周报 向嘉豪(2025-08-18)

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。通过系统性的硬件级分析，成功创建了 Table.V GPU Hardware Profiling Metrics，展示了从基准实现到 ATA 再到 ATA+FLP 的渐进式性能改进。完成了对审稿意见 p1.2、p2.3 和 p3.2 的详细技术回应，提供了计算利用率、内存利用率、缓存行为和 SM 占用率等关键硬件指标的定量分析。

下周计划: 完成论文最终修订整合工作, 目标是在下周内完成第三篇论文的完整修订并准备重新提交。

1 GPU 硬件级剖析完成

1.1 NVIDIA Nsight Compute 性能剖析实施

本周的核心工作集中于使用 NVIDIA Nsight Compute 进行全面的 GPU 硬件级性能剖析，这是回应审稿意见 p1.2 的关键要求。通过系统性的剖析方法学，我们获得了详细的硬件资源利用率数据，量化了 Thread-Adaptive allocation 和 Function-Level Parallelism 两种核心技术在 GPU 硬件层面的优化效果。

剖析方法学和数据收集 使用 NVIDIA Nsight Compute 工具对 RTX 4090 平台进行了三个阶段的性能剖析：基准实现（Wang et al. 2025）、应用 ATA 技术后的优化版本，以及最终的 ATA+FLP 完整优化版本。剖析过程覆盖了计算单元利用率、内存子系统性能、缓存行为、SM 占用率等关键硬件指标，确保了硬件级验证的全面性和准确性。

Table.V 创建和硬件指标量化 成功创建了 GPU Hardware Profiling Metrics 表格，展示了渐进式优化的量化效果。计算利用率从基准的 68.3% 提升至 ATA 的 78.9%，最终达到 ATA+FLP 的 84.7%。内存利用率相应地从 72.1% 改进至 81.2% 再到 87.4%。内存带宽利用率从 584.7 GB/s 提升至 758.9 GB/s，实现了 29.8% 的内存子系统效率改进。

1.2 审稿意见 p1.2、p2.3、p3.2 回应完成

p1.2 GPU 剖析指标分析回应

针对审稿意见 p1.2 关于 GPU 剖析指标的要求，我们提供了全面的硬件级评估数据。SM 占用率分析显示了从基准的 63.4% 到 ATA 优化的 74.8% 再到最终 ATA+FLP 的 81.3% 的显著改进。平均每 SM 的 warp 数量从 48.2 增加至 62.4，直接关联了观察到的 1.16×吞吐量改进，证明了我们的架构优化如何转化为可测量的硬件资源利用率提升。

L2 缓存效率验证展示了从 76.2% 到 89.1% 命中率的改进，验证了我们在 GPU 内存层次结构中的内存访问优化策略。这一缓存行为的改进直接支持了我们关于内存访问模式优化的技术声明，为硬件级验证提供了定量证据。

p2.3 硬件级评估与缓存/占用率数据

完成了审稿意见 p2.3 要求的硬件级评估工作，提供了详细的缓存行为、占用率和执行分解指标。端到端执行分解数据显示了跨计算单元的渐进式利用率改进：计算利用率从 68.3% 提升至 84.7%，内存利用率从 72.1% 推进至 87.4%。内存吞吐量效率在优化实现中达到 76.5%，证明了对 RTX 4090 GDDR6X 内存子系统的有效利用。

p3.2 内存影响和可扩展性分析

针对审稿意见 p3.2 关于大规模内存影响的关注，我们进行了全面的内存分析工作。内存利用率效率从基准的 72.1% 进步至优化版本的 87.4%，同时内存带宽利用率从 584.7 GB/s 增加至 758.9 GB/s。

可扩展性特征分析特别体现在 L2 缓存优化结果中，显示了从 76.2% 到 89.1% 的命中率改进。这一增强的缓存效率减少了主内存压力，实现了更有效的扩展行为。我们的 Thread-Adaptive 分配策略通过优化的线程到内存映射解决了内存争用问题，而 Function-Level Parallelism 通过战略性共享内存利用减少了内存访问延迟。