PR #21503 完整报告

sgl-project/sglang

Opt jit qknorm_across_heads cuda kernel

合并时间：2026-03-27 13:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21503

执行摘要

本次 PR 优化了 sglang 中用于 query 和 key 归一化的 JIT CUDA 内核，通过分离处理 q 和 k 以减少寄存器压力和共享内存使用，在 H200 上实现寄存器使用减半、共享内存减半、占用率翻倍，微基准测试显示性能提升最高达 15%。

功能与动机

优化动机源于旧内核在一个 CTA 中同时处理 q 和 k，导致状态过多，影响 GPU 占用率。PR body 引用："The old kernel handled both q and k inside one CTA, which kept too much state live at the same time"。新内核旨在减少 live state，遵循 PR #18073 的思路，提升硬件利用率和推理速度。

实现拆解

关键改动文件：python/sglang/jit_kernel/csrc/elementwise/qknorm_across_heads.cuh。

- 共享内存优化：从 __shared__float shared_memory[64]; 减少到 __shared__float shared_memory[32];，缓冲区减半。
- 工作分离：引入 grid.y = 2，使用 const bool is_q = blockIdx.y == 0; 判断处理 q 或 k，分离计算路径。
- 变量重构：将原 v_q、v_k、v_q_weight、v_k_weight 等合并为 v_data 和 v_weight，减少寄存器使用。
- 性能指标：寄存器 / 线程从 48 减少到 26，共享内存 / 块从 256B 减少到 128B，占用率从 45.25% 提升至 88.17%。

代码示例：

```
// 旧内核：同时处理 q 和 k
__shared__ float shared_memory[64]; // 双路缓冲区
// 新内核：分离处理
__shared__ float shared_memory[32];
const bool is_q = blockIdx.y == 0;
```

评论区精华

- HydraQYH提问："How many Stall Long Scoreboards are there here?" BBuf回应：通过代码更新，Stall Long Scoreboards 从 3% 减少到 1%，低延迟形状性能提升 1-5%。

- DarkSharpness建议："If the register pressure is not very heavy, we can try to load the weight in advance (e.g. before sqr sum) which can overlap computation with weight loading by leveraging ILP." BBuf采纳并更新代码，展示优化效果。

风险与影响

- 风险：代码重构可能引入归一化计算错误，需依赖现有测试验证；共享内存减少可能影响大批次稳定性；性能提升高度依赖 H200 硬件，其他 GPU 效果未知。
- 影响：用户享受轻微速度提升；系统资源利用率提高；团队获得 CUDA 优化实践案例，促进后续性能调优。

关联脉络

本次 PR 直接参考 PR #18073，延续内核优化脉络。近期 PR 如 #20562（优化 LoRA 性能）和 #20606（修复 NSA 内核），显示 sglang 项目持续关注性能优化和内核调优，本次变更符合这一演进方向。