谷典典

gudiandian1998@163.com · https://gudiandian.github.io

研究方向: 机器学习系统

- 集群管理：GPU 集群调度、深度学习任务调度、深度学习平台建设等
- 分布式机器学习：分布式深度学习训练、大模型训练优化等

教育背景

北京大学, 计算机学院, 计算机软件与理论, *博士研究生* 2020.9 - 2025.6(预计)

导师: 刘譞哲 长聘教授 指导教师: 金鑫 长聘副教授

永赢基金奖学金, 斯伦贝谢奖学金, 北京大学三好学生, 北京大学优秀科研奖

北京大学, 信息科学技术学院, 计算机科学与技术, *理学学士* 2016.9 - 2020.6

国家奖学金 (Top 2%), 休斯顿校友会奖学金, 华为奖学金, 北京大学本科生科研优秀项目, 北京市普通高等学校优秀毕业生, 北京大学优秀毕业生, 北京大学三好学生

实习经历

字节跳动 | Seed Foundation 机器学习系统, 机器学习系统研发实习生 2024.5 至今

- *工程项目：DiT 模型的长序列训练.* 针对 DiT 模型的特点设计长序列训练方案，提升可支持的训练序列长度和训练性能，并对更大规模集群下的长序列训练性能进行预言。

上海人工智能实验室 | AI 训练与计算部门, 模型训练研发实习生 2023.7-2024.1

- *科研项目: 面向长序列训练的大模型训练框架 LoongTrain(ASPLOS'25 在投).* 针对已有序列并行方案可扩展性差、通信效率低的弱点，结合 Head 并行和 Context 并行方法，并使用“双层环状通信”充分利用节点上全部网卡资源；在此基础上，分析不同配置和放置策略的性能影响；通过 ZeRO 和选择性梯度检查点等方法进一步提升长文训练的端到端性能。(项目链接)

微软亚洲研究院 | 系统与网络研究组, 全职研究实习生 2021.4-2022.3

- *科研项目: 服务器无感知的弹性分布式深度学习模型训练系统 ElasticFlow(发表于 ASPLOS'23).* 为开发者抽象了一组“服务器无感知”(serverless) 的 API，屏蔽大量资源配置细节，从而大幅度降低了模型训练任务的开发复杂度; 通过调度算法和作业放置算法保证尽可能多的模型训练任务可以在预期截止时间内完成，为开发者提供性能保障。(项目链接)

谷歌 (北京), 工程实习生 2018.7-2018.9

- *工程项目: 路线搜索场景下的谷歌搜索页面优化.* 在用户搜索词条包含从某地出发到另一地点的路线的含义时，优化谷歌搜索结果页面，在搜索结果第一条以卡片形式展示谷歌地图相关路线信息。

校内项目

工程项目: 基于 Kubernetes 的 MLOps 平台.(平台 demo) 主要参与者

- 参与平台的开发和部署。该平台可以自动完成算力用量评估、运行环境配置、调度和部署，为开发者简化了深度学习任务的部署流程。
- 目前平台已在北京大学计算中心和东南大学网络与信息中心内部部署测试。平台集成了包括 ElasticFlow 在内的多个学术界领先的深度学习训练调度算法和优化方法，支持在 NVIDIA 系列 GPU 和华为 Ascend 系列 NPU 上运行作业，后续还将为实验室更多技术提供落地验证平台。

科研项目: 分布式机器学习程序错误的实证分析——发表于 TOSEM'23.(项目链接) 独立完成

- 从 GitHub 和 StackOverflow 平台收集开发者进行分布式机器学习程序开发过程中遇到的难点和程序错误，对这些难点、错误和错误的解决方案进行分类。
- 为开发者、机器学习框架维护者和研究者提出建议和启示。

科研项目: 面向深度学习模型训练的低碳集群调度系统 GreenFlow——TPDS'24 在投. 独立完成

- 对深度学习训练作业的能耗建模，利用能耗模型和数据中心电网的实时碳密度数据，对集群中深度学习模型训练作业使用的资源进行调整，从而在集群碳排放量不超过预算的同时优化深度学习训练作业的性能。
- 该项目实现了首个既考虑作业性能也考虑碳排放的深度学习作业调度器。

论文发表

已发表论文（六篇，其中三篇 CCF-A 一作/导师一作本人二作，一篇 CCF-B 一作）：

- **[ASPLOS'23, CCF-A, 体系领域四大顶会之一]** Diandian Gu, Yihao Zhao, Yinmin Zhong, Yifan Xiong, Zhenhua Han, Peng Cheng, Fan Yang, Gang Huang, Xin Jin, Xuanzhe Liu. *ElasticFlow: An Elastic Serverless Training Platform for Distributed Deep Learning*.
- **[TOSEM'23, CCF-A]** Xuanzhe Liu, Diandian Gu, Zhenpeng Chen, Jinfeng Wen, Zili Zhang, Yun Ma, Haoyu Wang, Xin Jin. *Rise of Distributed Deep Learning Training in the Big Model Era: From a Software Engineering Perspective*. (导师一作，本人二作)
- **[计算机学报'22, CCF-A]** 谷典典, 石屹宁, 刘譞哲, 吴格, 姜海鸥, 赵耀帅, 马郓. 基于元算子的深度学习框架缺陷检测方法.
- **[ICSE'22, CCF-A]** Changlin Liu, Hanlin Wang, Tianming Liu, Diandian Gu, Yun Ma, Haoyu Wang, Xusheng Xiao. *Promal: Precise Window Transition Graphs for Android via Synergy of Program Analysis and Machine Learning*.
- **[TWEB'20, CCF-B]** Yun Ma, Ziniu Hu, Diandian Gu, Li Zhou, Qiaozhu Mei, Gang Huang, Xuanzhe Liu. *Roaming Through the Castle Tunnels: An Empirical Analysis of Inter-app Navigation of Android Apps*.
- **[CIKM'19, CCF-B]** Diandian Gu, Ziniu Hu, Shangchen Du, Yun Ma. *LinkRadar: Assist the Analysis of Inter-app Page Links via Transfer Learning*.

在投论文（三篇 CCF-A 一作在投）：

- **[TPDS'24, CCF-A]** Diandian Gu, Yihao Zhao, Peng Sun, Xin Jin, Xuanzhe Liu. *GreenFlow: A Carbon Efficient Scheduler for Deep Learning Workloads*.
- **[ASPLOS'25, CCF-A]** Diandian Gu, Peng Sun, Qinghao Hu, Ting Huang, Xun Chen, Yingtong Xiong, Guoteng Wang, Qiaoling Chen, Shangchun Zhao, Jiarui Fang, Yonggang Wen, Tianwei Zhang, Xin Jin, Xuanzhe Liu. *LoongTrain: Efficient Training of Long-Sequence LLMs with Head-Context Parallelism*.
- **[软件学报'24, CCF-A]** 谷典典, 金鑫, 刘譞哲. 基于云边协同的深度学习作业调度方法.

学生工作

- **常任班长**, 北京大学计算机学院软件研究所 20 级博士班 (46 人) 2020.9 至今
- **带班辅导员**, 北京大学信息科学技术学院 2020 级 8 班 (62 人) 2020.9-2024.6
- **团支书**, 北京大学信息科学技术学院 2016 级 4 班 (65 人) 2017.9-2020.6
- **课程助教**, 分布式机器学习: 理论与系统 2021 秋
- **课程助教**, 软件工程 2020 春、2020 秋

技术能力

- **编程语言**: C、C++、Python、Java、JavaScript、LaTeX、Markdown
- **英语能力**: CET-4 665 CET-6 645 GRE: 151+169+3.5
- **知识/技能**: 深度学习系统、机器学习; CUDA 编程;Kubernetes、Docker;Git、Github 开源协作