

PR #51273 完整报告

vllm-project/vllm

[ROCm][CI] Update AITER AR+RMS e2e fusion counts for final-norm coverage

合并时间: 2026-08-07 02:05

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51273>

执行摘要

- 一句话: 修正 AITER AR+RMS e2e 融合计数以覆盖最终 norm
- 推荐动作: 不值得精读, 改动机械 (2 行测试期望值)。值得注意的设计点: 功能 PR 必须同步所有相关测试期望; e2e 期望值与单元测试需配套维护; ROCm 特有路径的覆盖依赖平台条件断言, 修改时需谨慎。

功能与动机

PR #50802 新增 AiterAllreduceFusedAddRMSNormOutputOnlyPattern, 使 AITER AllReduce+RMSNorm 融合能覆盖最终 `model.norm`, 但只更新了单元测试, 未同步 `tests/compile/fusions_e2e/models.py` 中的 e2e 期望计数, 导致 AITER e2e 计数差一。本 PR 修复该遗漏, 将 llama3_8b 和 qwen3_a3b 的 `aiter_ar_rms_fusion` 从 `n_layers * 2` 提升到 `n_layers * 2 + 1`, 与 `ar_rms_fusion` 对齐。

实现拆解

1. 修改入口: `tests/compile/fusions_e2e/models.py` 中的 ModelFusionInfo 数据定义, 该文件集中维护 torch.compile 融合 e2e 测试的期望匹配值。
2. 将 llama3_8b 的 `aiter_ar_rms_fusion` 期望值由 `n_layers * 2` 改为 `n_layers * 2 + 1`, 与 `ar_rms_fusion` 对齐, 记录最终 `model.norm` 的融合。
3. 对 qwen3_a3b 做相同修改。
4. 验证方式: 运行 `pytest tests/compile/fusions_e2e/test_tp2_ar_rms.py::test_tp2_ar_rms_fusions`, 覆盖 llama3_8b 和 qwen3_a3b 两个模型; CUDA-only 路径不参与 `aiter_ar_rms_fusion` 断言, 该字段仅在 ROCm 上生效。无配置、schema 或部署配套改动。

关键文件:

- `tests/compile/fusions_e2e/models.py` (模块 融合测试; 类别 test; 类型 test-coverage) : 唯一修改的文件, 更新 AITER AR+RMS 融合的 e2e 期望计数, 是 PR #50802 的配套测试修正。

关键符号: 未识别

关键源码片段

tests/compile/fusions_e2e/models.py

唯一修改的文件，更新 AITER AR+RMS 融合的 e2e 期望计数，是 PR #50802 的配套测试修正。

```
# tests/compile/fusions_e2e/models.py 中 llama3_8b 与 qwen3_a3b 的融合期望定义
# 本次修改将 aiter_ar_rms_fusion 期望值从 n_layers * 2 调整为 n_layers * 2 + 1,
# 使其与 ar_rms_fusion 对齐，覆盖 PR#50802 引入的最终 norm 融合能力。
```

```
llama3_8b = ModelFusionInfo(
    model_name="meta-llama/Llama-3.1-8B-Instruct",
    matches=lambda n_layers: Matches(
        ar_rms_fusion=n_layers * 2 + 1,
        aiter_ar_rms_fusion=n_layers * 2 + 1, # 修改点：与 ar_rms_fusion 保持一致
        sequence_parallel=n_layers * 2 + 1,
        async_tp=n_layers * 4,
    ),
)

qwen3_a3b = ModelFusionInfo(
    model_name="Qwen/Qwen3-30B-A3B",
    matches=lambda n_layers: Matches(
        norm_rope_fusion=n_layers,
        ar_rms_fusion=n_layers * 2 + 1,
        aiter_ar_rms_fusion=n_layers * 2 + 1, # 修改点：与 ar_rms_fusion 保持一致
        sequence_parallel=n_layers * 2 + 1,
        async_tp=n_layers * 2,
    ),
)
```

评论区精华

该 PR 无实质性技术讨论。claude[bot] 提示来自 fork 的 PR 自动审查被禁用；AndreasKaratzas 直接批准合并，未留下评论。

- fork PR 自动 review 禁用 (other): 无技术讨论；AndreasKaratzas 直接批准合并。

风险与影响

- 风险：变更集中在 tests/compile/fusions_e2e/models.py 的 e2e 期望值。潜在风险是若 AITER 融合实现实际未覆盖最终 model.norm，修改后 ROCm CI 会因计数不匹配而失败；但 PR #50802 已实现 AiterAllreduceFusedAddRMSNormOutputOnlyPattern，且新期望值与 ar_rms_fusion 一致，因此风险极低。该 PR 不触碰任何推理运行时代码，无性能、安全或兼容性影响。
- 影响：对用户无影响；对推理系统无影响；对团队的影响是补全 ROCm CI 的测试覆盖，消除 PR #50802 遗留的计数不一致，避免后续 AITER 相关改动出现假阳性或假阴性。影响面限定在 ROCm 平台上的 e2e 融合测试。
- 风险标记：仅测试期望变更，ROCm 专用路径

关联脉络

- PR #50802 [ROCm] Fix AITER all-reduce fusion coverage: 本 PR 修复了该 PR 未同步的 e2e 融合期望计数，属于其配套测试修正。