

PR #54823 完整报告

vllm-project/vllm

[CI] Remove MRV2-specific tests

合并时间: 2026-09-02 05:15

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/54823>

执行摘要

- 一句话: 移除冗余的 MRV2 专属 CI 测试, 将相关测试整合至现有分组。
- 推荐动作: 这是一个值得精读的清理性 PR。它清晰地展示了项目在架构演进 (V2 runner 成为默认) 后, 如何对 CI 基础设施和测试策略进行适应性调整。关键的设计决策在于: 通过删除专属配置文件和修改 skipif 逻辑的默认值来体现“默认”语义, 并通过测试迁移确保覆盖无损。对于关注 CI 效率、测试策略或 vLLM 内部演进的工程师有参考价值。

功能与动机

PR body 明确指出: 'These are redundant now that MRV2 is the default for all tests.'

核心动机是 V2 模型运行器已成为默认选项, 因此任何强制启用 V2

(`VLLM_USE_V2_MODEL_RUNNER=1`) 的测试分组和相关环境变量设置都失去了独立存在的意义, 需要进行清理和整合。

实现拆解

1. 删除冗余 CI 配置: 完全删除 `.buildkite/test_areas/model_runner_v2.yaml` 文件。该文件定义了多个独立的 CI 测试组 (如 Core、Examples、Distributed、Pipeline Parallelism、Spec Decode), 每个都在执行命令前设置 `export VLLM_USE_V2_MODEL_RUNNER=1`。删除此文件直接移除了这些冗余的测试任务。
2. 调整测试默认行为: 修改 `tests/v1/distributed/test_pp_dp_v2.py` 中的 `pytestmark` 跳过条件。将 `os.environ.get("VLLM_USE_V2_MODEL_RUNNER", "0") != "1"` 改为 `os.environ.get("VLLM_USE_V2_MODEL_RUNNER", "1") == "0"`。这使得测试在 V2 为默认值时自动运行, 只有当显式禁用 V2 时才跳过, 符合当前默认策略。
3. 整合测试到现有分组: 在 `.buildkite/test_areas/distributed.yaml` 中进行两处关键调整:
 - a. 在 `:nvidia: (L4) Distributed DP Basic` 测试组中, 新增了 `tests/v1/e2e/general/test_sharded_sampling.py` 和 `tests/v1/e2e/spec_decode/test_sharded_sampling.py` 的 `source_file_dependencies` 和执行命令 (`pytest -v -s v1/e2e/general/test_sharded_sampling.py` 等), 并将超时时间从 35 分钟增加到 45 分钟以容纳新增测试。
 - b. 在 `:nvidia: (L4) Pipeline + Context Parallelism` 测试组中, 新增了 `tests/v1/distributed/test_pp_dp_v2.py` 的依赖和执行命令 (`pytest -v -s v1/distributed/test_pp_dp_v2.py`)。这确保了原本在 `model_runner_v2.yaml` 中运行的测试不会丢失。

关键文件：

- .buildkite/test_areas/model_runner_v2.yaml（模块 CI 配置；类别 config；类型 deletion）：整个文件被删除，是本次清理的核心对象，移除了所有 MRV2 专属的 CI 测试任务定义。
- tests/v1/distributed/test_pp_dp_v2.py（模块 并行测试；类别 test；类型 test-coverage）：修改了测试的默认跳过条件，将行为从“需要显式启用 V2”改为“默认运行，除非显式禁用”，是本次变更在测试代码层面的关键体现。
- .buildkite/test_areas/distributed.yaml（模块 CI 配置；类别 config；类型 configuration）：整合了被移除的 V2 专属测试（如 sharded_sampling 和 pp_dp_v2），确保测试覆盖在清理后得以保留。

关键符号：未识别

关键源码片段

tests/v1/distributed/test_pp_dp_v2.py

修改了测试的默认跳过条件，将行为从“需要显式启用 V2”改为“默认运行，除非显式禁用”，是本次变更在测试代码层面的关键体现。

```
# 变更前：只有显式设置为 '1' 才运行，否则跳过 pytestmark = [ pytest.mark.skipif(
    os.environ.get("VLLM_USE_V2_MODEL_RUNNER", "0") != "1",
    reason="VLLM_USE_V2_MODEL_RUNNER=1 required", ), # ... ] # 变更后：默认运行，
只有显式设置为 '0' 才跳过 pytestmark = [ pytest.mark.skipif(
os.environ.get("VLLM_USE_V2_MODEL_RUNNER", "1") == "0", reason="V2 model
runner disabled via VLLM_USE_V2_MODEL_RUNNER=0", ), # ... ] 此变更使测试行为与“V2 是默认值”的前提保持一致。
```

评论区精华

本 PR 的讨论较少。自动化审查机器人 (claude[bot]) 指出因 PR 来自 fork 而禁用了自动审查，需维护者手动触发。另一位审查者 sfeng33 给予了批准，但未留下具体评论。核心变更（删除冗余配置和调整测试默认值）逻辑清晰，旨在响应 MRV2 成为默认的架构演进，因此未引发显著争议。

- MRV2 测试策略演进 (design): 变更被批准 (sfeng33 APPROVED)，确认删除专属配置文件并迁移测试的方案。

风险与影响

- 风险：
 1. 测试覆盖调整风险：整合测试时可能遗漏某些场景或依赖关系。但作者在第二个提交中明确将“未覆盖的测试移动到其他组”，降低了此风险。
 2. 环境变量默认值变更：test_pp_dp_v2.py 中跳过条件的变更（从“需要显式启用”变为“显式禁用才跳过”）可能影响其他依赖此逻辑的测试或配置，但此影响范围有限，仅限于该测试文件自身。

3. CI 流水线稳定性：删除配置文件和修改 YAML 可能会因语法或引用错误导致 CI 失败，但变更内容相对简单，风险较低。 - 影响：用户影响：无直接影响。系统影响：简化了 CI 配置，减少了测试任务数量，可能略微加快 CI 总运行时间。团队影响：降低了维护成本，统一了 V2 作为默认运行器的测试策略。开发者不再需要在 MRV2 专属测试和通用测试之间做区分。 - 风险标记：测试覆盖率调整，环境变量默认值改变

关联脉络

- PR #54752 [CI] Shard distributed model jobs above the 24h P90 threshold: 同属 CI 构建配置优化，目标是降低测试耗时，与本 PR 的 CI 清理工作属于同一演进脉络（提升 CI 效率与可维护性）。
- PR #52353 [CI] Split nightly MTP acceptance tests: 同样通过调整 CI 配置来优化测试执行（分片），与本 PR 共同体现了对 CI 流水线持续优化的趋势。