PR #50304 完整报告

vllm-project/vllm

[CI][ROCm] Fix AMD nightly distributed regressions

合并时间：2026-07-30 06:02

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50304

执行摘要

- 一句话：修复 AMD 夜间分布式测试回归：回滚过早的 AG/RS 测试并稳定 shutdown 内存检查
- 推荐动作：建议立即合并。该 PR 无争议、风险低，且为保持 CI 稳定性所必需。值得关注的是，进程内 shutdown 测试的平台差异化处理可作为未来类似平台适配的参考模式。

功能与动机

PR body 指出两个 CI 构建失败：自定义 all-reduce 测试因缺少 custom_all_gather 接口全部失败（见 Buildkite 链接），以及进程内 shutdown 测试在 MI300 上达到 2.01 GiB 但超出 2 GiB 阈值导致超时。这些失败阻碍了 AMD nightly CI 的通过。

实现拆解

1. 回滚过早的自定义 AG/RS 测试：在 tests/distributed/test_custom_all_reduce.py 中删除两个测试函数 test_sp16_dispatches_only_to_mnnvl_lamport 和 test_custom_collectives_world_size_four，它们由 PR#50089 引入但依赖尚未合入的 Python 自定义 all-gather/reduce-scatter API。同时移除相关导入（SimpleNamespace、Mock、custom_all_reduce 等）。保留现有的 all-reduce 测试（graph_allreduce、eager_allreduce 和参数化 test_custom_allreduce），这些已在之前的 nightly 中通过。
2. 稳定进程内 shutdown 测试：在 tests/v1/shutdown/test_delete.py 的 test_llm_delete_inprocess 中，为 ROCm 平台添加 wait_for_gpu_memory_to_clear 的 threshold_ratio=0.01（利用已有的 ROCm 4 GiB 空闲内存下限）和 timeout_s=SHUTDOWN_TEST_TIMEOUT_SEC // 2（内部超时，确保失败在子进程超时前报告）。CUDA 平台保持不变。

关键文件：

- tests/distributed/test_custom_all_reduce.py（模块 自定义规约；类别 test；类型 test-coverage；符号 test_sp16_dispatches_only_to_mnnvl_lamport, test_custom_collectives_world_size_four）：删除 117 行过早引入的自定义 all-gather/reduce-scatter 测试，恢复现有的 all-reduce 测试框架，修复 7/7 回归项。
- tests/v1/shutdown/test_delete.py（模块 关闭流程；类别 test；类型 test-coverage；符号 test_llm_delete_inprocess）：为 ROCm 平台增加内存检查阈值和内部超时，修复 MI300 上进程内 shutdown 测试的超时失败。

关键符号：test_llm_delete_inprocess, test_sp16_dispatches_only_to_mnnvl_lamport (deleted), test_custom_collectives_world_size_four (deleted)

关键源码片段

tests/v1/shutdown/test_delete.py

为 ROCm 平台增加内存检查阈值和内部超时，修复 MI300 上进程内 shutdown 测试的超时失败。

```
# tests/v1/shutdown/test_delete.py ( 部分 )

@create_new_process_for_each_test("fork" if current_platform.is_cuda() else "spawn")
@pytest.mark.timeout(SHUTDOWN_TEST_TIMEOUT_SEC)
@pytest.mark.parametrize("model", MODELS)
@pytest.mark.parametrize("send_one_request", [False, True])
def test_llm_delete_inprocess(
    monkeypatch,
    model: str,
    send_one_request: bool,
) -> None:
    """Test that VllmRunner frees GPU memory in in-process (no MP) mode."""
    with monkeypatch.context() as m:
        m.setenv("VLLM_ENABLE_V1_MULTIPROCESSING", "0")

        with VllmRunner(model) as vllm_model:
            if send_one_request:
                vllm_model.generate(
                    ["Hello my name is"],
                    SamplingParams(max_tokens=1),
                )

        wait_for_gpu_memory_to_clear(
            devices=[0],
            threshold_bytes=SHUTDOWN_TEST_THRESHOLD_BYTES,
            # 激活 helper 的 ROCm 空闲运行时下限。VllmRunner 在退出上下文时
            # 已执行稳定内存等待。
            threshold_ratio=0.01 if current_platform.is_rocm() else None,
            # 在外部 pytest 超时之前让子进程内失败。
            timeout_s=(
                SHUTDOWN_TEST_TIMEOUT_SEC // 2
                if current_platform.is_rocm()
                else SHUTDOWN_TEST_TIMEOUT_SEC
            ),
        )
```

评论区精华

无 review 讨论。PR 由 mgoin 直接批准，claude[bot] 自动评论因 fork 而跳过审查。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试文件：删除未就绪的测试用例和调整平台特定的内存检查参数。不触及生产代码、内核或配置。唯一潜在风险是若未来自定义 AG/RS API 合入后，删除的测试无法自动回归，但开发人员需在后续 PR 中重新添加。
- 影响：影响范围：仅影响 ROCm CI 的 distirbuted 测试套件。CI 从 7/7 失败变为预期通过（对于该部分测试）。CUDA 测试不受影响。团队可立即恢复 nightly 构建的绿色状态。
- 风险标记：仅测试变更，平台特定路径

关联脉络

- PR #50089 [Feature] Add custom all-gather/reduce-scatter tests: 引入被回滚的测试，导致 ROCm 回归。
- PR #45195 [Feature] In-process shutdown test: 引入被调整的进程内 shutdown 测试。
- PR #49270 [BugFix] Expose strict 2 GiB threshold on MI300: 暴露 MI300 上阈值严格性问题，导致本次调整。
- PR #50000 [Kimi K3] Umbrella PR: 包含自定义 all-gather/reduce-scatter 集成（CUDA-only），PR body 提及作为回滚理由之一。