

PR #27589 完整报告

sgl-project/sglang

Unskip Marlin NVFP4 tests

合并时间: 2026-06-17 02:58

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27589>

执行摘要

- 一句话: 取消 Marlin NVFP4 测试的 CI 跳过标记
- 推荐动作: 建议合入。这是一个低风险、高回报的测试恢复变更, 能增强 NVFP4 Marlin 后端的 CI 覆盖。未来可考虑将超时测试移到 nightly 以减少 PR 等待时间。

功能与动机

PR body 引用之前的 PR #27161, 指出测试因编译超时被跳过; 现在基础条件改善后希望恢复测试以保障质量。

实现拆解

1. 移除 pytest/unittest skip 装饰器: 在 test/registered/jit/test_gptq_marlin.py、test/registered/jit/test_moe_wna16_marlin.py 和 test/manual/models/test_nvidia_nemotron_3_nano_archived.py 中共移除 6 行 @pytest.mark.skip 或 @unittest.skip 标记。
2. 保留硬件约束跳过条件: 仅移除通用的编译超时跳过, 保留 @pytest.mark.skipif(not (is_sm80_supported() or is_sm90_supported())) 等 CUDA 架构过滤, 确保测试只在支持的 GPU 上执行。
3. 覆盖三类测试: GPTQ Marlin 的 NVFP4 变换与数值精度测试、MoE WNA16 Marlin 的 fused 内核测试 (含非门控 ReLU2 和 padded intermediate 场景)、以及 Nemotron-3-Nano 模型级 NVFP4 Marlin 端到端评测。

关键文件:

- test/registered/jit/test_gptq_marlin.py (模块 内核测试; 类别 test; 类型 test-coverage) : 移除 test_nvfp4_marlin_support_and_scale_transforms_sm80_sm90 和 test_nvfp4_marlin_dense_matches_dequant_reference 的 @pytest.mark.skip, 恢复 NVFP4 变换和数值精度测试。
- test/registered/jit/test_moe_wna16_marlin.py (模块 内核测试; 类别 test; 类型 test-coverage) : 移除三个 fused MoE 测试的 @pytest.mark.skip, 包括非门控 ReLU2、padded intermediate 和数值测试。
- test/manual/models/test_nvidia_nemotron_3_nano_archived.py (模块 模型评测; 类别 test; 类型 test-coverage) : 移除模型级端到端测试的 @unittest.skip, 使 NVFP4 Marlin 路径在 CI 中接受 lm-eval 验证。

关键符号: `test_nvfp4_marlin_support_and_scale_transforms_sm80_sm90`,
`test_nvfp4_marlin_dense_matches_dequant_reference`,
`test_fused_marlin_moe_non_gated_relu2`, `test_fused_marlin_moe_nvfp4_non_gated_padded_intermediate_launches`, `test_fused_marlin_moe_nvfp4_non_gated_padded_numeric_test`

关键源码片段

`test/registered/jit/test_gptq_marlin.py`

移除 `test_nvfp4_marlin_support_and_scale_transforms_sm80_sm90` 和 `test_nvfp4_marlin_dense_matches_dequant_reference` 的 `@pytest.mark.skip`, 恢复 NVFP4 变换和数值精度测试。

```
# 变更前:
# @pytest.mark.skip(reason="Skip, test pass locally but compiling takes too long in CI")
# 变更后: 直接移除, 保留 skipif
@pytest.mark.skipif(
    not (is_sm80_supported() or is_sm90_supported()),
    reason="NVFP4 Marlin fallback tests require CUDA SM8X/SM9X",
)
@pytest.mark.parametrize("dtype", [torch.float16, torch.bfloat16])
def test_nvfp4_marlin_support_and_scale_transforms_sm80_sm90(dtype):
    # ... 测试逻辑不变
    pass
```

评论区精华

PR 无 review 评论, 仅作者和 CI 机器人交互。作者两次尝试 `/rerun-test` 修改测试路径, 最终成功定位到正确的测试文件路径 (`test/registered/jit/test_gptq_marlin.py` 等)。之后合并 main 分支以引入修复 (可能指 PR #27605), 并通过 `/tag-and-rerun-ci` 触发完整流水线。

- 暂无高价值评论线程

风险与影响

- 风险: 低风险。变更仅移除跳过标记, 测试本身已在本地验证通过, 且保留硬件过滤条件不会在不适配的 GPU 上执行。即使测试再次超时也不会影响生产功能。
- 影响: CI 流程: Marlin NVFP4 相关测试将重新在 CI 中运行, 可能增加 Stage B 测试总时长, 但有助于尽早发现编译或量化路径的回归。用户: 无直接影响。团队: 确保 Marlin 后端的 NVFP4 量化路径持续得到验证。
- 风险标记: 测试回归风险低, 可能增加 CI 耗时

关联脉络

- PR #27161 (未提供, 根据 PR body 引用): 该 PR 的初始合并因编译超时被回退, 当前 PR 正是在其基础上取消跳过标记。
- PR #27605 (未提供, 根据作者评论): 作者合并 main 以引入该修复, 期望解决超时问题。