

PR #50222 完整报告

vllm-project/vllm

[CI] Fix MXFP8 MOE backend selection tests on gfx942

合并时间: 2026-07-29 17:17

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50222>

执行摘要

- 一句话: 修复 gfx942 上 MXFP8 MoE 后端选择测试
- 推荐动作: 该 PR 修复了测试逻辑, 建议合入以确保 CI 稳定性。值得关注的是, 它揭示了不同 GPU 硬件 (gfx942 vs gfx950) 对 MXFP8 后端的支持差异, 以及测试需要模拟平台特性的最佳实践。

功能与动机

修复因平台差异导致的测试失败。PR #49747 的测试在 MI300 (gfx942) 上未通过, 因为该平台不支持 `supports_mx()`, 且原生 Triton MXFP8 后端仅适用于 gfx950。作者在 PR body 中说明: "These failures were overlooked as local testing of #49747 was done on MI350 and not on MI300."

实现拆解

1. 修改导入: 将 `mxfp8_native_moe` 中的 `Mxfp8NativeTritonExperts` 替换为 `mxfp8_emulation_moe` 中的 `Mxfp8EmulationTritonExperts`, 因为 gfx942 应使用模拟后端。
2. 调整 `test_gfx950_picks_aiter`: 新增 `patch(f"{_AITER_MOD}.current_platform.supports_mx", return_value=True)` 上下文管理器, 以模拟 gfx950 上的 MX 支持, 确保测试在 MI300 上也能正确通过。
3. 重命名并重写 `test_gfx942_picks_triton`: 改为 `test_gfx942_picks_emulation`; 将 `_flydsl_installed(False)` 替换为 `patch(f"{_AITER_MOD}.current_platform.supports_mx", return_value=False)`, 并断言 `Fp8MoeBackend.EMULATION` 和 `Mxfp8EmulationTritonExperts`, 因为 gfx942 应降级到模拟后端。

关键文件:

- `tests/kernels/moe/test_mxfp8_aiter_backend_selection.py` (模块测试; 类别 test; 类型 test-coverage; 符号 `test_gfx942_picks_triton`, `test_gfx942_picks_emulation`): 唯一修改的文件, 修复 gfx942 上的测试失败, 涉及导入变更、测试函数修改和新增模拟上下文。

关键符号: `test_gfx950_picks_aiter`, `test_gfx942_picks_emulation`

关键源码片段

tests/kernels/moe/test_mxfp8_aiter_backend_selection.py

唯一修改的文件，修复 gfx942 上的测试失败，涉及导入变更、测试函数修改和新增模拟上下文。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
"""MXFP8 MoE backend selection for the AITER FlyDSL kernel (gfx950).

GPU-free: mocks the platform (gfx950) and the ``flydsl`` package check, then
exercises the oracle so the FlyDSL backend is auto-picked when usable (including
under expert parallelism, since apply() forwards the expert_map as aiter's
expert_mask) and skipped (native fallback) when the device/package is missing.
"""

import dataclasses
from unittest.mock import patch

import pytest

from vllm.platforms import current_platform

if not current_platform.is_rocm():
    pytest.skip("This test can only run on ROCm.", allow_module_level=True)

# ... (imports)

# 关键变更：将 gfx942 的回退后端从 Mxfp8NativeTritonExperts 改为
Mxfp8EmulationTritonExperts
from vllm.model_executor.layers.fused_moe.experts.mxfp8_emulation_moe import (
    Mxfp8EmulationTritonExperts,
)

# ... (helper functions)

def test_gfx950_picks_aiter():
    """Auto-select on real ROCm hardware with flydsl usable -> FlyDSL wins."""
    # NOTE: Fp8MoeBackend.AITER_MXFP8 does not require VLLM_ROCM_USE_AITER=1
    # 新增 patch: 模拟 gfx950 的 MX 支持，使测试在 gfx942 (MI300) 上也能通过
    with (
        patch(f"{_AITER_MOD}.current_platform.supports_mx", return_value=True),
        _flydsl_installed(True),
    ):
        backend, experts_cls = select_mxfp8_moe_backend(_config())
        assert backend is Fp8MoeBackend.AITER_MXFP8
        assert experts_cls is AiterMxfp8Experts

def test_gfx942_picks_emulation():
```

```
"""flydsl unusable (e.g. gfx942, no FlyDSL support) -> native Triton
dot_scaled backend wins instead."""
# 直接模拟 supports_mx = False, 而非依赖 _flydsl_installed
with patch(f"{_AITER_MOD}.current_platform.supports_mx", return_value=False):
    backend, experts_cls = select_mxfp8_moe_backend(_config())
# 预期后端改为 EMULATION, 因为 gfx942 不支持原生 Triton MXFP8
assert backend is Fp8MoeBackend.EMULATION
assert experts_cls is Mxfp8EmulationTritonExperts
```

评论区精华

无 review 评论, 仅 automated review 和 maintainer approval。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改一个测试文件，使测试正确反映不同 GPU 平台的后端选择行为。
如果 Mxfp8EmulationTritonExperts 类不存在或导入路径错误，测试会立即失败。
- 影响：影响范围仅限于 ROCm 平台（gfx942/MI300）上的 MXFP8 MoE 后端选择测试。修复后，CI 中相关测试可在 MI300 上正确运行，避免假阳性失败。
- 风险标记：暂无

关联脉络

- PR #49747 [Model] Add MXFP8 MoE backend selection tests: 本 PR 修复了 #49747 引入的测试失败，是 #49747 的后续修复。
- PR #46123 [Misc] Track MXFP8 MoE backend selection: PR body 中提及的关联跟踪 PR。