

PR #49732 完整报告

vllm-project/vllm

[ROCm][CI] Skip three torchao tests of gfx950 until `torchao==0.18` is released

合并时间: 2026-07-27 17:36

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49732>

执行摘要

- 一句话: 跳过 gfx950 上 torchao 在线量化三项测试
- 推荐动作: 该 PR 内容明确, 无技术争议, 可正常合入。建议在 torchao 0.18.0 正式发布后, 尽快升级依赖并移除该 skipif 逻辑。作为临时性测试兼容性修复, 无需精读。

功能与动机

在 AMD MI350 (gfx950) 上, torchao 0.17.0 的 `_float8_dynamic_activation_float8_weight_transform` 断言只允许 CUDA<=8.9 或 MI300, 导致 `test_online_quant_config_dict_json`、`test_online_quant_config_file`、`test_reload_weights` 三个测试因 `AssertionError` 失败。上游 torchao main 分支已添加 `is_MI350()` 支持, 预计随 0.18.0 发布, 本 PR 在这些测试上设置临时跳过, 避免 CI 阻塞。

实现拆解

1. 新增导入: 在 `tests/quantization/test_torchao.py` 文件顶部添加 `importlib.metadata`、`packaging.version` 导入, 用于精确获取 torchao 版本。
2. 定义平台辅助函数: 通过条件导入 `vllm.platforms.rocm.on_gfx950`, 并在非 ROCm 环境提供返回 `False` 的 `fallback` 函数, 确保跨平台兼容。
3. 添加版本常量: 定义 `TORCHAO_VERSION_0_18_AVAILABLE`, 通过 `importlib.metadata.version("torchao")` 获取实际安装版本并与 0.18.0 比较。
4. 添加跳转装饰器: 为三个测试函数 `test_online_quant_config_dict_json`、`test_online_quant_config_file`、`test_reload_weights` 分别添加 `@pytest.mark.skipif(on_gfx950() and not TORCHAO_VERSION_0_18_AVAILABLE, reason="requires torchao>=0.18.0 on gfx950")`, 仅在 gfx950 且 torchao<0.18.0 时跳过。
5. 本次变更纯属测试配置调整, 无任何生产代码逻辑变动。

关键文件:

- `tests/quantization/test_torchao.py` (模块 测试; 类别 `test`; 类型 `test-coverage`; 符号 `on_gfx950`): 唯一的变更文件, 为三项测试添加平台和版本条件跳转, 避免 gfx950 上因 torchao 版本低导致的失败。

关键符号: `on_gfx950`

关键源码片段

tests/quantization/test_torchao.py

唯一的变更文件，为三项测试添加平台和版本条件跳转，避免 gfx950 上因 torchao 版本低导致的失败。

```
import importlib.metadata
import importlib.util

import pytest
import torch
from packaging import version

from vllm.model_executor.model_loader import get_model_loader
from vllm.platforms import current_platform

# 仅在 ROCm 平台导入真实的 on_gfx950，非 ROCm 直接返回 False 确保跨平台兼容
if current_platform.is_rocm():
    from vllm.platforms.rocm import on_gfx950
else:

    def on_gfx950() -> bool:
        return False

DEVICE_TYPE = current_platform.device_type
DTYPE = ["bfloat16"]

TORCHAO_AVAILABLE = importlib.util.find_spec("torchao") is not None
# 精确判断 torchao 版本是否 >= 0.18.0，避免因 nightly/dev 版本号解析错误
TORCHAO_VERSION_0_18_AVAILABLE = TORCHAO_AVAILABLE and version.parse(
    importlib.metadata.version("torchao")
) >= version.parse("0.18.0")

# 三个测试函数的跳转装饰器相同，此处以 test_online_quant_config_dict_json 为例
@pytest.mark.skipif(not TORCHAO_AVAILABLE, reason="torchao is not available")
@pytest.mark.skipif(
    on_gfx950() and not TORCHAO_VERSION_0_18_AVAILABLE,
    reason="requires torchao>=0.18.0 on gfx950", # 仅 gfx950 且 torchao<0.18.0 时跳过
)
def test_online_quant_config_dict_json(vllm_runner, enable_pickle):
    ... # 测试体保持不变
```

评论区精华

主要讨论围绕 reviewer tjanaa 的疑问：该问题是否应在所有 ROCm 平台出现？作者 fxmarty-amd 解释，MI300 已在 torchao 0.17.0 的 `is_MI300()` 断言中得到支持，而 gfx950 是新增硬件尚未被上游 torchao 识别。讨论很快达成共识，方案被接受。没有争议或未解决问题。

- MI350 专属问题还是全 ROCm 共通 (question): 仅 gfx950 受影响, MI300 因已包含在 torchao is_MI300() 断言中而正常通过。

风险与影响

- 风险: 风险极低。变更仅影响测试用例的 skipif 条件, 且条件中明确限定 on_gfx950() 和 not TORCHAO_VERSION_0_18_AVAILABLE 同时满足才会跳过。当 torchao 0.18.0 发布后, 测试将自动恢复运行。若未来 gfx950 有相似问题需更新版本下限, 可能会遗漏, 但属于可预期的小风险。
- 影响:
 - 影响范围: 仅限 AMD MI350/gfx950 硬件上的三项 torchao 量化测试。
 - 影响程度: 低。CI 不再因这些测试失败而报红, 其他硬件和测试不受影响。
 - 后续: torchao 0.18.0 发布后, 需要升级 vllm 依赖并移除本跳转逻辑。
 - 风险标记: 临时方案, 需跟踪上游版本

关联脉络

- 暂无明显关联 PR