

PR #27837 完整报告

sgl-project/sglang

[AMD] Register 3 JIT kernel unit tests for AMD CI

合并时间: 2026-06-18 14:28

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27837>

执行摘要

- 一句话: 为 AMD CI 注册 3 个 JIT 单元测试
- 推荐动作: 值得快速合入。该 PR 本身是增量改进, 但背后的测试分层和跨平台注册模式值得关注: 通过 `register_amd_ci` 和 `register_cuda_ci` 统一管理 CI 注册, 降低维护成本。

功能与动机

扩大 AMD CI 对 JIT kernel 的测试覆盖, 确保这些 kernel 在 MI325 (gfx942) 上正确运行。PR body 详细说明了验证结果和排除其他候选测试的原因。

实现拆解

1. 注册 AMD CI: 在 3 个测试文件 (`test_clamp_position.py`、`test_resolve_future_token_ids.py`、`test_rmsnorm_hf.py`) 的导入中添加 `register_amd_ci`, 并调用 `register_amd_ci(est_time=..., suite="jit-kernel-unit-test-amd")`, 保留原有 CUDA 注册。
2. 调整超时: 在 `pr-test-amd.yml` 和 `pr-test-amd-rocm720.yml` 中将 JIT kernel 测试步骤的 `timeout-minutes` 从 10 分钟提升至 30 分钟, 同时统一调整其他 `sgl-kernel` 测试步骤的超时。
3. 排除 `moe_align`: 初始版本注册了 4 个测试, 但在与 main 分支同步后, `test_moe_align_block_size` 因 ROCm 上的 4392 个参数化用例全部失败且超时被移除, 等待后续 kernel 移植。

关键文件:

- `test/registered/jit/test_clamp_position.py` (模块 JIT Kernel; 类别 test; 类型 test-coverage) : 核心变更之一: 导入 `register_amd_ci` 并注册 AMD CI, 增量为 +2 行。
- `test/registered/jit/test_resolve_future_token_ids.py` (模块 JIT Kernel; 类别 test; 类型 test-coverage) : 核心变更之一: 注册 AMD CI, 增量为 +2 行。
- `test/registered/jit/test_rmsnorm_hf.py` (模块 JIT Kernel; 类别 test; 类型 test-coverage) : 核心变更之一: 注册 AMD CI, 增量为 +2 行。
- `.github/workflows/pr-test-amd.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : CI 配置变更: 统一将 JIT kernel 测试和其他 `sgl-kernel` 测试步骤的超时从 10/14/20 分钟提升至 30 分钟。

- `.github/workflows/pr-test-amd-rocm720.yml` (模块 CI 配置; 类别 `infra`; 类型 `infrastructure`) : 同一超时调整的 ROCm 720 版本, 保持一致性。

关键符号: 未识别

关键源码片段

`test/registered/jit/test_clamp_position.py`

核心变更之一: 导入 `register_amd_ci` 并注册 AMD CI, 增量为 +2 行。

```
import sys

import pytest
import torch

from sglang.jit_kernel.clamp_position import clamp_position_cuda
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 注册 NVIDIA CI 套件 (保持不变)
register_cuda_ci(est_time=12, suite="base-b-kernel-unit-1-gpu-large")
register_cuda_ci(est_time=120, suite="nightly-kernel-1-gpu", nightly=True)
# 新增: 注册 AMD CI 套件, 使其在 AMD 平台上自动运行
register_amd_ci(est_time=12, suite="jit-kernel-unit-test-amd")

def _reference_clamp_position(seq_lens):
    return torch.clamp(seq_lens - 1, min=0).to(seq_lens.dtype)
```

`test/registered/jit/test_resolve_future_token_ids.py`

核心变更之一: 注册 AMD CI, 增量为 +2 行。

```
import sys

import pytest
import torch

from sglang.jit_kernel.resolve_future_token_ids import resolve_future_token_ids_cuda
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 注册 NVIDIA CI 套件 (保持不变)
register_cuda_ci(est_time=9, suite="base-b-kernel-unit-1-gpu-large")
register_cuda_ci(est_time=120, suite="nightly-kernel-1-gpu", nightly=True)
# 新增: 注册 AMD CI 套件
register_amd_ci(est_time=9, suite="jit-kernel-unit-test-amd")
```

`test/registered/jit/test_rmsnorm_hf.py`

核心变更之一: 注册 AMD CI, 增量为 +2 行。

```
from sglang.jit_kernel.rmsnorm_hf import (
    is_supported_rmsnorm_hf_hidden_size,
```

```
    rmsnorm_hf,
)
from sglang.jit_kernel.utils import get_ci_test_range
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 注册 NVIDIA CI 套件（保持不变）
register_cuda_ci(est_time=30, suite="base-b-kernel-unit-1-gpu-large")
register_cuda_ci(est_time=120, suite="nightly-kernel-1-gpu", nightly=True)
# 新增：注册 AMD CI 套件
register_amd_ci(est_time=30, suite="jit-kernel-unit-test-amd")
```

评论区精华

PR 无 review 评论，只有 HaiShaw 的 approve。但 PR body 详细记录了排除其他测试的原因：test_moe_align_block_size 因 ROCm 不兼容，其余测试因 CUDA-only 头文件或 fp8 问题被排除。

- test_moe_align_block_size 排除 (other): PR author 决定不注册该测试，等待 kernel 端 ROCm 移植完成后行注册。
- 超时时间对齐 (design): 已通过 CI 配置调整解决。

风险与影响

- 风险：风险低。仅涉及测试注册和 CI 配置调整，不修改核心逻辑。但需注意：1) test_moe_align_block_size 的排除意味着 AMD CI 中 MoE 相关测试缺失；2) 超时调整到 30 分钟可能掩盖个别测试的执行时间异常。
- 影响：影响范围限于 AMD CI 流程：新增 3 个测试集（总测试用例数：clamp_position 64、resolve_future_token_ids 64、rmsnorm_hf 43），确保这些 kernel 在 MI325 上持续验证。CI 超时放宽后，构建和测试更稳定，减少因超时导致的假阳性失败。
- 风险标记：缺少 MoE 测试覆盖

关联脉络

- PR #27644 [JIT kernel tests] Move into test/registered/jit and register for CUDA: 本 PR 延续了 #27644 的工作，将已在 CUDA 上注册的 JIT kernel 测试扩展到 AMD 平台。