PR #35116 完整报告

sgl-project/sglang

[SM120] flash_mla: allocate the page-split buffer outside inference mode

合并时间：2026-08-25 08:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35116

执行摘要

- 一句话：SM120 flash_mla 分页缓冲改在推断模式外分配
- 推荐动作：值得精读，但仅面向关注 SM120 内核和 CUDA graph 相关机制的工程师。核心设计决策是保持惰性缓冲区分配与使用模式一致性，并修复不对称保护。可借鉴其注释说明写法，但实现简单，无需深度分析。

功能与动机

作者在 PR body 中明确说明 _split_kv_pages_to_64 中存在两个惰性分配的持久缓冲区（buf 和 mbuf），但只有 mbuf 在分配时被 torch.inference_mode(False) 保护。buf 同样会在 CUDA graph 捕获期间被写入，若在 autotune 推理模式下首次分配，则会在捕获时因修改推理张量而触发运行时错误。该修复基于 @moxcat 在 SM120 上运行 DSV4 时的失败报告，当时修复覆盖了两个缓冲区，但 buf 的修复在后续代码中丢失。

实现拆解

1. 定位文件 python/sglang/kernels/ops/attention/flash_mla_sm120.py 中的 _split_kv_pages_to_64 函数。
2. 将 buf = torch.empty(...) 的分配操作包裹在 with torch.inference_mode(False): 上下文中，确保缓冲区在正常模式下分配，可被后续 CUDA graph 捕获修改。
3. 保持缓冲区形状、缓存键、生命周期和填写逻辑不变，仅改变分配时的模式。
4. 无测试、配置或部署配套改动。

关键文件：

- python/sglang/kernels/ops/attention/flash_mla_sm120.py（模块 内核模块；类别 infra；类型 infrastructure；符号 _split_kv_pages_to_64）：修改了 _split_kv_pages_to_64 中 buf 缓冲区的分配模式，修复潜在运行时错误，是核心变更文件。

关键符号：_split_kv_pages_to_64

关键源码片段

python/sglang/kernels/ops/attention/flash_mla_sm120.py

修改了 _split_kv_pages_to_64 中 buf 缓冲区的分配模式，修复潜在运行时错误，是核心变更文件。

```
# 关键片段：`_split_kv_pages_to_64` 的惰性缓冲区分配部分
key = f"flash_mla_sm120_split:{dev}"
buf = buffers.get(key)
if buf is None or buf.shape[0] < num_dst_pages:
    # 该缓冲区可能在 autotune（推理模式）下首次分配，
    # 但会在 CUDA graph 捕获（非推理模式）期间被再次写入，
    # 因此必须确保分配在非推理模式下进行，以避免后续修改时出错。
    with torch.inference_mode(False):
        buf = torch.empty(
            num_dst_pages,
            _BYTES_PER_DST_PAGE_PADDED,
            dtype=torch.uint8,
            device=dev,
        )
    buffers[key] = buf
out = buf[:num_dst_pages]
```

评论区精华

- @moxcat 在评论中提供了详细复现与分析：原始配置仍然存在，已在 v0.5.16 backport 栈和当前 main 上复现。发现缺失的拆分缓冲区崩溃是“写入者类型”问题，而非分配模式问题；arm 排序是真实存在但与分支相关。
- 作者指出相似模式同样存在于 sglang/srt/layers/moe/moe_runner/flashinfer_cutedsl.py，但本次仅修复 flash_mla 一处。
- 缺失修复的复现与验证 (question): @moxcat 的复现证实了修复的必要性，但无法提供跨分支一致的 before/after 用例。

风险与影响

- 风险：
  - 低风险：仅将分配操作移入 torch.inference_mode(False)，不改变缓冲区逻辑。
  - 潜在回归：若分配确实在非推理模式下发生，则此改动为 no-op；若在推理模式下发生，则修复了潜在崩溃。
  - 风险确认：该修复未能提供可复现的 before/after 用例，但基于代码对称性和 @moxcat 的历史报告，修复方向正确。
- 影响：
  - 影响范围：仅影响 SM120 平台上的 flash_mla 内核路径，具体是 DSV4 等模型在 autotune+CUDA graph 场景下的稳定性。
  - 对用户：消除 SM120 上潜在的运行时崩溃，提升部署稳定性。
  - 对团队：作为 #29927（SM120 支持）的前置依赖，虽不在其范围内，但被其依赖。
  - 风险标记：缺少可复现用例，无测试覆盖

关联脉络

- PR #29927 [SM120] enablement work (未提供标题): 本 PR 从 #29927 拆分而来，且 #29927 依赖它完成 SM120 支持。