杨凌云

更新于 2024 年 9 月 20 日

邮箱: lyangbk@cse.ust.hk **GitHub**: mental2008 **LinkedIn**: lingyun-yang-b2881b18a
手机号: (+86) 135-0284-6103 **Office**: BDI 101, UC, 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，主要研究大规模数据中心中的资源管理问题，尤其聚焦于 AI 基础设施。过去近 4 年中，我在阿里巴巴开展了深入的产学研合作。我的研究工作主要包括两个方面：（一）如何提高 AI/GPU 集群的资源利用效率；（二）如何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
分布式系统实验室（ADSL），博士 2020 – 至今
◇ 博士生导师：王威教授 (预计于 *2025* 年秋季毕业)

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）

实习经历

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

缓解 GPU 资源碎片化问题

◇ 量化分析了大规模 GPU 共享后集群中普遍存在的资源碎片问题，提出了创新的碎片梯度下降调度算法，相较于最优的策略可以显著减少 49% 的 GPU 碎片。研究成果已被系统顶级会议 ATC '23 接收。
◇ 针对 DLRM 推理场景，设计了一种 GPU-CPU 分离式异构架构，通过 RDMA 高性能网络拉远部署计算图，消除资源不匹配。实验表明可有效减少 53% CPU 碎片和 27% GPU 碎片。在季节性流量高峰期（如双 11 购物节），从训练集群借用 GPU 服务器最多可节省 90% GPU。
◇ 开发了 ParaSet，一种在 Kubernetes 上运行的 best-effort 工作负载，可根据集群实时资源可用性自适应调整实例数和资源需求，高效填充碎片化资源。

大规模 GPU 作业混部与资源超卖

◇ 为了进一步提高大规模 GPU 集群的资源利用率，支持大规模 GPU 共享与不同优先级（如 latency-sensitive、best-effort）作业的混部，日均运行超过 1 万个共享 GPU 容器。设计并实现了单机管控 agent 和全局 controller。Agent 周期性上报资源使用情况，动态分配 GPU 资源给容器。Controller 管理集群账本，计算可超卖资源量，为集群调度器提供调度决策指引。

构建高效、低成本的 AI 模型推理系统

高效的文生图推理系统

◇ 针对最新的文生图模型，提出了 SwiftDiffusion，一个高效整合了 diffusion 模型和 add-on 模块（如 ControlNets、LoRAs）的推理系统，包括多个创新设计如：服务化 ControlNet、异步 LoRA 加载。在保障图片质量的前提下，能够最多降低 7.8× 的推理时延、提高 1.6× 的吞吐量。

AI 推理服务的自动化配置

◇ 针对 AI 推理服务的集群部署，开发了 Morphling，一个基于 Kubernetes 的自动化配置框架，结合元学习和贝叶斯优化算法，可快速搜索到最优的资源配置和运行时参数。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“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

[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%)*

[C1] ◇ 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%)*

[I2] ◇ 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.

[I1] ◇ **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.*

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, SCUT
◇ 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

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

Misc　论文清单与笔记：https://paper.lingyunyang.com