# PR #39655 完整报告

vllm-project/vllm

fix(lmcache): correct store for cached requests and num_scheduled_tokens in lmcache_mp_connector.py

合并时间：2026-04-13 11:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39655

## 执行摘要

- 一句话：修复 LMCache MP 连接器中缓存请求的 KV 存储逻辑和令牌计数错误。
- 推荐动作：该 PR 值得精读，特别是对于涉及 LMCache 和 KV 连接器模块的开发者。关注点：1. 如何正确处理缓存请求的增量令牌计数；2. LMCache 命中块在存储计算中的纳入逻辑，体现了 KV 存储的边界处理设计。

## 功能与动机

根据 PR 描述，修复两个与缓存请求的 KV 存储行为相关的 bug：1. 在 _process_cached_requests 中，num_new_tokens 错误地使用了 cached_reqs.num_computed_tokens[idx] 而非 scheduler_output.num_scheduled_tokens[request_id]，导致与 _process_new_requests 使用的增量 num_scheduled_tokens 不一致；2. min_available_blocks 的上界计算未包含 num_lmcache_hit_blocks，导致上界过低，可能跳过应暂存存储的块。

## 实现拆解

修改了 vllm/distributed/kv_transfer/kv_connector/v1/lmcache_mp_connector.py 中的两个函数：1. 在 GetStoreMetadata 函数中，将 computed_blocks 的计算从 tracker.num_scheduled_tokens // vllm_block_size 改为 tracker.num_scheduled_tokens // vllm_block_size + tracker.num_lmcache_hit_blocks，并更新 min_available_blocks 使用 computed_blocks；2. 在 _process_cached_requests 函数中，将 num_new_tokens 的赋值从 cached_reqs.num_computed_tokens[idx] 改为 scheduler_output.num_scheduled_tokens[request_id]。

关键文件：

- vllm/distributed/kv_transfer/kv_connector/v1/lmcache_mp_connector.py（模块 distributed/kv_transfer）：唯一修改的文件，包含 LMCache MP 连接器的核心逻辑，修复了 KV 存储的关键 bug。

关键符号：GetStoreMetadata, _process_cached_requests

## 评论区精华

review 中 gemini-code-assist[bot] 指出 GetStoreMetadata 函数中有一行冗余注释片段（'# the num_stored_blocks'），建议删除以保持代码清晰；ApostaC 批准了 PR 并感谢修复。讨论焦点在于代码清晰度而非技术争议，无未解决疑虑。

- 冗余注释片段 (style): PR 已合并，但未明确回应是否删除该注释；从 patch 看注释片段未被移除。

## 风险与影响

- 风险：风险较低：1. 变更集中在单个文件的两个函数，影响范围有限；2. 修复的是逻辑错误，可能引入回归的风险较小，但需确保测试覆盖缓存请求和 LMCache 命中场景；3. 无性能或安全风险；4. 兼容性无影响。
- 影响：对系统影响：修复了 KV 存储逻辑，确保缓存请求的令牌计数和块存储上界计算正确，避免潜在的数据不一致或存储跳过问题。对用户影响：间接提升推理稳定性和正确性，但无直接功能变更。对团队影响：代码更清晰，维护性提升。
- 风险标记：逻辑错误修复，缺少测试覆盖确认

## 关联脉络

- PR #39354 [KVConnector][NIXL] Organize NIXL connector into its own directory: 同属 kv-connector 模块，涉及 KV 连接器的重构和组织。
- PR #38709 [Core][Metrics] Remove vllm:prompt_tokens_recomputed metric: 同属核心模块，涉及缓存和指标相关的逻辑调整。
- PR #37688 [HMA] [KVEvent] Enable GPU-side KV events for HMA: 同属 KV 相关功能，涉及 KV 事件和存储逻辑。