

PR #31321 完整报告

sgl-project/sglang

[Bugfix] Release Mamba cache after PP dynamic chunk profiling

合并时间: 2026-07-16 02:32

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31321>

执行摘要

- 一句话: 修复 PP 动态分块 profiling 时 Mamba 缓存泄漏
- 推荐动作: 推荐合并。这是一个典型的资源泄漏 bugfix, 改动量小、逻辑清晰、复现路径明确, 且通过了 PP 相关 CI 测试。

功能与动机

PP 动态分块 profiling 阶段会构造多达 128 个合成 prefill 请求, 对于混合 Mamba 模型, 每个请求会同时分配 KV 缓存和 Mamba 缓存槽位。但之前的 cleanup 只调用了 `free(req)` 释放请求池条目和 KV 缓存, 未调用 `free_mamba_cache`, 导致 Mamba 缓存槽位持续被占用, 最终可能达到 `--max-mamba-cache-size` 的上限, 使动态分块被禁用或 server warmup 失败。PR body 中明确给出了复现命令和错误日志。

实现拆解

1. 在 `python/sglang/srt/managers/scheduler_pp_mixin.py` 的 `profile_and_init_predictor` 方法中, 定位到 profiling 循环结束后的 cleanup 区域 (第 717-724 行)。
2. 在释放 KV 缓存之后、调用 `self.req_to_token_pool.free(req)` 之前, 插入对 `req.mamba_pool_idx is not None` 的判断: 若不为 `None`, 则调用 `self.req_to_token_pool.free_mamba_cache(req)` 释放该 profiling 请求占用的 Mamba 缓存槽位。
3. 确保释放顺序正确: 先释放 KV 缓存和 Mamba 缓存, 再释放请求池条目, 避免悬挂指针或重复释放。
4. 该改动仅影响 profiling 初始化阶段的 cleanup 逻辑, 不影响稳态推理路径。

关键文件:

- `python/sglang/srt/managers/scheduler_pp_mixin.py` (模块 调度器; 类别 source; 类型 core-logic; 符号 `profile_and_init_predictor`): 核心改动文件, 在 PP 动态分块 profiling cleanup 路径中增加了 Mamba 缓存释放逻辑。

关键符号: `profile_and_init_predictor`, `free_mamba_cache`

关键源码片段

`python/sglang/srt/managers/scheduler_pp_mixin.py`

核心改动文件，在 PP 动态分块 profiling cleanup 路径中增加了 Mamba 缓存释放逻辑。

```
# 以下代码位于 profile_and_init_predictor 方法的 profiling 循环末尾
# 此处是在释放 KV 缓存后、释放请求池条目之前，增加了对 Mamba 缓存的释放
# Release KV and Mamba cache
if req.req_pool_idx is not None:
    kv_indices = self.req_to_token_pool.req_to_token[
        req.req_pool_idx, : req.extend_range.end
    ]
    self.token_to_kv_pool_allocator.free(kv_indices)
# 新增：如果该请求分配了 Mamba 缓存槽位，则释放
if req.mamba_pool_idx is not None:
    self.req_to_token_pool.free_mamba_cache(req)
self.req_to_token_pool.free(req)
req.kv = None
```

评论区精华

该 PR 的 review 过程非常简洁，未产生实质性技术讨论。ShangmingCai 在 Issue 评论中执行了 `/rerun-group pp` 触发 PP 相关 CI 重跑，结果通过。ShangmingCai 最终给予了 LGTM 批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅 3 行，位于 profiling 循环 cleanup 区域内，条件判断 `req.mamba_pool_idx is not None` 保证了仅在混合 Mamba 模型场景下触发，且释放顺序符合资源管理约定。未引入新的依赖或状态变更。主要风险在于若将来 `free_mamba_cache` 实现有变（如添加引用计数），此处未考虑，但当前实现是直接的槽位释放，无副作用。
- 影响：影响范围限定于 PP 动态分块 + 混合 Mamba 模型的初始化阶段。修复后，profiling 期间 Mamba 缓存不再泄漏，动态分块能正确完成初始化，server 正常启动。对稳态推理性能无影响。对非 Mamba 模型或非 PP 配置无任何行为变化。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR