PR #44687 完整报告

vllm-project/vllm

[CI/Build] Skip test_use_trtllm_attention on non-CUDA platforms

合并时间：2026-06-11 06:18

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44687

执行摘要

- 一句话：TRT-LLM 测试在非 CUDA 平台跳过
- 推荐动作：该 PR 是简单的测试维护，无精读必要。可关注其统一的跳过模式，未来类似测试可参考。

功能与动机

避免在非 CUDA 平台（如 ROCm、CPU）上运行 TRT-LLM 相关测试，因为 TRT-LLM 仅支持 CUDA。该变更由 @AndreasKaratzas 在 #43232 中提出作为额外清理。

实现拆解

1. 导入调整：在 tests/kernels/attention/test_use_trtllm_attention.py 顶部增加 from vllm.platforms import current_platform。
2. 模块级跳过：在导入之后、测试函数之前插入 if not current_platform.is_cuda(): pytest.skip("TRTLLM attention is only supported on CUDA platforms.", allow_module_level=True)。
3. 模式统一：与同级文件 test_trtllm_kvfp8_dequant.py 已有的跳过逻辑保持一致。

关键文件：

- tests/kernels/attention/test_use_trtllm_attention.py（模块 测试文件；类别 test；类型 test-coverage）：唯一变更文件，添加了模块级 CUDA 平台跳过。

关键符号：未识别

关键源码片段

tests/kernels/attention/test_use_trtllm_attention.py

唯一变更文件，添加了模块级 CUDA 平台跳过。

```
# tests/kernels/attention/test_use_trtllm_attention.py
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project

from unittest.mock import patch

import pytest
```

```
import torch

from vllm.platforms import current_platform
from vllm.utils.flashinfer import (
    can_use_trtllm_attention,
    supports_trtllm_attention,
    use_trtllm_attention,
)

# 模块级跳过：非 CUDA 平台直接跳过整个文件
if not current_platform.is_cuda():
    pytest.skip(
        "TRTLLM attention is only supported on CUDA platforms.",
        allow_module_level=True,
    )

MODEL_CONFIGS = {
    "Llama-3-70B": dict(num_qo_heads=64, num_kv_heads=8),
    # ... 其余配置
}
```

评论区精华

无实质性讨论。审核者 @AndreasKaratzas 和 @yewentao256 均批准，仅提及需要从 main 合并。

- 平台跳过方法 (design)：已合并 main，PR 被批准。

风险与影响

- 风险：无显著风险。变更仅限于测试文件，仅影响非 CUDA 平台上的测试执行。正确性风险低。
- 影响：影响范围小。仅对 tests/kernels/attention/test_use_trtllm_attention.py 文件生效，非 CUDA 平台（如 ROCm、CPU）运行测试套件时将跳过该文件，避免不必要的失败。
- 风险标记：低风险，仅测试文件变更

关联脉络

- PR #43232 Unrelated Issue: PR 提及该变更是 #43232 讨论中 out-of-scope 清理的结果。