杨凌云

lyangbk@cse.ust.hk | lingyunyang.com | 论文笔记 | (+86) 135-0284-6103 | 香港

教育背景

香港科技大学 **香港**
博士，计算机科学与工程系；博士生导师：**王威** 2020 年 9 月 – 至今
- 研究领域：集群管理、模型推理系统、资源调度

华南理工大学 **广东省广州市**
本科，计算机科学与技术全英创新班（GPA 3.82/4） 2016 年 9 月 – 2020 年 7 月
- 国家奖学金、校级一等奖学金

研究项目

我关注于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，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）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。

SwiftDiffusion：高效的文生图模型推理系统 (review, 2407.02031)
***Lingyun Yang**, Suyi Li, 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*
- 构建了文生图推理系统 SwiftDiffusion，高效整合了 diffusion 模型和 add-on 模块（如 ControlNets、LoRAs）。
- 基于对 50w 条生产请求 trace 的特征研究，提出多个系统设计如：服务化 ControlNet、异步 LoRA 加载、latent 并行计算。
- 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。

Prism：针对深度学习推荐模型的 GPU 分离式推理系统 (review)
***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*
- 针对在线推荐服务的推理场景，设计了一种 GPU-CPU 分离式异构架构，通过 RDMA 网络拉远部署计算图，消除资源不匹配；支持拓扑感知调度、基于资源消耗的切图策略。
- 在日常调度场景下可减少 **53%** CPU 碎片和 **27%** GPU 碎片；双 11 期间借调训练集群的 GPU 服务器最多可节省 **90%** GPU。

FGD：大规模集群中 GPU 资源碎片的量化及调度策略优化 (ATC'23)
***Lingyun Yang**, Qizhen Weng, Yinghao Yu, Wei Wang, Xiaochuan Tang, Guodong Yang, Liping Zhang* hkust-adsl/kubernetes-scheduler-simulator
- 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，并提出了创新的碎片梯度下降调度算法。
- 相较于最优的调度策略可以显著减少 **49%** 的 GPU 碎片。

GPU 共享：大规模 GPU 集群的资源共享和超卖
- 实现大规模集群的 GPU 分时复用，以及不同优先级（如 latency-sensitive、best-effort）作业的混部，日均运行超过 **1w** 个共享 GPU 容器。
- 设计并实现了单机管控 agent 和全局 controller。Agent 周期性上报节点的资源使用情况，动态分配 GPU 资源给容器。中心侧的 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。

Morphling：针对通用 AI 推理服务的自动化参数配置框架 (SoCC'21)
***Lingyun Yang**, Luping Wang, Yinghao Yu, Wei Wang, Bo Li, Xianchao Sun, Jian He, Liping Zhang* kubedl-io/morphling
- 作为阿里巴巴开源的 KubeDL 其中的一个独立子项目，成为云原生计算基金会（CNCF）sandbox 项目。
- 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（如批处理大小）。

工作经验

阿里巴巴集团 **浙江省杭州市**
研究实习生，集群管理团队；主管：**余英豪** 2020 年 12 月 – 至今
- 完成多个研究和工程项目，包括 Morphling、GPU 共享、FGD、Prism、SwiftDiffusion（具体如上所述）。

微软亚洲研究院 **北京市**
研究实习生，创新工程组 2019 年 7 月 – 2020 年 6 月
- 优化模型结构以提高人脸识别算法的鲁棒性；获"明日之星"杰出实习生奖。

技术栈

- **编程语言：** Golang, C++, Python, JavaScript, asynchronous, multithread, multiprocess, distributed, RDMA
- **机器学习：** PyTorch, TensorFlow, Numpy, Matplotlib, HuggingFace
- **全栈开发：** Web Frontend, Backend, SQL, Grafana, Docker, Kubernetes, Git, CI/CD