PR #22213 完整报告

sgl-project/sglang

Fix streaming session busy check double-counting; add compat CI tests

合并时间：2026-04-12 16:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22213

执行摘要

- 一句话：修复流式会话繁忙内存检查重复计数问题，避免虚假泄漏断言。
- 推荐动作：建议精读该 PR，重点关注 SessionSlot 中 is_active 字段的设计决策，以及如何平衡内存检查准确性与重试幂等性。同时，留意提交历史中的迭代过程，理解内存管理复杂性和后续问题追踪。

功能与动机

根据 PR body 描述，当 restore_to_req 从槽位借出 KV 到活动请求时，槽位的 req_pool_idx 有意不清除（以支持分块预填充重试的幂等性）。这导致 session_held_tokens() 和 _get_total_uncached_sizes() 同时计数相同的 KV 页面，触发了虚假的内存泄漏断言。

实现拆解

实现方案分两部分：在内存管理模块的 python/sglang/srt/mem_cache/session_aware_cache.py 中，为 SessionSlot 类添加 is_active 布尔字段，并在 save_from_req 和 restore_to_req 方法中设置和清除该字段，同时修改 session_held_tokens()、session_held_swa_tokens() 和 session_held_req_count() 方法，跳过活动槽位的计数。在测试模块的 test/registered/sessions/test_streaming_session.py 中，启用环境变量 SGLANG_ENABLE_STRICT_MEM_CHECK_DURING_BUSY=2 以激活严格内存检查，并添加 TestStreamingSessionMixedChunk 测试类验证混合块场景下的兼容性。

关键文件：

- python/sglang/srt/mem_cache/session_aware_cache.py（模块 mem_cache）：核心内存管理文件，添加 is_active 字段并修改 KV 计数逻辑，直接解决重复计数问题。
- test/registered/sessions/test_streaming_session.py（模块 test）：测试文件，启用严格内存检查并添加混合块测试变体，验证修复效果和兼容性。

关键符号：SessionSlot.save_from_req, SessionSlot.restore_to_req, session_held_tokens, session_held_swa_tokens, session_held_req_count

评论区精华

Review 评论为空，但从提交历史中可见关键讨论点：提交消息显示，初始修复后发现了重叠调度问题（commit 94509ae），导致在繁忙内存检查中仍需跳过 KV 转移的请求以避免双重计数

；后续提交（如 88541edd）移除了严格检查以处理预存在的不匹配问题；最终提交（83acdb9）跳过了 retract 变体，因为识别出真实的令牌泄漏问题（约 14 个令牌），留待后续工作（如 #21875）解决。结论是修复了重复计数核心问题，但其他相关问题被标记为单独追踪。

- 重复计数问题修复 (correctness): 通过添加 is_active 字段并在相关方法中跳过活动槽位，成功修复双重计数问题。
- 测试覆盖与迭代调整 (testing): 最终测试启用严格检查（level=2）并添加混合块变体，但跳过了 retract 相关测试以隔离已知问题。

风险与影响

- 风险：技术风险包括：1. 核心内存管理逻辑变更（session_aware_cache.py）可能引入回归，影响流式会话的 KV 缓存跟踪准确性。2. 测试依赖环境变量 SGLANG_ENABLE_STRICT_MEM_CHECK_DURING_BUSY，若未正确设置可能导致测试覆盖不全。3. 提交历史显示有迭代调整（如跳过 retract 变体），表明底层内存检查逻辑复杂，可能存在未覆盖的边缘情况。
- 影响：对用户：修复了虚假内存泄漏断言，提升流式会话功能的稳定性和可靠性。对系统：改进了繁忙内存检查的准确性，避免错误触发断言导致服务中断。对团队：增强了测试覆盖，提供混合块场景的兼容性验证，有助于后续开发和维护。
- 风险标记：核心路径变更，测试依赖环境变量，边缘情况覆盖不足

关联脉络

- PR #21499 Add SWA support for runtime busy memory check: 都涉及运行时繁忙内存检查逻辑，修改了相关内存管理模块（scheduler_runtime_checker_mixin.py），功能上有交叉。
- PR #22562 [mem] Flatten memory checkers into composable per-pool invariant checks: 都涉及内存检查器重构，修改了相同或相关的内存管理文件（如 scheduler_runtime_checker_mixin.py），共享 observability 和 scheduling 标签。