

PR #55660 完整报告

vllm-project/vllm

[CI] [Test] skip test_wna16_cuda_high_bit_skips_humming on non-CUDA platforms

合并时间: 2026-09-07 16:57

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/55660>

执行摘要

- 一句话: 为 INC WNA16 高位量化测试添加平台跳过条件, 修复 XPU CI 失败。
- 推荐动作: 这是一个简单、清晰的测试修复。对于负责维护量化模块、多平台 CI 或理解测试 mock 局限性的工程师有参考价值。代码变更清晰, 风险可控, 建议快速审阅。

功能与动机

根据 PR body 描述, 该测试在 XPU CI 运行器上失败, 错误为 `NotImplementedError: INC quantization with bits=4, sym=True is not supported`. 根本原因是 `AutoGPTQLinearMethod.__init__` (或其调用链中的 `check_marlin_supported()`) 会执行一个真实的、未被测试 mock 的设备能力检查。在真正的 XPU 主机上, 此检查会失败, 无论测试如何 mock 平台标志。因此, 该测试仅适用于在真正的 CUDA 硬件上运行。

实现拆解

本次变更仅涉及一个测试文件, 核心是增加一个平台跳过条件:

1. 变更入口: 在测试文件 `tests/quantization/test_auto_round.py` 中, 为函数 `test_wna16_cuda_high_bit_skips_humming` 添加装饰器。
2. 核心逻辑: 在 `@pytest.mark.parametrize("bits", [4, 8])` 装饰器之上, 新增 `@pytest.mark.skipif(not current_platform.is_cuda(), reason="This test only exercises the CUDA Marlin path.")`。这确保了当测试不在真实的 CUDA 硬件上执行时, 该测试会被跳过, 从而避免因未 mock 的设备检查而失败。
3. 配套改动: 无其他配套改动。此修改与该文件内其他 CUDA 专用测试的跳过模式 (例如 `MODELS` 参数化测试中的 `skipif` 标记) 完全一致。

关键文件:

- `tests/quantization/test_auto_round.py` (模块 量化测试; 类别 test; 类型 test-coverage)
: 这是本次变更的唯一文件, 包含需要修复的测试用例。

关键符号: 未识别

关键源码片段

`tests/quantization/test_auto_round.py`

这是本次变更的唯一文件, 包含需要修复的测试用例。

```

@pytest.mark.skipif(
    not current_platform.is_cuda(),
    reason="This test only exercises the CUDA Marlin path.",
)
@pytest.mark.parametrize("bits", [4, 8])
def test_wna16_cuda_high_bit_skips_humming(monkeypatch, bits) -> None:
    """4/8-bit int stays on the Marlin/GPTQ/AWQ path even on CUDA so a single
    model can mix high-bit Marlin and low-bit humming layers."""
    called = {"humming": False}

    # 以下 mock 仅能模拟平台标志，无法阻止 AutoGPTQLinearMethod 内部
    # 对真实设备能力的检查，因此该测试只能在真正的 CUDA 硬件上运行。
    monkeypatch.setattr(current_platform, "is_cuda", lambda: True)
    monkeypatch.setattr(current_platform, "is_xpu", lambda: False)
    monkeypatch.setattr(current_platform, "is_cpu", lambda: False)
    monkeypatch.setattr(
        "vllm.model_executor.layers.quantization.inc.schemes."
        "inc_wna16_scheme._build_humming_linear_method",
        lambda layer_config: called.update({"humming": True}),
    )

    method = INCWna16Scheme().get_linear_method(
        make_config(),
        object(),
        "model.layers.0.mlp.down_proj",
        make_layer_config(bits=bits),
    )

    assert called["humming"] is False
    assert isinstance(method, INCLinearMethod)

```

评论区精华

Review 讨论内容很少。Claude bot 提示这是一个 fork 提交，自动审查被禁用。提交者 chaojun-zhang 触发了 CI 运行。合作者 jikunshang 直接批准了该 PR。核心的技术决策和修复方案完全在 PR body 和 commit message 中阐述清楚。

- CI 测试失败修复 (testing): 解决方案是添加 `pytest.mark.skipif`，在非 CUDA 平台跳过此测试。

风险与影响

- 风险：风险极低。此变更仅修改了一个测试文件，添加了一个平台跳过条件。它不会影响任何生产代码或核心逻辑。唯一的风险是跳过条件可能过于宽泛，但 `reason` 字段明确说明了原因，且该测试本身的设计就决定了它只能在 CUDA 上验证特定路径。
- 影响：影响范围非常有限且积极。该 PR 修复了 Intel XPU 平台的 CI 流水线失败，提升了 CI 的可靠性。对用户和系统功能无直接影响。对团队而言，它消除了非 CUDA 平台上的一个误报测试失败。

- 风险标记：测试隔离不足

关联脉络

- PR #55453 [CI] Restore clean GPU state before OAI Triton MoE tests: 同为修复 CI 测试环境或隔离问题，提升测试可靠性。