PR #44368 完整报告

vllm-project/vllm

[ROCm][CI] Fix stale wvSplitK GEMM fallback test for N=5

合并时间：2026-06-03 11:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44368

执行摘要

- 一句话：修复 ROCm wvSplitK GEMM 回退测试的边界值
- 推荐动作：值得合并。虽然变更量小，但确保了测试与代码逻辑的一致性，避免了 CI 的虚假失败。

功能与动机

PR #40687 改变了 wvSplitK 的触发边界，使得原有回退测试失效（断言回退但实际未回退）。此 PR 确保测试与代码逻辑保持一致，正确验证边界条件。

实现拆解

1. 将 test_rocm_unquantized_gemm_gfx1x_n_gt_4_falls_back 重命名为 test_rocm_unquantized_gemm_gfx1x_n_gt_5_falls_back。
2. 将输入张量 x 的第一个维度从 5 改为 6，使得 n=6 > 5，确保回退路径被触发。
3. 添加注释说明 wvSplitK skinny-GEMM 处理 n 在 [1,5] 范围，n>5 必须回退。

关键文件：

- tests/model_executor/layers/test_rocm_unquantized_gemm.py（模块 ROCm 测试；类别 test；类型 test-coverage；符号 test_rocm_unquantized_gemm_gfx1x_n_gt_4_falls_back, test_rocm_unquantized_gemm_gfx1x_n_gt_5_falls_back）：修改了回退测试用例的输入维度和名称，以匹配 wvSplitK 边界条件变更。

关键符号：test_rocm_unquantized_gemm_gfx1x_n_gt_5_falls_back

关键源码片段

tests/model_executor/layers/test_rocm_unquantized_gemm.py

修改了回退测试用例的输入维度和名称，以匹配 wvSplitK 边界条件变更。

```
# 位于 tests/model_executor/layers/test_rocm_unquantized_gemm.py
# 变更前函数名：test_rocm_unquantized_gemm_gfx1x_n_gt_4_falls_back (n=5)
# 变更后函数名：test_rocm_unquantized_gemm_gfx1x_n_gt_5_falls_back (n=6)

def test_rocm_unquantized_gemm_gfx1x_n_gt_5_falls_back(monkeypatch):
    # wvSplitK skinny GEMM handles n in [1, 5] (see PR #40687); n > 5 must
```

```
    # fall back to torch.nn.functional.linear.
    x = torch.randn(6, 64, dtype=torch.float16) # 将 n 从 5 改为 6，确保超过边界
    weight = torch.randn(128, 64, dtype=torch.float16)

    monkeypatch.setattr(utils, "use_aiter_triton_gemm", lambda *args: False)
    monkeypatch.setattr(utils.envs, "VLLM_ROCM_USE_SKINNY_GEMM", True)
    monkeypatch.setattr("vllm.platforms.rocm.on_gfx1x", lambda: True)
    monkeypatch.setattr("vllm.platforms.rocm.on_gfx9", lambda: False)
    monkeypatch.setattr("vllm.platforms.rocm.on_gfx950", lambda: False)
    monkeypatch.setattr(utils, "num_compute_units", lambda: 120)

    wvsplitk_mock = MagicMock(side_effect=lambda w, x_view, _, __: x_view @ w.t())
    monkeypatch.setattr(utils.ops, "wvSplitK", wvsplitk_mock)
    llmm1_mock = MagicMock(side_effect=lambda w, x_view, _: x_view @ w.t())
    monkeypatch.setattr(utils.ops, "LLMM1", llmm1_mock)

    out = utils.rocm_unquantized_gemm_impl(x, weight, None)
    ref = torch.nn.functional.linear(x, weight, None)

    # 断言 wvSplitK 和 LLMM1 均未被调用，确保回退到 torch 的线性操作
    wvsplitk_mock.assert_not_called()
    llmm1_mock.assert_not_called()
    assert torch.allclose(out, ref, atol=1e-3, rtol=1e-3)
```

评论区精华

无。仅 reviewer AndreasKaratzas 批准并评论 'Nice catch'。

- 暂无高价值评论线程

风险与影响

- 风险：极低。仅修改测试用例的输入维度和名称，不涉及任何生产代码。
- 影响：仅影响 ROCm 平台下 wvSplitK GEMM 的测试覆盖。修复后，CI 能正确验证回退边界，防止后续的回归。
- 风险标记：暂无

关联脉络

- PR #40687 [ROCm] wvSplitK skinny-GEMM cutoff change: 此 PR 的变更（提高 cutoff 至 n<=5）导致原有测试失效，当前 PR 是对其的修复。