PR #21166 完整报告

sgl-project/sglang

[Not-Merge][AMD] GLM-5 performance optimization

合并时间：2026-04-12 14:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21166

执行摘要

本 PR 针对 GLM-5 模型在 AMD 平台（特别是 TP8 配置）下的 Aiter 稀疏注意力进行性能优化，通过填充头数以适配内核要求，并移除硬编码参数让内核内部优化。影响范围限于使用 Aiter 后端的 AMD 平台，能提升 GLM-5 推理速度，但需等待外部 Aiter 更新后方可合并。

功能与动机

优化 GLM-5 模型性能，解决在 AMD 平台 TP8 配置下（每设备头数 =8）运行 Aiter 稀疏注意力时，因内核要求头数为 16 的倍数而导致的性能问题。PR body 中明确目标为“Optimize GLM-5 model performance”，并通过填充和参数简化来适配内核约束。

实现拆解

主要修改集中在两个文件：

1. python/sglang/srt/layers/attention/nsa_backend.py：
   - 在 __init__ 中添加 need_pad_heads 和 head_repeat_factor 逻辑，判断是否需要填充。
   - 在 _forward_aiter 和 _forward_aiter_extend 中实现张量填充与还原：python if self.need_pad_heads: q_kernel = q.view(-1, layer.tp_q_head_num, layer.head_dim).repeat_interleave(self.head_repeat_factor, dim=1) o_kernel = q.new_empty((q.shape[0], layer.tp_q_head_num * self.head_repeat_factor, layer.v_head_dim)) 调用 mla_decode_fwd 后，通过 o = o_kernel[:, :: self.head_repeat_factor, :] 还原输出。
2. python/sglang/srt/layers/attention/nsa/nsa_indexer.py：
   - 删除 deepgemm.fp8_paged_mqa_logits 调用中的硬编码参数 ChunkK=128, TotalCuCount=256, WavePerEU=5，让内核内部配置优化参数。

评论区精华

review 讨论较少，但有两个关键点：

> gemini-code-assist[bot]：*"The padding and unpadding of query and output tensors are handled correctly within the _forward_aiter and _forward_aiter_extend methods. The changes are well-explained by comments and appear to be correct and efficient for the stated motivation."*

> HaiShaw：*"Merge later, see https://github.com/sgl-project/sglang/issues/21302"* 并说明等待 ROCm/aiter#2213 合并。

风险与影响

- 风险：填充操作引入额外张量开销，但为满足内核必要；移除硬编码参数依赖内核内部优化，需确保 Aiter 更新后配置最优。
- 影响：仅影响使用 Aiter 稀疏注意力的 AMD 平台，提升 GLM-5 TP8 配置性能，对其他模型或后端无影响。

关联脉络

- 与 PR #21986（AMD 平台简化 API）相似，都涉及移除硬编码 / 白名单以提升维护性。
- 与 PR #22361（Whisper 批量编码）和 #22567（Tokenizer 消除 $O(n^2)$ 复制）同属性能优化类别，反映团队持续关注关键路径性能调优。
- 依赖外部 Aiter 更新（ROCm/aiter#2213），需跟踪 issue #21302 以确定合并时机。