PR #39709 完整报告

vllm-project/vllm

[CI][Metrics] Fix local_cache_hit assertion after prompt tokens metrics updates

合并时间：2026-04-13 23:16

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39709

执行摘要

本 PR 修复了 MultiConnector 边缘案例测试中 local_cache_hit 指标的断言，以适配 PR #38709 移除 recomputed 令牌计数后的指标语义变更。当所有提示令牌都被缓存时，调度器将 num_cached_tokens 减 1，该减量现被 local_cache_hit 吸收，因此测试断言需相应调整。变更仅涉及测试文件，风险低，但揭示了指标计算可能需调度器根本修复。

功能与动机

- 动机：PR #38709 移除了 PromptTokenStats.update_from_output() 中的 recomputed 令牌计数，改变了指标语义，导致相关测试断言失效。具体而言，当所有提示令牌都被缓存（本地 +NIXL）时，调度器减少 num_cached_tokens by 1，该减量现在被 local_cache_hit 指标吸收。
- 目标：临时更新 MultiConnector 边缘案例测试的断言，以匹配新的指标语义，确保 CI 测试通过。

实现拆解

仅修改了测试文件 tests/v1/kv_connector/nixl_integration/test_multi_connector_edge_cases.py，调整两个测试函数的断言：

函数名	原断言	新断言	变更说明
test_full_decode_gpu_cache_hit_metrics	assert d["local_cache_hit"] == cached	assert d["local_cache_hit"] == cached - 1	反映 local_cache_hit 吸收减量后的值
test_partial_decode_gpu_cache_hit_metrics	assert d["local_cache_hit"] == cached	assert d["local_cache_hit"] == cached - 1	同上

评论区精华

review 中无实质性讨论，但关联 Issue 评论提供了关键洞察：

> markmc: "Thanks for the quick fix. I don't expect #37460 to restore the old behaviour If we want to account 'correctly' for these recomputed tokens, we need the scheduler to report those metrics correctly rather than try to infer it :+1:"

这表明指标计算的根本问题可能需调度器修复，而非仅测试调整，暗示当前变更为临时方案。

风险与影响

- 技术风险：
  1. 测试断言调整可能未完全覆盖指标语义变更的所有场景，导致测试覆盖不足。
  2. 依赖 PR #38709 的指标变更，若该变更有误，本测试调整可能引入错误预期。
- 影响分析：
  - 对用户：无直接影响，仅测试调整。
  - 对系统：修复测试失败，维护 CI 稳定性。
  - 对团队：避免 CI 阻塞，但需关注指标计算的长期正确性。

关联脉络

- 直接关联：本 PR 修复由 PR #38709（移除 recomputed 令牌计数）导致的测试断言失败，两者在指标语义变更上紧密耦合。
- 潜在关联：PR #37460 可能涉及调度器修复，关联 Issue 评论暗示其可能不恢复旧行为，但需调度器正确报告指标。
- 模块关联：同属 kv-connector 模块的 PR #39655 也涉及缓存和令牌计数逻辑，反映该模块近期在缓存指标和连接器逻辑上的持续优化。
- 演进趋势：从近期历史 PR 看，v1 和 kv-connector 标签频繁出现，表明该模块在 v1 版本中活跃，涉及缓存、连接器和指标计算的迭代改进。