PR #31181 完整报告

sgl-project/sglang

[Hicache][1/2]Support Mamba branching in Unified Radix Cache with HiCache

合并时间：2026-07-25 19:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31181

执行摘要

- 一句话：支持 HiCache 下 Mamba 分支的缓存命中
- 推荐动作：建议精读，尤其是 mamba_component.py 中 finalize_match_result 的 Mamba 分支点计算逻辑。该设计将 Full KV 命中长度与 Mamba 状态解耦，体现了良好的层次化缓存设计原则。

功能与动机

Unified Radix Cache 不支持 Mamba 分支，导致在某些分支场景下 Mamba 缓存无法命中（来自 PR body）。

实现拆解

1. 数据结构扩展：在 base_prefix_cache.py 的 MatchResult 中新增 full_kv_hit_length: int 字段，独立记录完整的 Full KV 命中长度，用于后续计算 Mamba 分支点。
2. 前缀匹配跟踪：在 unified_radix_cache.py 的 _match_prefix_helper 方法中，遍历 RadixTree 节点时累加 prefix_len，最终返回 full_kv_hit_length。返回值扩展为五元组，新增 _match_post_processor 参数传递该值。
3. Mamba 分支点计算：在 mamba_component.py 的 finalize_match_result 中，删除原先基于 value_chunks 的 HiCache 排除逻辑，改用 full_kv_hit_length 和 mamba_boundary_len（device 命中 + host 命中）计算 mamba_branching_seqlen。分支点取 full_kv_hit_length 向下对齐到 mamba_cache_chunk_size，且仅当该值大于 mamba_boundary_len 时才设置。
4. 单元测试：在 test_unified_radix_cache_unittest.py 中新增两个测试用例，分别验证 Device Full KV 命中（Mamba value 缺失但 Full KV 存在）和 Host Full KV 命中（Full KV 备份到 Host）场景下的分支长度正确，以及分支状态可重用。

关键文件：

- python/sglang/srt/mem_cache/base_prefix_cache.py（模块 缓存层；类别 source；类型 core-logic；符号 MatchResult）：定义了 MatchResult 数据结构，新增 full_kv_hit_length 字段，作为 Mamba 分支长度计算的依据。
- python/sglang/srt/mem_cache/unified_radix_cache.py（模块 缓存层；类别 source；类型 core-logic；符号 match_prefix, _match_prefix_helper, _match_post_processor）：修改 match_prefix 和 _match_prefix_helper，跟踪并返回 full_kv_hit_length，为

Mamba 分支计算提供数据。

- python/sglang/srt/mem_cache/unified_cache_components/mamba_component.py（模块 缓存组件；类别 source；类型 core-logic；符号 finalize_match_result）：核心变更：优化 Mamba 分支点计算逻辑，改为基于 full_kv_hit_length 和 mamba_boundary_len 计算，消除了对 HiCache 的特殊排除。
- test/registered/unit/mem_cache/test_unified_radix_cache_unittest.py（模块 缓存测试；类别 test；类型 test-coverage；符号 test_mamba_branching_seqlen_uses_device_full_hit_under_hicache, test_mamba_branching_from_host_full_is_reusable_after_insert）：新增两个测试验证 HiCache 下 Device/Full Host 的 Mamba 分支长度正确且可重用。

关键符号：finalize_match_result, _match_prefix_helper, _match_post_processor, zero_match_result

关键源码片段

python/sglang/srt/mem_cache/unified_cache_components/mamba_component.py

核心变更：优化 Mamba 分支点计算逻辑，改为基于 full_kv_hit_length 和 mamba_boundary_len 计算，消除了对 HiCache 的特殊排除。

```
def finalize_match_result(
    self,
    result: MatchResult,
    params: MatchPrefixParams,
    value_chunks: list[torch.Tensor],
    best_value_len: int,
) -> MatchResult:
    cow_mamba = params.cow_mamba
    req = params.req
    last_node = result.best_match_node

    # 计算当前 Mamba 状态的有效边界长度（设备上命中的 + 主机上命中的）
    mamba_boundary_len = len(result.device_indices) + result.host_hit_length

    # 基于完整的 Full KV 命中长度计算对齐后的 Mamba 分支点
    # 分支点是 full_kv_hit_length 向下对齐到 chunk 大小后的位置，
    # 且必须大于当前 Mamba 边界才会真正设置。
    aligned_seqlen = (
        result.full_kv_hit_length // self.mamba_cache_chunk_size
    ) * self.mamba_cache_chunk_size
    branching_seqlen = (
        aligned_seqlen if aligned_seqlen > mamba_boundary_len else None
    )

    mamba_value = last_node.component_data[self.component_type].value
    if cow_mamba and mamba_value is not None:
        # COW 逻辑保持不变，此处省略其他代码
```

```
    ...
return result
```

评论区精华

> Reviewer hzh0425 在 unified_radix_cache.py line 962 提问：'should we only record the device kv hit length?' 指出 full_kv_hit_length 累加的是遍历中的所有节点，包括 host，可能应该仅记录 device 命中。最终实现选择了记录完整的 Full KV 命中长度（包括 device 和 host），因为分支点计算需要结合 host_hit_length，因此需要包含 host 部分。该讨论已解决。

- full_kv_hit_length 是否应该只记录 device 命中？ (design): 作者最终选择了记录完整的 Full KV 命中长度（包括 device 和 host），因为 Mamba 分支计算需要结合 host_hit_length，因此需要包含 host 部分。该讨论已解决。

风险与影响

- 风险：新增 full_kv_hit_length 字段会影响 MatchResult 的序列化（若有），但当前仅用于内存计算，无序列化场景。匹配路径上增加累加操作，但前缀遍历已存在，性能开销可忽略。Mamba 分支计算逻辑从原先基于 value_chunks 改为基于 full_kv_hit_length，可能改变现有非 HiCache 场景的行为，但代码中通过条件判断确保非 HiCache 下行为一致（full_kv_hit_length 在非 HiCache 时仍会累加，但 mamba_boundary_len 在无 host 时等同于 device 长度，分支计算仍正确）。测试覆盖了 device 和 host 两种场景，但未覆盖混合多层 HiCache 的完整链路，且需依赖后续 PR 完成 L2 插入支持。
- 影响：用户：启用 HiCache 的 Mamba 模型（如 Qwen3-Next-80B）在共享前缀场景下的缓存命中率和吞吐量显著提升。系统：修改了 Radix Cache 的核心匹配路径，但保持向后兼容；新增的 full_kv_hit_length 字段仅供 Mamba component 使用，不影响其他组件。团队：后续需要完成第二部分 L2 插入支持，预计将进一步扩展缓存层次。
- 风险标记：核心路径变更，新字段匹配结构，测试覆盖有限

关联脉络

- 暂无明显关联 PR