PR #21624 完整报告

sgl-project/sglang

[HiCache] fix: Clone host indices to avoid memory leak

合并时间：2026-04-02 08:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21624

执行摘要

本 PR 修复了 SGLang 中 HiCache 模块的内存泄漏问题，通过克隆 host indices 避免长期引用旧存储。变更仅涉及两处代码修改，但显著提升了系统在启用分层缓存时的稳定性和资源管理能力。

功能与动机

动机：根据 Issue #19332，当启用 L3 KV 缓存（HiCache）时，SGLang 在 TCP 环境下显示内存使用持续增加，甚至导致 OOM。根本原因是 HiCache 在存储 alloc() 视图时，保留了旧 host-pool backing storage 的引用，导致内存泄漏。PR body 中提供了复现步骤，显示内存从 77g 增长到 102g。

实现拆解

核心修改点：

- 文件 python/sglang/srt/mem_cache/hi_mamba_radix_cache.py：
  - 在 write_backup 函数中，将 node.host_value = host_indices 改为 node.host_value = host_indices.clone()。
  - 在 mamba_backup_commit 函数中，同样添加 .clone()。
- 文件 python/sglang/srt/mem_cache/hiradix_cache.py：
  - 在 write_backup 函数中，将 node.host_value = host_indices 改为 node.host_value = host_indices.clone()。

关键逻辑：通过克隆 host indices，确保缓存节点持有独立副本，避免共享引用导致的长期内存占用。

评论区精华

review 讨论：

- gemini-code-assist[bot] 评论：" 克隆 host indices 以防止共享引用副作用。"
- xiezhq-hermann 批准变更。

Issue 反馈：dcosmos 确认修复有效，解决了 #19332 中的问题。

风险与影响

风险分析：

- 性能开销：.clone() 可能引入额外内存分配和拷贝，轻微影响性能，但权衡修复内存泄漏是必要的。
- 回归风险：变更范围小，回归风险低，但建议部署后监控内存使用。

影响评估：

- 用户影响：修复了启用 HiCache 时的内存泄漏，提升系统稳定性，减少 OOM 发生。
- 系统影响：改善资源管理，适用于长时间运行的高并发工作负载。

关联脉络

相关 PR：

- PR #21884：移除 HiRadixCache 中的 TTL 硬钉功能，同样涉及缓存管理优化。
- PR #21764：修复 HiCache 缓存命中统计，属于同一功能线的改进。

演进趋势：近期 HiCache 相关 PR 频繁，显示团队在持续优化缓存性能和可靠性，本 PR 是解决内存泄漏的关键一环。