# PR #40767 完整报告

vllm-project/vllm

[CI][AMD]BugFix] Fix deadlock occuring in test_moe_layer

合并时间：2026-04-25 21:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40767

## 执行摘要

- 一句话：修复 MI300 上 MoE 测试死锁问题
- 推荐动作：此 PR 值得精读，展示了在多 worker 并行测试中处理非确定性缓存回收的一种简洁有效方法：使用分布式 barrier 保持 worker 同步，而非复杂的 GC 操作。

## 功能与动机

test_moe_layer 测试在 MI300 上执行时死锁，原因是 WeakValueDictionary 缓存的 DeepEP buffers 被垃圾回收器非确定性回收，导致部分 worker 的通信变成空操作而其他 worker 无限等待，最终触发 DeepEP error: CPU recv timeout。

## 实现拆解

在 tests/kernels/moe/test_moe_layer.py 中 _parallel_worker 函数的 finally 块末尾添加 torch.distributed.barrier()。

1. 定位问题：死锁源于 base_device_communicator.py 中 WeakValueDictionary 缓存 DeepEP buffers，GC 非确定性回收弱引用导致部分 worker 的 all2all_manager 调用无响应。
2. 插入同步点：在 finally 子句（已处理 DeepEP 管理器清理）之后添加全局 barrier，迫使所有 worker 在进入下一个子测试前完成缓存清理，确保通信状态一致。
3. 仅修改测试文件：不涉及生产代码，改动量小，风险低。

关键文件：

- tests/kernels/moe/test_moe_layer.py（模块 MoE 层；类别 test；类型 test-coverage）：唯一修改的文件，在 finally 块末尾添加 torch.distributed.barrier()，修复死锁。

关键符号：未识别

## 关键源码片段

### tests/kernels/moe/test_moe_layer.py

唯一修改的文件，在 finally 块末尾添加 torch.distributed.barrier()，修复死锁。

```
# tests/kernels/moe/test_moe_layer.py 关键片段
            finally:
                # 只在 DeepEP 后端下清理 all2all_manager
```

```
            if test_config.backend in {
                "deepep_low_latency",
                "deepep_high_throughput",
            }:
                torch.accelerator.synchronize()
                all2all_manager = get_ep_group().device_communicator.all2all_manager
                if all2all_manager is not None:
                    all2all_manager.destroy()
        total = total + 1
        # 所有 worker 同步，确保上一个 subtest 的清理完成
        # 避免因 WeakValueDictionary 缓存被非确定性回收导致的死锁
        torch.distributed.barrier()
```

## 评论区精华

gemini-code-assist[bot] 曾建议通过禁用垃圾回收（gc.disable()）来修复，并警告该做法可能导致 OOM。rasmith 回应：“这仅用于测试。我尝试将 gc 操作移至测试循环内，但不起作用。”随后 rasmith 改用更优的 barrier 方案。最终 yewentao256 批准了该修改。

- 禁用 GC 导致 OOM 风险 (performance): rasmith 尝试后表示不可行，最终改用 barrier 方案，避免了 GC 操作。

## 风险与影响

- 风险：仅修改测试文件，不涉及生产代码，风险较低。barrier 可能引入轻微额外延迟，但测试场景下可忽略。没有 OOM 风险。
- 影响：直接影响：test_moe_layer 测试在 ROCm/MI300 平台上不再死锁，12 passed, 20 skipped, 17 warnings in 349.84s。间接影响：为类似的并行测试中因缓存弱引用回收导致的不一致问题提供了可复现的修复模式。
- 风险标记：测试环境死锁，弱引用非确定性

## 关联脉络

- PR #38503 [ROCm][Engine] Fix GPU memory leaks in engine shutdown and test workaround for async KV prefix cache reset: 同为 ROCm 平台修复，且涉及测试中的资源清理，共享相似上下文。
- PR #40640 [Refactor] Remove unused dead code: 修改了同一仓库下的 MoE 相关测试文件，但关联性较弱。