PR #45302 完整报告

vllm-project/vllm

[ROCm][CI] fix fp8 support for test_deepep_moe

合并时间：2026-06-12 13:16

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45302

执行摘要

- 一句话：修复 ROCm 上 FP8 MOE 测试的 dtype 硬编码
- 推荐动作：该 PR 是典型的平台兼容性修复，值得 ROCm 开发者关注其模式：使用 current_platform.fp8_dtype() 替代硬编码的 dtype。对于其他测试中的类似硬编码问题，可参考此模式进行修复。

功能与动机

ROCm 平台使用 float8_e4m3fnuz 而非 NVIDIA 的 float8_e4m3fn，测试中硬编码的 dtype 导致 FP8 相关断言和初始化失败，影响 CI 测试 kernels/moe/test_deepep_moe.py::test_deep_ep_moe 的正常运行。

实现拆解

1. 在 tests/kernels/moe/test_deepep_moe.py 中新增导入 from vllm.platforms import current_platform。
2. 将 make_weights、TestTensors.make、build_expert_map、torch_moe_impl、_deep_ep_moe 等多个函数中所有 torch.float8_e4m3fn 的引用替换为 current_platform.fp8_dtype()。
3. 将测试参数化列表 DTYPES 中的 torch.float8_e4m3fn 替换为 current_platform.fp8_dtype()。
4. 所有修改仅涉及测试文件，无核心源码变更。

关键文件：

- tests/kernels/moe/test_deepep_moe.py（模块 MoE 内核；类别 test；类型 test-coverage；符号 make_weights, TestTensors.make, build_expert_map, torch_moe_impl）：唯一变更文件，修复了所有 FP8 dtype 硬编码问题。

关键符号：未识别

关键源码片段

tests/kernels/moe/test_deepep_moe.py

唯一变更文件，修复了所有 FP8 dtype 硬编码问题。

```
# tests/kernels/moe/test_deepep_moe.py（关键变更片段）
```

```
from vllm.platforms import current_platform # 新增导入，用于获取平台相关的 FP8 dtype

def make_weights(e, n, k, dtype):
    ...
    # 之前：assert dtype == torch.float8_e4m3fn
    assert dtype == current_platform.fp8_dtype() # 动态适配平台
    ...

class TestTensors:
    @staticmethod
    def make(config: TestConfig, low_latency_mode: bool) -> "TestTensors":
        # 之前：assert config.dtype in [torch.bfloat16, torch.float8_e4m3fn]
        assert config.dtype in [torch.bfloat16, current_platform.fp8_dtype()]
        # 之前：token_dtype = torch.bfloat16 if config.dtype == torch.float8_e4m3fn else config.
        dtype
        token_dtype = (
            torch.bfloat16
            if config.dtype == current_platform.fp8_dtype()
            else config.dtype
        )
        ...

# 测试参数化列表也做了同样替换
# 之前：DTYPES = [torch.bfloat16, torch.float8_e4m3fn]
DTYPES = [torch.bfloat16, current_platform.fp8_dtype()]
```

评论区精华

无 review 评论。审核者 AndreasKaratzas 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅限于测试文件，且替换逻辑通过平台抽象层 current_platform.fp8_dtype() 保证了向后兼容性，NVIDIA 平台下返回 torch.float8_e4m3fn，行为不变。
- 影响：影响范围限于 ROCm CI 中的 deepep MOE 测试。修复后该测试可在 ROCm 上正常执行，不影响现有 NVIDIA 平台行为。
- 风险标记：暂无

关联脉络

- PR #46176 [ROCm] Use vLLM's fp8 quant max in AITER hipBLASLt accuracy test: 同为 ROCm FP8 兼容性修复，展示了类似的平台适配模式。