Yunlong Cheng

Shanghai Jiao Tong University, Shanghai, 200240, China

Phone: (+86) 17621813353 Email: aweftr@sjtu.edu.cn Homepage: aweftr.github.io

EDUCATION

Shanghai Jiao Tong University (SJTU), Ph.D. Program 2021/09 - Present

Major: Computer Science and Technology, Advisor: Xiaofeng Gao, GPA: 3.77/4.0

Research Interests: Cloud Computing, Data Mining, Prediction, and Scheduling Algorithms

Shanghai Jiao Tong University (SJTU), Bachelor's Degree 2017/09 - 2021/06

Major: Computer Science and Technology

Overall GPA: 89.67 / 100 (3.85 / 4.30), Major GPA: 89.75 / 100 (3.85 / 4.30), Rank: 28/148

PUBLICATIONS

- **Yunlong Cheng**, Hongji Dong, Tianyao Shi, Xiaofeng Gao and Guihai Chen, An Interference-Aware QoS Violation Alleviation Framework for Multi-Tenancy Public Clouds, 2025. (Preparing for submission to TPDS)
- Jiaxi Wu*, **Yunlong Cheng***, Jibin Wang, Wenquan Yang, Xiaofeng Gao and Guihai Chen, Online Virtual Machine Provisioning under Uncertainty: A Robust Approach to Ultimate Resource Utilization, submitted to IEEE Cybernetics, 2025. (Under review)
- **Yunlong Cheng**, Tin Ping Chan, Jiadong Chen and Xiaofeng Gao, PADMA: Prediction-Aware Dynamic VM Scheduling in Multi-NUMA Clouds with Transformer-based PPO, submitted to INFOCOM, 2026. (Under review)
- Hongji Dong, **Yunlong Cheng**, Tin Ping Chan, Xiaofeng Gao and Guihai Chen, TRACE: A Targeted Recommender for VM Assignment in Cloud Environment, IEEE International Conference on Cluster Computing (CLUSTER), Edinburgh, Scotland, September 2-5, 2025. (To Appear)
- Jiale Zhang, Yang Luo, **Yunlong Cheng**, Leixia Wang, Xiaofeng Gao, Xiaochun Yang and Guihai Chen, Timely Watchers: Cost-Effective Schedule for Urban Sensor Patrols, The 19th International Conference on Wireless Artificial Intelligent Computing Systems and Applications (WASA), Tokyo, Japan, June 24-26, 2025. (**Best Paper Award**)
- **Yunlong Cheng*,** Tianyao Shi*, Xiuyuan Wei, Yulong Song, Xiaofeng Gao, Zhipeng Bian, and Zhenli Sheng, Predicting Enterprise Users' Consuming Potential for Cloud Services, In 30th International Conference on Database Systems for Advanced Applications (DASFAA), Singapore, May 26-29, 2025.
- Tin Ping Chan, **Yunlong Cheng**, Yizhan Zhu, Xiaofeng Gao and Guihai Chen, Symmetry-Preserving Architecture for Multi-NUMA Environments (SPANE): A Deep Reinforcement Learning Approach for Dynamic VM Scheduling, IEEE International Conference on Computer Communications (INFOCOM), London, United Kingdom, May 19-22, 2025.
- Yucen Gao, **Yunlong Cheng**, Chan Tin Ping, Xiaofeng Gao and Guihai Chen, RL with Balanced Reward and Masking Mechanism for Multi-NUMA Virtual Machine Scheduling, Conference on Ubiquitous Information Management and Communication (IMCOM), Thailand, January 3-5, 2025.
- **Yunlong Cheng**, Xiuqi Huang, Zifeng Liu, Jiadong Chen, Xiaofeng Gao, Zhen Fang and Yongqiang Yang, FEDGE: An Interference-Aware QoS Prediction Framework for Black-Box Scenario in IaaS Clouds with Domain Generalization, The 38th International Parallel & Distributed Processing Symposium (IPDPS), San Francisco, California, May 27-31, 2024.
- Tianyao Shi, Yingxuan Yang, **Yunlong Cheng**, Xiaofeng Gao, Zhen Fang and Yongqiang Yang. Alioth: A Machine Learning Based Interference-Aware Performance Monitor for Multi-Tenancy Applications in Public Cloud. In 37th International Parallel and Distributed Processing Symposium (IPDPS), IEEE, St. Petersburg, Florida USA, May 15-19, 2023.
- **Yunlong Cheng**, Hao Zhou, Xiaofeng Gao, Jiaqi Zheng and Guihai Chen, Optimizing Incremental SDN Upgrades for Load Balancing in ISP Networks, Theoretical Computer Science (TCS), 962, p.113927, 2023.
- **Yunlong Cheng**, Hao Zhou, Xiaofeng Gao, Jiaqi Zheng and Guihai Chen. Incremental SDN Deployment to Achieve Load Balance in ISP Networks. In Algorithmic Aspects in Information and Management (AAIM), Guangzhou, China, August 13–14, 2022.
- Xiuqi Huang, **Yunlong Cheng**, Xiaofeng Gao and Guihai Chen. TEALED: A Multi-Step Workload Forecasting Approach Using Time-Sensitive EMD and Auto LSTM Encoder-Decoder. In Database Systems for Advanced Applications: 27th International Conference (DASFAA), Virtual Event, April 11–14, 2022.

RESEARCH EXPERIENCES

Serving LLMs over Heterogeneous GPUs and Network 2025/04 – 2025/07

- Abstract heterogeneous GPU clusters as directed, weighted graphs, where nodes correspond to GPU instances and edges represent network connections. Formulate LLM serving as a Max-Flow problem, incorporating both model placement across nodes and dynamic pipeline assignment of requests, with the objective of maximizing LLM serving throughput.

QoS-aware Scheduling Technology for Public Clouds 2023/06 – 2024/12

- Schedule the virtual machines (VMs) while considering the interference-aware QoS prediction method in both offline and online cloud scheduling environments. Enhance overall resource utilization in the cloud while maximizing user satisfaction.
- Optimize VM scheduling through symmetric reinforcement learning, online packing algorithms, and robust optimization, ensuring that the scheduling algorithms perform well when predictions are accurate and remain robust when predictions are inaccurate.

Interference-Aware QoS Prediction for Public Cloud Service 2021/01 – 2022/07

- Design an interference-aware data collection framework aimed at optimizing QoS prediction; automatically select the most important features using Stochastic Gates, an embedded feature selection method, to reduce the overhead of data collection.
- Propose a framework based on multi-domain Maximum Mean Discrepancy and adversarial denoising autoencoder to predict QoS degradation of co-located VMs in black-box scenario within IaaS clouds, ensuring broad generalizability.

Workload Forecasting for System Service 2019/10 – 2020/05

- Improve empirical mode decomposition method to process workload curve.
- Optimize workload forecasting using an Encoder-Decoder architecture and Neural Architecture Search methods.

Learning Index and Query Optimization for Database Systems 2019/07 – 2019/09

- Collect and process query and insert data from database systems; propose an innovative method using Deep Deterministic Policy Gradient algorithm in reinforcement learning to dynamically predicts and constructs database indexes.

AWARDS

- Merit Student of SJTU (Top 1% in department, 113 people in total) 2021 – 2022
- First-Class Postgraduate Academic Scholarships in Shanghai Jiao Tong University 2021 – 2024
- Second Prize, The Chinese Mathematics Competitions (Top 20%) 2019/12
- Second Prize, Contemporary Undergraduate Mathematical Contest in Modeling (Top 20%) 2018/11
- Class C Scholarship in Shanghai Jiao Tong University (Top 20% in department, 148 people in total) 2018 – 2019
- Class B Scholarship in Shanghai Jiao Tong University (Top 10% in department, 148 people in total) 2017 – 2018

ACADEMIC SERVICES

- International Computing and Combinatorics Conference (COCOON 2024) volunteer 2024/08
- International Conference on Data Mining (ICDM 2023) volunteer 2023/12
- External Reviewer of Journals and Conferences
  - Journal of Parallel and Distributed Computing (JPDC), Theoretical Computer Science (TCS), IEEE/ACM Transactions on Networking (TNET), IEEE Transactions on Network and Service Management (TNSM), CCF National Database Conference (NDBC), IEEE International Conference on Big Data (IEEE BigData), European Conference on Artificial Intelligence (ECAI)

SKILLS AND HOBBIES

- Programming Languages: Python, C/C++, MATLAB, Lua, CUDA, OpenMP, MPI
- Hobbies: Table tennis, Badminton, Coffee, FPV drone, Photography