Lingyun Yang (杨凌云)

Updated July 23, 2024

Email: lyangbk@cse.ust.hk **GitHub**: mental2008 **LinkedIn**: stephenyang1999
Phone: (+86) 135 0284 6103 **Office**: BDI 101, UC, HKUST **Web**: https://www.lingyunyang.com/

Profile

Ph.D. Candidate
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Research Interests

I am broadly interested in resource management for large-scale data centers, especially in improving resource efficiency for large-scale AI/GPU clusters and optimizing model serving systems.

Education

Hong Kong University of Science and Technology (HKUST)
Department of Computer Science and Engineering
Ph.D. in Computer Science and Engineering 2020 – Present
◇ Advisor: Prof. Wei Wang

South China University of Technology (SCUT)
School of Computer Science and Engineering
B.Eng. in Computer Science and Technology 2016 – 2020
◇ Studied at All-English Innovation Class (GPA: 3.82/4)

Professional Experience

Alibaba Group Hangzhou, China
Research Intern, Cluster Management Group Dec. 2020 – Present
◇ Mentor: Dr. Yinghao Yu

Mitigate GPU Resource Fragmentation

◇ **TBD.** GPU disaggregation for recommendation services.
◇ [C3] Formally defined *statistical GPU resource fragments* and proposed the *fragmentation gradient descent* scheduling algorithm to reduce resource fragmentation. Large-scale trace evaluations show that our scheduling policy can significantly improve GPU allocation rate by 3% compared to state-of-the-art policies.
◇ Developed *ParaSet*, a best-effort workload on Kubernetes that can dynamically adjust the number of instances and resource requirements based on the real-time resource availability in the cluster. It aims to fill resource fragments in the cluster. It is integrated into KubeDL for internal use.

Large-Scale GPU Sharing in Production

◇ Enabled *large-scale GPU sharing* in production clusters, with over 10k shared GPU containers running on a daily basis. Support the co-location of GPU tasks with different priorities (e.g., *latency-sensitive*, *best-effort*).

◇ This was a multi-team collaborative project. I was responsible for the design and implementation of the single-node agent and the centralized controller. The agent periodically collects and reports resource usage metrics, as well as dynamically allocates GPU resources to containers. The controller calculates potential overcommitment of resources and provides scheduling guidance to the scheduler.

Efficient Diffusion Model Serving System

◇ **TBD.**

Auto-Configuration for AI Serving Service

◇ [C1] Developed Morphling, an open-source auto-configuration framework for AI serving on Kubernetes. It combines meta-learning and bayesian optimization to quickly find the optimal configuration. Internally, it was widely used for automated recommendation of container resource specifications. [code]

Microsoft Research Asia (MSRA) Beijing, China

Research Intern, Innovation Engineering Group (IEG) Jul. 2019 – Jun. 2020

◇ Research on model robustness, face recognition, attention mechanisms, knowledge distillation, and neural architecture search.

Publications

** denotes co-first authors*

[C3] Qizhen Weng*, **Lingyun Yang***, Yinghao Yu, Wei Wang, Xiaochuan Tang, Guodong Yang, Liping Zhang, "Beware of Fragmentation: Scheduling GPU-Sharing Workloads with Fragmentation Gradient Descent," in the *Proceedings of USENIX Annual Technical Conference* (**ATC '23**), Boston, MA, USA, July 2023.

[C2] 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 Management in Alibaba Clusters: The Good, the Bad, and the Ugly," in the *Proceedings of ACM Symposium on Cloud Computing* (**SoCC '22**), San Francisco, CA, USA, November 2022.

[C1] Luping Wang*, **Lingyun Yang***, Yinghao Yu, Wei Wang, Bo Li, Xianchao Sun, Jian He, and Liping Zhang, "Morphling: Fast, Near-Optimal Auto-Configuration for Cloud-Native Model Serving," in the *Proceedings of ACM Symposium on Cloud Computing* (**SoCC '21**), Seattle, WA, USA, November 2021.

Preprint

[P1] Suyi Li, **Lingyun Yang***, Xiaoxiao Jiang, Hanfeng Lu, Zhipeng Di, Weiyi Lu, Jiawei Chen, Kan Liu, Yinghao Yu, Tao Lan, Guodong Yang, Lin Qu, Liping Zhang, Wei Wang, "SwiftDiffusion: Efficient Diffusion Model Serving with Add-on Modules," *arXiv preprint arXiv:2407.02031*, 2024.

Honors and Scholarships

◇ Postgraduate Scholarship 2020 – Present, HKUST

◇ Star of Tomorrow Internship Award of Excellence Jul. 2020, MSRA

◇ Merit Student & Excellent Student Cadre Nov. 2019, SCUT

◇ National Scholarship Oct. 2019, China

◇ Silver Medal, ICPC China Xian National Invitational Contest May 2019

◇ First Prize, 17th Guangdong Collegiate Programming Contest May 2019

◇ Silver Medal, 37Games Cup Programming Contest — Apr. 2019
◇ Gold Medal, SCUT ACM Programming Contest — Apr. 2019
◇ Bronze Medal, ACM-ICPC Asia Xuzhou Regional Contest — Oct. 2018
◇ Silver Medal, 1st Xiao Mi Collegiate Programming Contest — Sept. 2018
◇ Gold Medal, SCUT ACM Programming Contest — Apr. 2018
◇ The First Prize Scholarship — Nov. 2017, SCUT
◇ Bronze Medal, ACM-ICPC Asia Xian Regional Contest — Oct. 2017
◇ Gold Medal, 12th China Youth Robot Competition — Jul. 2012
◇ Champion, RoboCup Youth Robot World Cup, China Division — Mar. 2012

Academic Services

Artifact Evaluation Committee
◇ SIGCOMM (2024), HPCA (2024)
◇ SOSP (2023), OSDI (2023), ATC (2023), MLSys (2023)

External Reviewer
◇ INFOCOM (2022, 2023, 2024)
◇ ICDCS (2023), APSys (2021), MSN (2021), Qshine (2020)

Student Helper
◇ APNet (2023), ICMLC & ICWAPR (2018)

Teaching Activities

Hong Kong University of Science and Technology
Teaching Assistant, Department of Computer Science and Engineering
◇ CSIT6000O: Advanced Cloud Computing (Spring 2022, Spring 2023)
◇ COMP4651: Cloud Computing and Big Data Systems (Spring 2021, Fall 2021, Spring 2024)
◇ COMP3511: Operating Systems (Fall 2023)

Other Experience

ACM-ICPC Competition Group
Group Member & Team Leader — 2016 – 2019
◇ Coach: Prof. Chuhua Xian
◇ Major domains: Dynamic Programming, Number Theory, Data Structure, etc.

Machine Learning & Cybernetics Research Group
Undergraduate Research Assistant — 2017 – 2019
◇ Advisor: Prof. Patrick Chan
◇ Projects: Fundus Stitching, Tableware Recognition, and NN Visualization.

Tencent Innovation Club
Vice Chairman — 2018 – 2019
◇ Led the largest student club in SCUT CSE, sponsored by Tencent.

ByteDance Summer Camp — Beijing, China
Camper, Algorithm track — Aug. 2019
◇ Mentor: Dr. Yibo Zhu
◇ Totally 150 participants selected from more than 6k candidates.

Skills

Programming Languages: Golang, C++, Python
Toolkits: Kubernetes, Git, LaTeX, Linux Shell, Qt, MySQL, MarkDown

Languages: English (fluent), Mandarin (Native speaker)

Miscellaneous Play basketball & badminton & squash, workout at the gym, foodie.
My paper reading notes are available at https://paper.lingyunyang.com/.