PR #35105 完整报告

sgl-project/sglang

Revert "[AMD] [GLM5] Fuse shared-expert append into aiter grouped-topk (skip per-layer append kernel)"

合并时间：2026-08-17 14:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35105

执行摘要

- 一句话：回滚 #31323，恢复逐层 append，解除 CI 阻塞
- 推荐动作：该 PR 无需精读，其价值在于揭示回滚决策与 CI 治理的关系。建议关注原 PR #31323 的后续演进，如果问题修复后希望重新引入该优化，应补充更全面的测试覆盖和 CI 验证。

功能与动机

原 PR #31323 的 CI 出现失败，回滚以解除阻塞。评论中 hnyls2002 明确写道「Revert this to unblock the CI」，并 @HaiShaw 指出这是为了恢复 CI 健康状态。

实现拆解

1. 删除持久化缓冲区相关定义：在 python/sglang/srt/layers/moe/topk.py 中移除 _get_aiter_topk_fuse_shared_max_tokens、_get_aiter_topk_fuse_shared_buf 函数及配套的 _AITER_TOPK_FUSE_SHARED_MAX_TOKENS_CAP 等全局变量。
2. 简化 biased_grouped_topk_gpu aiter 分支：移除 _shared_fuse 判断和 fused_shared_experts_scaling_factor 参数，恢复为每次调用直接分配 (token, topk) 的临时 topk_weights/topk_ids，并直接交给 aiter_biased_grouped_topk 写入 routed 列后返回。
3. 恢复 _post_process_topk_ids 无条件 append：删除对 TopK 张量宽度的形状检查，恢复 aiter 路径下对 fused_append_shared_experts 的无条件调用，逐层追加共享专家。
4. 无测试与配置配套：本次回滚未包含测试文件更新，也未调整服务器参数或环境变量。

关键文件：

- python/sglang/srt/layers/moe/topk.py（模块 专家路由；类别 source；类型 core-logic；符号 biased_grouped_topk_gpu, _post_process_topk_ids, _get_aiter_topk_fuse_shared_max_tokens, _get_aiter_topk_fuse_shared_buf）：唯一变更文件，回滚了 #31323 的共享专家融合逻辑，恢复逐层 append 路径。

关键符号：biased_grouped_topk_gpu, _post_process_topk_ids

关键源码片段

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

唯一变更文件，回滚了 #31323 的共享专家融合逻辑，恢复逐层 append 路径。

```
# 回滚后的 aiter 路径：共享专家不再预填充，逐层 append 恢复
def biased_grouped_topk_gpu(...):
    ...
    elif _use_aiter:
        token = gating_output.shape[0]
        device = gating_output.device
        # 直接分配临时输出张量，不再复用持久化缓冲区
        topk_weights = torch.empty((token, topk), dtype=torch.float32, device=device)
        topk_ids = torch.empty((token, topk), dtype=torch.int32, device=device)
        # aiter 内核只写 routed 列，共享专家列由后续 append 填充
        aiter_biased_grouped_topk(
            gating_output,
            correction_bias.to(dtype=gating_output.dtype),
            topk_weights,
            topk_ids,
            num_expert_group,
            topk_group,
            renormalize,
            routed_scaling_factor if routed_scaling_factor is not None else 1.0,
        )
        return topk_weights, topk_ids
```

评论区精华

唯一讨论来自关联 Issue #31323 的评论，作者 hnyls2002 直接提出「Revert this to unblock the CI」，没有展开技术争议或替代方案讨论。仓库维护者 HaiShaw 被 @ 但未在本 PR 内留下 review 评论。

- 回滚以解除 CI 阻塞 (other): 直接回滚合并，恢复原逻辑，CI 恢复。

风险与影响

- 风险：本 PR 为纯回滚，风险较低。主要影响是撤销 #31323 在 AMD/GLM-5.2 aiter 非 EP 路径的性能优化（原 PR 实测提升约 0.4%～1.4%），恢复逐层 append 内核后性能回退幅度很小。需要注意：若未来重新合入 #31323，必须重新验证 CI；回滚也可能与后续基于该功能的其他改动产生合并冲突。由于未附带新增测试，对兼容性没有额外保障。
- 影响：影响范围限定在 AMD GPU（gfx950/MI355X）上 GLM-5.2 使用 aiter grouped-topk 且非 EP 的路径；其他架构、EP 模式及 CUDA 路径不受影响。对用户而言，性能会有微小回退，但功能行为不变；对团队而言，CI 阻塞被解除，主干恢复稳定。
- 风险标记：AMD 特定路径回滚，未附带测试变更，性能小幅回退

关联脉络

- PR #31323 [AMD] [GLM5] Fuse shared-expert append into aiter grouped-topk (skip per-layer append kernel): 本 PR 的精确回滚目标，撤销其全部改动。