PR #39938 完整报告

vllm-project/vllm

[CI Bug] fix flaky test test_fewer_blocks_with_hma[google/gemma-3-1b-it-512]

合并时间：2026-04-16 05:20

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39938

执行摘要

- 一句话：修复 NIXL 连接器 HMA 测试中因 GPU 内存利用率设置过高导致的偶发性失败。
- 推荐动作：该 PR 变更简单直接，无需精读。对于工程师，可关注其作为解决 CI 不稳定性问题的范例：通过微调资源相关参数（如内存利用率）来适应测试环境波动，而非修改核心逻辑。

功能与动机

根据 PR body 中引用的 Buildkite CI 失败日志，测试 test_fewer_blocks_with_hma[google/gemma-3-1b-it-512] 因内存不足而失败。具体错误为：Free memory on device cuda:0 (10.93/22.05 GiB) on startup is less than desired GPU memory utilization (0.5, 11.02 GiB)。这表明测试设置的 GPU 内存利用率（0.5）计算出的请求内存（11.02 GiB）略高于 CI 环境中的实际空闲内存（10.93 GiB），导致 ValueError。修复目的是降低此利用率阈值，使测试在内存波动下更稳定。

实现拆解

1. 定位并调整测试配置：修改文件 tests/v1/kv_connector/unit/test_nixl_connector_hma.py 中的 test_fewer_blocks_with_hma 函数。将 llm_kwargs 字典中的 gpu_memory_utilization 键值从 0.5 改为 0.47。
2. 影响分析：此调整仅影响测试执行时的内存分配计算，不改变生产代码逻辑。降低利用率意味着测试初始化时请求的内存减少，从而避免因 CI 环境内存波动导致的偶发性失败。
3. 测试配套：本次变更仅涉及测试文件，无其他源码、配置或部署改动。

关键文件：

- tests/v1/kv_connector/unit/test_nixl_connector_hma.py（模块 连接器测试；类别 test；类型 test-coverage；符号 test_fewer_blocks_with_hma）：这是唯一被修改的文件，包含不稳定的测试函数，直接修复了 CI 失败问题。

关键符号：test_fewer_blocks_with_hma

关键源码片段

tests/v1/kv_connector/unit/test_nixl_connector_hma.py

这是唯一被修改的文件，包含不稳定的测试函数，直接修复了 CI 失败问题。

```
@pytest.mark.parametrize("model_name, sw_size", [("google/gemma-3-1b-it", 512)])
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": 0.47, # 从0.5调整为0.
        47，降低内存利用率阈值，避免CI环境内存波动导致测试失败
        "kv_transfer_config": kv_transfer_config,
        "max_model_len": 2048,
        # NOTE: Make sure HMA is enabled
        "disable_hybrid_kv_cache_manager": False,
        "max_num_batched_tokens": 1024,
        "enable_prefix_caching": False,
        "block_size": block_size,
    }
    # ... 后续测试逻辑保持不变
```

评论区精华

review 评论较少，主要结论是认可变更。gemini-code-assist[bot] 指出：“This adjustment likely fine-tunes memory allocation for the test environment.” tlrmchlsmth 直接批准。无争议点或未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅影响单个测试函数的配置参数，不涉及核心业务逻辑、性能或安全。潜在风险是如果阈值降得过低（如远低于 0.47），可能掩盖真实的内存不足问题，但当前从 0.5 微调至 0.47 幅度很小，旨在解决环境波动，不影响测试意图。
- 影响：影响范围：仅限于 CI 测试稳定性。影响程度：低。修复后，该特定测试在内存受限的 CI 环境中应不再偶发失败，提高测试套件的可靠性。对用户、系统或团队无其他影响。
- 风险标记：测试环境依赖

关联脉络

- PR #39724 [Bugfix][NIXL] Fix _logical_to_kernel_block_ids conversion for non-mamba models: 两者都涉及 NIXL 连接器的测试修复，且修改了同一测试文件（test_nixl_connector_hma.py），属于同一功能模块的持续维护。

- PR #39596 [Mooncake] Fix mixed MLA+Eagle block-size validation: 同属 kv-connector 模块的 bugfix，反映了该模块测试稳定性的持续优化。