PR #23327 完整报告

sgl-project/sglang

Skip mamba_pool_idx revert for session requests in _get_new_batch_prefill_raw

合并时间：2026-04-22 22:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23327

执行摘要

- 一句话：修复调度器中会话请求的 Mamba 池索引双重释放问题。
- 推荐动作：该 PR 值得精读，尤其是对于涉及会话管理和 Mamba 模型调度的开发者。关注点在于理解会话槽位生命周期与调度器批次管理之间的交互，以及如何通过简单的属性检查避免复杂的资源管理冲突。

功能与动机

根据 PR 描述，先前在 #21404 中引入的回滚路径会释放会话持有的 Mamba 槽位，当请求无法加入批次时，这会导致会话后续关闭时发生双重释放（实际工作负载中观察到 mamba_leaked=-9）。流式会话请求的 mamba_pool_idx 是通过 SessionSlot.restore_to_req 从会话槽恢复的，该槽位有意在多个轮次中保持相同索引。无条件释放 req.mamba_pool_idx 会使槽位的引用悬空，因此需要跳过会话请求的回滚。

实现拆解

1. 定位问题代码：在 python/sglang/srt/managers/scheduler.py 文件的 _get_new_batch_prefill_raw 方法中，找到负责回滚未加入批次请求的 Mamba 池索引释放的逻辑块。
2. 添加会话检查：修改条件判断，在原有检查（请求未加入批次且 mamba_pool_idx 不为空）的基础上，增加 not getattr(req, "session", None) 条件，以识别请求是否属于会话。
3. 更新注释说明：在代码中添加注释，解释仅释放在本批次中新分配的槽位（而非来自会话的预存槽位），因为会话持有的槽位有其独立的生命周期，在此处释放会导致双重释放。
4. 测试验证：通过现有 CI 测试和手动测试流式会话工作负载，验证修复后不再出现 mamba_leaked=-9 异常。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 _get_new_batch_prefill_raw）：这是唯一的变更文件，包含了调度器核心逻辑中修复双重释放问题的关键修改。

关键符号：_get_new_batch_prefill_raw

关键源码片段

python/sglang/srt/managers/scheduler.py

这是唯一的变更文件，包含了调度器核心逻辑中修复双重释放问题的关键修改。

```
# 在调度器的 _get_new_batch_prefill_raw 方法中，处理请求未成功加入批次时的回滚逻辑
added = len(adder.can_run_list) > 0 and req is adder.can_run_list[-1]
if (
    not added
    and req.mamba_pool_idx is not None
    and not getattr(req, "session", None) # 关键新增：仅当请求不属于会话时才执行释放
):
    # 释放 Mamba 池索引，避免内存泄漏
    self.tree_cache.req_to_token_pool.mamba_pool.free(
        req.mamba_pool_idx.unsqueeze(-1)
    )
    req.mamba_pool_idx = None
```

评论区精华

PR 的评论主要集中在 CI 流程和代码风格上，而非技术讨论。作者 ispobock 要求修复 lint 问题，并使用 /tag-and-rerun-ci 和 /rerun-test 命令触发特定会话相关测试的运行。测试结果显示所有相关测试通过，表明修复未引入回归。没有出现关于设计或实现细节的深度技术讨论。

- 代码风格与 CI 验证 (style): 作者提交修复后，CI 测试通过，确认变更符合代码规范且未破坏现有功能。

风险与影响

- 风险：回归风险：修改位于调度器的核心批次构建路径，如果条件判断逻辑错误，可能导致非会话请求的 Mamba 池索引泄漏（未释放）或会话请求的意外释放。但变更仅增加了一个属性检查，且测试通过，风险较低。 性能影响：新增的 getattr 调用可能引入微小开销，但位于批次构建的非关键路径，影响可忽略。 兼容性：无破坏性变更，完全向后兼容。
- 影响：对系统的影响：修复了特定场景下的内存泄漏问题，提升了系统在流式会话工作负载下的稳定性和资源管理正确性。 对用户的影响：终端用户可能不会直接感知，但避免了潜在的服务中断或性能下降。 对团队的影响：明确了会话槽位与调度器回滚逻辑的生命周期边界，为后续类似功能开发提供了参考。
- 风险标记：核心路径变更

关联脉络

- PR #21404 未知（根据 PR 描述引用）：本 PR 修复了 #21404 中引入的回滚路径导致的 Mamba 池索引双重释放问题。