

# PR #48219 完整报告

vllm-project/vllm

[CI] split tests to reduce CI time

合并时间: 2026-07-12 04:00

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48219>

## 执行摘要

- 一句话: 调整 CI 并行度减少测试时间
- 推荐动作: 该 PR 为纯粹的 CI 性能优化, 无代码逻辑变更, 值得合并。建议关注合并后 CI 实际耗时变化, 可为进一步优化提供参考。

## 功能与动机

PR body 明确所述目的是 "Split tests to reduce CI time", 即通过并行化测试执行来降低 CI 总耗时。

## 实现拆解

1. 修改 `.buildkite/test_areas/models_multimodal.yaml`:
  - 为 "Multi-Modal Processor (CPU)" 任务增加 `parallelism: 4` 配置, 使该任务在 4 个并行节点上分片运行。
  - 同时更新该任务的 `commands`, 在 `pytest` 命令中添加  
`--num-shards=$$BUILDKITE_PARALLEL_JOB_COUNT`  
`--shard-id=$$BUILDKITE_PARALLEL_JOB` 参数以支持分片。
  - 将任务标签追加 `%N` 后缀以区分并行实例。
2. 修改 `.buildkite/test_areas/models_basic.yaml`:
  - 将 "Basic Models Tests (Extra Initialization)" 任务的 `parallelism` 从 2 提升到 4。
  - 该步骤无需修改 `commands`, 因为命令已使用 `--num-shards` 和 `--shard-id`, 仅需增加并行节点数。

不涉及任何 Python 源代码、测试用例或依赖变更。

关键文件:

- `.buildkite/test_areas/models_multimodal.yaml` (模块 CI 配置; 类别 `config`; 类型 `configuration`): 为 CPU 多模态处理器测试增加并行分片配置, 是减少 CI 时间的核心变更之一。
- `.buildkite/test_areas/models_basic.yaml` (模块 CI 配置; 类别 `config`; 类型 `configuration`): 增加基础模型额外初始化测试的并行度, 从 2 提升到 4, 缩短该步骤耗时。

关键符号: 未识别

## 评论区精华

无实质 review 讨论。claude[bot] 自动评论表示由于是 fork 发起，禁用自动审查；khluu 直接批准。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低。变更仅为 CI 配置调整，不涉及任何业务逻辑。主要关注点：
  - 并行度增加可能对 CI 资源（GPU/CPU 节点）造成更大压力，但 Buildkite 流水线通常会根据可用资源自动调度，且已有其他任务使用类似的并行配置。
  - 理论上测试分片可能导致部分分片负载不均衡，但 pytest 的 --shard-id 按测试用例散列分配，通常均匀。
- 影响：
  - CI 执行时间：预期 "Multi-Modal Processor (CPU)" 任务（原超时 125 分钟）和 "Basic Models Tests (Extra Initialization)" 任务（原超时 100 分钟）的 wall-clock 时间将大致减半（并行度翻倍），从而缩短整体 CI 流水线时长。
  - CI 资源消耗：并行节点数增加，总计算资源消耗略有上升，但测试内容不变，总 CPU/GPU 时间基本持平。
  - 开发者体验：更快获得 CI 反馈，提升开发效率。
  - 风险标记：CI 资源竞争

## 关联脉络

- 暂无明显关联 PR