PR #25473 完整报告

sgl-project/sglang

fix(overlap): skip empty future interval for dp attention idle ranks

合并时间：2026-05-16 17:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25473

执行摘要

- 一句话：跳过 DP attention 空闲 rank 的空区间存储
- 推荐动作：建议尽快合入，属于明确的小 bugfix，补全了边界情况处理，确保与相关方法的一致性。

功能与动机

在 DP attention 中，空闲 rank 的 future interval 为空 slice，如果直接赋值会导致无效的 tensor 操作。PR 描述明确说明需要将 store_to_map_for_new_batch 中的空 slice guard 镜像到 store_to_map 的 spec_algo.is_none() 分支。

实现拆解

1. 在 python/sglang/srt/managers/overlap_utils.py 的 store_to_map 方法中，当 self.spec_algo.is_none() 时，在获取 intv 后增加 if self.is_empty_slice(intv): return 检查，如果 intv 为空则直接返回，不执行后续的 self.token_ids_buf[intv] = batch_result.next_token_ids。
2. 该 guard 与 store_to_map_for_new_batch 方法中的已有逻辑完全一致，只是补全了前者遗漏的分支。

关键文件：

- python/sglang/srt/managers/overlap_utils.py（模块 调度器；类别 source；类型 core-logic；符号 store_to_map）：修复核心位置：为 DP attention idle rank 的空 future interval 增加 guard，避免无效的 tensor 赋值。

关键符号：store_to_map

关键源码片段

python/sglang/srt/managers/overlap_utils.py

修复核心位置：为 DP attention idle rank 的空 future interval 增加 guard，避免无效的 tensor 赋值。

```
def store_to_map(self, future_indices: FutureIndices, batch_result: GenerationBatchResult):
    if self.spec_algo.is_none():
        intv = future_indices.interval
```

```
            # 新增 guard：与 store_to_map_for_new_batch 保持一致，
            # 空区间（空闲 DP rank）无需存储任何内容，直接返回
            if self.is_empty_slice(intv):
                return
            self.token_ids_buf[intv] = batch_result.next_token_ids
        else:
            draft_input: EagleDraftInput = batch_result.next_draft_input
            self.store_to_map_for_new_batch(future_indices, draft_input)
```

评论区精华

无实质性讨论；自动化机器人 gemini-code-assist[bot] 仅给出正面评价，作者通过 /rerun-test 触发 DP attention 测试并已通过。

- 暂无高价值评论线程

风险与影响

- 风险：低风险：仅增加了提前返回 guard，且逻辑与同一文件中已有的 store_to_map_for_new_batch 方法完全对称，设计上无风险。
- 影响：仅影响 DP attention 空闲 rank 的执行路径，避免空 slice 写入 token_ids_buf，防止潜在的内存越界或无效操作。其他场景不受影响。
- 风险标记：边界情况修复

关联脉络

- PR #25380 [Disagg] Fix MegaMoE topk_ids dtype mismatch and FakeKVManager missing kv_args: 关联 issue 场景：同样涉及 DP attention 空闲 rank 的处理，本 PR 是补充遗漏的空 slice guard。
- PR #25444 Bundle Scheduler rank/size fields into a frozen ParallelState: 同为调度器模块重构，涉及并行状态管理。