周宇航

手机：(+86) 18851822106 · **邮箱：**yuhangzhou@smail.nju.edu.cn

个人主页：https://njuzyh.github.io · **住址：江苏省南京市**

教育背景

南京大学，计算机学院，博士（硕博连读） 2021.09 - 至今

- **导师**：田臣教授
- **研究方向**：机器学习系统、数据中心网络

南京大学，计算机科学与技术系，本科 2016.09 - 2020.06

实习经历

华为计算产品线，实习生 2023.01 - 2024.12

- 大模型训练/推理的性能建模与优化
- 大模型训练容错

鹏城实验室，实习生 2022.06 - 2022.12

- 基于鹏城云脑 2 的大模型训练性能 profiling 和瓶颈分析

发表论文

Squeezing Operator Performance Potential for the Ascend Architecture. ASPLOS'25 **CCF-A**

- **第一作者**

Accelerating Model Training on Ascend Chips: An Industrial System for Profiling, Analysis and Optimization. ATC'25 **CCF-A**

- **第一作者**

Chameleon: Adaptive Fault Tolerance for Distributed Training via Real-time Policy Selection. INFOCOM'26 **CCF-A**

- **第一作者**

PushBox: Making Use of Every Bit of Time to Accelerate Completion of Data-parallel Jobs. IEEE TPDS **CCF-A**

- **合作作者**

MEET: rack-level pooling based load balancing in datacenter networks. IEEE TPDS **CCF-A**

- **合作作者**

项目经历

基于昇腾芯片架构的算子性能建模与优化 2024.03 - 2024.10

- 提出“组件”抽象，描述 Ascend 芯片中计算与数据传输单元的串并行执行特性。
- 设计并实现面向 Ascend 架构的组件级 Roofline 模型，显著提升瓶颈分析的准确性。
- 完成多个真实的算子优化案例，支持训练/推理场景，涵盖 11 个模型，算子加速比高达 2.15 倍。
- 本人负责核心思想提出、建模实现、部分实验验证及论文撰写，论文已被 ASPLOS'25 接收。

大模型训练性能瓶颈定位与分析 2023.01 - 2024.12

- 提出新的 profiling 机制，包括轻量级 monitor 和细粒度 profiler，采集完整的性能指标应对性能波动。
- 设计分层瓶颈分析框架，先进行算子间并行分析，再深入算子内分析瓶颈原因，全面准确地识别并行、I/O、CPU、计算和通信瓶颈。
- 基于 135 个典型瓶颈案例，开发出优化工具 mstt advisor，自动检测瓶颈原因并提供优化建议。
- 本人负责分层瓶颈分析的设计，瓶颈成因分析总结，实验验证和撰写论文，论文已被 ATC'25 接收。

大模型训练容错 2025.01 - 2025.07

- 提出自适应容错，实时策略选择实现训练的高效故障恢复，解决现有无备份方法的性能局限。
- 构建统一性能模型，结合解析建模与 profiling，预估执行时间和内存占用，为策略选择提供理论支撑。
- 启发式执行计划搜索，考虑并行度、数据分配、模型层划分等因素，在控制开销的同时快速寻优。
- 将权重传输抽象为二分图匹配问题，将非对称同步通信建模为图着色问题，降低状态迁移开销和单步执行时间。
- 本人负责 Chameleon 系统的核心设计、实验验证及论文撰写，论文已被 INFOCOM'26 接收。

云数据中心网络数据分析及建模 2021.10 - 2022.09

- 基于 NS3 仿真平台对 DLRM 模型训练通信进行建模，发现 NCCL 路径选择对训练性能有显著影响。
- 全面分析 NCCL 参数空间和对通信的影响，结合随机森林和模拟退火算法实现自动化 NCCL 配置。
- 本人主要负责 NS3 通信建模、自动化 NCCL 配置核心算法的提出、部分 NCCL 通信实验验证等工作。

荣誉奖励

博士生国家奖学金 2025.10

南京大学优秀研究生标兵 2025.12