PR #26945 完整报告

sgl-project/sglang

[HiCache] feat: truncate mamba prefetch length to available host KV size

合并时间：2026-06-05 03:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26945

执行摘要

- 一句话：Mamba 预取长度截断至可用主机 KV 大小
- 推荐动作：值得关注，尤其是使用 HiCache + Mamba 模型的用户。设计上参考了 HiRadixCache 的做法，属于鲁棒性改进。建议精读 prefetch_from_storage 方法中的截断逻辑。

功能与动机

PR body 明确指出：当 HiMambaRadixCache 从存储预取时，主机 KV 分配可能在内存压力下失败，导致整个预取被放弃。需要镜像 HiRadixCache 的行为：分配失败时，将预取长度截断到页对齐的可用主机大小并重试，使得部分预取仍能发生，而不是完全跳过。

实现拆解

1. 修改文件python/sglang/srt/mem_cache/hi_mamba_radix_cache.py 中 prefetch_from_storage 方法。
2. 原始逻辑：调用 _alloc_with_evict 尝试分配主机 KV 内存，若返回 None 则立即释放保护节点并返回，放弃预取。
3. 新逻辑：在 _alloc_with_evict 失败后，查询 mem_pool_host.available_size() 获取当前可用主机 KV 大小，计算页对齐后的 prefetch_length，如果小于 prefetch_threshold 则放弃预取，否则截断 new_input_tokens 并改用 mem_pool_host.alloc（无驱逐）直接分配，然后继续后续的 mamba 槽位分配和预取操作。
4. 影响：仅修改了预取长度调整和替代分配路径，其余逻辑保持不变。Mamba 槽位分配 (mamba_prefetch_alloc) 不受影响，因为主机槽位是单次分配。

关键文件：

- python/sglang/srt/mem_cache/hi_mamba_radix_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，修改了 prefetch_from_storage 方法，在主机 KV 分配失败时增加截断重试逻辑。

关键符号：prefetch_from_storage

关键源码片段

python/sglang/srt/mem_cache/hi_mamba_radix_cache.py

唯一变更文件，修改了 prefetch_from_storage 方法，在主机 KV 分配失败时增加截断重试逻辑。

```
def prefetch_from_storage(
    self,
    req_id: str,
    last_host_node: TreeNode,
    new_input_tokens: List[int],
    last_hash: Optional[str] = None,
    prefix_keys: Optional[List[str]] = None,
):
    prefetch_length = len(new_input_tokens) - (
        len(new_input_tokens) % self.page_size
    )
    new_input_tokens = new_input_tokens[:prefetch_length]
    if (
        not self.enable_storage
        or prefetch_length < self.prefetch_threshold
        or self.cache_controller.prefetch_rate_limited()
    ):
        return

    self._protect_host_node(last_host_node, protect_mamba=False)

    # 尝试分配主机 KV 内存，带驱逐
    host_indices = self._alloc_with_evict(
        self.cache_controller.mem_pool_host,
        prefetch_length,
        self.evict_host,
    )
    if host_indices is None:
        # 主机内存不足时，截断预取长度到页对齐的可用大小
        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:
            self._release_host_node(last_host_node, release_mamba=False)
            return
        new_input_tokens = new_input_tokens[:prefetch_length]
        # 直接分配，不再驱逐（已失败过一次）
        host_indices = self.cache_controller.mem_pool_host.alloc(prefetch_length)

    if host_indices is None:
        self._release_host_node(last_host_node, release_mamba=False)
        return

    # 后续 mamba 槽位分配与预取流程不变
    extra_pools = self.mamba_prefetch_alloc(new_input_tokens, last_hash)
    # ... 省略后续代码
```

评论区精华

Review 中 hzh0425 询问：“此修复是否应该放在 UnifiedRadixTree 中？”作者 alphabetc1 回应：“我检查了，UnifiedRadixTree 已经支持预取截断。”表明设计决策是保持 HiMambaRadixCache 的特化处理，而非统一抽象。

- 是否应在 UnifiedRadixTree 中修复 (design): 作者回答 UnifiedRadixTree 已经支持预取截断，无需统一修复。

风险与影响

- 风险：低风险。变更仅影响主机 KV 分配失败的回退路径，且截断逻辑与 HiRadixCache 一致。可能的风险：若 available_size() 返回值过小（例如接近 0），截断后的 prefetch_length 可能小于阈值导致放弃，与预期一致。没有新增测试，回归风险可控。
- 影响：影响范围：仅 HiCache 场景下 Mamba 模型的内存预取行为。用户在主机内存紧张时可获得部分预取，减少完全未命中的情况，提升 Mamba 模型部署的鲁棒性。不影响非 Mamba 模型或非 HiCache 路径。
- 风险标记：缺少测试覆盖

关联脉络

- PR #27258 [HiSparse PD & PP]Fix HiSparse compatibility with PP decode: 同为 HiCache 相关 bugfix，涉及相似的内存管理模块。
- PR #27046 [HiCache] fix PD L3 cache hit details from decode responses: 同为 HiCache 修复，涉及缓存命中报告，与本 PR 共享 HiCache 上下文。