PR #40597 完整报告

vllm-project/vllm

[Bugfix][CI] Fix `v1/kv_connector/unit/test_nixl_connector_hma.py::test_fewer_blocks_with_hma`

合并时间：2026-04-22 21:21

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40597

PR 分析报告：修复 NixlConnector HMA 测试的 flaky 失败

执行摘要

本 PR 修复了 test_fewer_blocks_with_hma 测试在 CI 中因 GPU 内存残留导致的间歇性失败问题。通过降低内存利用率、添加显式清理和调整配置参数，确保测试在单进程环境下稳定运行，属于典型的测试稳定性修复，不影响生产代码。

功能与动机

该测试在 CI 中持续失败，错误信息显示 GPU 空闲内存不足（4.51/22.05 GiB）低于期望的 GPU 内存利用率（0.47，10.36 GiB）。作者在 PR body 中指出，之前的修复（PR #39938）未解决核心问题，推测是由于 VLLM_ENABLE_V1_MULTIPROCESSING=0 环境变量导致内存清理不彻底，残留内存被其他测试占用。因此，本 PR 尝试手动刷新内存，并调整配置以避免内存不足。

实现拆解

变更仅涉及一个测试文件，具体拆解如下：

1. 导入依赖调整：在文件头部新增 import gc 和 import torch，为后续内存清理提供工具支持。
2. 配置参数优化：在 test_fewer_blocks_with_hma 函数中，调整了 LLM 初始化参数：
   - 将 gpu_memory_utilization 从 0.47 降低至 0.3，减少内存需求。
   - 新增 max_num_seqs: 1，限制并发序列数。
   - 将 max_num_batched_tokens 从 1024 提升至 2048，避免因 token 数限制影响测试逻辑。
3. 显式内存清理：在 run_test_and_cleanup 函数中，于 LLM 初始化前添加以下代码，强制清理残留内存：

tests/v1/kv_connector/unit/test_nixl_connector_hma.py

唯一变更文件，包含测试逻辑和配置调整，直接解决 flaky 失败问题。

关键源码片段

tests/v1/kv_connector/unit/test_nixl_connector_hma.py

唯一变更文件，包含测试逻辑和配置调整，直接解决 flaky 失败问题。

```
def test_fewer_blocks_with_hma(monkeypatch, model_name, sw_size):
    """Test that a prefill instance returns fewer "remote blocks" for the SWA groups
    when sequence exceeds the sliding window.
    """
    kv_transfer_config = KVTransferConfig(
        kv_connector="NixlConnector",
        kv_role="kv_both",
    )
    block_size = 16
    llm_kwargs = {
        "model": model_name,
        "enforce_eager": True,
        "gpu_memory_utilization": 0.3, # 从 0.47 降低至 0.3，减少内存需求
        "kv_transfer_config": kv_transfer_config,
        "max_model_len": 2048,
        "max_num_seqs": 1, # 新增参数，限制并发序列数
        "disable_hybrid_kv_cache_manager": False,
        "max_num_batched_tokens": 2048, # 从 1024 提升至 2048，避免 token 数限制
        "enable_prefix_caching": False,
        "block_size": block_size,
    }

    monkeypatch.setenv("VLLM_ENABLE_V1_MULTIPROCESSING", "0")

    def run_test_and_cleanup():
        gc.collect() # 强制 Python 垃圾回收，清理残留对象
        torch.accelerator.empty_cache() # 清空 GPU 缓存，释放显存
        llm = LLM(**llm_kwargs)
        try:
            run_hma_test(llm) # 执行实际测试逻辑
        finally:
            llm.llm_engine.engine_core.shutdown()

    run_test_and_cleanup()
```

评论区精华

Review 讨论较少，主要总结如下：

- gemini-code-assist[bot]：简要概括了变更内容，指出调整了配置参数并添加了内存清理，无进一步反馈。
- markmc：直接批准了 PR，未提出异议或深入讨论。

无争议点或未解决疑虑，变更被视为直接修复。

风险与影响

风险分析：

- 低风险，变更仅影响测试文件，不涉及生产代码逻辑。

- 内存清理操作是标准做法，但若清理不足，测试可能仍会间歇性失败（作者已预留使用 clean_gpu_memory_between_tests 的备选方案）。
- 配置参数调整可能使测试覆盖场景略有变化，但仍在合理范围内。

影响分析：

- 对用户无直接影响，纯测试修复。
- 提升 CI 稳定性，减少 flaky 测试导致的构建失败。
- 对团队减少维护负担，确保 kv-connector 模块测试可靠性。

关联脉络

- PR #39938：根据 PR body 提及，之前的修复未彻底解决问题，本 PR 是对其的补充或替代。
- PR #38453：同属 kv-connector 模块，涉及 HMA 相关功能，可能共享测试环境或逻辑，体现了该模块在持续演进中。
  - 从近期历史 PR 看，kv-connector 和 v1 标签频繁出现，表明这是 vLLM v1 版本中活跃开发的子系统，测试稳定性修复是维护常态。