PR #46650 完整报告

vllm-project/vllm

[AMD][CI] Fix Pipeline + Context Parallelism test group

合并时间：2026-06-25 11:23

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46650

执行摘要

- 一句话：修复 ROCm 上 Pipeline+Context Parallelism 测试失败
- 推荐动作：值得合并，修复了 CI 阻塞问题，设计简洁，平台兼容性好。

功能与动机

修复 ROCm CI 中 test_pp_cudagraph[FLASH_ATTN-2-JackFram/llama-160m] 测试失败。FLASH_ATTN 后端在 ROCm 上不可用，FlashAttentionImpl.forward() 会因 get_flash_attn_version() 返回 None 触发 AssertionError。

实现拆解

1. 修改参数化列表：在 tests/distributed/test_pp_cudagraph.py 中，将 ATTN_BACKEND 参数化从固定 ["FLASH_ATTN"] 改为条件表达式：ROCm 上为 [None]，否则为 ["FLASH_ATTN"]。
2. 调整函数签名：ATTN_BACKEND 类型从 LiteralString 改为 str | None，以支持 None 值。
3. 条件添加参数：仅当 ATTN_BACKEND 不为 None 时才添加 --attention-backend 参数。在 ROCm 上不传递后端参数，由平台自动选择 TRITON_ATTN 后端。
4. 移除未使用导入：删除了不再需要的 LiteralString 导入，新增 current_platform 导入用于平台检测。

关键文件：

- tests/distributed/test_pp_cudagraph.py（模块 测试脚本；类别 test；类型 test-coverage；符号 test_pp_cudagraph）：唯一的变更文件，修改测试参数化逻辑，根据平台选择注意力后端，修复 ROCm 上的测试失败。

关键符号：test_pp_cudagraph

关键源码片段

tests/distributed/test_pp_cudagraph.py

唯一的变更文件，修改测试参数化逻辑，根据平台选择注意力后端，修复ROCm上的测试失败。

```
# tests/distributed/test_pp_cudagraph.py
import pytest
from vllm.platforms import current_platform
```

```
from ..utils import compare_two_settings, create_new_process_for_each_test


@pytest.mark.parametrize(
    "PP_SIZE, MODEL_NAME",
    [
        (2, "JackFram/llama-160m"),
    ],
)
@pytest.mark.parametrize(
    "ATTN_BACKEND",
    # ROCm 上不强制指定后端，由平台自动选择 TRITON_ATTN
    [None] if current_platform.is_rocm() else ["FLASH_ATTN"],
)
@create_new_process_for_each_test()
def test_pp_cudagraph(
    PP_SIZE: int,
    MODEL_NAME: str,
    ATTN_BACKEND: str | None,
):
    cudagraph_args = [
        "--dtype",
        "float16",
        "--pipeline-parallel-size",
        str(PP_SIZE),
        "--distributed-executor-backend",
        "mp",
    ]
    # On ROCm, defer to the platform attention selector instead of forcing a backend.
    if ATTN_BACKEND is not None:
        cudagraph_args.append(f"--attention-backend={ATTN_BACKEND}")

    eager_args = cudagraph_args + ["--enforce-eager"]

    compare_two_settings(MODEL_NAME, eager_args, cudagraph_args)
```

评论区精华

PR 无 review 讨论，仅有 claude bot 自动评论和 AndreasKaratzas 的批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅限于测试文件，且通过平台检测保护，不影响 CUDA 行为。ROCm 上未指定后端时，平台会默认选择 TRITON_ATTN，该后端已被验证支持 CUDA 图捕获。
- 影响：仅影响 ROCm CI 中的 test_pp_cudagraph 测试，修复了测试失败，使其在 ROCm 上能正常通过。CUDA 用户无影响。

- 风险标记：测试专用变更，平台条件分支

关联脉络

- PR #46548 [ROCm] Fix OOB During Model Warmup With ROCM_ATTN and MRV2: 同为 ROCm 相关注意力后端修复，体现对 ROCm 平台的持续 CI 稳定性改进。
- PR #46636 [ROCm] Begin Deprecation Window for CUDA_VISIBLE_DEVICES on ROCm: 同为 ROCm 平台适配 PR，显示该项目对 ROCm 支持的持续投入。