

谭邵杰

943648187@qq.com | shaojiemike.top | GitHub | Google Scholar



华为 Ascend 训练开发工程师，中国科学技术大学计算机科学与技术硕士。聚焦 Ascend NPU 上的大模型训练推理优化、多模态模型迁移、分布式并行、RL 后训练与量化加速，并具备体系结构与多节点 CPU/GPU/NPU 优化背景。

教育经历

中国科学技术大学 | 计算机科学与技术 (计算机体系结构) · 硕士 2021.09–2024.06
中国科学技术大学 | 计算机科学与技术 · 学士 2017.09–2021.06

工作与研究经历

华为 Ascend 训练开发部 | AI 训练推理与框架性能优化 2024 H2–至今

- 优化 PyTorch / torch_npu 训练链路，覆盖细粒度 CPU 绑核、算子 Lazy Init、参数寻优和 L2 Cache；完成 OpenSoraPlan、DeepSeek-VL2、GLM-4.1V 的 Megatron / FSDP 适配。
- 参与 DeepSeek-R1、Qwen3 大 EP 推理与 Attention INT8 量化；适配 UniVL、verl + Qwen2.5-VL 及 DanceGRPO，将生成模型 RL 分数由 0.4 提升至 0.8。
- 通过 GMM 和后处理高效计算，使 Intern-S1 Pro 推理性能翻倍 (技术报告)；通过 GDN Triton 算子和框架优化，使 Intern-S2 Preview (Qwen3.5-35B) SFT 从 0.12 倍提升至 0.72 倍 H200 (技术报告)。

华为 2012 实验室 | 暑期实习 · 存内计算 PIM 任务调度研究 2023.07–2023.09

- 构建编译时性能模型，通过局部性聚类与连接性聚类减少数据移动。
- 结合并行度、访存端口压力等静态特征，在 PIM 与 CPU 间选择更优执行位置；仅依赖静态分析获得接近最优的调度结果。

中国科学技术大学 × 华为 | 鲲鹏 920 基本块吞吐量建模 2021.09–2022.11

- 开发基本块吞吐量预测框架与基准测试集，整理吞吐量、延迟和端口压力等指令描述信息。
- 结合鲲鹏 920 微架构改进 llvm-mca 在 AArch64 下的精度，成果发表于 IEEE HPCC 2022。

代表项目

浦江实验室 Intern-S2 Preview SFT 优化 | Qwen3.5-35B 2026.04–2026.06

- 定位 GDN 占开箱耗时 70%+，消除 causal_conv1d / KKT 路径的 transpose，将关键低效 Triton 实现替换或优化为 Ascend C，并扩展 chunk size 128。
- 优化 solve-tril、KKT、RMSNormGated、CV 核同步、绑核与 SP2 通信等算子和框架路径，使 SFT 从 0.12 倍提升至 0.72 倍 H200。

QuEST 多 GPU 量子门电路模拟 | ASC20-21 决赛一等奖 2020.11–2021.05

- 使用 CUDA-aware MPI、多 stream / buffer 与 cuBLAS，在 4 机 8 张 A100 上将规模从 30 拓展至 34 qubits，相较 CPU 系统节约 200+ GB 内存。

大数据串行应用的 CPU 多节点优化 | IPCC 2022 全国三等奖 · 5/130+ 2022.07–2022.11

- 结合循环展开、向量化、OpenMP 与 MPI 优化数据布局、混合精度和 PSRS，在 AMD EPYC 7452 上实现双节点 10,000 倍、单节点最高 2,800,000 倍加速。

技术能力

- **AI 框架**: PyTorch、torch_npu、Megatron、FSDP、verl、vLLM；**算子与系统**: Triton、Ascend C、Ascend NPU 训练推理、多模态、RL、Expert Parallel、Attention INT8 量化。
- **编程与并行**: C、C++、Python、Shell、MPI、OpenMP、CUDA、cuBLAS；**体系结构**: AArch64、CPU 微架构、PIM、性能建模、缓存与带宽优化。

荣誉与论文

- 中国科学技术大学优秀毕业生 (2024)；ASC20-21、ISC21 世界大学生超级计算机竞赛决赛一等奖。
- 论文: S. Tan et al., “Uncovering the Performance Bottleneck of Modern HPC Processor,” CCF Trans. HPC 2024；Q. Jiang, S. Tan, et al., “A³PIM,” DATE 2024；“Quantifying Throughput of Basic Blocks,” IEEE HPCC 2022。

AI 实践与工程自动化

积极拥抱 AI，以项目加速理解与自动化开发: AIHisTrendFuture(在线)梳理技术演进, DevTracking 跟踪进展, IPD-Teams 组织有状态 Agent Teams; AutoResearch 与 ProgramModeling 探索性能建模驱动优化自动化。

职业愿景

以第一性原理洞察本质，以系统创新放大智能价值。