PR #42326 完整报告

vllm-project/vllm

[AMD] skip machete tests for rocm

合并时间：2026-05-13 20:11

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42326

执行摘要

- 一句话：跳过 ROCm machete 测试
- 推荐动作：此 PR 是简单但必要的平台兼容性修复，无需深入审查。

功能与动机

AMD 的 gfx942 和 gfx950 虽然有设备能力 90，但实际与 machete (CUTLASS W4A16) 内核不兼容，需要在测试中跳过以避免错误。PR body 明确指出 'gfx942 and gfx950 have device capability 90 but aren't compatible with Machete'。

实现拆解

仅修改 tests/quantization/test_cutlass_w4a16.py 第 17 行的条件判断，从原来的 `if not current_platform.has_device_capability(90)` 扩展为 `if not current_platform.has_device_capability(90) or current_platform.is_rocm()`，当平台为 ROCm 时也跳过测试。

关键文件：

- tests/quantization/test_cutlass_w4a16.py（模块 测试；类别 test；类型 test-coverage）：核心变更文件，增加 ROCm 平台跳过逻辑，防止在不兼容条件下执行测试。

关键符号：未识别

关键源码片段

tests/quantization/test_cutlass_w4a16.py

核心变更文件，增加 ROCm 平台跳过逻辑，防止在不兼容条件下执行测试。

```
# ... 文件头部注释和导入

from vllm.platforms import current_platform

# 增加 is_rocm() 条件：ROCm 平台虽可能满足 device capability 90，
# 但实际与 Machete 内核不兼容，必须跳过。
if not current_platform.has_device_capability(90) or current_platform.is_rocm():
    pytest.skip(
        "Machete W4A16 requires Hopper (sm_90).",
```

```
        allow_module_level=True,
    )

# 后续导入和测试代码保持不变 ...
```

评论区精华

无实质讨论，review 均直接批准。机器人评论确认变更合理。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试跳过条件，不影响任何生产代码。若未来 AMD GPU 支持 machete，需记得移除此条件。
- 影响：影响范围仅限于 ROCm 平台上运行 cutlass W4A16 测试用例的 CI，不会再因不兼容内核导致测试失败。
- 风险标记：仅测试变更

关联脉络

- PR #42411 [ROCm] Run AITER RMSNorm pad fusion before AR RMS fusion: 同样涉及 ROCm 平台的兼容性调整，属于同一维护方向。