PR #44046 完整报告

vllm-project/vllm

[ROCm][CI] Stabilize memory-release in the Hybrid model generation tests

合并时间：2026-06-04 22:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44046

执行摘要

- 一句话：用上下文管理器稳定 ROCm Hybrid 模型生成测试
- 推荐动作：该 PR 虽然只改动测试文件，但体现了良好的测试资源管理实践：使用上下文管理器确保资源释放，以及平台特定的等待策略。值得 CI 和测试维护者阅读，同样的模式可推广到其他类似的测试场景。

功能与动机

在 ROCm CI 中，Hybrid 模型生成测试（尤其是 APC 相关测试）经常出现间歇性内存释放失败（见 Buildkite 构建 #68818、#68975）。测试中多个 VllmRunner 实例的生命周期管理不当，导致 GPU 内存未及时释放，进而影响后续测试。PR 通过强制上下文管理和内存稳定等待来消除这类竞态条件。

实现拆解

1. 导入新依赖：在文件头部添加 from contextlib import contextmanager, nullcontext 和 from tests.utils import wait_for_gpu_memory_to_clear，为上下文管理和内存等待提供基础。
2. 新增 ROCm 内存稳定函数 _wait_for_rocm_memory_to_settle：该函数检查当前平台是否为 ROCm，若是则调用 wait_for_gpu_memory_to_clear 以 0.01 的阈值等待最多 120 秒，确保所有 GPU 内存释放。
3. 创建上下文管理器 _owned_vLLM_runner：使用 @contextmanager 装饰器，在 with 块中创建 VllmRunner 实例，并在 finally 块中调用内存稳定函数，保证退出时释放资源。
4. 重构 _get_vLLM_output 函数：将原有的直接 vllm_runner(**kwargs) 创建逻辑改为使用上下文管理器：如果 vllm_model 为 None，则通过 _owned_vLLM_runner 创建新 runner；否则使用 nullcontext 保持现有模型。确保所有 runner 实例都被正确管理。
5. 更新 APC 测试用例：将 test_apc_multiple_prompts_all_cached_outputs、test_apc_multiple_prompts_block_align_alignment、test_apc_multiple_prompts_partial_cached_outputs 等测试中显式使用 _owned_vLLM_runner 上下文管理器替换直接 _get_vLLM_output 调用，以确保在多次测试间强制内存回收。

关键文件：

- tests/models/language/generation/test_hybrid.py（模块 Hybrid 测试；类别 test；类型 test-coverage；符号 _wait_for_rocm_memory_to_settle, _owned_vLLM_runner, _get_vLLM_output）：唯一变更文件，新增内存管理和上下文管理辅助函数，重构测试逻辑

关键符号：_wait_for_rocm_memory_to_settle, _owned_vLLM_runner, _get_vLLM_output

关键源码片段

tests/models/language/generation/test_hybrid.py

唯一变更文件，新增内存管理和上下文管理辅助函数，重构测试逻辑

```
def _wait_for_rocm_memory_to_settle() -> None:
    # 仅在 ROCm 平台上执行
    if not current_platform.is_rocm():
        return

    num_gpus = current_platform.device_count()
    if num_gpus == 0:
        return

    # 等待所有 GPU 内存释放，阈值 0.01，超时 120 秒
    wait_for_gpu_memory_to_clear(
        devices=list(range(num_gpus)),
        threshold_ratio=0.01,
        timeout_s=120,
    )


@contextmanager
def _owned_vLLM_runner(vllm_runner, kwargs):
    # 使用上下文管理器创建并管理 VllmRunner 实例
    try:
        with vllm_runner(**kwargs) as runner:
            yield runner
    finally:
        # 退出时强制等待 ROCm 内存稳定
        _wait_for_rocm_memory_to_settle()
```

评论区精华

- JartX 在 Issue 评论中确认效果：在不同 ROCm GPU 上运行通过所有五个 APC 测试，并观察到日志输出 "Done waiting for free GPU memory ..." 在引擎切换间正常工作，验证了内存等待机制的有效性。
- 确认内存等待机制有效 (testing): 批准变更，机制有效。

风险与影响

- 风险：

    - 平台兼容性：新增的 _wait_for_rocm_memory_to_settle 仅在 ROCm 平台生效，不影响 NVIDIA 或 CPU 等其他平台，风险较低。
    - 超时导致测试失败：内存等待设置了 120 秒超时，若 ROCm 环境内存释放异常缓慢，可能导致测试超时失败，但属于正常防御行为。
    - 依赖外部函数：依赖 tests.utils.wait_for_gpu_memory_to_clear，确保该函数在 utils 中已存在且稳定。
- 影响：
    - 用户：无直接影响，仅 CI 测试稳定性提升。
    - 系统：减少 ROCm CI 中 Hybrid 测试的间歇性失败，提高流水线可靠性。
    - 团队：降低了 ROCm 维护者排查内存相关问题的负担。
    - 风险标记：仅影响 ROCm 测试，新增上下文管理器，潜在超时风险

关联脉络

- PR #44255 [ROCm][CI] Specifying time outs for the lm eval models: 同样关注 ROCm CI 测试稳定性，通过设置超时避免挂起，本 PR 通过内存等待避免失败，都是提高 ROCm 测试可靠性。