PR #39730 完整报告

vllm-project/vllm

[ROCm][CI] Fix condition for `test_per_token_group_quant_fp8_packed`

合并时间：2026-04-15 00:14

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39730

执行摘要

该 PR 修复了 ROCm 平台上因新增 DeepGEMM 量化测试导致的 CI 失败问题，通过将测试跳过条件从 torch.cuda.is_available() 改为 current_platform.is_cuda() 并更新跳过原因，确保测试在非 CUDA 平台上被正确跳过。这是一个针对测试基础设施的小幅调整，风险极低，主要影响跨平台 CI 稳定性。

功能与动机

PR body 明确指出：自 PR #39547 引入 test_per_token_group_quant_fp8_packed 测试后，该测试在 ROCm 平台上失败，因为测试的功能（DeepGEMM）在 ROCm 上不可用。因此需要更新 pytest.skip 条件，使该测试在非 CUDA 平台上被跳过，以维护 CI 的稳定性。

实现拆解

仅修改了 tests/kernels/quantization/test_per_token_group_quant.py 文件：

1. 导入平台抽象层：添加 from vllm.platforms import current_platform。
2. 更新跳过条件：将 @pytest.mark.skipif(not torch.cuda.is_available(), reason="CUDA not available") 改为 @pytest.mark.skipif(not current_platform.is_cuda(), reason="DeepGEMM not available on this platform")。

关键代码变更：

```
from vllm.platforms import current_platform

@pytest.mark.skipif(
    not current_platform.is_cuda(), reason="DeepGEMM not available on this platform"
)
def test_per_token_group_quant_fp8_packed(...):
    ...
```

评论区精华

review 中只有一条实质性讨论：

> AndreasKaratzas: "Can we also modify the reason? I see that this is a DeepGEMM test. So probably let's update this and mention something like 'DeepGEMM is not supported on this platform'"

作者在后续 commit 中采纳了该建议，将跳过原因从通用的 "CUDA not available" 更新为更准确的 "DeepGEMM not available on this platform"，体现了对测试意图的清晰传达。

风险与影响

风险分析：

- 仅修改测试装饰器，不影响任何生产代码逻辑，回归风险几乎为零。
- 使用 current_platform.is_cuda() 比 torch.cuda.is_available() 更符合项目平台抽象层设计，理论上能更准确地检测 CUDA 平台。

影响分析：

- 对终端用户无直接影响，仅影响内部 CI 流程。
- 修复了 ROCm 平台 CI 失败问题，确保跨平台测试套件稳定性。
- 推广了平台抽象层 current_platform 的使用，有助于统一项目中的平台检测逻辑。

关联脉络

- 关联 PR #39547：根据 PR body，本 PR 修复的测试是在 PR #39547 中新增的，该测试引入了 DeepGEMM 功能，但在 ROCm 平台上不可用，导致 CI 失败。
- 技术趋势：从 torch.cuda.is_available() 到 current_platform.is_cuda() 的变更，反映了项目在多平台支持（如 ROCm、XPU）背景下，对平台检测基础设施的持续标准化努力，这与近期多个涉及平台适配的 PR（如 #39776、#38061）一脉相承。