PR #38137 完整报告

vllm-project/vllm

[ROCm][CI] Fix AITER state leak in shared_fused_moe_routed_transform test

合并时间：2026-03-27 00:26

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38137

执行摘要

该 PR 修复了 ROCm 平台上 AITER 状态在测试间的泄露问题，通过在 cleanup_dist_env_and_memory 函数中集成状态重置逻辑，并调整相关测试文件的环境变量设置，确保跨测试环境独立性和结果一致性，避免 CI 误报失败。

功能与动机

PR body 明确指出，问题源于 use_rocm_aiter monkeypatch 仅在 True 时运行，导致当 False 测试在 True 测试后运行时，继承了陈旧的 AITER 状态（来自 rocm_aiter_ops 类变量），这造成了使用不同内核（AITER vs Triton）的模块产生完全不同的结果（max_diff=0.277）。因此，修复动机是确保测试隔离性，防止环境变量泄露影响 ROCm MoE 测试的准确性。

实现拆解

实现分为三个关键部分：

1. 测试文件修正：在 tests/kernels/moe/test_shared_fused_moe_routed_transform.py 中，简化 test_routed_input_transform_inside_vs_outside 函数的条件，移除 use_rocm_aiter 标志，改为始终在 ROCm 平台设置环境变量 VLLM_ROCM_USE_AITER 和 VLLM_ROCM_USE_AITER_MOE，并导入 rocm_aiter_ops。
2. 核心清理增强：在 vllm/distributed/parallel_state.py 的 cleanup_dist_env_and_memory() 函数中添加代码，当检测到 ROCm 平台时，调用 rocm_aiter_ops.refresh_env_variables() 来重置类变量，确保每次测试后 AITER 状态与当前环境匹配。
3. 环境变量泄露修复：在 tests/kernels/moe/test_routing_simulator.py 中，将直接赋值 envs.environment_variables[env_name] = lambda s=strategy: s 改为使用 monkeypatch.setitem，避免全局字典突变导致的跨测试泄露。

评论区精华

Review 讨论较少，主要由机器人评论和简单批准构成。gemini-code-assist[bot] 评论指出："It simplifies a conditional check by removing the use_rocm_aiter flag, making the monkeypatch.setenv calls and import of rocm_aiter_ops dependent solely on current_platform.is_rocm()." 这表明变更聚焦于代码简化。yewentao256 批准表示："LGTM, thanks for the work!"，无进一步技术辩论。

风险与影响

风险方面：修改 cleanup_dist_env_and_memory() 可能影响所有依赖此函数的测试，但新增代码仅针对 ROCm 平台，且 refresh_env_variables() 调用是幂等的，风险可控。在测试文件中使用 monkeypatch.setitem 修复了全局变量泄露，但需检查其他测试是否类似问题。影响方面：主要提升 CI 稳定性，减少 ROCm 测试的 flakiness，对最终用户透明，团队将受益于更可靠的测试反馈。

关联脉络

从近期历史 PR 分析，本 PR 与 #38161（修复 ROCm GPTQ 测试随机失败）和 #38167（修复 ROCm mock 参数顺序）高度相关，它们共同构成对 ROCm CI 测试套件的持续改进。这些 PR 均使用‘rocm’和‘ci’标签，反映出团队在加强 AMD 硬件支持方面的努力，尤其是针对 MoE 和量化相关测试的健壮性提升。