2025 - Aug. 2025

- Resource allocation for inference serving in GPU cluster.
- Request scheduling for LLM applications.

University of Macau
Research Assitant — Feb. 2022 - Dec. 2024

- Resource allocation algorithms for large-scale clusters with performance guarantees.
- Resource assignment for serverless applications
- Request scheduling for LLM applications.

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
Visting Student — Jul. 2017 - Sep. 2019

- Distributed system resources management, including microservices and DL distributed training.
- Resource allocation algorithms for large-scale clusters with performance guarantees.
- Anomaly detection algorithm for large-scale clusters.
- Server cluster administrator, manages more than 140 physical servers, and provides experimental environment and services including Docker, GPU, Spark, Kubernetes, etc.

Work Experience

Alibaba Group, Zhejiang, China
Academic Intern — Jun. 2020 - Feb. 2023

- Participate in the functional development of Alibaba's infrastructure platform.

Honors and Awards

Dean's award for academic performance, SIAT, Chinese Academy of Sciences — Jan. 2021, 2024

Outstanding student, University of Chinese Academy of Sciences, — Jan. 2022

Outstanding student, SIAT, Chinese Academy of Sciences — Feb. 2019

	Outstanding graduate, Zhejiang University	Jul. 2018
SERVICES	Teaching Assistant of Computer Network, SIAT, Chinese Academy of Sciences	2021
SKILLS	Python, GO, Git, Latex, Java Kubernetes, Microservice, Serverless, Docker	