Lingyun Yang (杨凌云)

Updated December 17, 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/

Research Interests

I have a broad interest in resource management for large-scale data centers / AI infrastructure. Specifically, my research focuses on: (a) improving *resource efficiency* for AI/GPU clusters; (b) building *efficient* and *low-cost* AI model serving systems. My first-authored papers have been published in top-tier systems conferences (NSDI, ATC, SoCC).

Education

Hong Kong University of Science and Technology (HKUST)
Department of Computer Science and Engineering
Ph.D. in Computer Science and Engineering — Sep. 2020 – Present
◇ Advisor: Prof. Wei Wang — (*expected to graduate in Fall 2025*)

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

Publications

** denotes co-first authors, sort in alphabetical order*

Refereed Papers in Conference Proceedings

[C4] **Lingyun Yang**, Yongchen Wang, Yinghao Yu, Qizhen Weng, Jianbo Dong, Kan Liu, Chi Zhang, Yanyi Zi, Hao Li, Zechao Zhang, Nan Wang, Yu Dong, Menglei Zheng, Lanlan Xi, Xiaowei Lu, Liang Ye, Guodong Yang, Binzhang Fu, Tao Lan, Liping Zhang, Lin Qu, Wei Wang, "GPU-Disaggregated Serving for Deep Learning Recommendation Models at Scale," in the *Proceedings of the 22nd USENIX Symposium on Networked Systems Design and Implementation* (**NSDI '25**), Philadelphia, PA, USA, April 2025. *(**CCF-A**, acceptance rate: 55/401=13.7%)*

[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. *(**CCF-A**, acceptance rate: 65/353=18.4%)*

[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. *(**CCF-B**, acceptance rate: 38/155=24.5%)*

[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. ***(CCF-B**, acceptance rate: 46/145=31.7%)*
In submission / Preprint

[I1] 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.

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

2024 – 2024 **Efficient Text-to-Image Diffusion Model Serving with Add-on Modules**
◇ Developed SwiftDiffusion, a system that efficiently generates high-quality images with stable diffusion models and add-on modules (i.e., ControlNets and LoRAs).
◇ Analyzed usage patterns of various add-on modules (ControlNets, LoRAs) based on **500k** request traces from production text-to-image services.
◇ Incorporated several novel system designs, e.g., ControlNet-as-a-Service, async bounded LoRA loading, latent parallelelism for CFG computation.
◇ Achieved up to **7.8**× in serving latency and **1.6**× in throughput.
◇ One paper is in submission and under review [arXiv].

2022 – 2023 **GPU-Disaggregated Serving for Deep Learning Recommendation Models at Scale**
◇ Proposed a GPU-disaggregated DLRM serving system to eliminate *resource mismatch* and meet *elastic* demand; leveraged RDMA network to offload computation on separate GPU and CPU nodes; resource-aware graph partitioning; topology-aware scheduling.
◇ In daily scenarios (e.g., a crowded GPU cluster with > 90% allocation rate), reduced CPU fragments by **53%** and GPU fragments by **27%**. In the Double 11 Shopping Festival, saved up to **90%** of GPUs when loaning GPU servers from training clusters.
◇ One paper was accepted by **NSDI '25**.

2022 – 2023 **Resource Fragmentation Analysis and Optimization for GPU-Sharing Clusters**
◇ Formally quantified *statistical GPU resource fragments* and proposed the *fragmentation gradient descent* scheduling algorithm to reduce resource fragmentation. Our scheduling policy can significantly reduce *unallocated* GPUs by up to **49%** compared to state-of-the-art policies. [code] [trace]
◇ Developed ParaSet, a *best-effort* workload on Kubernetes that dynamically adjusts 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 and is integrated into KubeDL for internal use.
◇ One paper was accepted by **ATC '23**.

2021 – 2022 **Large-Scale GPU Sharing and Overcommitment in Production**

◇ Enabled *large-scale GPU sharing and overcommitment* in production clusters, with *tens of thousands* of *shared* GPU containers running daily. Support the co-location of GPU tasks with different priorities (e.g., *latency-sensitive*, *spot*, *best-effort*).

◇ Specifically, I designed and implemented the *node-level* agent and the *cluster-level* controller. The agent periodically collects and reports resource usage metrics and *dynamically* allocates GPU resources to containers. The controller calculates potential resource overcommitment and provides scheduling guidance to the cluster scheduler.

2020 – 2021 **Fast, Near-Optimal Auto-Configuration for Cloud-Native Model Serving**

◇ Developed Morphling, an auto-configuration framework for AI serving on Kubernetes; combined *meta-learning* and *bayesian optimization* to quickly find the *optimal* resource configuration (e.g., CPU cores, GPU timeshare, GPU memory, GPU type) and runtime parameters (e.g., batch size). [code]

◇ It was widely used in Alibaba for automated recommendation of container resource specifications; part of Alibaba's open-sourced KubeDL, a CNCF sandbox project.

◇ One paper was accepted by **SoCC '21**.

Microsoft Research Asia (MSRA) Beijing, China

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

◇ Mentors: Lewei Lu & Chong Li

◇ Designed a novel pooling layer to enhance the robustness of image recognition models; built Neural Architecture Search (NAS) workflow for face recognition tasks; implemented various attention modules for face recognition models on CNTK.

◇ Star of Tomorrow Internship Award of Excellence.

Awards

◇ Postgraduate Scholarship	2020 – 2024, 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 Xi'an 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 (2022–2025), ICDCS (2023), APSys (2021), MSN (2021), Qshine (2020)

Student Volunteer

◇ 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/Fall 2021, Spring 2024)

◇ COMP3511: Operating Systems (Fall 2023)

Other Experience

ACM-ICPC Competition Group SCUT

Group Member & Team Leader 2016 – 2019

◇ Coach: Prof. Chuhua Xian

◇ Major domains: Dynamic Programming, Number Theory, Data Structure, etc.

Machine Learning & Cybernetics Research Group SCUT

Undergraduate Research Assistant 2017 – 2019

◇ Advisor: Prof. Patrick Chan

◇ Projects: Fundus Stitching, Tableware Recognition, and NN Visualization.

Tencent Innovation Club SCUT, CSE

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, Javascript

Toolkits: Kubernetes, Docker, Grafana, Git, LaTeX, SQL, MarkDown

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

Misc

Play basketball & badminton & squash, workout at the gym, foodie.

My paper reading notes are available at https://paper.lingyunyang.com/.