PR #23387 完整报告

sgl-project/sglang

[HiCache][SPEC] fix: empty key after page alignment in match_prefix

合并时间：2026-04-29 05:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23387

执行摘要

- 一句话：修复 page 对齐后空 key 索引越界
- 推荐动作：可直接合并。改动简洁清晰，修复了一个明确的边界条件 bug。建议后续补充针对空 key 或短 key 的单元测试，以防未来重构引入类似问题。

功能与动机

修复 HiCache 前缀匹配中，当输入 key 短于 page_size 时，page_aligned() 截断后 key 为空但仍进入 _match_prefix_helper 导致索引越界的 bug。PR body 明确指出 'When the input key is shorter than page_size, page_aligned() truncates it to empty' 以及 'an empty post-alignment key would proceed into _match_prefix_helper and cause index errors'。

实现拆解

1. 抽取空匹配结果函数：在 match_prefix 方法内新增 empty_match_result() 局部函数，封装返回空匹配结果的逻辑（MatchResult 包含空 tensor 和根节点）。
2. 调整 guard 顺序：将 self.disable 提前单独检查并返回空结果。然后将 key = params.key 和 maybe_to_bigram_view 移回原来位置，但关键改动是将 page_aligned() 调用移到 len(key) == 0 检查之前。
3. 合并空 key 处理：如果 page_aligned() 后 len(key) == 0，直接返回空匹配结果，避免进入 _match_prefix_helper。
4. 清理无效变量：删除未使用的 page_aligned_len 局部变量。
5. 不影响测试：仅涉及 hiradix_cache.py 一个文件，+9/-4 行，无测试配套变更。

关键文件：

- python/sglang/srt/mem_cache/hiradix_cache.py（模块 HiCache；类别 source；类型 core-logic；符号 empty_match_result）：唯一修改的文件，包含所有核心逻辑变更：新增 empty_match_result 函数、调整空 key 检查顺序、删除未使用变量。

关键符号：match_prefix, empty_match_result

关键源码片段

python/sglang/srt/mem_cache/hiradix_cache.py

唯一修改的文件，包含所有核心逻辑变更：新增 empty_match_result 函数、调整空 key 检查顺序、删除未使用变量。

```
# python/sglang/srt/mem_cache/hiradix_cache.py

    def match_prefix(self, params: MatchPrefixParams):
        # 预先创建空 tensor，用于返回空匹配结果
        empty_value = torch.empty((0,), dtype=torch.int64, device=self.device)

        # 新增：封装返回空匹配结果的逻辑，减少重复代码
        def empty_match_result():
            return MatchResult(
                device_indices=empty_value,
                last_device_node=self.root_node,
                last_host_node=self.root_node,
                host_hit_length=0,
            )

        # 先检查 disable 标志，如果 disabled 则直接返回空结果
        if self.disable:
            return empty_match_result()

        # 获取 key，并可能转为 bigram 视图（用于 EAGLE 推测解码）
        key = params.key
        key, _ = key.maybe_to_bigram_view(self.is_eagle)
        # 关键修复：将 page_aligned 移到空检查之前；
        # 当 key 短于 page_size 时，page_aligned() 会将其截断为空字符串
        key = key.page_aligned(self.page_size)
        # 检查对齐后是否为空，避免后续 _match_prefix_helper 索引越界
        if len(key) == 0:
            return empty_match_result()

        # 执行实际的前缀匹配
        value, last_node = self._match_prefix_helper(self.root_node, key)
        if value:
            value = torch.cat(value)
        else:
            value = empty_value

        # 处理被逐出的节点（host 层命中）
        host_hit_length = 0
        last_host_node = last_node
        while last_node.evicted:
            host_hit_length += len(last_node.host_value)
            last_node = last_node.parent
        while not last_host_node.backuped:
            last_host_node = last_host_node.parent

        return MatchResult(
```

```
            device_indices=value,
            last_device_node=last_node,
            last_host_node=last_host_node,
            host_hit_length=host_hit_length,
        )
```

评论区精华

无 reviewer 评论或讨论。PR 获得两位 reviewer 的 approve，合并流程顺畅。作者通过 /rerun-test 多次触发 CI，并成功通过 HiCache 相关单元测试（test_radix_cache_unit.py、test_unified_radix_cache_unittest.py 以及多个 test_hicache_* 测试）。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅 13 行且逻辑直观：将 page 对齐提前到空 key 检查之前，并抽取公共返回逻辑。不改变任何外部接口或行为语义，仅修复了边缘条件下的缺陷。但缺少专门针对空 key 情况的单元测试，回归依赖现有测试覆盖。
- 影响：影响范围仅限 HiCache 前缀匹配功能，在 key 短于 page_size 的极端场景下避免索引越界崩溃。对正常长度 key 的行为无影响。对其他模块（如推理、预取）无影响。
- 风险标记：缺少空 key 专项单元测试，边界条件修复

关联脉络

- PR #23631 [HiCache][SPEC] fix: normalize storage prefetch key: 同属 HiCache SPEC 修复系列，都涉及 key 处理逻辑。