Qizhen Weng

Ph.D. in Computer Science and Engineering
Systems Researcher in AI Infrastructure
Shanghai AI Laboratory, Xuhui District, Shanghai, China

Email: qwengaa@cse.ust.hk
Web: https://qzweng.github.io

Research Interests

My interests cover **AI infrastructure**, **Machine Learning Systems** and **Cloud Computing**, especially on **large-scale model training**, **inference**, and **fine-tuning**.

Education

Hong Kong University of Science and Technology, Hong Kong, China
Ph.D. *(HKPFS)*, Computer Science and Engineering, GPA: 4.2/4.3 Sep. 2017 – Dec. 2022

University of California, Berkeley, CA, US
Exchange student, GPA: 4.0/4.0 Jun. 2015 – Aug. 2015

Shanghai JiaoTong University, Shanghai, China
B.Eng. *(Shanghai Outstanding Graduates)*, Cyber Security, GPA: 3.8/4.3 Sep. 2013 – Jun. 2017

Experiences

Shanghai AI Laboratory, Shanghai, China
Systems Researcher, Nov. 2022 – present
· Large Language Model Training, Inference, and Fine-Tuning (InternLM 2)
· Efficient Massive Low-Rank Adapters Serving for LLM Inference (arXiv '24)

Alibaba CTO Line & Alibaba Cloud, Hangzhou, China
Research Intern, Jun. 2020 – Oct. 2022
· Defragment Resources in Heterogeneous GPU Cluster (ATC '23)
· Characterize AI Workloads in Production Clusters and Improve Scheduling (NSDI '22)
· Simulate and Improve ML Job Scheduling (released codes and traces)

Computer Science and Engineering Dept. in HKUST, Hong Kong, China
Ph.D. Candidate, Sep. 2017 – Dec. 2022
· Schedule Applications in Shared Clusters with Reinforcement Learning (SC '20)
· Coordinate Workers for More Efficient Distributed Model Training (SoCC '20, TCC '21)
· Characterize Dataflow Computation Performance with Learning Methods (APSys '19)

Publications

Suyi Li, Hanfeng Lu, Tianyuan Wu, Minchen Yu, **Qizhen Weng**, Xusheng Chen, Yizhou Shan, Binhang Yuan, and Wei Wang.
"CaraServe: CPU-Assisted and Rank-Aware LoRA Serving for Generative LLM Inference," in the *arXiv preprint arXiv:2401.11240*, Jan. 2024.

Qizhen Weng*, Lingyun Yang*(co-first author), Yinghao Yu, Wei Wang, Xiaochuan Tang, Guodong Yang, and Liping Zhang.
"Beware of Fragmentation: Scheduling GPU-Sharing Workloads with Fragmentation Gradient Descent," in the Proc. *USENIX ATC '23*, Jul. 2023.

Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding.
"MLaaS in the Wild: Workload Analysis and Scheduling in Large-Scale Heterogeneous GPU Clusters," in the Proc. *USENIX NSDI '22*, Apr. 2022.

Yongkang Zhang, Yinghao Yu, Wei Wang, Qiukai Chen, Jie Wu, Zuowei Zhang, Jiang Zhong, Tianchen Ding, **Qizhen Weng**, Lingyun Yang, Cheng Wang, Jian He, Guodong Yang, and Liping Zhang.
"Workload Consolidation in Alibaba Clusters: The Good, the Bad, and the Ugly," in the Proc. *ACM SoCC '22*, Nov. 2022.

Chen Chen, **Qizhen Weng**, Wei Wang, Baochun Li, and Bo Li.
"Accelerating Distributed Learning in Non-Dedicated Environments," in the *IEEE TCC '21*, Jul. 2021.

Luping Wang*, **Qizhen Weng*(co-first author)**, Wei Wang, Chen Chen, and Bo Li.
"Metis: Learning to Schedule Long-Running Applications in Shared Container Clusters at Scale," in the Proc. *IEEE/ACM SC '20*, Nov. 2020.

Chen Chen, **Qizhen Weng**, Wei Wang, Baochun Li, and Bo Li.
"Semi-Dynamic Load Balancing: Efficient Distributed Learning in Non-Dedicated Environments," in the Proc. *ACM SoCC '20*, Oct. 2020.

Huangshi Tian, **Qizhen Weng**, and Wei Wang.
"Towards Framework-Independent, Non-Intrusive Performance Characterization for Dataflow Computation," in the Proc. *ACM APSys '19*, Aug. 2019.

Chen Chen, **Qizhen Weng**, Wei Wang, Baochun Li and Bo Li.
"Fast Distributed Deep Learning via Worker-adaptive Batch Sizing," poster paper in the Proc. *ACM SoCC '18*, Oct. 2018.

Yinghao Yu, Wei Wang, Jun Zhang, **Qizhen Weng**, and Khaled Ben Letaief.
"OpuS: Fair and Efficient Cache Sharing for In-Memory Data Analytics," in the Proc. *IEEE ICDCS '18*, Jul. 2018.

Awards and Fellowships

Hong Kong PhD Fellowship
Research Grants Council (RGC) of Hong Kong, 2017 - 2020

Shanghai Outstanding Graduates (2%)
Shanghai Jiao Tong University, 2017

Cyber-Security Scholarship
China Internet Development Foundation (CIDF), 2016

China Aerospace Science and Technology Corporation Scholarship
China Aerospace Science and Technology Corporation (CASC), 2015

Skills

Python, Golang, C/C++, PyTorch, TensorFlow,
Kubernetes, Ray, DeepSpeed, ColossalAI, Spark

References

Wei Wang, Associate Professor
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Web: https://www.cse.ust.hk/~weiwa/