PR #46409 完整报告

vllm-project/vllm

[ROCm][CI]Fix test_concat_and_cache_mla_rope_fused on ROCm

合并时间：2026-06-27 12:38

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46409

执行摘要

- 一句话：修复 ROCm 上 MLA RoPE 融合测试的浮点精度失败
- 推荐动作：值得快速浏览以了解 ROCm 上 fp16 数值不一致的典型解决方案。不涉及系统架构变更，测试改动清晰。

功能与动机

该测试在 ROCm CI 上因 fp16 数值不匹配而失败。PR body 指出根本原因是 torch-native 参考（forward_native）隐式将 fp16 升格为 fp32，而融合 CUDA kernel 在原生 fp16 中运行。PR 引用了 @mawong-amd 的深入分析（PR#32408）。

实现拆解

1. 新增 fixture default_vllm_config：在 tests/kernels/core/test_rotary_embedding_mla_cache_fused.py 中引入一个 pytest fixture，在 ROCm 上设置 VllmConfig 启用 +rotary_embedding custom op，并设置环境变量 VLLM_ROCM_USE_AITER=1 和 VLLM_ROCM_USE_AITER_TRITON_ROPE=1，然后刷新 AITER ops 的环境变量缓存。非 ROCm 路径返回空配置。这样参考实现使用 AITER triton rope，数值精度与融合 kernel 一致。
2. 调整断言容差：在测试的主体部分，针对 ROCm 且 neox_style=True 的情况，对 fp8 路径的 rtol 从 0.1 放宽到 0.15（约一个 e4m3 ULP，~12.5%），对非 fp8 路径的 kv_cache 对比增加宽松的 atol=1e-3、rtol=1e-3（标准 fp16 边界）。其他路径保持原 CUDA 默认容差。

关键文件：

- tests/kernels/core/test_rotary_embedding_mla_cache_fused.py（模块 测试；类别 test；类型 test-coverage；符号 default_vllm_config）：唯一被修改的文件，包含新增 fixture 和容差调整，是整个 PR 的核心。

关键符号：default_vllm_config, test_concat_and_cache_mla_rope_fused

关键源码片段

tests/kernels/core/test_rotary_embedding_mla_cache_fused.py

唯一被修改的文件，包含新增 fixture 和容差调整，是整个 PR 的核心。

```
# tests/kernels/core/test_rotary_embedding_mla_cache_fused.py

import pytest
from vllm.platforms import current_platform


@pytest.fixture
def default_vllm_config(monkeypatch):
    """Enable the AITER triton rope on ROCm for fp16-consistent numerics.

    The fused CUDA kernel runs native fp16 while forward_native upcasts to
    fp32, so on ROCm we route through the AITER triton rope (+rotary_embedding)
    to match. Its env gates are cached at import, hence refresh_env_variables().
    """
    from vllm._aiter_ops import rocm_aiter_ops
    from vllm.config import CompilationConfig, VllmConfig, set_current_vllm_config

    is_rocm = current_platform.is_rocm()
    if is_rocm:
        config = VllmConfig(
            compilation_config=CompilationConfig(custom_ops=["+rotary_embedding"])
        )
    else:
        config = VllmConfig()
    try:
        with monkeypatch.context() as m, set_current_vllm_config(config):
            if is_rocm:
                m.setenv("VLLM_ROCM_USE_AITER", "1")
                m.setenv("VLLM_ROCM_USE_AITER_TRITON_ROPE", "1")
                rocm_aiter_ops.refresh_env_variables()
            yield config
    finally:
        if is_rocm:
            rocm_aiter_ops.refresh_env_variables()


# 在测试函数中，针对 ROCm neox-style 放宽容差（片段）
# ...
rocm_neox = current_platform.is_rocm() and is_neox_style
if kv_cache_dtype == "fp8":
    # ... fp8 转换后
    torch.testing.assert_close(
        result_temp, expected_temp, atol=0.001, rtol=0.15 if rocm_neox else 0.1
    )
elif rocm_neox:
    torch.testing.assert_close(kv_cache, ref_kv_cache, atol=1e-3, rtol=1e-3)
else:
    torch.testing.assert_close(kv_cache, ref_kv_cache)
```

评论区精华

无 review 评论；tjtanaa 直接批准了 PR。讨论主要体现在 PR body 中，引用了 @mawong-amd 在 PR#32408 中的深度分析。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更完全在测试文件内，不涉及生产代码。放宽的容差仅针对 ROCm neox-style 场景，且已分析为 Triton FMA 与融合 kernel 的已知较小偏差（约一个 e4m3 ULP）；非 ROCm 行为不变。若未来 AITER 或融合 kernel 数值行为改变，可能导致测试假通过，但代码中已有明确的 rocm_neox 条件，易于维护。
- 影响：仅影响 ROCm 平台上的相同测试。修复了 CI 失败，使 MLA RoPE 融合 kernel 在 ROCm 上得到有效测试覆盖。非 ROCm 平台无影响。团队收益是更清洁的 CI 状态。
- 风险标记：仅测试变更，放宽容差可能掩盖回归

关联脉络

- PR #32408 in-depth analysis of fp16 numerics in fused MLA kernels: PR body 引用 @mawong-amd 的分析，为本次修改提供了根本原因依据。
- PR #46758 [ROCm][CI TG] refactor and fix deepep_moe test group: 同为 ROCm 测试修复，展示了 AMD CI 持续改善测试稳定性的努力。
- PR #46859 [Hardware][AMD][CI] Fix Kernels Quantization test timeout: 同属 AMD CI 测试修复系列，体现团队对 ROCm 测试稳定性的持续投入。