PR #23611 完整报告

sgl-project/sglang

[AMD] Optimize MiniMax-M2.5 - use aiter biased_grouped_topk for sigmoid scoring in MoE routing

合并时间：2026-04-25 13:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23611

执行摘要

- 一句话：AMD MoE routing 使用 aiter 内核，性能提升 35%
- 推荐动作：建议批准并合并，性能提升明显且经过充分验证，代码改动小、风险可控。

功能与动机

MiniMax-M2.5 模型使用带 correction bias 的 sigmoid 评分进行 MoE 路由，现有 topk_sigmoid 实现在 AMD GPU 上性能不够理想。aiter.biased_grouped_topk 内核将每次调用从 ~9.3µs 降低至 ~6µs，减少约 35% 的路由开销，从而提升整体推理吞吐。

实现拆解

1. 修改 MoE 路由核心函数：在 python/sglang/srt/layers/moe/topk.py 的 fused_topk 函数中，当 scoring_func == "sigmoid" 时新增条件分支：若 _use_aiter 为 True 且 correction_bias 不为 None，则调用 aiter_biased_grouped_topk 替代原有的 topk_sigmoid。
2. 封装 aiter 内核调用：调用时传入 gating_output、correction_bias（转换为与 gating_output 相同 dtype）、预分配的 topk_weights 和 topk_ids，并设置 num_expert_group=1、topk_group=1、need_renorm=renormalize，以匹配 sigmoid scoring 语义。
3. 保留回退路径：如果 _use_aiter 为 False 或 correction_bias 为 None，仍然走原来的 topk_sigmoid 路径，确保兼容性。

关键文件：

- python/sglang/srt/layers/moe/topk.py（模块 MoE 路由；类别 source；类型 core-logic）：MoE 路由核心实现，新增 aiter 内核调度分支以替换 sigmoid scoring 路径

关键符号：未识别

关键源码片段

python/sglang/srt/layers/moe/topk.py

MoE 路由核心实现，新增 aiter 内核调度分支以替换 sigmoid scoring 路径

```
# python/sglang/srt/layers/moe/topk.py
```

```
# 在 fused_topk 函数中，elif scoring_func == "sigmoid" 分支处新增 AMD 优化路径
elif scoring_func == "sigmoid":
    if _use_aiter and correction_bias is not None:
        # 使用 AMD aiter 库的 biased_grouped_topk ASM 内核
        # 该内核比通用 topk_sigmoid 快约 35%（6μs vs 9.3μs）
        aiter_biased_grouped_topk(
            gating_output,
            correction_bias.to(dtype=gating_output.dtype),
            topk_weights,
            topk_ids,
            num_expert_group=1,
            topk_group=1,
            need_renorm=renormalize,
        )
    else:
        # 回退到通用 sigmoid topk 实现
        topk_sigmoid(
            topk_weights,
            topk_ids,
            gating_output,
            renormalize,
            correction_bias,
        )
```

评论区精华

本次 PR 仅包含 1 个 commit，reviewer HaiShaw 直接批准，无额外评论区交锋。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 功能风险：新增分支仅在 _use_aiter and correction_bias is not None 时生效，其他路径保持不变，回归风险低。
  - 精度风险：PR 提供了 GSM8K 精度对比（93.3% vs 93.4%），无退化。
  - 兼容性风险：只影响 AMD GPU + sigmoid scoring + correction_bias 非 None 的场景，不影响现有软 max 路径或非 AMD 硬件。
  - 性能风险：仅正向影响，无回归。
- 影响：
  - 用户侧：使用 AMD GPU 运行 MiniMax-M2.5 模型的用户将获得 2-2.4% 吞吐提升，延迟降低。
  - 系统侧：MoE 路由内核调用更高效，可能轻微降低 GPU 占用，无明显副作用。
  - 团队侧：变更集中在 1 个文件，逻辑清晰，维护成本低。
  - 风险标记：暂无

关联脉络

- PR #23620 [AMD] Optimize MiniMax-M2.5 - enable fused Triton kernel for FP8 KV cache write in aiter decode path: 同为 MiniMax-M2.5 的 AMD 优化，关注 aiter 在解码路径的应用
- PR #23568 Parakeet nemotron encoder: 同属模型优化系列，涉及新模型支持