PR #24244 完整报告

sgl-project/sglang

[Bug] Size mamba mappings from req pool, not mamba pool

合并时间：2026-05-02 06:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24244

执行摘要

- 一句话：修复 Mamba 映射张量大小错误
- 推荐动作：值得精读，虽然改动小但揭示了内存池大小不匹配的潜在风险。设计上应确保索引张量与索引源（请求池）大小一致。

功能与动机

HybridReqToTokenPool._init_mamba_pool 使用 size 参数（代表 mamba 池大小）分配 req_index_to_mamba_index_mapping 等映射张量，但这些张量实际按 req_pool_idx 索引。当用户设置 --max-mamba-cache-size 小于 max_running_requests，或默认比例自动计算出的值低于请求池大小时，alloc() 中的 self.req_index_to_mamba_index_mapping[select_index] 会以超出 mamba 池大小的索引越界写入，引发静默内存损坏。

实现拆解

1. 重命名参数消除歧义：在 memory_pool.py 中，将 _init_mamba_pool 方法的 size 参数重命名为 mamba_size，明确其用途是 mamba 状态池大小。同时更新 __init__ 中的调用处，将 size=mamba_size 改为 mamba_size=mamba_size。
2. 修正映射张量大小：在 _init_mamba_pool 中，从 self.req_to_token.shape[0] 获取请求池大小 req_pool_size，并用于初始化 req_index_to_mamba_index_mapping 和 req_index_to_mamba_ping_pong_track_buffer_mapping 的第一个维度大小。
3. 同步修改 decode.py：在 python/sglang/srt/disaggregation/decode.py 的 HybridMambaDecodeReqToTokenPool.__init__ 中，将调用 _init_mamba_pool 时的实参名从 size= 改为 mamba_size=，以匹配签名变更。

关键文件：

- python/sglang/srt/mem_cache/memory_pool.py（模块 内存池；类别 source；类型 core-logic；符号 _init_mamba_pool, HybridReqToTokenPool.init）：核心修复文件：修正 _init_mamba_pool 中映射张量大小来源，并重命名参数。
- python/sglang/srt/disaggregation/decode.py（模块 PD 分离；类别 source；类型 core-logic；符号 HybridMambaDecodeReqToTokenPool.init）：decode 路径中调用 _init_mamba_pool 时需同步重命名实参名。

关键符号：_init_mamba_pool, HybridReqToTokenPool.init, HybridMambaDecodeReqToTokenPool.init

关键源码片段

python/sglang/srt/mem_cache/memory_pool.py

核心修复文件：修正 _init_mamba_pool 中映射张量大小来源，并重命名参数。

```
def _init_mamba_pool(
    self,
    mamba_size: int, # Renamed from 'size' to avoid confusion with req pool size
    mamba_spec_state_size: int,
    cache_params: BaseLinearStateParams,
    mamba_layer_ids: List[int],
    device: str,
    enable_mamba_extra_buffer: bool,
    speculative_num_draft_tokens: int = None,
):
    self.mamba_pool = MambaPool(
        size=mamba_size,
        spec_state_size=mamba_spec_state_size,
        cache_params=cache_params,
        mamba_layer_ids=mamba_layer_ids,
        device=device,
        enable_memory_saver=self.enable_memory_saver,
        speculative_num_draft_tokens=speculative_num_draft_tokens,
    )
    self.mamba_map = {layer_id: i for i, layer_id in enumerate(mamba_layer_ids)}

    self.device = device # Note: parent __init__ already sets this, considered redundant
    # Indexed by req_pool_idx, so size from the req pool buffer
    # (self.req_to_token.shape[0]), not from the mamba state pool size.
    req_pool_size = self.req_to_token.shape[0]
    self.req_index_to_mamba_index_mapping: torch.Tensor = torch.zeros(
        req_pool_size, dtype=torch.int32, device=self.device
    )
    if enable_mamba_extra_buffer:
        self.req_index_to_mamba_ping_pong_track_buffer_mapping: torch.Tensor = (
            torch.zeros(
                (req_pool_size, self.mamba_ping_pong_track_buffer_size),
                dtype=torch.int32,
                device=self.device,
            )
        )
```

python/sglang/srt/disaggregation/decode.py

decode 路径中调用 _init_mamba_pool 时需同步重命名实参名。

```
# Inside HybridMambaDecodeReqToTokenPool.__init__
self._init_mamba_pool(
    mamba_size=effective_mamba_size, # Renamed from 'size='
    mamba_spec_state_size=size + pre_alloc_size,
```

```
        cache_params=cache_params,
        mamba_layer_ids=mamba_layer_ids,
        device=device,
        enable_mamba_extra_buffer=self.enable_mamba_extra_buffer,
        speculative_num_draft_tokens=speculative_num_draft_tokens,
    )
```

评论区精华

review 中 gemini-code-assist[bot] 指出 self.device = device 在父类 __init__ 中已赋值，建议删除该冗余赋值并进一步重构（移除 device 参数，直接使用 self.device）。但作者未采纳该建议，仅修复了核心 bug。

- 冗余 self.device 赋值及参数重构 (design): 作者未采纳该建议，仅修复核心 bug。

风险与影响

- 风险：风险很低，仅 9 行变更。主要风险是 self.req_to_token 在调用 _init_mamba_pool 时可能尚未初始化（若子类覆盖了父类初始化顺序），但根据代码执行路径，_init_mamba_pool 在 super().__init__() 之后调用，而 req_to_token 在 ReqToTokenPool.__init__ 中分配，因此安全。另外，decode.py 中 mamba_spec_state_size 参数仍使用旧值 size + pre_alloc_size，未受影响。
- 影响：影响范围限于使用 Mamba 模型且满足触发条件的场景（--max-mamba-cache-size < max_running_requests 或自动计算值不足）。修复后这些场景将不再发生静默内存越界，行为正确。对非 Mamba 模型无影响。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #23696 [BugFix][HiMamba] Fix host-protected node deletion in HiMamba tombstone del: 同为 HiMamba 相关 bugfix，涉及 mamba 缓存管理。
- PR #19746 [P/D disagg] - support decode side radix cache: 引入了 decode 端 mamba 池和 HybridMambaDecodeReqToTokenPool，本 PR 修复了该处的一个潜在 bug。