

PR #47482 完整报告

vllm-project/vllm

[ROCm][CI] Adding extract hs 2gpu

合并时间: 2026-07-03 08:59

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47482>

执行摘要

- 一句话: 调整 ROCm CI 中 Extract Hidden States 测试的 GPU 分配和超时
- 推荐动作: 本 PR 为一次小的 CI 配置修正, 无源码逻辑变更, 不值得精读。但可作为 CI 配置与测试需求对齐的典型案例参考。

功能与动机

PR body 明确指出: 底层 TP=2 的提取测试已经需要两个 GPU, 但 CI 配置中的 agent_pool 仍为 `mi300_1` (单卡池), 导致资源不匹配; 同时 ROCm 平台上测试执行较慢, 需要更长的超时时间以降低误报率。

实现拆解

1. 更新 Buildkite 配置 (`.buildkite/test-amd.yaml`): 将 Extract Hidden States Integration 任务的 agent_pool 从 `mi300_1` 改为 `mi300_2`, 并新增 `num_gpus: 2` 字段, 显式声明需要 2 个 GPU。
2. 调整测试超时 (`tests/v1/kv_connector/extract_hidden_states_integration/test_extraction.py`): 在 `test_extract_hidden_states_tp2` 函数的 `@pytest.mark.timeout` 装饰器中, 使用 `current_platform.is_rocm()` 条件表达式, ROCm 平台 timeout 为 240 秒, 其他平台保持 60 秒。同时新增 `from vllm.platforms import current_platform` 导入。

关键文件:

- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 修改了 ROCm CI 中 Extract Hidden States Integration 任务的 agent_pool 和 num_gpus 配置, 是整个 PR 的核心配置变更。
- `tests/v1/kv_connector/extract_hidden_states_integration/test_extraction.py` (模块 提取测试; 类别 test; 类型 test-coverage): 调整了 `test_extract_hidden_states_tp2` 测试的超时时间, 使其在 ROCm 平台上有更宽松的超时限制。

关键符号: 未识别

关键源码片段

`tests/v1/kv_connector/extract_hidden_states_integration/test_extraction.py`

调整了 test_extract_hidden_states_tp2 测试的超时时间，使其在 ROCm 平台上有更宽松的超时限制。

```
# tests/v1/kv_connector/extract_hidden_states_integration/test_extraction.py
# 新增导入
from vllm.platforms import current_platform

# 测试超时装饰器变更：ROCM 平台允许 240 秒，其他平台保持 60 秒
@pytest.mark.timeout(240 if current_platform.is_rocm() else 60)
@multi_gpu_test(num_gpus=2)
@create_new_process_for_each_test()
def test_extract_hidden_states_tp2():
    """Test that hidden states extraction works with tensor_parallel_size=2."""
    # 测试体无变更
```

评论区精华

无实质性 review 讨论。PR 获得 khluu 批准，claude[bot] 自动注释指出这是一个 fork 的 PR。

- 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 CI 配置和测试超时参数，不修改任何核心逻辑。超时调整仅在 ROCm 平台生效，不会影响其他平台的测试行为。
- 影响：影响范围：仅限于 ROCm CI 中的 Extract Hidden States Integration 测试任务。影响程度：低。修正了资源分配错误，理论上可减少该测试在 ROCm 上的超时失败。其他平台无影响。
- 风险标记：暂无

关联脉络

- PR #43232 Xqa decode kernels: 同为 ROCm 相关 kv-connector 方向的 CI 改进