

PR #27781 完整报告

sgl-project/sglang

[CI] Move misplaced mhc kernel test into test/registered/kernels

合并时间: 2026-06-10 16:49

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27781>

执行摘要

- 一句话: 搬迁 mHC 内核测试到标准注册目录
- 推荐动作: 可直接合入。变更清晰、动机明确, 且已通过 1-gpu-h100 的实际运行验证。建议未来注意避免向顶层 tests/ 添加文件。

功能与动机

mHC kernel 测试文件是仓库中唯一位于顶层 tests/ 目录下的文件, 该目录不在 CI 扫描路径内, 导致测试从未被实际执行。PR 旨在将其归入 test/registered/kernels/ 并补全注册逻辑。

实现拆解

1. 文件搬迁: 将 tests/kernels/test_mhc_kernels.py 移动至 test/registered/kernels/test_mhc_kernels.py。
2. 添加 CI 注册: 在文件头部导入 register_cuda_ci 并调用 register_cuda_ci(est_time=30, stage="base-b", runner_config="1-gpu-large"), 使 run_suite.py 能够发现该测试。runner_config 从首次提交的 1-gpu-small 修正为 1-gpu-large, 因为 mHC kernel 内部的 TF32 prenorm GEMM 依赖 DeepGEMM, 而 DeepGEMM 在 SM120 (RTX 5090) 上不可用, 需使用 H100 (1-gpu-large)。
3. 添加主入口: 在文件末尾添加标准 if __name__ == "__main__" 块, 支持直接 python test_mhc_kernels.py 执行。
4. 清理: 删除已为空的顶层 tests/ 目录。

关键文件:

- test/registered/kernels/test_mhc_kernels.py (模块 测试覆盖; 类别 test; 类型 rename-or-move): 核心变更文件: 将之前被忽略的测试迁移至标准目录, 并添加了 CI 注册和主入口。

关键符号: 未识别

关键源码片段

test/registered/kernels/test_mhc_kernels.py

核心变更文件: 将之前被忽略的测试迁移至标准目录, 并添加了 CI 注册和主入口。

```
import pytest
```

```

import torch

import sglang.srt.layers.mhc as mhc
from sglang.srt.layers.mhc import mhc_fused_post_pre, mhc_post, mhc_pre
# 导入 CI 注册函数，使测试能被 run_suite.py 发现
from sglang.test.ci.ci_register import register_cuda_ci
# 注册为 CUDA CI 测试：预估 30 秒，base-b 阶段，需要 1-gpu-large 节点（H100，支持 DeepGEMM）
register_cuda_ci(est_time=30, stage="base-b", runner_config="1-gpu-large")

@pytest.mark.parametrize("hidden_size", [4096, 7168])
@pytest.mark.parametrize("num_tokens", [0, 1, 8, 17, 32, 64])
@pytest.mark.parametrize("use_norm", [False, True])
def test_mhc_fused_post_pre_matches_unfused(
    monkeypatch, hidden_size, num_tokens, use_norm
):
    # ... 测试体不变，验证 fused kernel 与 unfused 组合的数值一致性 ...
    pass

# 标准主入口，支持直接运行 python test_mhc_kernels.py
if __name__ == "__main__":
    import sys
    sys.exit(pytest.main([__file__]))

```

评论区精华

无人工 review 评论。仅 bot 自动回复，提示 Gemini Code Assist 即将停用。作者通过 `/rerun-test` 命令验证了在不同 GPU 节点上的执行结果：在 `1-gpu-5090`（SM120）上失败（DeepGEMM 不可用），在 `1-gpu-h100`（SM90）上成功，因此最终调整为 `1-gpu-large`。

- 测试运行配置确认 (other): 调整 `runner_config` 为 `1-gpu-large`，最终在 H100 上通过。

风险与影响

- 风险：此 PR 仅涉及测试文件迁移和 CI 注册，不修改任何源码逻辑，风险极低。唯一需关注的是 `runner_config` 选择是否正确：若未来 mHC kernel 不再依赖 DeepGEMM 或新 GPU 支持，可能需要调整。
- 影响：直接修复了 mHC kernel 测试被遗漏的问题，使其在 PR CI 中自动运行。对用户无感知，对开发流程的影响是增加了约 30 秒的 CI 耗时。
- 风险标记：仅影响测试覆盖，无源码变更

关联脉络

- PR #25976 Add mHC kernel with CI coverage: 原始添加该测试文件的 PR，当时将其放入了错误的目录。