# PR #48860 完整报告

vllm-project/vllm

[Bugfix] Prefix-cache metrics double-counted when a KV connector defers requests

合并时间：2026-07-21 22:02

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/48860

## 执行摘要

- 一句话：修复前缀缓存指标因连接器延迟被重复计数
- 推荐动作：建议合并此 PR，因为它修复了指标错误，设计清晰，测试完备。值得精读以理解调度器与缓存统计的交互，特别是 did_prefix_cache_lookup 标志和 record_prefix_cache_stats 的配合。

## 功能与动机

Issue #43736 报告了本地前缀缓存命中率在高负载下虚高。根本原因是当 KV 连接器返回 None（延迟请求）或分配槽位失败时，请求留在等待队列，但 get_computed_blocks 已在查找时记录了查询统计；下次重试时再次记录，导致同一个请求被计数多次。

## 实现拆解

1. 提取统一判断方法：在 KVCacheManager 中新增 prefix_cache_lookup_enabled(request)，集中判断是否应该执行缓存查找（基于 enable_caching 和 skip_reading_prefix_cache），替代 get_computed_blocks 中原有的内联条件。
2. 新增记录方法：在 KVCacheManager 中添加 record_prefix_cache_stats(request, num_hits)，内部调用 PrefixCacheStats.record，仅在 log_stats 且查找启用时执行。
3. 清理 get_computed_blocks：移除其中的统计记录逻辑，使其成为纯查找函数，只返回缓存块和命中信息。
4. 调度器改造：在 schedule 方法中，当 num_computed_tokens == 0 时设置 did_prefix_cache_lookup = True；在成功分配槽位后，若标志为 True 则调用 record_prefix_cache_stats。对混合 Mamba 模型的独立查找路径也做了相同调整，并新增缓存禁用时的早期返回。
5. 测试配套：更新 mock_kv 和 MockKVConnector（位于 tests/v1/core/utils.py 和 tests/v1/kv_connector/unit/utils.py）以支持 num_defers_before_matching 参数模拟延迟；添加四个新测试覆盖连接器延迟、抢占重录、缓存禁用和分配失败重试场景。

关键文件：

- vllm/v1/core/kv_cache_manager.py（模块 缓存管理器；类别 source；类型 core-logic；符号 prefix_cache_lookup_enabled, record_prefix_cache_stats）：核心变更：提取 prefix_cache_lookup_enabled 统一判断，新增 record_prefix_cache_stats 方法，从 get_computed_blocks 中移除统计记录。

- vllm/v1/core/sched/scheduler.py（模块 调度器；类别 source；类型 core-logic）：调度器调整：新增 did_prefix_cache_lookup 标志，在分配成功后才记录统计；对混合 Mamba 路径也做了相应处理。
- tests/v1/core/test_scheduler.py（模块 测试；类别 test；类型 test-coverage；符号 test_prefix_cache_query_not_inflated_by_connector_defer, test_preemption_re_records_prefix_cache_query, test_prefix_cache_stats_not_recorded_when_caching_disabled, test_prefix_cache_stats_counted_once_for_retried_then_scheduled_request）：新增四个测试用例，覆盖连接器延迟、抢占重录、缓存禁用和分配失败重试场景，确保统计行为正确。
- tests/v1/core/utils.py（模块 测试工具；类别 test；类型 test-coverage；符号 mock_kv）：更新 mock_kv 支持 num_defers_before_matching 参数，用于构造模拟延迟的连接器。
- tests/v1/kv_connector/unit/utils.py（模块 测试工具；类别 test；类型 test-coverage）：更新 MockKVConfig 和 MockKVConnector 以支持延迟模拟，使测试能够触发连接器延迟路径。

关键符号：prefix_cache_lookup_enabled, record_prefix_cache_stats, get_computed_blocks, schedule, mock_kv, get_num_new_matched_tokens, test_prefix_cache_query_not_inflated_by_connector_defer, test_preemption_re_records_prefix_cache_query, test_prefix_cache_stats_not_recorded_when_caching_disabled, test_prefix_cache_stats_counted_once_for_retried_then_scheduled_request

## 关键源码片段

### vllm/v1/core/kv_cache_manager.py

核心变更：提取 prefix_cache_lookup_enabled 统一判断，新增 record_prefix_cache_stats 方法，从 get_computed_blocks 中移除统计记录。

```
# kv_cache_manager.py 新增方法片段

    def prefix_cache_lookup_enabled(self, request: Request) -> bool:
        """Whether a local prefix cache lookup may be run for this request."""
        return self.enable_caching and not request.skip_reading_prefix_cache

    def record_prefix_cache_stats(self, request: Request, num_hits: int) -> None:
        # Don't count a request that skipped the cache lookup.
        if not self.log_stats or not self.prefix_cache_lookup_enabled(request):
            return
        assert self.prefix_cache_stats is not None
        self.prefix_cache_stats.record(
            num_tokens=request.num_tokens,
            num_hits=num_hits,
            preempted=request.num_preemptions > 0,
        )
```

get_computed_blocks 中原有的记录代码被替换为只调用 prefix_cache_lookup_enabled 早期返回，不再直接记录。

### vllm/v1/core/sched/scheduler.py

调度器调整：新增 did_prefix_cache_lookup 标志，在分配成功后才记录统计；对混合 Mamba 路径也做了相应处理。

```
# scheduler.py 中 schedule 方法的局部变更 num_external_computed_tokens = 0 load_kv_async = False connector_prefix_cache_queries, connector_prefix_cache_hits = 0, 0 did_prefix_cache_lookup = False # 新增标志 # Get already-cached tokens. if request.num_computed_tokens == 0: did_prefix_cache_lookup = True # 执行了查找 # ... 其余查找逻辑不变 ... # 在分配成功后，根据标志记录统计 if did_prefix_cache_lookup: self.kv_cache_manager.record_prefix_cache_stats( request, num_new_local_computed_tokens ) 注意：混合 Mamba 路径中原有的内联记录已移除，改为调用相同的 record_prefix_cache_stats。
```

### tests/v1/core/test_scheduler.py

新增四个测试用例，覆盖连接器延迟、抢占重录、缓存禁用和分配失败重试场景，确保统计行为正确。

```
# 测试用例片段（部分） deftest_prefix_cache_query_not_inflated_by_connector_defer(): """验证连接器延迟多次后，查询只记录一次。""" num_defers_before_matching = 3 scheduler = create_scheduler( enable_prefix_caching=True, use_kv_connector=mock_kv( matched_tokens=0, is_async=False, num_defers_before_matching=num_defers_before_matching, ), ) request = create_requests(num_requests=1, num_tokens=32, block_size=16)[0] scheduler.add_request(request) # Each deferred step re-runs the lookup but records nothing. for _ in range(num_defers_before_matching): assert not scheduler.schedule().scheduled_new_reqs output = scheduler.schedule() assert any(r.req_id == request.request_id for r in output.scheduled_new_reqs) stats = scheduler.kv_cache_manager.prefix_cache_stats assert stats is not None assert stats.requests == 1 assert stats.queries == request.num_tokens 其他测试类似，验证抢占重录、缓存禁用、分配失败重试场景。
```

## 评论区精华

njhill 在审批时补充了来自 #44431 的测试，确保缓存命中场景也被覆盖。

- 补充来自 #44431 的测试 (testing): 测试已补充

## 风险与影响

- 风险：主要风险是统计语义变更：此前从未被成功调度的请求也会被记录，现在则不会，可能影响依赖旧行为的监控；但这是预期的修复，使统计更准确。另一风险是

did_prefix_cache_lookup 标志在各路径上是否正确设置，尤其是混合 Mamba 模型的 find_longest_cache_hit_per_group 路径。测试已覆盖关键路径，降低回归风险。

- 影响：影响使用 V1 调度器和前缀缓存的用户，尤其是使用 KV 连接器（如 OffloadingConnector）的用户。指标将更准确反映实际缓存命中，不再因延迟或重试而虚高。无功能正确性影响。
- 风险标记：统计语义变更，核心路径变更，需确认监控适应新行为

# 关联脉络

- PR #43736 [Bug]: Local prefix cache hit rate can be over-counted when scheduling retries after KV allocation failure: 关联 Issue，报告了本 PR 修复的 bug
- PR #43822 fix: record prefix cache stats after allocation: 同一问题的早期修复尝试（已关闭）
- PR #44431 [Bugfix] Don't double-count local prefix cache stats on scheduling retries: 同一问题的更完整修复尝试（未合并），本 PR 采纳了其测试和部分思路
- PR #45202 [Bugfix] Don't double-count local prefix cache stats on scheduling retries: 同一问题的另一修复尝试（已关闭）