PR #49839 完整报告

vllm-project/vllm

[Test][ROCm] Account for gfx950 FP8 RMSNorm rounding

合并时间：2026-07-30 18:46

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49839

执行摘要

- 一句话：为 gfx950 放宽 FP8 RMSNorm 测试容差
- 推荐动作：此 PR 值得关注作为处理 FP8 精度差异的测试模式参考。设计上体现了 fp8_allclose 工具函数的灵活使用。

功能与动机

gfx950 的融合 RMSNorm 实现使用不同的归约树，在 FP16 输入且无残差加法时，归一化值可能落在 E4M3 相邻编码上，导致 FP8 量化结果相差 12.5%。此 PR 为此特定场景添加专用断言，避免测试误报。

实现拆解

1. tests/kernels/core/test_layernorm.py (test_fused_rms_norm_quant)：将 ROCm 分支中的通用异常容忍断言替换为 gfx950 专用检查。当运行在 gfx950、dtype=torch.float16、无残差时，使用 fp8_allclose（rtol=0.125, atol=2e-3）并验证最大 ULP 距离 ≤ 1；否则保留原异常容忍逻辑。
2. tests/kernels/core/test_fused_quant_layernorm.py (test_rms_norm)：在文件级检测 ON_GFX950 标志，在量化为 FP8、group_size=None、dtype=torch.bfloat16 时设置 use_gfx950_fp8_allclose，对未通过常规 allclose 检查的用例使用 fp8_allclose。
3. 导入调整：两个文件新增 from tests.kernels.utils import fp8_allclose。

关键文件：

- tests/kernels/core/test_layernorm.py（模块 层归一化；类别 test；类型 test-coverage；符号 test_fused_rms_norm_quant）：核心测试文件，新增 gfx950 专用 FP8 RMSNorm 断言逻辑
- tests/kernels/core/test_fused_quant_layernorm.py（模块 融合量化；类别 test；类型 test-coverage；符号 test_rms_norm）：第二个测试文件，为融合量化 RMSNorm 添加类似 gfx950 专用分支

关键符号：test_fused_rms_norm_quant, test_rms_norm

关键源码片段

tests/kernels/core/test_layernorm.py

核心测试文件，新增 gfx950 专用 FP8 RMSNorm 断言逻辑

```
# tests/kernels/core/test_layernorm.py (partial)
if current_platform.is_rocm():
    from vllm.platforms.rocm import on_gfx950

    if on_gfx950() and dtype == torch.float16 and not add_residual:
        # gfx950 融合归约树可能使归一化值跨越 E4M3 边界，
        # 允许 12.5% 相对误差（1 ULP），但最大 ULP 不超过 1
        assert fp8_allclose(out_quant_fused, out_quant, rtol=0.125, atol=2e-3)
        assert int(fp8_ulp_distance(out_quant_fused, out_quant).max()) <= 1
    else:
        # 原有逻辑：容忍少量孤立 FP8 异常值
        ulp = fp8_ulp_distance(out_quant, out_quant_fused)
        max_outliers = ulp.numel() // 100_000 + 8
        num_outliers = int((ulp > 0).sum().item())
        assert num_outliers <= max_outliers, (
            f"FP8 quant mismatch: {num_outliers} fp8 outliers "
            f"(allowed {max_outliers})"
        )
```

tests/kernels/core/test_fused_quant_layernorm.py

第二个测试文件，为融合量化 RMSNorm 添加类似 gfx950 专用分支

```
# tests/kernels/core/test_fused_quant_layernorm.py (partial)
ON_GFX950 = False
if current_platform.is_rocm():
    from vllm.platforms.rocm import on_gfx950
    ON_GFX950 = on_gfx950()

# ... 在 test_rms_norm 函数内部:
use_gfx950_fp8_allclose = (
    current_platform.is_rocm()
    and ON_GFX950
    and group_size is None
    and dtype == torch.bfloat16
    and quant_dtype == current_platform.fp8_dtype()
)

# 在原始 allclose 失败后的分支中:
if relax_block_rocm:
    # 原有逻辑
    ...
elif use_gfx950_fp8_allclose:
    # gfx950 归约树可能跨越 E4M3 边界
    ok = fp8_allclose(ops_out, ref_out, rtol=0.125, atol=2e-3)
    ok = ok and int(fp8_ulp_distance(ops_out, ref_out).max()) <= 1
else:
    # CUDA 及其他情况
    ...
```

评论区精华

审阅者 tjtanaa 建议在 test_layernorm.py 中将 from vllm.platforms.rocm import on_gfx950 放在使用处附近，并在 test_fused_quant_layernorm.py 中增加 ON_GFX950 全局变量初始化（含 or 分支）以安全处理非 ROCm 平台。作者采纳建议并更新。

- gfx950 import 放置位置 (style): 作者修改代码，采用了审阅者的建议。

风险与影响

- 风险：风险极低。仅修改测试断言逻辑，无生产代码变更。gfx950 专用分支仅在满足特定运行时条件时触发；其他架构和配置沿用原有逻辑。新增的 ON_GFX950 变量在非 ROCm 平台默认 False，不会意外生效。
- 影响：影响仅限于 ROCm gfx950 上的 FP8 RMSNorm 测试，减少因舍入差异导致的测试失败。对用户无直接影响，对开发者的 CI 可靠性有正面改进。
- 风险标记：影响仅限测试，只在 gfx950 特定条件下生效

关联脉络

- PR #50339 [FlexAttention] Avoid encoder block-mask compile explosion: 同为 ROCm 相关测试调整，但无直接联系
- PR #49937 [ROCm] Add AITER FP8 ViT encoder attention: 同为 ROCm FP8 精度处理，但不同模块
- PR #48257 [ROCm] [CI] Support cached K/V (key/value=None) in Triton prefix-prefill: 同为 ROCm 测试稳定性改进