

PR #55731 完整报告

vllm-project/vllm

[CI] reduce npu CI use time and add timeout

合并时间: 2026-09-07 22:51

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/55731>

执行摘要

- 一句话: 优化 Ascend NPU CI 流程, 移除耗时构建并增加测试超时。
- 推荐动作: 此 PR 是一个有价值的 CI 效率优化, 建议合并。其核心改动 (移除构建步骤、增加超时、缩减测试范围) 直接针对了耗时问题, 且改动集中、影响可控。值得关注的决策是: 1) 以依赖预置换取构建速度; 2) 以跳过失败测试换取 CI 快速通过。这些权衡在讨论中已得到维护者认可, 并计划后续修复。代码审阅者应确认 `timeout_in_minutes` 的设置值是否足够缓冲正常的测试波动。

功能与动机

根据 PR body 描述, 当前 NPU 测试用例平均运行时间超过 1 小时, 其中 Docker 容器内编译安装 vllm 和 vllm-ascend 耗时 40 分钟, vllm-interface 测试用例本身耗时超过 20 分钟。因此需要优化 CI 流程以减少总耗时。

实现拆解

1. 移除耗时构建步骤: 在 `.buildkite/scripts/hardware_ci/run-npu-test.sh` 的 Dockerfile 部分, 删除了安装 vllm 和 vllm-ascend 依赖及其源码安装的多个 RUN 指令。此变更假设 CI 环境已预置这些依赖或使用缓存, 从而将构建时间从约 40 分钟大幅缩减。
2. 添加测试超时: 在 `.buildkite/hardware_tests/ascend_npu.yaml` 配置文件中, 为 Ascend NPU Test 步骤新增 `timeout_in_minutes: 30` 字段。这为测试执行设置了硬性时间上限, 防止因个别测试卡住而无限期占用 CI 资源。
3. 缩减测试范围: 在 `.buildkite/scripts/hardware_ci/run-npu-test.sh` 的末尾, 将执行的 `pytest` 命令从 `tests/e2e/vllm_interface/` 目录改为仅运行 `tests/e2e/vllm_interface/test_vllm_pr_interface_compatibility.py` 单个测试文件。这跳过了已知失败的测试用例, 以聚焦于核心接口兼容性验证。
4. 配套设施: 本次变更仅涉及 CI 脚本和配置, 未修改任何业务源码或核心测试逻辑, 属于纯粹的 CI 基础设施优化。

关键文件:

- `.buildkite/scripts/hardware_ci/run-npu-test.sh` (模块 CI 脚本; 类别 `infra`; 类型 `infrastructure`): 包含 Dockerfile 构建流程和测试执行命令的核心改动, 是减少耗时和调整测试范围的关键文件。

- `.buildkite/hardware_tests/ascend_npu.yaml` (模块 NPU 测试配置; 类别 config; 类型 configuration): 定义了 NPU 测试的 CI 配置, 是添加超时限制的地方。

关键符号: 未识别

关键源码片段

`.buildkite/scripts/hardware_ci/run-npu-test.sh`

包含 Dockerfile 构建流程和测试执行命令的核心改动, 是减少耗时和调整测试范围的关键文件。

```
# 移除耗时的依赖安装和源码安装步骤, 依赖预置或缓存
# 以下是移除的安装 vllm 及其依赖的代码块 (示意, 非当前片段)
# COPY requirements/common.txt /workspace/vllm/requirements/common.txt
# RUN ... uv pip install -r requirements/common.txt ...
# COPY . .
# RUN ... uv pip install -e /workspace/vllm/ ...
# RUN ... uv pip install -r /workspace/vllm-ascend/requirements.txt ...
# RUN ... uv pip install -e /workspace/vllm-ascend/ ...

# 修改后的测试执行命令, 仅运行核心兼容性测试
pytest -v -s tests/e2e/vllm_interface/test_vllm_pr_interface_compatibility.py
```

评论区精华

在 review 过程中, 主要讨论围绕 CI 运行的触发和结果展开。维护者 [ZJY0516](#) 在批准 PR 并运行 CI 后, 指出了一个新的失败问题: `ModuleNotFoundError: No module named 'huggingface_hub'`。他决定先合并此 PR, 并让作者在后续 PR 中修复此依赖缺失问题。这表明优化后的流程暴露了一个新的环境配置缺陷。

- CI 运行触发与环境问题 (other): 维护者决定先合并此 PR, 让作者在后续 PR 中修复此环境依赖缺失问题。

风险与影响

- 风险:
 1. 环境依赖风险: 移除了构建步骤中显式安装依赖的命令。如果 CI 基础镜像未预装这些依赖 (如讨论中出现的 `huggingface_hub` 缺失), 或缓存策略失效, 将导致测试环境不完整而失败。
 2. 测试覆盖风险: 将测试范围从整个 `tests/e2e/vllm_interface/` 目录缩小到单个文件, 虽然解决了已知的失败问题, 但也可能掩盖其他新出现或潜在的兼容性问题。
 3. 超时设置合理性: 30 分钟的超时值是基于之前 20+ 分钟的测试时间设定的, 但未来若测试用例增多或性能下降, 可能需要调整此值。
 - 影响: 此变更主要影响 vllm 项目的 Ascend NPU CI 流水线。
 - 对团队: 显著缩短了 NPU 相关的 CI 反馈周期 (从约 1 小时降至 30 分钟内), 提升了开发者效率。同时, 超时机制增强了 CI 资源的可预测性和管理性。
 - 对系统: 降低了 CI 系统的负载和运行成本。
 - 对用户: 间接影响, 用户将更早获得与 NPU 相关的代码变更反馈。
 - 风险标记: 环境依赖缺失风险, 测试覆盖缩减

关联脉络

- 暂无明显关联 PR