杨凌云

更新于 August 27, 2024

邮箱: lyangbk@cse.ust.hk **GitHub**: mental2008 **LinkedIn**: lingyun-yang-b2881b18a
手机号: (+852) 135-0284-6103 **Office**: CYT 3007, HKUST **个人网站**: https://www.lingyunyang.com

个人简介

我目前是香港科技大学的在读博士生，预计将于 2025 年秋季毕业。我专注于研究在 AI 时代下大规模数据中心的资源管理问题，在阿里巴巴拥有近 4 年的工业界学术合作经历。具体而言，我的研究包括：（一）提高 AI/GPU 集群的资源利用效率；（二）构建高效且低成本的 AI 模型推理系统。

教育经历

香港科技大学（HKUST）
计算机科学与工程系
计算机科学，博士 2020 – 至今
◇ 导师：王威教授

华南理工大学（SCUT）
计算机科学与工程学院
计算机科学与技术全英创新班，本科 2016 – 2020
◇ 国家奖学金、校级一等奖学金（GPA: 3.82/4）

职业经历

阿里巴巴集团 中国，杭州
研究实习生, 集群管理团队 2020/12 – 至今
◇ 主管：余英豪

针对 AI/GPU 集群的资源效率优化

GPU 碎片资源整理

◇ 针对大规模混部后的 GPU 集群存在的资源碎片问题，提出了碎片的可量化指标。基于量化指标，相较于最优的调度算法可以提升 3% 的 GPU 分配率。研究成果已被系统顶级会议 ATC '23 接收。

大规模 GPU 作业混部与资源超卖

◇ 为了进一步提高集群 GPU 资源使用效率，我们支持了 GPU 作业的混部和分级的资源超卖。我设计并实现单机侧的管控 agent 和中心侧的 controller。日均 vGPU 容器数超过 10k。

构建高效、低成本的 AI 模型推理系统

◇ 针对 stable diffusion 等文生图模型，我们提出了 SwiftDiffusion。在保障图片质量的前提下，降低了最多 5x 的推理时延、提高了最多 2x 的推理吞吐。
◇
◇ 针对 AI 推理服务的部署，我们提出了开发了 Morphling，这是一个用于 Kubernetes 上 AI 服务的开源自动配置框架，结合了元学习和贝叶斯优化，可以快速找到最优容器配置（如资源申请量、运行时参数），在阿里巴巴内部广泛用于自动推荐容器资源规格。研究成果已被云计算顶级会议 SoCC '21 接收。

微软亚洲研究院（MSRA） 中国，北京
研究实习生, 创新工程组（IEG） 2019/07 – 2020/06
◇ 获得明日之星实习生项目的杰出实习生奖

学术论文

*表示共同第一作者，名字按字典序排列

◇ 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.

◇ **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," *under review*.

◇ 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.

◇ 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.

◇ 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.

获奖经历

◇ Postgraduate Scholarship — 2020 – 至今, HKUST
◇ Star of Tomorrow Internship Award of Excellence — 2020/07, MSRA
◇ 三好学生 & 优秀学生干部 — 2019/11
◇ 国家奖学金 — 2019/10
◇ 银奖，2019 年 ICPC 中国西安程序设计竞赛（全国邀请赛） — 2019/05
◇ 一等奖，第 17 届广东省程序设计竞赛 — 2019/05
◇ 银奖，2019 年三七互娱杯程序设计竞赛 — 2019/04
◇ 金奖，2019 年华南理工大学 ACM 程序设计竞赛 — 2019/04
◇ 铜奖，2018 年 ACM-ICPC 亚洲区域赛（徐州站） — 2018/10
◇ 银奖，第 1 届小米程序设计竞赛 — 2018/09
◇ 金奖，2018 年华南理工大学 ACM 程序设计竞赛 — 2018/04
◇ 校级一等奖学金 — 2017/11
◇ 铜奖，2017 年 ACM-ICPC 亚洲区域赛（西安站） — 2017/10
◇ 金牌，第 12 届全国青少年机器人竞赛 — 2012/07
◇ 冠军，2012 年 RoboCup 青少年机器人世界杯中国区选拔赛 — 2012/03

本科经历

ACM-ICPC 集训队
集训队成员 — 2016 – 2019
◇ 教练：冼楚华教授
◇ 研究领域：动态规划，数论，数据结构等。

机器学习研究小组
本科研究助理 — 2017 – 2019
◇ 导师：陈百基教授
◇ 完成研究项目：眼底图拼接，餐碟识别，神经网络可视化。

腾讯创新俱乐部
副主席 — 2018 – 2019
◇ 组织并管理腾讯与高校联合建立的学生俱乐部，ACM 竞赛部负责人。

技术栈	编程语言：Golang、C++、Python、Javascript 工具包：Kubernetes、PyTorch、Docker、Grafana、Git、LaTeX、SQL、Markdown