杨凌云

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 我目前就读于香港科技大学计算机系，博士研究方向为大规模数据中心中的资源管理问题，尤其聚焦于 AI 基础设施优化。过去 4 年在阿里巴巴开展了深入的产学研合作。我的研究主要围绕两个方面：（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。

职业经历 **阿里巴巴集团**，爱橙科技–技术风险效能部–集群管理–*AI* 系统工程 浙江省杭州市
技术专家（P7） 2025/10 – 至今
◇ 主管：余英豪博士

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 2020/09 – 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 2016/09 – 2020/07
◇ 国家奖学金、校级一等奖学金（GPA: 3.82/4）

学术论文 * 表示共同第一作者，名字按字典序排列

会议论文

[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%)*

arXiv 预印版

[I1] ◇ Suyi Li*, **Lingyun Yang***, Xiaoxiao Jiang, Hanfeng Lu, Dakai An, 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.

实习经历 **阿里巴巴集团** 浙江省杭州市
爱橙科技–技术风险效能部–集群管理–*AI* 系统工程

研究实习生 2020/12 – 至今

◇ 导师：余英豪博士

2024 – 2024 **针对文生图 AIGC 模型的高效推理系统**

◇ 系统分析了生产集群中文生图模型里各类 add-on 模块（ControlNets、LoRAs）的调用特征。

◇ 基于 50w 条推理请求 trace 分析，设计实现了 SwiftDiffusion 系统，高效整合了 diffusion 模型和 add-on 模块，系统创新包括：服务化 ControlNet、异步 LoRA 加载、并行 CFG 计算等。

◇ 在保证图片质量的同时，能够最多降低 7.8× 推理时延、提升 1.6× 吞吐量。

◇ 研究成果已投递 ATC '25，同行评议中 [arXiv]。

2022 – 2023 **针对在线推荐 DLRM 服务的 CPU-GPU 分离式推理系统**

◇ 系统评估了多种分离部署方案：拉远 PCIe 交换机、跨机 CUDA API 调用、跨机子图计算等。

◇ 针对深度学习推荐服务的在线推理场景，应用了一种通过 RDMA 网络拉远部署计算图的分离式推理架构，以消除作业需求和节点剩余资源的错配问题。

◇ 实验表明可有效减少日常调度场景中 53% CPU 碎片和 27% GPU 碎片。在季节性流量高峰期（如双 11 购物节），从训练集群借调 GPU 服务器最多可避免 90% GPU 碎片。

◇ 研究成果已被网络领域顶级会议 NSDI '25 接收。

2022 – 2023 **GPU 共享集群的资源碎片量化分析与调度策略优化**

◇ 量化分析了大规模 GPU 共享集群中的资源碎片问题，定义了统计意义上的资源碎片指标 [trace]。

◇ 提出了碎片梯度下降（FGD）调度算法，相较于最优的策略可以显著减少 49% 的 GPU 碎片 [code]。

◇ 开发了 ParaSet，一种在 Kubernetes 上运行的 *best-effort* 工作负载，根据集群实时资源可用性自适应调整实例数和资源需求，填充碎片化资源。

◇ 研究成果已被系统领域顶级会议 ATC '23 接收 [paper]。

2021 – 2022 **大规模 GPU 集群的作业混部与资源超卖**

◇ 支持了大规模的 GPU 资源共享、多优先级（如 latency-sensitive、spot、best-effort）作业的混部。

◇ 设计并实现了单机管控 agent 和全局 controller。Agent 周期性上报节点的资源使用情况，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。

◇ 实现日均运行超过 1 万个共享 GPU 容器，覆盖统一资源池全量的 GPU 机器。

2020 – 2021 **通用 AI 推理服务的自动化资源与参数配置**

◇ 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 [code]。

◇ 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（如 CPU 核数、GPU 显存、GPU 时间片、GPU 卡型）和运行时参数（batch size）。

◇ 研究成果已被云计算领域顶级会议 SoCC '21 接收 [paper]。

微软亚洲研究院（MSRA），创新工程组（*IEG*） 北京市

研究实习生 2019/07 – 2020/06

◇ 设计了一种新的池化层来提升图像识别模型的鲁棒性；构建了针对人脸识别任务的神经架构搜索（NAS）工作流；在 CNTK 框架上针对人脸识别模型应用了多种注意力模块。

◇ 获“明日之星”实习生项目杰出实习生奖。

获奖经历

◇ USENIX NSDI 2025 Student Grant Mar, 2025

◇ 博士研究生奖学金 2020 – 2024, HKUST

◇ “明日之星”实习生项目的杰出实习生奖 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（16/221） 2019/05

◇ 银奖，2019 年三七互娱杯程序设计竞赛（2/107） 2019/04

◇ 金奖，2019 年华南理工大学 ACM 程序设计竞赛（3/127） 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

◇ 二等奖，第 16 届广东省程序设计竞赛（31/214） 2018/05

◇ 金奖，2018 年华南理工大学 ACM 程序设计竞赛（7/94） 2018/04
◇ **校级一等奖学金**（10%） 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

学术服务 **Artifact Evaluation Committee**
◇ ACM SIGCOMM (2024), IEEE HPCA (2024)
◇ ACM SOSP (2023), USENIX OSDI (2023), USENIX ATC (2023), MLSys (2023)
外部审稿人
◇ IEEE INFOCOM (2022–2025), IEEE ICDCS (2023, 2025), APSys (2021), IEEE MSN (2021), EAI QShine (2020)
学生助理
◇ APNet (2023), ICMLC & ICWAPR (2018)

教学经历 **香港科技大学**
助教，计算机科学与工程系
◇ CSIT6000O: Advanced Cloud Computing 2022 春、2023 春
◇ COMP4651: Cloud Computing and Big Data Systems 2021 春、2021 夏、2024 春
◇ COMP3511: Operating Systems 2023 夏

其他经历 **ACM-ICPC 集训队** SCUT
集训队成员，教练：冼楚华教授 2016 – 2019
◇ 主要负责领域：动态规划、数论、数据结构、思维题等。
机器学习研究小组 SCUT
本科研究助理，导师：陈百基教授 2017 – 2019
◇ 完成多个计算机视觉的研究项目：眼底图拼接、餐碟识别、神经网络可视化等。
腾讯创新俱乐部 SCUT
副主席 2018 – 2019
◇ 组织并管理腾讯与华南理工大学联合建立的学生俱乐部，ACM 竞赛部负责人。
字节跳动暑期夏令营，算法组（录取率：2.5%，导师：朱亦博） 北京市，2019/08
◇ 通过流水线并行和模型并行来支持大规模 BERT 模型的分布式训练。

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