仲书璋

(+86) 182-6132-7246 · zsz@stu.pku.edu.cn

研究兴趣

高能效的人工智能

- 人工智能算法、系统与架构的跨层次协同优化

教育背景

北京大学, 中国 2023.09 – 至今

博士在读 集成电路科学与工程

北京航空航天大学, 中国 2019.09 – 2023.06

本科 计算机科学与技术 GPA 3.83/4.00 排名 20/205

科研经历

混合专家模型推理加速系统, ICCAD 2024, DAC 2025 2023.12 – 2024.12

- 一作 《AdapMoE: Adaptive Sensitivity-based Expert Gating and Management for Efficient MoE Inference》
- 优化 MoE 模型在边端推理时按需加载专家权重的访存瓶颈问题，最终实现 1.35× 的加速比
- 一作 《HybriMoE: Hybrid CPU-GPU Scheduling and Cache Management for Efficient MoE Inference》
- 优化 MoE 模型在异构平台的混合推理，最大化硬件资源利用率，最终实现 1.70× 的加速比

大模型并行解码加速算法, ICCAD 2024 2023.07 – 2023.11

- 一作 《ProPD: Dynamic Token Tree Pruning and Generation for LLM Parallel Decoding》
- 优化传统并行解码算法缺乏上下文关系造成的计算瓶颈问题，最终实现 3.2× 的加速比
- 动态词元树剪枝算法：依据前期层的判定结果对词元树进行剪枝，降低词元树验证的计算开销
- 动态词元树生成算法：提出硬件感知的词元树动态生成算法，动态调整词元树的大小和形状

面向内存约束平台的神经网络编译优化，ICCAD 2023 2022.07 – 2023.06

- 一作 《Memory-aware Scheduling for Complex Wired Networks with Iterative Graph Optimization》
- 神经网络峰值内存占用受到算子调度顺序影响，最优化调度顺序，实现 13.4× 的峰值内存占用降低
- 基于整数线性规划（ILP）的算子调度：基于 ILP 的神经网络调度公式，实现峰值内存占用最小化
- 内存感知的算子融合策略：提出迭代算子融合策略以简化网络计算图，同时保证调度结果最优性

获奖情况

ASC 世界大学生超级计算机竞赛 2021.11 – 2023.05

- 北航代表队队长，负责集群软硬件平台搭建、功耗控制、HPL 优化和集群运行策略调度
- 以全球第三名获得一等奖

奖学金

- 商汤奖学金（全国共三十名） 2022.12
- 北航五四奖章提名奖 2023.05

学生工作

《计算机组成原理》课程助教 2021.07-2022.01

- 负责上机测试出题，教程网站撰写

超算社团 2022.09-2023.06

- 创立北航超算社团，担任首任社长

社会实践

- 北航计算机学院学生会主席团成员 2021.11-2022.11