PR #49939 完整报告

vllm-project/vllm

[XPU][CI] Use platform device in InputBatch V2 test

合并时间：2026-07-27 15:28

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49939

执行摘要

- 一句话：测试中设备类型从硬编码 CUDA 改为平台动态获取
- 推荐动作：值得合并，是标准的多平台兼容修复。可推广至其他测试中类似的硬编码设备字符串。

功能与动机

PR 描述明确指出目的是 'Fix the InputBatch V2 test to use the current platform device instead of hard-coding CUDA'，解决测试在 Intel GPU 等平台上因设备类型硬编码为 CUDA 而失败的问题。

实现拆解

1. 在 tests/v1/worker/test_gpu_input_batch_v2.py 中，新增导入语句 from vllm.platforms import current_platform，引入平台工具函数。
2. 将第 10 行的硬编码 DEVICE = "cuda" 改为 DEVICE = current_platform.device_type，使测试自动适配当前运行平台的设备类型（如 cuda、xpu 等）。
3. 其他测试逻辑保持不变，仅调整设备来源，确保兼容性。

关键文件：

- tests/v1/worker/test_gpu_input_batch_v2.py（模块 输入批处理；类别 test；类型 test-coverage）：测试文件是唯一变更文件，通过使用 current_platform.device_type 替代硬编码 cuda 实现多平台兼容。

关键符号：未识别

关键源码片段

tests/v1/worker/test_gpu_input_batch_v2.py

测试文件是唯一变更文件，通过使用 current_platform.device_type 替代硬编码 cuda 实现多平台兼容。

```
# tests/v1/worker/test_gpu_input_batch_v2.py
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
"""Tests for the V2 model runner's InputBatch (vllm.v1.worker.gpu.input_batch)."""
```

```
import pytest
import torch

from vllm.platforms import current_platform # 新增导入，用于获取当前平台设备类型
from vllm.v1.worker.gpu.input_batch import InputBatch, InputBuffers

# 原来硬编码为 "cuda"，现在自动获取当前平台的设备类型
DEVICE = current_platform.device_type


@pytest.mark.parametrize(
    "num_reqs,num_tokens",
    [
        (256, 496), # remainder 240: previously gave the last request 241 tokens
        (128, 512), # no remainder
        (3, 8),
        (1, 7),
    ],
)
def test_make_dummy_distributes_remainder(num_reqs: int, num_tokens: int):
    """No dummy request may exceed ceil(num_tokens / num_reqs) tokens."""
    buffers = InputBuffers(
        max_num_reqs=num_reqs, max_num_tokens=num_tokens, device=torch.device(DEVICE)
    )
    batch = InputBatch.make_dummy(num_reqs, num_tokens, buffers)
    max_per_req = -(-num_tokens // num_reqs)
    assert batch.num_scheduled_tokens.sum() == num_tokens
    assert batch.num_scheduled_tokens.max() == max_per_req
    assert batch.num_scheduled_tokens.min() >= num_tokens // num_reqs
    assert (batch.num_scheduled_tokens[:-1] <= batch.num_scheduled_tokens[1:]).all()
```

评论区精华

无 review 讨论。claude[bot] 自动评论因来自 fork 跳过审查，jikunshang 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改一行别名定义和新增一行导入，未改变任何测试功能逻辑。需要确保 current_platform.device_type 在所有目标平台上返回有效设备类型字符串。
- 影响：直接解除了 InputBatch V2 测试在 Intel GPU (XPU) 等非 CUDA 平台上的阻塞，是 CI 多平台支持的增量改进。对 CUDA 平台无行为变化。
- 风险标记：平台工具函数依赖

关联脉络

- 暂无明显关联 PR