PR #39651 完整报告

vllm-project/vllm

[ROCm][CI] Removed stale tests and extended acceptance test

合并时间：2026-04-13 10:40

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39651

执行摘要

本次 PR 清理了 ROCm CI 配置中过时的 speculative decoding 测试套件，移除了 MI250、MI325、MI355 平台上的 7 个慢速测试，并将 MI325 的 acceptance 测试扩展为运行所有测试（而非仅慢速测试）。变更旨在精简 CI 执行，减少冗余测试时间，但对 speculative decoding 功能的测试覆盖率可能有所降低。DarkLight1337 已批准合并，表明团队认为清理合理。

功能与动机

根据 PR 描述，主要动机是“移除过时测试并将 acceptance 测试扩展到慢速和快速测试”。虽然没有关联 issue，但从变更内容推断，这些 speculative decoding 测试可能已不再适用或与其他测试重复，导致 CI 执行效率低下。清理旨在优化测试资源配置，加速 CI 流程。

实现拆解

仅修改了 CI 配置文件 .buildkite/test-amd.yaml，具体改动如下：

变更类型	硬件平台	具体调整
删除测试块	MI250、MI325、MI355	移除整个“V1 Speculative Decoding (slow)”测试步骤，包括 7 个子测试：test_eagle.py、test_extract_hidden_states.py、test_max_len.py、test_mtp.py、test_ngram.py、test_speculators_eagle3.py、test_tree_attention.py
修改测试命令	MI325	将 acceptance 测试命令从 pytest -v -s v1/spec_decode/test_acceptance_length.py -m slow_test 改为 pytest -v -s v1/spec_decode/test_acceptance_length.py，移除了慢速测试标记

变更集中在一个文件，逻辑简单：删除冗余测试块，调整单个测试的执行范围。

评论区精华

review 讨论较少，但 gemini-code-assist[bot] 的自动化评论指出了关键风险：

> “这些变更导致 speculative decoding 功能的测试覆盖率显著减少，特别是在 MI250 和 MI355 平台上没有添加替代测试，而在 MI325 平台上更新后的测试并未覆盖与已移除套件相同的逻辑。”

然而，DarkLight1337 直接批准了 PR，表明团队评估后认为这些测试确实过时或冗余，清理利大于弊。

风险与影响

风险：

1. 测试覆盖率降低：移除的 speculative decoding 测试可能包含独特验证逻辑，如果未在其他测试中覆盖，可能导致相关功能 bug 漏测。
2. 回归风险：如果这些测试实际上仍有效且必要，移除后可能无法及时发现 regression。

影响：

1. 对 CI 流程：减少测试执行时间，可能加速 PR 合并。
2. 对团队：需要确保 speculative decoding 功能仍有足够测试覆盖，可能需在其他地方补充测试。
3. 对用户：无直接影响，仅内部基础设施调整。

关联脉络

从近期历史 PR 看，本次 PR 与以下 PR 同属基础设施调整范畴：

- PR 39555：修复 ROCm CI 中多模态内存泄漏测试配置，同样关注测试稳定性。
- PR 39656：调整 XPU 依赖版本，涉及依赖管理。

这些 PR 共同反映了团队对 CI 配置的持续优化，特别是在 AMD 平台（ROCm）和基础设施维护方面。本次清理 speculative decoding 测试可能是更大规模测试重组的一部分，旨在提高 CI 效率。