Minchen Yu

Present Position	Assistant Professor School of Data Science The Chinese University of Hong Kong, Shenzhen	*Office*: Daoyuan 420c *Email*: yuminchen@cuhk.edu.cn *Web*: https://mincyu.github.io

Research Interests: Cloud computing, distributed systems

Education

Hong Kong University of Science and Technology, Hong Kong SAR, China
Department of Computer Science and Engineering

◊ **Ph.D.**, Computer Science and Engineering, August 2023
  ◊ *Dissertation:* "Towards Usable, Efficient Serverless Computing Systems"
  ◊ *Advisor:* Prof. Wei Wang

Nanjing University, Nanjing, Jiangsu, China
Software Institute

◊ **B.Eng.**, Software Engineering, July 2018
  ◊ NJU Outstanding Graduate Award

Honors and Awards

◊ Shenzhen Pengcheng Peacock Plan (Class-C)	2024
◊ HKUST RedBird Academic Excellence Award, HKUST	2023
◊ Best Paper Runner-Up Award, IEEE ICDCS	2021
◊ SENG Academic Award for Continuing PhD Students, HKUST	2020
◊ Huawei PhD Fellowship	2018 - 2021
◊ University Outstanding Graduate, Nanjing University	2018
◊ Chinese National Scholarship	2016
◊ Excellent Student Awards, Nanjing University	2015

Professional Experience

The Chinese University of Hong Kong, Shenzhen, Shenzhen, China
Assistant Professor (tenure-track) December 2023 – Present

Hong Kong University of Science and Technology, Hong Kong SAR, China
Post-Doctoral Fellow September – November 2023

Hong Kong University of Science and Technology, Hong Kong SAR, China
Research/Teaching Assistant January 2018 – August 2023

Alibaba Cloud, Hangzhou, China
Research Intern December 2021 – June 2023

Huawei Hong Kong Research Center, Hong Kong SAR, China
Research Intern October 2020 – March 2021

Morgan Stanley IT Department, Shanghai, China
Software Development Engineer (intern) July – September 2017

Publications

All publications are sorted in a reverse chronological order.

Refereed Papers in Conference and Workshop Proceedings

[C7] **Minchen Yu**, Tingjia Cao, Wei Wang, Ruichuan Chen, "Following the Data, Not the Function: Rethinking Function Orchestration in Serverless Computing," in the *Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI '23)*, Boston, MA, April 2023.

[C6] **Minchen Yu**, Tingjia Cao, Wei Wang, Ruichuan Chen, "Following the Data, Not the Function: Rethinking Function Orchestration in Serverless Computing," extended abstract in the *proceedings of the 3rd Workshop On Resource Disaggregation and Serverless Computing (WORDS'22)*, Nov, 2022.

[C5] **Minchen Yu**, Zhifeng Jiang, Hok Chun Ng, Wei Wang, Ruichuan Chen, Bo Li, "Gillis: Serving Large Neural Networks in Serverless Functions with Automatic Model Partitioning," in the *Proceedings of the 41st IEEE International Conference on Distributed Computing Systems (ICDCS'21)*, Virtual Conference, July 2021. (**Best Paper Runner Up**)

[C4] Huangshi Tian, **Minchen Yu**, Wei Wang, "CrystalPerf: Resource-Centric Performance Characterization for Dataflow Computation," in the *proceedings of USENIX Annual Technical Conference (ATC'21)*, Virtual Conference, July 2021.

[C3] **Minchen Yu**, Yinghao Yu, Yunchuan Zheng, Baichen Yang, Wei Wang, "RepBun: Load-Balanced, Shuffle-Free Cluster Caching for Structured Data," in the *proceedings of IEEE INFOCOM'20*, Virtual Conference, July 2020.

[C2] Chengliang Zhang, **Minchen Yu**, Wei Wang, Feng Yan, "MArk: Exploiting Cloud Services for Cost-Effective, SLO-Aware Machine Learning Inference Serving," in the *proceedings of USENIX Annual Technical Conference (ATC'19)*, Renton, WA, July 2019.

[C1] Huangshi Tian, **Minchen Yu**, Wei Wang, "Continuum: A Platform for Cost-Aware, Low-Latency Continual Learning," in the *proceedings of ACM Symposium on Cloud Computing (SoCC'18)*, Carlsbad, CA, October 2018.

Refereed Journal Articles

[J2] **Minchen Yu**, Tingjia Cao, Wei Wang, Ruichuan Chen, "Pheromone: Restructuring Serverless Computing with Data-Centric Function Orchestration," accepted in *IEEE/ACM Transactions on Networking (ToN)*, Oct 2024.

[J1] Chengliang Zhang, **Minchen Yu**, Wei Wang, Feng Yan, "Enabling Cost-Effective, SLO-Aware Machine Learning Inference Serving on Public Cloud," in *IEEE Transactions on Cloud Computing (TCC)*, June 2020.

Patents

[P2] Ao Wang, **Minchen Yu**, Wei Wang, Dong Chen, "Data processing method based on server non-perception calculation and electronic equipment," China Patent CN117499494A, 2023.

[P1] Ao Wang, **Minchen Yu**, "Method, equipment and storage medium for processing function call request," China Patent CN116501485A, 2023.

Preprints

[E5] Kaiyu Huang, Hao Wu, Zhubo Shi, Han Zou, **Minchen Yu**, Qingjiang Shi, "SpecServe: Efficient and SLO-Aware Large Language Model Serving with Adaptive Speculative Decoding," in *arXiv preprint arXiv:2503.05096*.

[E4] **Minchen Yu***, Rui Yang*, Chaobo Jia, Zhaoyuan Su, Sheng Yao, Tingfeng Lan, Yuchen Yang, Yue Cheng, Wei Wang, Ao Wang, Ruichuan Chen, "λScale: Enabling Fast Scaling for Serverless Large Language Model Inference," in *arXiv preprint arXiv:2502.09922*. (*Co-first)

[E3] Hao Wu, Junxiao Deng, **Minchen Yu**, Yue Yu, Yaochen Liu, Hao Fan, Song Wu, Wei Wang, "FaaSTube: Optimizing GPU-oriented Data Transfer for Serverless Computing," in *arXiv preprint arXiv:2411.01830*.

[E2] Suyi Li, Hanfeng Lu, Tianyuan Wu, **Minchen Yu**, Qizhen Weng, Xusheng Chen, Yizhou Shan, Binhang Yuan, Wei Wang, "CaraServe: CPU-Assisted and Rank-Aware LoRA Serving for Generative LLM Inference," in *arXiv preprint arXiv:2401.11240*.

[E1] **Minchen Yu**, Ao Wang, Dong Chen, Haoxuan Yu, Xiaonan Luo, Zhuohao Li, Wei Wang, Ruichuan Chen, Dapeng Nie, Haoran Yang, "FaaSwap: SLO-Aware, GPU-Efficient Serverless Inference via Model Swapping," in *arXiv preprint arXiv:2306.03622*.

External Grants

[G2] Principal Investigator, "End-to-End Optimizations for Large Language Model Inference," Huawei Gifted Fund, 2024-27 (amount: 600,000 CNY).

[G1] Principal Investigator, "Towards Unified, High-Performance Data Passing Framework for Heterogeneous Serverless Functions," CCF-Huawei Populus Grove Fund, 2024-25 (amount: 300,000 CNY).

Professional Services

Membership in Program Committee

- The 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 2025)
- The 45th IEEE International Conference on Distributed Computing Systems (ICDCS 2025) Cloud Computing Track

Reviewer for Journal Manuscript Submissions

- IEEE/ACM Transactions on Networking (ToN)

◊ IEEE Transactions on Parallel and Distributed Systems (TPDS)

◊ IEEE Transactions on Cloud Computing (TCC)

External Reviewer for Conference Manuscript Submissions

IEEE INFOCOM, IEEE ICNP, IEEE ICDCS, IEEE/ACM IWQoS, IEEE GLOBECOM

Invited Talks (2024-Present)

[T4] "FaaSwap: Usable, Resource-Efficient Serverless AI inference", Longgang District Government AI Application and Data Elements Solution Conference, Shenzhen, 30 December, 2024.

[T3] "Towards Usable, Efficient Serverless AI Inference Systems", The 9th Chinese Youth Congress on Artifical Intelligence, Shenzhen, 21-22 December, 2024.

[T2] "GRouter: Efficient, Unified Data Passing for Serverless Inference", Huawei Sentosa Software Technology Summit, Singapore, August 2024.

[T1] "Towards Usable, Efficient Serverless Computing Systems", Huawei Strategy and Technology Workshop (STW), Shenzhen, May 2024.

Graduate Supervision

Current Students in the Doctor of Philosophy Program

Name	*Duration of Study*
Ye Wang	September 2024 – Present
Kaiyu Huang (Co-supervised with Prof. Qingjiang Shi)	January 2024 – Present

Ph.D. Thesis Defense Committees

Renwen Ma, 2025
Lele Li, 2024

Ph.D. Thesis Proposal Defense Committees

Yuxuan Liu, 2024

Teaching Experience

The Chinese University of Hong Kong, Shenzhen (Instructor)

◊ CSC4303: *Network Programming* — Spring 2025

◊ CSC4160: *Cloud Computing* — Fall 2024

◊ CSC4303/CSC6203: *Network Programming* — Spring 2024

Hong Kong University of Science and Technology (Teaching Assistant)

◊ COMP4651: *Cloud Computing and Big Data Systems* — Spring 2021, Spring 2022

◊ COMP3511: *Operating System* — Spring 2019