PR #47782 完整报告

vllm-project/vllm

[Core] Preserve Marconi caching with selective hybrid cache retention

合并时间：2026-07-14 04:24

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47782

执行摘要

- 一句话：Mamba/SWA 共享前缀缓存保留优化
- 推荐动作：值得精读。PR 展示了如何在不改变上层接口的情况下，通过扩展内部数据结构（Request.shared_prefix_boundary）和统一分块停止条件，将 Marconi 缓存与稀疏 retention 策略优雅融合。_mamba_block_aligned_split 的重写是理解 vLLM 调度器核心逻辑的绝佳范例。建议关注 Mooncake 连接器的后续统一适配以及跨组共享前缀边界的精确性校验。

功能与动机

PR #37898 实现了 Marconi 缓存（arxiv 2411.19379），但仅支持 align 模式下的共享前缀缓存，未与 PR #43447 / #45845 引入的 retention-interval 稀疏缓存集成。这导致当用户设置 retention interval = 0 以优化长上下文 /agent 会话时，系统提示等共享前缀的缓存会被稀疏策略丢弃，无法跨请求复用。本 PR 在稀疏缓存层保留共享前缀边界，使零 retention interval 也能受益于隐式缓存的系统提示。

实现拆解

1. Request 新增共享前缀边界属性：在 vllm/v1/request.py 中为 Request 类增加 shared_prefix_boundary 字段（默认 0），记录需要保留的共享前缀块对齐位置，该值在整个调度步骤中保持稳定。
2. Cache Manager 返回边界：KVCacheManager.get_computed_blocks 的返回值从 (blocks, tokens) 扩展为 (blocks, tokens, shared_prefix_boundary)。新值由 shared_prefix_boundary = num_new_computed_tokens + num_uncached（若 num_uncached > 0） 计算，其中 num_uncached 来自 Coordinator.find_longest_cache_hit 的新返回。
3. Coordinator 暴露 uncached 公共前缀：HybridKVCacheCoordinator.find_longest_cache_hit 修改为返回 (blocks, hit_length, num_uncached_common_prefix_tokens)。当某一 attention 组的命中长度大于协调后的命中长度时，差值即为 uncached 公共前缀。
4. 调度器重写分块对齐逻辑：Scheduler._mamba_block_aligned_split 完全重写，从条件分枝变为统一计算最早停止点集合（下一块边界、最后缓存位置、partial-tail 边界、共享前缀边界），取其中最小正值作为分块结束点。共享前缀边界被计算为 start + (request.shared_prefix_boundary - start) // block_size * block_size，仅当边界落在当前分块内时生效。

5. 缓存层保留边界可达块：SingleTypeKVCacheManager.cache_blocks 将 request.shared_prefix_boundary（若非零）加入 reachable_boundaries 列表，传递给 reachable_block_mask。SWAReachableBlockMask.reachable_block_mask 方法接受 reachable_boundaries 参数，为每个边界保留其尾部 need 块，防止稀疏保留策略跳过它们。
6. 分布式连接器适配：mooncake/store/coordinator.py 更新 _reachable_masks 调用，以 (num_prompt_tokens - 1,) 作为 reachable_boundaries 传入。
7. 测试覆盖：tests/v1/core/test_prefix_caching.py 新增三个测试用例：test_mamba_reachable_block_mask_pins_shared_prefix（验证共享前缀被标记为保留）、test_mamba_shared_prefix_survives_zero_retention（验证零 retention 下共享前缀存活）、test_mamba_shared_prefix_reuse_under_zero_retention（验证后续请求复用该前缀）以及对应的 SWA 测试用例。同时适配已有测试以处理三元素返回值。

关键文件：

- vllm/v1/core/sched/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 _mamba_block_aligned_split）：核心调度器，_mamba_block_aligned_split 完全重写，将 Marconi 共享前缀边界纳入统一的块对齐停止条件，是保证缓存保留的关键决策点。
- vllm/v1/core/kv_cache_manager.py（模块 缓存管理层；类别 source；类型 core-logic；符号 get_computed_blocks）：get_computed_blocks 返回值扩展，新增 shared_prefix_boundary 计算，该边界是下游调度器和缓存层做出保留决策的依据。
- vllm/v1/core/single_type_kv_cache_manager.py（模块 单类型缓存；类别 source；类型 core-logic；符号 cache_blocks, reachable_block_mask）：cache_blocks 开始考虑 shared_prefix_boundary，并将其加入 reachable_boundaries 传递给 reachable_block_mask，使稀疏保留组（SWA/Mamba）不会丢弃该边界处的可达块。
- vllm/v1/core/kv_cache_coordinator.py（模块 缓存协调器；类别 source；类型 core-logic；符号 find_longest_cache_hit）：find_longest_cache_hit 返回值扩展，增加 num_uncached_common_prefix_tokens，用于计算 shared_prefix_boundary。对 HybridKVCacheCoordinator 的影响最大。
- vllm/v1/request.py（模块 请求模型；类别 source；类型 data-contract；符号 Request）：新增 shared_prefix_boundary 属性（整数类型，默认 0），作为跨步骤跟踪的共享前缀边界状态。
- tests/v1/core/test_prefix_caching.py（模块 前缀缓存测试；类别 test；类型 test-coverage；符号 test_mamba_reachable_block_mask_pins_shared_prefix, retained, test_mamba_shared_prefix_survives_zero_retention, cached_mamba_blocks）：大规模的测试适配和新增，验证共享前缀在 Mamba 和 SWA 下的保留与复用，是证明 PR 正确性的关键。
- vllm/distributed/kv_transfer/kv_connector/v1/mooncake/store/coordinator.py（模块 分布式连接器；类别 source；类型 core-logic；符号 _reachable_masks）：分布式连接器需要同步更新 _reachable_masks 以使用新的 reachable_boundaries 参数签名，否则远程缓存保留会失效。
- tests/v1/core/prefix_cache/test_partial_prefix_cache_hits.py（模块 偏前缀缓存测试；类别 test；类型 test-coverage）：适配函数签名变更，测试用例需要解包三元素返回值。

- vllm/v1/simple_kv_offload/manager.py（模块 简单卸载；类别 source；类型 core-logic）：微小调整适配新参数。

关键符号：_mamba_block_aligned_split, get_computed_blocks, find_longest_cache_hit, cache_blocks, reachable_block_mask, _reachable_masks

评论区精华

- chunked prefill 兼容性：@ivanium 询问 shared_prefix_boundary 是否需要在当前调度分片中检查，njhill 回复已包含在 _mamba_block_aligned_split 的停止条件中，分片调度自然会处理。
- Mooncake 连接器更新：@ivanium 建议作为 follow-up，njhill 在当前 PR 中同步更新了 coordinator.py 的 _reachable_masks 调用。
- shared_prefix_boundary 所有权争议：@tdoublep 提问为何将该值作为 request 属性而非函数参数，njhill 解释需要跨调度步骤跟踪且 request 对象已在多处被传递，避免额外字典管理更简洁。
- 先前 Marconi 缓存被禁用：@ivanium 指出修改前 _mamba_block_aligned_split 不消耗 num_uncached_common_prefix_tokens，导致 Marconi 逻辑失效。njhill 通过彻底重写该函数，将共享前缀边界纳入统一停止条件集修复。
- 函数签名统一建议：@ivanium 建议用 reachable_boundaries: Sequence[int] 代替分别传递 num_prompt_tokens 和 shared_prefix_boundary，被采纳。
  - chunked prefill 兼容性检查 (correctness): @njhill 确认在 _mamba_block_aligned_split 中会检查边界是否在 start 和 end 之间，且调度器自然分片，不会出现越界。
  - Mooncake 连接器更新范围 (design): @njhill 选择在当前 PR 中包含更新，以保证一致性。
  - shared_prefix_boundary 所有权的设计选择 (design): 维护为 request 属性，避免额外字典管理。
  - Marconi 缓存逻辑失效修复 (correctness): @njhill 通过彻底重写该函数，将共享前缀边界纳入统一停止条件集修复。
  - 函数签名统一 (design): 被采纳，简化了接口。

风险与影响

- 风险：
  1. 调度器重写回归风险（vllm/v1/core/sched/scheduler.py）：_mamba_block_aligned_split 完全重写，虽然简化了逻辑但改变了分块收敛的行为。Eagle 模式下的边界处理（last_cache_position -= block_size）需要确保与新的停止条件集合不冲突。
  2. 分布式协调器同步风险（vllm/distributed/kv_transfer/kv_connector/v1/mooncake/store/coordinator.py）：仅更新了 _reachable_masks 的参数，但 KVCacheCoordinator 基类的 find_longest_cache_hit 返回值变更（增加 num_uncached）可能影响其他分布式实现。目前其他远程协调器未适配，若被调用会类型错误。

3. shared_prefix_boundary 计算准确性（vllm/v1/core/kv_cache_manager.py）：shared_prefix_boundary = num_new_computed_tokens + num_uncached if num_uncached else 0 在多个 cache group 的协调中可能产生偏差，若 num_uncached 为 0 则边界不生效，可能遗漏需要保留的共享前缀。
4. 测试覆盖不足：新增测试主要在单节点单元测试层面，缺乏大规模多请求并发和分布式场景的集成测试，边缘条件（如同时使用 Eagle + Mamba + SWA）未覆盖。 - 影响：用户影响：使用 Mamba 或 SWA 层（如 DeepSeek、Qwen 等）的推理服务在长上下文和 agent 场景下首 token 延迟将明显下降，系统提示缓存不再被 retention interval 丢弃，且该优化无需修改模型配置或用户代码。系统影响：每个请求新增一个整数属性 shared_prefix_boundary，内存开销可忽略。调度器中分块停止计算由多个 if-elif 分支简化为候选集取最小值，理论上降低 CPU 开销。缓存层 reachable_block_mask 的额外计算仅在 reachable_boundaries 非空时发生，对纯 Full Attention 模型无影响。团队影响：后续需要为所有分布式 KVCacheCoordinator 子类（如远程调用实现）同步返回值签名变更。社区贡献者的 #47491 被此 PR 替代关闭。

- 风险标记：核心路径变更，调度器重写，多组协调器返回值变更，分布式同步风险，Eagle 模式边界条件

关联脉络

- PR #37898 [RFC] Marconi: Efficient Prefix Caching for Mamba-style Models: 基础 PR，实现了 Marconi 缓存（对齐模式），本 PR 在其上扩展 retention 保留。
- PR #43447 [RFC] Retention-based sparse KV cache for sliding window attention: 引入 retention-interval 稀疏缓存的起点，本 PR 为其增加共享前缀保留支持。
- PR #45845 Extend retention-based caching to Mamba states: 将 retention 机制扩展到 Mamba，本 PR 进一步确保共享前缀不被丢弃。
- PR #46384 [Core] Partial prefix cache hit for hybrid models: 引入偏前缀缓存命中机制，本 PR 使用了其偏尾边界作为停止条件之一。
- PR #47491 [BugFix] Preserve Mamba shared prefix cache under retention interval 0: 外部 contributor 的尝试，此 PR 更完整地解决了同一问题并被关闭。