赵汉宇

158-1003-6083 (电话/微信)
zhaohanyu1994@gmail.com (邮件)
博士五年级 · 北京大学分布式系统组 https://zhypku.github.io/ (主页)

研究方向

分布式系统，机器学习系统，云计算：博士主要研究方向为集群级机器学习系统的设计与实现，尤其关注其中的经典分布式系统问题和技术，如资源调度、数据存储等。

教育经历

- **北京大学** 理学博士 计算机系统结构，导师：代亚非教授 2016.9 – 2021.7（预期）
- **武汉大学** 工学学士 仿真科学与技术 2012.9 – 2016.6

实习经历

- **微软亚洲研究院** 全职研究实习生 系统组，导师：张权路博士 2017.11 – 2020.7（共 33 月）

主要项目

- **HiveD：多租户 GPU 集群调度系统** ***OSDI '20***（*OSDI* 北大首篇）· 第一作者 2018.6 – 2020.7

  关键词：资源共享，GPU 拓扑，开源，落地部署，K8s，Go，Python

  传统的多租户 GPU 集群为租户预留一定的 GPU 数量（quota）作为资源保障。然而深度学习任务的性能与 GPU 拓扑紧密相关，仅预留 quota（数量）无法反映租户应得的资源拓扑，使得租户任务的性能得不到保障，破坏用户在共享集群中的体验。HiveD 使用全新的资源抽象，以精确定义每个租户的资源拓扑，并设计了资源的动态分配机制，为租户任务提供拓扑和性能保证。https://github.com/microsoft/hivedscheduler

  – 分析微软内部集群 trace，揭示了用 quota 共享 GPU 资源带来的异常现象：任务在共享集群中的性能（拓扑）可能比其租户的私有集群中（不共享）更差，或等待时间更长，破坏了资源共享的意义
  – 提出 cell 的抽象，将租户的资源虚拟化为私有集群（数量 + 拓扑），并设计 Buddy Cell Allocation 算法以动态分配 cell，理论证明了该算法保证了任务在私有集群中的 GPU 拓扑在共享集群中可以被满足
  – 设计多优先级 cell 分配机制和硬件容错机制，以及和现有调度算法的兼容能力，保证资源利用率和调度效率
  – 基于 K8s 实现调度器，具有组调度（Gang Scheduling）、优先级抢占、容错、重配置等实用特性
  – 调度器与微软 OpenPAI 平台深度整合，在微软内部部署超过 9 个月，管理多个集群、超过 1000 块 GPU

- **DARE：训练数据存储与调度系统** 论文准备中 · 第一作者 2019.12 – 今

  关键词：缓存策略，性能建模，优化算法，Alluxio

  云上训练平台往往采取计算与存储相分离的设计，而由于数据规模和训练速度的提升，计算和存储集群间有限的网络带宽逐渐成为性能瓶颈。DARE 为云上训练场景设计了新型的数据缓存策略，并利用深度学习的任务特点对其性能进行建模，从而对缓存、带宽分配以及任务调度进行联合优化，提升任务性能和资源利用率。

  – 设计“均匀缓存”策略，结合深度学习任务均匀、反复、随机的数据访问特点，解决了传统策略的缓存失效（如 LRU）、带宽浪费（如 Belady-MIN）等问题
  – 利用均匀缓存下的均匀带宽开销和深度学习的执行稳定性，建立任务性能关于其缓存、带宽分配的预测模型
  – 利用性能模型，建模出缓存、带宽分配和任务调度的联合最优化问题，并设计启发式优化算法
  – 基于 Alluxio 实现的原型系统已初步在微软内部集群部署

- **SDPaxos：半分散式状态机副本协议** ***SoCC '18*** · 第一作者 2016.10 – 2017.5

  关键词：分布式一致性，Paxos，跨地域副本，Go，开源

  SDPaxos 是一种为跨地域副本设计的一致性协议。它采用“半分散式”设计，将指令的复制分散化，而将指令的排序中心化，同时解决了传统的纯中心或纯分散式设计带来的性能问题。https://github.com/zhypku/SDPaxos

  – 观察到中心式协议（如 Multi-Paxos，Raft）的负载不均衡问题，以及分散式协议（如 Mencius，EPaxos）由额外的分布式协调开销产生的性能下降问题，提出“半分散式”设计，同时解决二者的性能问题

– 将指令的复制和排序分离为两个独立的 Paxos 流程，将两阶段并行化实现 1 轮延迟，理论证明协议的正确性
– 部署在 EC2 上的实验证明 SDPaxos 相比中心式协议提升性能 6 倍，相比分散式协议提升性能 1.7 倍

主要论文

[1] **HiveD: Sharing a GPU Cluster for Deep Learning with Guarantees**
Hanyu Zhao, Zhenhua Han, Zhi Yang, Quanlu Zhang, Fan Yang, Lidong Zhou, Mao Yang, Francis C.M. Lau, Yuqi Wang, Yifan Xiong, Bin Wang
14th USENIX Symposium on Operating Systems Design and Implementation (**OSDI '20**)

[2] **SDPaxos: Building Efficient Semi-Decentralized Geo-replicated State Machines**
Hanyu Zhao, Quanlu Zhang, Zhi Yang, Ming Wu, Yafei Dai
ACM Symposium on Cloud Computing 2018 (**SoCC '18**)

[3] **Don't Miss Any Piece of Knowledge: In-Network Mutual Learning with Sketch Side Branches**
Yunteng Luan, **Hanyu Zhao**, Zhi Yang, Yafei Dai
arXiv preprint (1911.09418)

[4] **ScheD2: Scheduling Deep Learning Training via Deep Reinforcement Learning**
Yunteng Luan, Xukun Chen, **Hanyu Zhao**, Zhi Yang, Yafei Dai
IEEE Global Communications Conference 2019 (**GlobeCom '19**)

[5] **Gandiva: Introspective Cluster Scheduling for Deep Learning**
Wencong Xiao, Romil Bhardwaj, Ramachandran Ramjee, Muthian Sivathanu, Nipun Kwatra, Zhenhua Han, Pratyush Patel, Xuan Peng, **Hanyu Zhao**, Quanlu Zhang, Fan Yang, Lidong Zhou
13th USENIX Symposium on Operating Systems Design and Implementation (**OSDI '18**)

[6] **Scheduling CPU for GPU-based Deep Learning Jobs** (Poster)
Wencong Xiao, Zhenhua Han, **Hanyu Zhao**, Xuan Peng, Quanlu Zhang, Fan Yang
ACM Symposium on Cloud Computing 2018 (**SoCC '18**)

[7] **Building Efficient and Available Distributed Transaction with Paxos-based Coding Consensus**
Shenglong Li, Quanlu Zhang, Zhi Yang, **Hanyu Zhao**, Yafei Dai
IEEE INFOCOM WKSHPS DCPerf 2018

主要奖励

- **北京大学优秀科研奖** 2019.12
- **北大天网-秒针创新奖学金** 2018.12
- **SoCC '18 Student Scholarship** 2018.10
- **武汉大学优秀毕业生** 2016.6
- **武汉大学三好学生** 2014.11, 2013.11
- **武汉大学唇舌烽火辩论赛亚军** 2014.11

社会活动

- **武汉大学校辩论队主力队员** 2013.11 – 2016.6
- **武汉大学计算机学院辩论队队长** 2013.11 – 2014.11

专业能力

- **语言**：C, Go, Python, C++, Java, LaTeX, Shell, Markdown
- **系统**：Linux, TensorFlow, Kubernetes, Docker, PyTorch, Hadoop
- **知识/技能**：分布式系统，机器学习，调度算法，一致性协议；Git，GitHub 开源协作，英文写作