PR #44289 完整报告

vllm-project/vllm

[XPU] skip unapplied UT in test_gpu_model_runner.py

合并时间：2026-06-04 08:48

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44289

执行摘要

- 一句话：修复 XPU 下 GPU 模型 runner 测试跳过条件
- 推荐动作：此 PR 是典型的平台兼容性修复，技术价值不高，但保证了 CI 流水线在 XPU 上的稳定性。建议快速合并，无需深入审查。感兴趣的读者可留意测试跳过条件的未来调整。

功能与动机

XPU 平台（如 Intel GPU）不支持 FlashInfer attention 后端，导致原标记为仅在 ROCm 跳过的测试在 XPU 上运行失败。此变更将跳过条件从 ' 如果为 ROCm 则跳过 ' 改为 ' 如果不是 CUDA 则跳过 '，以覆盖所有非 NVIDIA GPU 平台，保持 CI 的稳定性。

实现拆解

该 PR 仅修改一个测试文件，分两处调整 @pytest.mark.skipif 装饰器的条件。

1. 修改 test_hybrid_attention_mamba_tensor_shapes 的跳过条件：将 current_platform.is_rocm() 替换为 not current_platform.is_cuda()。
2. 修改 test_mamba_cache_raises_when_max_num_seqs_exceeds_blocks 的跳过条件：同样将 current_platform.is_rocm() 替换为 not current_platform.is_cuda()。

这两项调整均位于 tests/v1/worker/test_gpu_model_runner.py，不涉及其他源码、配置或基础设施变更。由于 FlashInfer attention 后端仅在 NVIDIA CUDA 上受支持，此变更使得在 XPU、Intel GPU 等非 CUDA 平台上自动跳过这些测试，从而避免不兼容导致的测试失败。

关键文件：

- tests/v1/worker/test_gpu_model_runner.py（模块 测试；类别 test；类型 test-coverage）：唯一修改的文件，调整了两个测试函数的跳过条件以适应 XPU 等非 CUDA 平台。

关键符号：test_hybrid_attention_mamba_tensor_shapes, test_mamba_cache_raises_when_max_num_seqs_exceeds_blocks

关键源码片段

tests/v1/worker/test_gpu_model_runner.py

唯一修改的文件，调整了两个测试函数的跳过条件以适应 XPU 等非 CUDA 平台。

```
# tests/v1/worker/test_gpu_model_runner.py
```

```
# 将原来的 is_rocm() 改为 not is_cuda()，使得在 CUDA 之外的平台上跳过这些测试
# 因为 FlashInfer attention backend 仅支持 NVIDIA CUDA。
@pytest.mark.skipif(
    not current_platform.is_cuda(),
    reason="Attention backend FLASHINFER is only supported on CUDA.",
)
def test_hybrid_attention_mamba_tensor_shapes():
    """..."""
    pass

@pytest.mark.skipif(
    not current_platform.is_cuda(),
    reason="Attention backend FLASHINFER is only supported on CUDA.",
)
def test_mamba_cache_raises_when_max_num_seqs_exceeds_blocks():
    """..."""
    pass
```

评论区精华

该 PR 无任何 review 评论，仅由合并者 jikunshang 直接批准。无公开的技术讨论或权衡。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试跳过条件，不影响任何生产代码逻辑。主要风险是：如果未来 FlashInfer 后端在其他非 CUDA 平台（如 ROCm、AMD）上得到支持，该跳过条件可能过于宽泛，导致相关测试被错误跳过。不过当前 FlashInfer 官方仅支持 CUDA，因此此条件在未来一段时间内是合适的。
- 影响：影响范围极小，仅影响 test_gpu_model_runner.py 中两个测试函数在非 CUDA 平台上的执行行为：
  - XPU/Intel GPU: 测试被跳过，避免 FlashInfer 不兼容导致的失败。
  - ROCm: 之前这些测试会因 is_rocm() 返回 True 而跳过；变更后，is_cuda() 返回 False，因此 not is_cuda() 为 True，同样跳过，行为一致。
  - CUDA: is_cuda() 返回 True，not is_cuda() 为 False，测试正常执行，与之前相同。

因此，对用户和系统的影响是消除了 XPU 平台上的 CI 误报，提升了 CI 可靠性。

- 风险标记：测试跳过条件可能过于宽泛

关联脉络

- 暂无明显关联 PR