PR #35931 完整报告

vllm-project/vllm

[Bugfix][LMCache][KVConnector] fix potential memory leak in LMCache multiprocess mode

合并时间：2026-03-08 05:52

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/35931

执行摘要

本 PR 修复了 LMCache 多进程模式下因未释放查找锁导致的内存泄漏问题，通过新增锁释放逻辑确保 LRU 驱逐策略正常工作，防止内存无限增长，提升系统稳定性。

功能与动机

修复 LMCache 多进程模式中潜在的内存泄漏。PR body 描述，当 vLLM 计算自身而非从 LMCache 检索的缓存块时，相应的查找锁未被释放，导致 LMCache 无法驱逐旧条目，内存使用持续增加。该问题在长时间运行后可能导致系统内存耗尽，影响生产环境可靠性。

实现拆解

修改集中在 vllm/distributed/kv_transfer/kv_connector/v1/lmcache_mp_connector.py 文件的 update_state_after_alloc 函数：

- 新增逻辑检查 tracker.num_lmcache_hit_blocks，若有命中块则处理锁释放。
- 根据 condition 标志决定释放范围：若无检索需求，释放所有锁定块；否则，仅释放 vLLM 计算的块的锁。
- 边界不对齐问题由 free_lookup_locks 函数内部通过 floor division 处理。

关键代码片段：

```
if tracker.num_lmcache_hit_blocks > 0:
    if not condition:
        free_end = tracker.num_lmcache_hit_blocks * self.vllm_block_size
    else:
        free_end = tracker.num_vllm_hit_blocks * self.vllm_block_size
    if free_end > 0:
        self.scheduler_adapter.free_lookup_locks(
            token_ids=list(tracker.all_token_ids),
            start=0,
            end=free_end,
            request_id=request.request_id,
        )
```

评论区精华

- 拼写错误修正：ApostaC 指出 "Bounday" 应为 "Boundary"，作者已修复。

- 日志消息优化：ApostaC 建议改进日志消息为 "Free locks of tokens %d-%d since it is cached by vLLM."，作者采纳。
- 未解决疑虑：maobaolong 提出 " 前缀块被驱逐后，后续块是否成为孤儿？"，此问题未进一步讨论，可能暗示设计风险。

风险与影响

- 风险：锁释放逻辑错误可能导致数据竞争或进一步内存问题；maobaolong 的问题指出潜在缓存一致性问题，需后续验证。
- 影响：正面影响使用 LMCache 多进程模式的用户，防止内存泄漏，增强系统可靠性和性能，支持大规模部署。

关联脉络

与近期 PR 相关：

- PR 37160 引入 SimpleCPUOffloadConnector，同属 KV 连接器模块，展示团队在 KV 缓存卸载方面的持续优化。
- PR 38659 标准化 KV 缓存相关逻辑，反映代码库的清理趋势。 本修复是 LMCache 功能完善的一部分，支持系统稳定性和性能提升。