

ZHANG, Yongkang

(Last updated on: June 28th, 2025)

Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong S.A.R., China

yzhangne@cse.ust.hk
(852)96745094/(86)17386243151
<https://ykzhang1999.github.io/>

RESEARCH AREAS

Areas: Cloud Computing; Containers; Resource Management; GPU Virtualization.

Focus: High-performance, resource-efficient GPU cloud platforms.

EDUCATION

Hong Kong University of Science and Technology

Hong Kong S.A.R., China

Ph.D. in Computer Science and Engineering

Sep. 2021 - Present

GPA: 3.77 / 4.30; HKPFS Awardee

Thesis Supervisors: Prof. WANG, Shuai and Prof. CHU, Xiaowen

Wuhan University

Wuhan, Hubei, China

B.Eng. in Computer Science and Technology

Sep. 2017 - Jun. 2021

GPA: 3.98 / 4.00; GPA Ranking: 2 / 334; Excellent Undergraduate Thesis

Thesis: Idle Memory Reclamation and Overcommitment on Cloud

Thesis Supervisor: Prof. ZHANG, Huyin

INDUSTRIAL EXPERIENCE

Alibaba Cloud

Hangzhou, Zhejiang, China

Research Intern of Cluster Management Group, Cloud Native Division

Oct. 2020 - Jul. 2021

Mentor: HE, Jian

Microsoft Research Asia

Beijing, China

Research Intern of Networking Research Group

Jul. 2020 - Oct. 2020

Mentors: Dr. CHENG, Wenxue and Dr. CHENG, Peng

PUBLICATIONS

Conferences

- 2025 **Yongkang Zhang**, Haoxuan Yu, Chenxia Han, Cheng Wang, Baotong Lu, Yunzhe Li, Zhifeng Jiang, Yang Li, Xiaowen Chu, and Huaicheng Li, “SGDRC: Software-Defined Dynamic Resource Control for Concurrent DNN Inference on NVIDIA GPUs,” in *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (ACM PPoPP ’25)*, Las Vegas, NV, March 2025. (**Acceptance Rate: 20.1% = 38/189**)
- 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 Proceedings of ACM Symposium on Cloud Computing (ACM SoCC ’22)*, San Francisco, CA, November 2022. (**Acceptance Rate: 24.5% = 38/155**)

PATENTS

Method, Apparatus, Device, and Storage Medium for Allocating GPU VRAM Channels. *China Patent (Under substantive examination). Application No.: CN119938331A.*

ACADEMIC SERVICES

Reviewer / External Reviewer: IEEE Transactions on Cloud Computing, IEEE Internet of Things Journal, ACM ChinaSys (2024)

Artifact Evaluation Committee: IEEE HPCA (2024), ACM CCS (2025), ACM KDD (2025), ACM ASPLOS (2026), USENIX FAST (2026)

Shadow Program Committee: ACM EuroSys (2026)

TEACHING

Teaching Assistant: Cloud Computing and Big Data Systems (HKUST, 2022 & 2023), Computer Organization (HKUST, 2025)

SKILLS

Language: Chinese - Mandarin (Mother tongue); English (TOEFL: 113 / 120; CET-6: 683 / 710).

Programming: C++ / C, Go, Rust, Python, Java, Verilog HDL, Tensorflow, PyTorch

AWARDS

Awards Obtained in the Ph.D. Program

UGC Research Travel Grant, Research Office, HKUST	2025
RedBird Ph.D. Scholarship, School of Engineering, HKUST	2021 & 2022
Hong Kong Ph.D. Fellowship (<i>Only 300 Awardees in HK</i>), University Grant Council	2021 - 2025

Awards Obtained in the Undergraduate Program

Excellent Undergraduate Thesis, Wuhan University	2021
Sensetime Scholarship (Runner-up), Sensetime Group	2019
National Scholarship, The Ministry of Education	2018
The First Class Scholarship, Wuhan University	2018

Awards Obtained in Olympiad in Informatics (Organized by China Computer Federation)

Silver Medal, China Team Selection Competition	2016
Silver Medal, Winter Camp of National Olympiad in Informatics	2016
Bronze Medal, National Olympiad in Informatics	2016
Bronze Medal, Asia-Pacific Informatics Olympiad (China District)	2015 & 2016
First Prize, National Olympiad in Informatics in Provinces	2014 & 2015