

PR #43236 完整报告

vllm-project/vllm

[ROCm][CI] add warmup to mem_util test before measurement

合并时间: 2026-05-22 11:17

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43236>

执行摘要

- 一句话: 为 CI 内存测试添加 warmup, 修复 ROCm 基线波动
- 推荐动作: 该 PR 价值专一且明确, 值得快速合并。虽然实现简单, 但体现了对平台差异的细致处理, 测试稳定性维护的好范例。

功能与动机

ROCm 上 PyTorch 上下文初始化开销较大, 导致 `test_memory_profiling` 测试中的非 torch 内存增量超过 5% 容差限, CI 测试失败。本 PR 通过预热步骤将初始化开销纳入基线, 从而获得更准确的测量结果。

实现拆解

1. 在 `tests/utils/_test_mem_utils.py` 的 `test_memory_profiling` 函数中, 于 `baseline_snapshot` 创建之前, 新增 3 行代码: 创建一个小 `tensor _warmup = torch.zeros(1, device="cuda")`, 然后 `del _warmup`, 最后调用 `torch.accelerator.empty_cache()` 释放缓存。
2. 目的是触发 PyTorch 内部对 CUDA/ROCm 上下文的初始化 (如 `stream`、`cuBLAS handle` 等), 使其内存开销包含在基线快照中, 避免后续测量非 torch 内存增量时被计入。
3. 该变更仅影响单测文件, 无源码、配置或部署配套改动。

关键文件:

- `tests/utils/_test_mem_utils.py` (模块 内存测试; 类别 `test`; 类型 `test-coverage`): 唯一被修改的文件, 在 `test_memory_profiling` 函数中增加了 CUDA/ROCm 上下文的预热步骤, 解决 ROCm CI 测试稳定性问题。

关键符号: `test_memory_profiling`

关键源码片段

`tests/utils/_test_mem_utils.py`

唯一被修改的文件, 在 `test_memory_profiling` 函数中增加了 CUDA/ROCm 上下文的预热步骤, 解决 ROCm CI 测试稳定性问题。

```
# 位于函数 test_memory_profiling() 内, 在 handle1 分配之后、baseline_snapshot 之前
# 预热 PyTorch 的 CUDA/ROCm 上下文, 使其内部初始化开销
```

```
# (streams, cuBLAS handles 等) 被包含在基线中,  
# 从而不会夸大非 torch 增量, 该增量在 ROCm 上比 CUDA 更大  
_warmup = torch.zeros(1, device="cuda")  
del _warmup  
torch.accelerator.empty_cache() # 清空缓存, 释放预留内存  
  
baseline_snapshot = MemorySnapshot() # 后续基线快照将包含已稳定的上下文
```

评论区精华

Reviewer tjtanana 质问是否 CUDA 也需要预热 ("so cuda also needs this?")。提交者 divakar-amd 回应: CUDA 同样存在相同开销, 但幅度较小, 现有 5% 容差限足以通过, 而 ROCm 上开销较大导致失败。结论是 CUDA 侧无实际影响, 但预热步骤无害。

- CUDA 是否也需要预热 (question): divakar-amd 解释 CUDA 上也有相同开销, 但幅度小, 现有 5% 容差足以通过; 预热步骤对 CUDA 无害。

风险与影响

- 风险: 风险极低。变更仅涉及单测文件, 新增 3 行代码, 逻辑简单明确。唯一潜在风险是预热步骤可能小幅增加测试执行时间 (约几毫秒), 可忽略不计。
- 影响: 直接影响仅限 CI 中的 test_memory_profiling 测试, 使其在 ROCm 上稳定通过。不影响任何线上逻辑、API 或系统行为。团队收益是减少 CI 中的假阳性失败, 提升开发效率。
- 风险标记: 低风险

关联脉络

- PR #42988 [Perf] zeros -> empty to remove additional fill: 同样涉及 torch.zeros 调整, 但上下文完全不同。