PR #23496 完整报告

sgl-project/sglang

[session] fix mamba pool leak in StreamingSession.release_session + plumb idle leak check

合并时间：2026-05-02 11:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23496

执行摘要

- 一句话：修复 StreamingSession 释放时 mamba 池泄露
- 推荐动作：建议合并。该 PR 修复了一个明确的资源泄露问题，并且代码质量良好，遵循了已有的 session_held_* 模式。如果可能，后续可添加对应的单元测试，但并非阻塞条件。

功能与动机

StreamingSession.release_session 释放 KV token 和请求池槽位，但未释放槽位持有的 mamba 池状态（mamba_pool_idx + mamba_ping_pong_track_buffer）。HybridReqToTokenPool.alloc 为每个新请求批量分配一个 mamba_pool_idx，若启用额外缓冲区特性还会分配多个 ping-pong 跟踪缓冲区条目。SessionSlot.save_from_req 接管了所有这些资源，因此每次会话关闭都会永久泄漏 1 + ping_pong_size 个 mamba 槽位。对于依赖短生命周期会话（例如在 mamba 骨干上进行全双工 / 流式推理）的混合 SSM 模型，池会逐渐被耗尽，直至最大并发崩溃。另：在同类池检查中，_check_mamba_pool 为会话持有的槽位传递了硬编码 0，一旦有会话打开并持有 mamba 状态，空闲不变性检查每次都会触发假阳性泄露警告。

实现拆解

1. streaming_session.py：新增 _free_slot_mamba(slot) 辅助方法，通过 mamba_pool.free() 返回 mamba_pool_idx 和 mamba_ping_pong_track_buffer，若底层池无 mamba_pool 属性则无操作；在 release_session 末尾调用。新增 session_held_mamba_slots(active_pool_idxs) 方法，遵循已有 session_held_* 约定，排除当前批次中拥有请求的槽位。
2. base_prefix_cache.py：添加默认返回 0 的 session_held_mamba_slots 存根。
3. unified_radix_cache.py：添加对 self.session.session_held_mamba_slots 的传递。
4. scheduler_runtime_checker_mixin.py：添加 _session_held_mamba_slots() 辅助方法，传递给 _check_pool_invariant 替代硬编码 0。

关键文件：

- python/sglang/srt/session/streaming_session.py（模块 会话管理；类别 source；类型 core-logic；符号 session_held_mamba_slots, _free_slot_mamba）：核心改动：新增 _free_slot_mamba 和 session_held_mamba_slots，修复泄露及空闲检查入口。

- python/sglang/srt/managers/scheduler_runtime_checker_mixin.py（模块 调度器；类别 source；类型 core-logic；符号 _session_held_mamba_slots）：新增 _session_held_mamba_slots 辅助方法，并替换 _check_mamba_pool 中的硬编码 0，修复假阳性泄露警告。
- python/sglang/srt/mem_cache/base_prefix_cache.py（模块 内存管理；类别 source；类型 core-logic；符号 session_held_mamba_slots）：添加默认 session_held_mamba_slots 存根，保持接口一致性。
- python/sglang/srt/mem_cache/unified_radix_cache.py（模块 内存管理；类别 source；类型 core-logic；符号 session_held_mamba_slots）：添加对 self.session.session_held_mamba_slots 的传递调用。

关键符号：_free_slot_mamba, session_held_mamba_slots, _session_held_mamba_slots

关键源码片段

python/sglang/srt/session/streaming_session.py

核心改动：新增 _free_slot_mamba 和 session_held_mamba_slots，修复泄露及空闲检查入口。

```
# python/sglang/srt/session/streaming_session.py

def _free_slot_mamba(self, slot: SessionSlot) -> None:
    """Return a session slot's mamba pool state to the allocator."""
    # 通过 getattr 安全获取 mamba_pool，兼容无 mamba 池的配置
    mamba_pool = getattr(self.req_to_token_pool, "mamba_pool", None)
    if mamba_pool is None:
        return
    # 释放 mamba_pool_idx：每个会话槽位持有 1 个标量索引
    if slot.mamba_pool_idx is not None:
        mamba_pool.free(slot.mamba_pool_idx.unsqueeze(0))
        slot.mamba_pool_idx = None
    # 释放 ping-pong 跟踪缓冲区：可能为 None 或多个条目
    if slot.mamba_ping_pong_track_buffer is not None:
        mamba_pool.free(slot.mamba_ping_pong_track_buffer)
        slot.mamba_ping_pong_track_buffer = None

def session_held_mamba_slots(self, active_pool_idxs: Optional[set] = None) -> int:
    """Total mamba_pool entries held by session slots (mamba_pool_idx +
    mamba_ping_pong_track_buffer). Excludes slots whose owning req is
    currently in the batch -- those slots are counted via the normal
    alloc/free paths (same convention as the sibling ``session_held_*``
    accessors).
    """
    total = 0
    for slot in self.slots.values():
        # 如果请求正在批次中，其 mamba 槽位通过常规路径统计，此处跳过
        in_batch = (
            active_pool_idxs is not None and slot.req_pool_idx in active_pool_idxs
```

```
            )
            if in_batch:
                continue
            if slot.mamba_pool_idx is not None:
                total += slot.mamba_pool_idx.numel()
            if slot.mamba_ping_pong_track_buffer is not None:
                total += slot.mamba_ping_pong_track_buffer.numel()
        return total
```

评论区精华

作者在 review 评论中指出注释过长，并进行了修剪。没有其他实质性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更集中在会话释放和内存检查路径，影响范围限于使用 mamba 池的混合 SSM 模型。_free_slot_mamba 通过 getattr 检查 mamba_pool 属性，对不含 mamba 池的配置无影响。缺少单元测试覆盖，但作者提到已在内部验证。
- 影响：直接影响使用 StreamingSession 和 mamba 池的功能（如全双工 / 流式推理），避免池耗尽导致并发崩溃。修复空闲检查误报，提升监控准确性。对非混合 SSM 模型无影响。
- 风险标记：缺少测试覆盖，内部验证未开源

关联脉络

- PR #24244 [Bug] Size mamba mappings from req pool, not mamba pool: 同为 mamba 池相关的 bug 修复，修改了相同领域的内存池代码。
- PR #23696 [BugFix][HiMamba] Fix host-protected node deletion in HiMamba tombstone del: 同为 HiMamba 缓存的 bug 修复，涉及 mamba 池生命周期管理。