PR #42945 完整报告

vllm-project/vllm

[Bugfix][KV Offload] count appended GPU blocks in store group_sizes

合并时间：2026-05-18 19:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42945

执行摘要

- 一句话：修复 KV Offload 计数 bug，1 行代码变动
- 推荐动作：建议快速合并。该 PR 虽小但精准，修复了一个真实触发断言的 bug，值得关注其背后的窗口跨越边界场景设计约束。

功能与动机

修复 when a block straddles the sliding-window edge, the count of appended GPU blocks in group_sizes exceeded the actual number in src_block_ids, triggering a sum(group_sizes) == len(block_ids) assertion failure in GPULoadStoreSpec.

实现拆解

1. 在 _build_store_jobs 的 for 循环复制块时，将 num_group_blocks 从一次性加 block_size_factor（第 713 行 base）改为每次成功追加一个子块后自增 1（第 722 行 head）。
2. 同时移除旧的 num_group_blocks += block_size_factor 语句（第 713 行 base）。
3. 该变更仅影响 vllm/distributed/kv_transfer/kv_connector/v1/offloading/scheduler.py 中的 _build_store_jobs 方法，未涉及测试、配置或部署。

关键文件：

- vllm/distributed/kv_transfer/kv_connector/v1/offloading/scheduler.py（模块 KV 卸载调度器；类别 source；类型 core-logic）：修复了 _build_store_jobs 中 num_group_blocks 计数逻辑，确保与 src_block_ids 长度一致。

关键符号：未识别

关键源码片段

vllm/distributed/kv_transfer/kv_connector/v1/offloading/scheduler.py

修复了 _build_store_jobs 中 num_group_blocks 计数逻辑，确保与 src_block_ids 长度一致。

```
def _build_store_jobs(self):
    # ... 前面的逻辑不变 ...
    for idx, offload_key in ...:
        if offload_key not in keys_to_store:
```

```
            continue
        offloaded_block_idx = start_block_idx + idx
        gpu_block_idx = offloaded_block_idx * block_size_factor
        # 修复前：num_group_blocks += block_size_factor // 即使有块被跳过也全量加
        for i in range(block_size_factor):
            block_id = block_ids[gpu_block_idx + i]
            if block_id == 0:
                # 滑动窗口外，跳过，不计数
                assert start_gpu_block_idx is None
                continue
            elif start_gpu_block_idx is None:
                start_gpu_block_idx = gpu_block_idx + i
            src_block_ids.append(block_id)
            num_group_blocks += 1 # 修复后：每追加一个块加一次，保证与实际计数一致
            if is_sliding_window:
                sliding_window_block_ids.append(block_id)
            else:
                non_sliding_window_block_ids.append(block_id)
    group_sizes.append(num_group_blocks)
    # ... 后续不变 ...
```

评论区精华

gemini-code-assist[bot] 的自动审查确认 bug 根因并说明修复正确，但未发现人工讨论。orozery 快速批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：只有 1 行删除 + 1 行添加，核心逻辑明确是计数方式修正，回退也简单。影响范围仅限于 _build_store_jobs 中 group_sizes 的构造，不会影响其他路径。
- 影响：修复了 KV Offload 在 hybrid-attention + sliding window 场景下的断言崩溃，使该特性可正常使用。影响仅作用于 KV 卸载调度器，不涉及用户 API、模型推理结果或性能。
- 风险标记：无测试覆盖

关联脉络

- PR #41233 [Bugfix][Hybrid][NemotronH] Fix mamba_cache_mode=all + speculative decoding crash: 涉及类似的 hybrid-attention 滑动窗口与块处理逻辑。