周报 - GPU 硬件级剖析完成

2025-08-19

概述

本周研究摘要

本周使用 NVIDIA Nsight Compute 完成了全面的 GPU 硬件级性能剖析工作：

- 创建 Table.V GPU Hardware Profiling Metrics：展示从基准到 ATA 再到 ATA+FLP 的渐进式性能改进
- 完成审稿意见 p1.2、p2.3、p3.2 技术回应：提供计算利用率、内存利用率、缓存行为等关键硬件指标的定量分析

审稿意见技术回应

NVIDIA Nsight Compute 性能剖析方法学

剖析平台和工具配置：

- RTX 4090 GPU 平台
- NVIDIA Nsight Compute 专业剖析工具
- 覆盖计算单元利用率、内存子系统性能、缓存行为、SM 占用率

三阶段剖析设计：

- 基准实现（Wang et al. 2025）
- ATA 技术优化版本
- ATA+FLP 完整优化版本

p1.2 GPU 剖析指标分析回应

SM 占用率量化结果:

- 基准：63.4%
- ATA 优化：74.8%
- ATA+FLP：81.3% (+17.9% 改进)

warp 利用率验证:

- 平均每 SM warp 数量：48.2 → 62.4
- 直接关联 1.16×吞吐量改进
- 证明架构优化转化为可测量的硬件资源利用率提升

p2.3 硬件级评估与缓存数据

L2 缓存效率验证：

- 基准命中率：76.2%
- 优化后命中率：89.1% (+12.9%)
- 验证 GPU 内存层次结构中的内存访问优化策略

p3.2 内存影响和可扩展性分析

内存利用率效率分析：

- 内存利用率：72.1% → 87.4% (+15.3%)
- 内存带宽利用率：584.7 GB/s → 758.9 GB/s

可扩展性特征验证：

- L2 缓存优化：76.2% → 89.1% 命中率
- 减少主内存压力，实现更有效的扩展行为

总结

下周任务

- 完成论文最终修订整合工作
- 目标：下周内完成第三篇论文的完整修订并准备重新提交

老师评语

是这个月底提交修改稿?

是月底提交