PR #25991 完整报告

sgl-project/sglang

[HiCache] fix: truncate prefetch key on degraded allocation

合并时间：2026-06-03 22:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25991

执行摘要

- 一句话：修复 HiCache 降级分配时 prefetch 键截断错误
- 推荐动作：值得精读，因其展示了退化路径中细微但关键的 Bug 修复模式。鼓励为此分支编写单元测试。

功能与动机

当主机内存不足、prefetch_length 被迫缩小后，原代码截断原始 new_input_tokens 而非 prefetch_key，导致后续使用错误的键进行预取，可能触发键长度不匹配或无效查找。PR 描述指出“Crop prefetch_key when HiRadixCache falls back to a smaller host allocation for L3 prefetch”正是该修复的核心动机。

实现拆解

1. 变量修正：将 avaliable_size 拼写修正为 available_size（改进可读性）。
2. 截断对象调整：在退化分配分支（if host_indices is None）中，将 new_input_tokens = new_input_tokens[:prefetch_length] 改为 prefetch_key = prefetch_key[:prefetch_length]，因为 RadixKey 支持截断操作，且后续使用的都是 prefetch_key。
3. 空指针防御：在 host_indices = self.cache_controller.mem_pool_host.alloc(prefetch_length) 之后检查返回是否为 None，若是则释放保护、提前返回，防止空索引导致崩溃。
4. 文件：仅修改了 python/sglang/srt/mem_cache/hiradix_cache.py 中的 prefetch_from_storage 方法，无其他文件变更。

关键文件：

- python/sglang/srt/mem_cache/hiradix_cache.py（模块 缓存层；类别 source；类型 core-logic；符号 prefetch_from_storage）：唯一变更文件，核心预取逻辑所在，修复退化分配中键截断 Bug。

关键符号：prefetch_from_storage

关键源码片段

python/sglang/srt/mem_cache/hiradix_cache.py

唯一变更文件，核心预取逻辑所在，修复退化分配中键截断 Bug。

```
# python/sglang/srt/mem_cache/hiradix_cache.py
# 方法内退化分配分支（host_indices is None 后的降级路径）
if host_indices is None:
    # 修正前：`new_input_tokens = new_input_tokens[:prefetch_length]`
    # 修正后：直接截断 `prefetch_key`，因为 RadixKey 支持切片
    available_size = self.cache_controller.mem_pool_host.available_size()
    prefetch_length = available_size - (available_size % self.page_size)
    if prefetch_length >= self.prefetch_threshold:
        prefetch_key = prefetch_key[:prefetch_length]
        host_indices = self.cache_controller.mem_pool_host.alloc(
            prefetch_length
        )
        # 新增空指针检查：若仍分配失败则释放并返回
        if host_indices is None:
            last_host_node.release_host()
            return
    else:
        last_host_node.release_host()
        return
```

评论区精华

作者 alphabetc1 在自评中指出："RadixKey supports truncation, so writing it this way is OK." 确认了使用 [:prefetch_length] 截断 prefetch_key 的可行性。审核者 xiezhq-hermann 批准了该 PR，未提出异议。

- RadixKey 截断可行性确认 (design): 确认方法可行，无需额外转换。

风险与影响

- 风险：变更范围很小（+6/-3），仅影响主机内存不足时的降级预取路径，回归风险低。但缺乏对应的单元测试覆盖该退化分支，若未来 RadixKey 的切片语义发生变化，此代码可能静默失效。
- 影响：对用户无直接可感知影响，但可避免在主机内存紧张场景下的预取键长度错误，提升 HiCache 的健壮性。影响范围限定于启用了 HiCache 且主机内存分配可能失败的部署环境。
- 风险标记：缺少测试覆盖

关联脉络

- PR #25395 [UnifiedTree] Add CP sync: 与 HiCache 预取机制相关，均涉及 hiradix_cache.py 或同属 HiCache 功能线。
- PR #27049 docs: add DeepSeek-V4 EPLB Waterfill tips: 无直接关联，但同属仓库近期活跃开发，示明围绕 DeepSeek 和 KV 缓存的持续改进。