

PR #45967 完整报告

vllm-project/vllm

[ROCm][CI] skip test_double_aiter_rms_quant_fusion

合并时间: 2026-06-22 05:53

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45967>

执行摘要

- 一句话: 跳过因编译器优化失效的 ROCm 测试
- 推荐动作: 该 PR 属于临时性测试适配, 技术深度较低, 无需精读。但值得注意其展示了编译器优化如何使 tests 与 code 产生脱节, 团队可考虑在 fusion pass 设计时增加更鲁棒的测试模式 (如强制保留算子或使用独立子图)。

功能与动机

PR body 指出单元测试 `test_double_aiter_rms_quant_fusion` 在 ROCm 上持续失败: 预期 `DoubleQuant` 融合模式应匹配 (`matched_count == 1`), 实际结果为 0。分析发现 PyTorch 编译器进行了优化, 将两次 `rocm_aiter_group_fp8_quant` 调用中的一次删除, 使得原本设计为 1-to-2 扇出的 fusion pass 无法触发。作者认为该测试用例在编译器优化后已不再适用, 因此跳过。

实现拆解

1. 跳过失效测试: 在 `tests/compile/passes/test_double_aiter_rms_quant_fusion.py` 中, 将 `@pytest.mark.skipif(not is_aiter_found_and_supported(), ...)` 替换为 `@pytest.mark.skip(reason="...")`, 无条件跳过测试; 同时移除 `is_aiter_found_and_supported` 的导入。
2. CI 配置增强: 在 `.buildkite/test_areas/pytorch.yaml` 中为 PyTorch Compilation Passes Unit Tests 步骤添加 `mirror.amd` 配置, 指定 `device: mi300_1`、超时 180 分钟, 并依赖 `image-build-amd`, 确保该测试在 AMD GPU 上正确执行。

关键文件:

- `tests/compile/passes/test_double_aiter_rms_quant_fusion.py` (模块 测试适配; 类别 test; 类型 test-coverage): 核心变更文件, 跳过因编译器优化而失效的测试; 移除不再需要的导入和 `skipif` 条件。
- `.buildkite/test_areas/pytorch.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 为 PyTorch Compilation Passes 测试步骤添加 AMD GPU 镜像配置, 确保在 `mi300_1` 上运行。

关键符号: 未识别

关键源码片段

tests/compile/passes/test_double_aiter_rms_quant_fusion.py

核心变更文件，跳过因编译器优化而失效的测试；移除不再需要的导入和 skipif 条件。

```
# 跳过因编译器优化而失效的测试
@pytest.mark.skip(
    reason="Skipping for now because pytorch compiler removes one the two quant ops"
)
def test_double_aiter_rms_fp8_group_quant_fusion(
    model_cls: type[torch.nn.Module],
    monkeypatch: pytest.MonkeyPatch,
) -> None:
    # 测试逻辑保持不变（但永远不会执行）
    ...
```

评论区精华

PR 无 review 评论，仅有一条来自 [AndreasKaratzas](#) 的批准。提交记录显示作者在独立提交后执行了三次 `main` 分支合并，最终获得批准并入。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅涉及跳过一项已知因编译器行为变更而失效的单元测试，并增强 CI 配置。不会影响生产代码逻辑。注意：若未来编译器行为再次改变（即不再消除冗余算子），需重新启用该测试。
- 影响：影响范围：仅影响 ROCm 环境的 CI 测试流程。影响程度：低。消除了持续失败噪声，确保其他测试得以正常运行。新增的 AMD 镜像配置使测试能在正确硬件上执行。
- 风险标记：测试跳过，编译器行为依赖

关联脉络

- 暂无明显关联 PR