

Minchen Yu

PRESENT POSITION	PhD Candidate Dept. of Computer Science and Engineering Hong Kong University of Science and Technology	Phone: +86-15651721686 Email: myuaj@cse.ust.hk Web: https://mincyu.github.io
EDUCATION	Hong Kong University of Science and Technology , Hong Kong SAR, China <i>Department of Computer Science and Engineering</i> ◇ Ph.D. , Computer Science and Engineering September 2018 – August 2023 (estimated) ◇ <i>Supervisor</i> : Prof. Wei Wang Nanjing University , Nanjing, Jiangsu, China <i>Software Institute</i> ◇ B.Eng. , Software Engineering September 2014 – July 2018 ◇ NJU Outstanding Graduate Award	
HONORS AND AWARDS	◇ HKUST RedBird Academic Excellence Award, HKUST	2023
	◇ Student Travel Grant, USENIX NSDI	2023
	◇ Best Paper Runner-Up Award, IEEE ICDCS	2021
	◇ SENG Academic Award for Continuing PhD Students, HKUST	2020
	◇ Huawei PhD Fellowship, HKUST	2018 - 2021
	◇ University Outstanding Graduate, Nanjing University	2018
	◇ National Scholarship (top 2%)	2016
	◇ People's Scholarship for Academic Excellence, Nanjing University	2015, 2017
	◇ Excellent Student Awards, Nanjing University	2015
PROFESSIONAL EXPERIENCE	Hong Kong University of Science and Technology , Hong Kong SAR, China <i>Research/Teaching Assistant</i> January 2018 – Present Alibaba Cloud , Hangzhou, China <i>Research Intern</i> December 2021 – Present Huawei Hong Kong Research Center , Hong Kong SAR, China <i>Research Intern</i> October 2020 – March 2021 Morgan Stanley IT Department , Shanghai, China <i>Software Development Engineer (intern)</i> July – September 2017	
PUBLICATIONS	Conference Papers (in reverse chronological order) [C7] Minchen Yu , Ao Wang, Dong Chen, Haoxuan Yu, Xiaonan Luo, Zhuohao Li, Wei Wang, Ruichuan Chen, Dapeng Nie, Haoran Yang, "FaaSv2: SLO-Aware, GPU-Efficient Serverless	

Inference via Model Swapping,” in *arXiv preprint arXiv:2306.03622*, June 2023. (under review)

[C6] **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.

[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.

Workshop Paper

[W1] **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.

Journal Article

[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.

TEACHING EXPERIENCE

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