PR #29959 完整报告

sgl-project/sglang

[DSA][GLM5.2] Index Share for MHA

合并时间：2026-07-04 17:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29959

执行摘要

- 一句话：GLM5.2 MHA 路径跳过 DSA indexer 实现索引共享
- 推荐动作：值得精读，展示了如何通过提取方法安全地跳过不必要的 indexer 调用，是一种常见的性能优化模式。需要注意 should_run_indexer 实现中的边界条件（如 is_nextn 回退）的合理性。

功能与动机

根据 issue #29951，prefill 运行 MHA（无 CUDA graphs）时 DSA indexer 在每个层都被调用，但 skip-topk 共享层的 indexer 结果不会被使用，因此无必要。PR 旨在跳过这些层的 indexer，减小开销。

实现拆解

1. 提取 should_run_indexer 方法：在 DeepseekMLAForwardMixin 中添加 should_run_indexer(prev_topk_indices) 方法，封装原有的 not self.skip_topk or (self.is_nextn and prev_topk_indices is None) 逻辑。
2. 替换 MLA 前向中的调用点：在 forward_absorb_prepare 中的两个 indexer 调用处（一个在 alt_stream decode 分支内，一个在非 decode 分支内），将原有的内联条件替换为 self.should_run_indexer(prev_topk_indices)，并移除内联注释。
3. 替换 MHA 前向中的调用点：在 forward_normal_prepare 中，indexer 调用被 if self.should_run_indexer(): 包裹（无参数），使得 MHA 路径在 skip_topk 层不再运行 indexer。
4. 测试与验证：无新增测试文件；CI 测试 test_deepseek_v32_indexcache.py 通过；作者提供 profiler trace 确认 eager prefill MHA 路径下 indexer 只出现在非共享层。

关键文件：

- python/sglang/srt/models/deepseek_common/attention_forward_methods/forward_mla.py（模块 ML A 前向；类别 source；类型 data-contract；符号 should_run_indexer）：核心变更文件：添加 should_run_indexer 方法并修改两处 indexer 调用点，统一逻辑。
- python/sglang/srt/models/deepseek_common/attention_forward_methods/forward_mha.py（模块 MHA 前向；类别 source；类型 logic）：MHA 前向路径：用 should_run_indexer 包裹 indexer 调用，实现跳过。

关键符号：should_run_indexer

评论区精华

- 审核者 Fridge003要求附上 profiler 截图，以确认 indexer 在 MHA 路径被正确跳过。作者上传了 trace 文件，并说明 eager prefill 路径中所有 78 层 decoder 均调用 MHA 路径，但只有预期中的非共享层产生了 indexer，验证了实现正确。
- 用户 tobeprozy提问“精度下降这么多？”但未进一步澄清，可能针对其他 PR，本 PR 未回复。
- 请求 profiler 截图验证 (correctness): 作者提供 trace 文件并解释 eager prefill 路径中 indexer 只在非共享层运行，验证正确。

风险与影响

- 风险：
  - 回归风险：低。should_run_indexer 逻辑与原有内联条件完全一致，仅提取为方法。遗漏 update 特定条件时可能出错，但未修改条件本身。
  - 正确性风险：如果 should_run_indexer 在非 skip 层错误返回 False，会导致 indexer 未运行，后续 MLA 读取空 KV cache，触发错误。但现有条件覆盖了所有情况。
  - 测试风险：无新增单元测试，但 CI 中有集成测试验证 indexcache。
- 影响：
  - 性能影响：对 GLM5.2 模型在 MHA prefill 路径（不使用 CUDA graph）时，共享层的 indexer 和 KV 写入被跳过，减少计算量和 memory traffic，提升吞吐。
  - 兼容性：完全透明，API 和模型输出不变。
  - 影响范围：仅限于使用 DSA 的 GLM5.2 模型。
  - 风险标记：核心路径变更（需注意）

关联脉络

- PR #29951 [Feature] [GLM5.2] Index share reuse with MHA: 关联的 feature request issue，本 PR 实现了该功能。