

PR #32298 完整报告

sgl-project/sglang

[CI] Fix XPU platform test on machines without the XPU sgl-kernel op

合并时间: 2026-07-24 16:42

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32298>

执行摘要

- 一句话: 修复 XPU platform 测试在无 XPU 设备上失败
- 推荐动作: 建议所有维护者合入此修复, 以解除 CI 阻塞。该 PR 本身无需精读, 但其所用的 `patch.object + create=True` 模式值得在跨平台测试中推广, 以避免因可选依赖缺失导致的导入错误。

功能与动机

PR #31949 添加了 XPU 平台测试后, `base-a-test-cpu (3)` 在每个 PR 上都会失败:

`test_default_get_device_capability_uses_xpu` 使用

`@patch("torch.ops.sgl_kernel.query_device.default")`, 该 op 仅在 XPU 构建的 sgl-kernel 中注册, 在 CPU/CUDA 机器上导入时抛出 `ModuleNotFoundError`。PR body 明确指出该问题阻塞了许多 PR。

实现拆解

1. 定位根因: 在 `test/registered/unit/platforms/test_platform_interface.py` 的 `TestXpuDeviceMixin.test_default_get_device_capability_uses_xpu` 中, 装饰器 `@patch("torch.ops.sgl_kernel.query_device.default", return_value=(9, 0))` 在测试加载时就会尝试导入 `torch.ops.sgl_kernel`, 而非 XPU 机器的 `sgl_kernel` 缺失, 导致 `ModuleNotFoundError`。
2. 改用 `patch.object + create=True`: 移除装饰器, 在测试函数内部使用 `patch.object(torch.ops.sgl_kernel, "query_device", fake_query_device, create=True)` 进行模拟。`patch.object` 直接修改命名空间对象的属性, 不会触发模块导入, `create=True` 允许在属性不存在时创建。
3. 新增 `torch.xpu.current_device()` 模拟: `XpuSRTPlatform.get_device_capability` 内部调用了 `torch.xpu.current_device()`, 该函数在无 XPU 设备时也会失败。因此同时添加 `patch("torch.xpu.current_device", return_value=0)`。
4. 使用 `with` 语句管理上下文: 将两个 `patch` 放入 `with` 语句中, 确保模拟作用域仅限于测试断言。

关键文件:

- `test/registered/unit/platforms/test_platform_interface.py` (模块 平台接口; 类别 `test`; 类型 `test-coverage`; 符号 `test_default_get_device_capability_uses_xpu`): 单文件变更

，修复了 XPU 平台测试在无 XPU 设备环境下的导入失败问题。

关键符号：test_default_get_device_capability_uses_xpu

关键源码片段

test/registered/unit/platforms/test_platform_interface.py

单文件变更，修复了 XPU 平台测试在无 XPU 设备环境下的导入失败问题。

```
# 变更后的 test_default_get_device_capability_uses_xpu
# 使用 patch.object 并设置 create=True 避免导入失败

def test_default_get_device_capability_uses_xpu(self):
    # torch.ops.sgl_kernel.query_device 仅在 XPU 构建的 sgl-kernel 中注册，
    # 因此不能使用装饰器 @patch("torch.ops.sgl_kernel.query_device.default")
    # 因为它在测试加载时就会尝试导入模块，在没有 sgl-kernel 的机器上会失败。
    # 改用 patch.object 并设置 create=True 来绕过此问题。
    # 同时，torch.xpu.current_device() 也需要 XPU 设备，因此也 mock 掉。
    base = XpuSRTPlatform()
    fake_query_device = MagicMock()
    fake_query_device.default.return_value = (9, 0)
    with (
        patch("torch.xpu.current_device", return_value=0),
        patch.object(
            torch.ops.sgl_kernel, "query_device", fake_query_device, create=True
        ),
    ):
        self.assertEqual(base.get_device_capability(0), DeviceCapability(9, 0))
    fake_query_device.default.assert_called_once_with(0)
```

评论区精华

没有实质性 review 讨论。审核者 ShangmingCai 批准了该 PR，并评论 "Nice fix, this is blocking many PRs."，确认了其重要性。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试代码，不修改任何生产逻辑。模拟对象与原接口一致，测试逻辑不变。唯一可能的风险是如果未来 XpuSRTPlatform.get_device_capability 的实现不再调用 torch.xpu.current_device() 或 torch.ops.sgl_kernel.query_device，则该测试会变为无效，但这属于正常的测试维护范畴。
- 影响：影响范围：仅影响 CI 中的一个测试文件，但影响面大，因为该测试的失败阻塞了所有 PR 的 CI 流程。影响程度：修复后，所有 PR 的 base-a-test-cpu (3) 测试将不再因该原因失败，CI 绿线恢复。对用户无影响，对开发者体验有正面影响。
- 风险标记：仅测试变更，解除 CI 阻塞，低风险

关联脉络

- PR #31949 [Intel GPU] Add XPU Platform support: 本 PR 修复了 #31949 引入的 XPU 测试在非 XPU 机器上的失败问题，两者直接关联。