PR #46161 完整报告

vllm-project/vllm

[CI] Add TP=4 requirement to `test_mixed_precision_model_accuracies`

合并时间：2026-06-23 07:19

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46161

执行摘要

- 一句话：为混合精度测试加上 TP=4 装饰器
- 推荐动作：PR 较为琐碎，仅作测试稳健性提升，不值得精读。值得注意的点是团队在 review 中推荐复用已有装饰器而非重复造轮子，体现了良好的代码复用习惯。

功能与动机

该测试在 CI runner 上因设备不足而失败，PR body 引用 CI 构建链接说明了问题。测试需要 tensor_parallel_size=4，但 runner 未提供足够的 GPU。

实现拆解

1. 导入装饰器：在 tests/quantization/test_mixed_precision.py 头部新增 from tests.utils import multi_gpu_only。
2. 添加装饰器：在 test_mixed_precision_model_accuracies 函数上增加 @multi_gpu_only(num_gpus=4)，使其在 GPU 数量少于 4 时自动跳过。
3. 移除手动检查：根据 review 评论，将最初的手动 device_count 检查替换为现有的装饰器，保持代码一致。

关键文件：

- tests/quantization/test_mixed_precision.py（模块 量化测试；类别 test；类型 test-coverage）：唯一变更文件，添加了导入 multi_gpu_only 和装饰器，确保测试在设备不足时自动跳过。

关键符号：test_mixed_precision_model_accuracies

关键源码片段

tests/quantization/test_mixed_precision.py

唯一变更文件，添加了导入 multi_gpu_only 和装饰器，确保测试在设备不足时自动跳过。

```
# 文件：tests/quantization/test_mixed_precision.py
# 新增导入：复用已有的 multi_gpu_only 装饰器，统一跳过策略
from tests.utils import (
    multi_gpu_only,
)
```

```
# 在函数定义前添加装饰器，要求至少 4 个 GPU，否则 pytest 跳过
@multi_gpu_only(num_gpus=4)
def test_mixed_precision_model_accuracies(model_name: str, accuracy_numbers: dict):
    results = lm_eval.simple_evaluate(
        model="vllm",
        model_args=EvaluationConfig(model_name).get_model_args(),
        tasks=list(accuracy_numbers.keys()),
        batch_size=8,
    )
    # ... 验证逻辑不变
```

评论区精华

Reviewer AndreasKaratzas建议使用 tests/utils.py 中已有的 multi_gpu_only 装饰器，而不是手动编写设备计数检查。作者 fxmarty-amd采纳建议并在第二次提交中修复。最终两位 reviewer 均批准。

- 使用现有装饰器替代手动设备计数检查 (style): 作者采纳并更新代码，使用 @multi_gpu_only(num_gpus=4)。

风险与影响

- 风险：无技术风险。变更仅为测试跳过条件的改进，不涉及任何生产代码逻辑。
- 影响：对用户无影响。对 CI 系统：确保在 AMD GPU 数量不足 4 的 runner 上该测试被正确跳过，避免误报失败。影响范围限定于 ROCm CI。
- 风险标记：暂无

关联脉络

- PR #46141 [ROCm][CI] Query total device memory via amdsmi to avoid HIP init: 同为 AMD ROCm CI 改进，涉及测试环境适配。