王梓博

（+86）186-7680-9140 · wangzb@smail.nju.edu.cn · https://wangzb.site

教育背景

南京大学，计算机科学与技术系，本科 2018.9–2022.6

- 学分绩：4.40/5 排名：**32/164**
- 荣誉奖项：人民奖学金（2019, 2020, 2021）、优秀毕业生

南京大学，计算机学院，博士 2022.9–至今

- 导师：田臣
- 研究方向：机器学习系统优化
- 荣誉奖项：博士研究生校长特别奖学金（2022）

发表论文

Using Analytical Performance/Power Model and Fine-Grained DVFS to Enhance AI Accelerator Energy Efficiency. ASPLOS'25 (CCF-A) 第一作者

Squeezing Operator Performance Potential for the Ascend Architecture. ASPLOS'25(CCF-A) 合作作者

Accelerating Model Training on Ascend Chips: An Industrial System for Profiling, Analysis and Optimization. ATC'25 (CCF-A) 合作作者

项目经历

分布式机器学习系统优化 2021.9–2023.5

- 以推荐模型为对象，测量分布式机器学习作业训练训练过程中模型的性能指标、集合通信的性能及训练服务器的性能，结合训练服务器上硬件的限制，分析模型训练过程中可能存在的性能瓶颈；
- 结合本领域的前沿工作，提出优化方案，用机器学习+模拟退火的方法为 NCCL 集合通讯库探索最优参数；
- 负责核心思想提出、代码实现、实验验证及论文撰写，相关内容发表在 APNet'23。

基于昇腾 NPU 的 AI 加速器训练能耗优化 2023.8–2024.10

- 构建了白盒性能模型，首次揭示并证明了 AI 算子的执行周期数与频率之间存在凸分段线性关系，将模型平均预测误差分别控制在 1.96%；
- 构建了高精度功耗模型，通过在功耗模型中引入温度依赖项，将模型的平均预测误差控制在 4.62%；
- 设计并实现了一套 AI 加速器端到端能效优化框架，利用昇腾 NPU 提供的细粒度 DVFS 能力在 GPT-3 等任务上，将核心功耗降低 13.44%，芯片功耗降低 4.95%，性能损失<2% ；
- 负责核心思想提出、建模实现、部分实验验证及论文撰写，相关内容发表在 ASPLOS'25。

面向动态图框架的训练显存优化 2023.10–2025.4

- 设计了首个考虑并且能适应动态算子序列的内存交换框架 SmartSwap，支持训练超过硬件显存 4 倍的大模型，解决了同类技术在 PyTorch 等框架下的失效问题；
- 对内存交换技术应用的流程进行了端到端的优化，设计了轻量级的 profiler 以实现持续的算子序列监控，通过多特征模糊匹配解决动态图框架下策略应用的难题，并对夸流内存复用机制进行优化以攻克主机侧性能瓶颈；
- 负责核心思想提出、部分代码实现、部分实验验证及论文撰写，论文在投。

实习经历

华为计算产品线，实习生，杭州 2023.8–2025.4

- 大模型训练的建模与系统优化

鹏城实验室，实习生，深圳 2022.7–2022.9

- 基于鹏城云脑 2 的大模型训练性能 profiling 和瓶颈分析

阿里云-开放平台-企业 IT 治理-身份管理，后端开发实习生，杭州 2021.7–2021.9

- 基于 Spring Boot 框架进行后端开发