PR #46260 完整报告

vllm-project/vllm

[ROCm][Test] Fix stale test_gfx950_moe MXFP4 oracle tests

合并时间：2026-06-23 23:07

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46260

执行摘要

- 一句话：修复 gfx950 MXFP4 MoE 测试失效
- 推荐动作：建议审核并通过。该 PR 是纯测试修复，无产品风险，应尽快合并以恢复 gfx950 CI 测试覆盖。值得学习的设计点包括：使用 unittest.mock.patch 桩化全局配置 fixture 的方法，以及测试与产品 API 同步维护的重要性。

功能与动机

PR body 指出 gfx950 MXFP4 MoE oracle 测试因产品 API 演进发生三次独立破坏，测试已无法运行。目的是恢复这些测试，确保它们能够正确验证 MXFP4 后端分发行为。

实现拆解

1. 修复 FusedMoEConfig 参数：在 _make_w4a4_moe_config 中将 intermediate_size_per_partition=256 改为 intermediate_size=256，因为该字段已成为必填项，而 intermediate_size_per_partition 在 __post_init__ 中派生。
2. 添加 vLLM config 桩 fixture：新增 mxfp4_oracle_config fixture，使用 unittest.mock.patch 模拟 get_current_vllm_config 返回 model_config.quantization_config=None，避免因缺少真实配置上下文导致的 AssertionError/AttributeError。
3. 更新回退行为断言：将原 test_w4a4_raises_without_aiter_and_no_moe_backend 重命名为 test_w4a4_falls_back_to_triton_unfused_without_aiter，断言期望从 NotImplementedError 改为 Mxfp4MoeBackend.TRITON_UNFUSED，以匹配 #45896 引入的变化。
4. 为所有需要配置上下文的测试注入 fixture：除回退测试外，test_w4a4_dispatches_to_aiter 和 test_w4a4_dispatches_to_emulation_with_moe_backend 也添加了 mxfp4_oracle_config 参数。此外，测试执行结果确认：AITER 关闭时 2 通过 1 跳过，AITER 开启时 2 通过 1 跳过。

关键文件：

- tests/quantization/test_gfx950_moe.py（模块 测试；类别 test；类型 test-coverage；符号 mxfp4_oracle_config, test_w4a4_dispatches_to_aiter, test_w4a4_raises_without_aiter_and_no_moe_backend, test_w4a4_falls_back_to_triton_unfused_without_aiter）：唯一变更文件，修复了所有因

产品 API 变化导致的测试失效

关键符号：mxfp4_oracle_config, test_w4a4_dispatches_to_aiter, test_w4a4_raises_without_aiter_and_no_moe_backend, test_w4a4_falls_back_to_triton_unfused_without_aiter, test_w4a4_dispatches_to_emulation_with_moe_backend, _make_w4a4_moe_config

关键源码片段

tests/quantization/test_gfx950_moe.py

唯一变更文件，修复了所有因产品 API 变化导致的测试失效

```
from unittest.mock import patch

@pytest.fixture
def mxfp4_oracle_config():
    """Stub the config the oracle reads (``model_config.quantization_config``)
    so backend dispatch resolves without a real model / user override."""
    # 模拟 get_current_vllm_config 返回 quantization_config = None
    # 避免 select_mxfp4_moe_backend 因缺少真实配置而报错
    with patch(
        "vllm.model_executor.layers.fused_moe.oracle.mxfp4.get_current_vllm_config"
    ) as mock_get_config:
        mock_get_config.return_value.model_config.quantization_config = None
        yield


def _make_w4a4_moe_config(moe_backend: str = "auto") -> FusedMoEConfig:
    from vllm.model_executor.layers.fused_moe.activation import MoEActivation
    # 现在使用必填的 intermediate_size 代替已移除的 intermediate_size_per_partition
    return FusedMoEConfig(
        num_experts=8,
        experts_per_token=2,
        hidden_dim=256,
        intermediate_size=256, # 必填字段，intermediate_size_per_partition 由 __post_init__ 派生
        num_local_experts=8,
        num_logical_experts=8,
        moe_parallel_config=FusedMoEParallelConfig.make_no_parallel(),
        activation=MoEActivation.SILU,
        in_dtype=torch.bfloat16,
        device="cuda",
        routing_method=RoutingMethodType.Renormalize,
        moe_backend=moe_backend,
    )


# 测试用例均使用 mxfp4_oracle_config fixture
@pytest.mark.skipif(not ROCM_GFX950, reason="Requires GFX950 (mi355x)")
@pytest.mark.skipif(not ROCM_AITER_AVAILABLE, reason="Requires AITER enabled")
```

```
def test_w4a4_dispatches_to_aiter(mxfp4_oracle_config):
    # 省略实现，仅展示 fixture 注入方式
    pass


@pytest.mark.skipif(not ROCM_GFX950, reason="Requires GFX950 (mi355x)")
@pytest.mark.skipif(ROCM_AITER_AVAILABLE, reason="Test requires AITER disabled")
def test_w4a4_falls_back_to_triton_unfused_without_aiter(mxfp4_oracle_config):
    # 无 AITER 时，预期回退到 TRITON_UNFUSED 而非抛异常
    config = _make_w4a4_moe_config()
    backend, experts_cls = select_mxfp4_moe_backend(
        config, activation_key=kMxfp4Dynamic
    )
    assert backend == Mxfp4MoeBackend.TRITON_UNFUSED
    assert experts_cls is not None
```

评论区精华

仅有审核人 tjtanaa 的批准评论 "LGTM. Let's check AMD CI"，表示认可并计划运行 CI。无其他讨论或未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险：该 PR 仅修改测试文件，不涉及产品代码，回归风险极低。修复后的测试在 MI355 硬件上通过，但可能存在对其他 ROCm 平台（如 gfx942）的兼容性风险：新增的 mxfp4_oracle_config fixture 未作平台跳过，若在其他平台上运行，可能因缺少 on_gfx950 等条件而失败（但测试本身已通过 pytest.mark.skipif 跳过）。此外，patch 桩化了全局配置，如果未来 get_current_vllm_config 行为变化，测试可能再次失效。
- 影响：影响范围仅限于单个测试文件 tests/quantization/test_gfx950_moe.py，属于对 gfx950 硬件的 MXFP4 MoE 测试修复。恢复了三个关键测试的正确性和可执行性，确保它们能随产品 API 变化持续运行。对用户无感知，对团队的意义是保证 CI 中 gfx950 相关测试的可靠性，防止测试无声跳过或失败。
- 风险标记：测试独占变更，无产品风险

关联脉络

- PR #41566 [Quantization] Rework quantization_config to use QuantKey and allow for activation override: 该 PR 引入的 get_current_vllm_config 读取导致测试因缺少配置上下文而失败，本 PR 通过 fixture 桩化修复。
- PR #45896 MiniMax-M3-MXFP4 work (推断): PR body 提到该 PR 改变了无 AITER 时 MXFP4 的回退行为，从抛 NotImplementedError 改为使用 TRITON_UNFUSED。