

PR #29197 完整报告

sgl-project/sglang

[AMD] Register 5 JIT kernel unit tests for AMD nightly CI

合并时间: 2026-06-26 04:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29197>

执行摘要

- 一句话: 为 5 个 JIT 单元测试注册 AMD 夜间 CI
- 推荐动作: 建议精读 PR body 中的验证表格和 CI 运行链接, 了解哪些测试被确认兼容及资源消耗。这对于 AMD 平台开发者或有 AMD 部署需求的用户有参考价值。

功能与动机

为了将已在 AMD GPU 上验证通过的 JIT 内核单元测试纳入 AMD 夜间 CI, 确保 AMD 平台的持续正确性。PR body 明确指出这些测试在 ROCm 7.0 和 7.2 上全部通过, 且是 #28967 的下一批。

实现拆解

1. 导入 register_amd_ci: 在 5 个测试文件中, 将导入语句从 `from sglang.test.ci.ci_register import register_cuda_ci` 改为 `from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci`。
2. 添加注册调用: 在每个文件已有的 CUDA 注册之后, 添加一行 `register_amd_ci(est_time=..., suite="nightly-amd-kernel-1-gpu", nightly=True)`, 其中 `est_time` 根据各测试的实际运行时间设定 (15-30 秒不等)。
3. 无需 workflow 修改: 夜间 AMD CI 套件 `nightly-amd-kernel-1-gpu` 已在之前的 PR 中配置于 `nightly-test-amd.yml` 和 `nightly-test-amd-rocm720.yml`, 因此本次变更无需修改 CI 工作流文件。

关键文件:

- `test/registered/jit/deepseek_v4/test_c128_v2.py` (模块 JIT 测试; 类别 test; 类型 test-coverage) : DeepSeek V4 c128 压缩内核测试, 加入 AMD CI 注册, 确保 AMD 平台对此关键内核的回归测试。
- `test/registered/jit/deepseek_v4/test_c4_v2.py` (模块 JIT 测试; 类别 test; 类型 test-coverage) : DeepSeek V4 c4 压缩内核测试, 与 `test_c128_v2` 类似的注册变更。
- `test/registered/jit/diffusion/test_group_norm_silu.py` (模块 JIT 测试; 类别 test; 类型 test-coverage) : GroupNorm+SiLU 融合内核测试, 注册 AMD CI。
- `test/registered/jit/diffusion/test_qwen_image_modulation.py` (模块 JIT 测试; 类别 test; 类型 test-coverage) : Qwen-Image 调制内核测试, 注册 AMD CI。

- test/registered/jit/diffusion/test_varlen_pack_pad.py (模块 JIT 测试; 类别 test; 类型 test-coverage) : 变长打包 / 填充内核测试, 注册 AMD CI。

关键符号: 未识别

关键源码片段

test/registered/jit/deepseek_v4/test_c128_v2.py

DeepSeek V4 c128 压缩内核测试, 加入 AMD CI 注册, 确保 AMD 平台对此关键内核的回归测试。

```
from __future__ import annotations

import sys
from typing import Tuple, Union

import pytest
import torch
import triton

from sglang.jit_kernel.dsv4 import compress_forward
from sglang.jit_kernel.tests.deepseek_v4.common import (
    LegacyContext,
    PagedContext,
    make_legacy_context,
    make_paged_context,
    make_state_pool,
    to_seq_extend,
)
# 同时导入 register_amd_ci 和 register_cuda_ci
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 保留原有 CUDA 注册
register_cuda_ci(est_time=30, suite="base-b-kernel-unit-1-gpu-large")
register_cuda_ci(est_time=30, suite="nightly-kernel-1-gpu", nightly=True)
# 新增 AMD 夜间 CI 注册
register_amd_ci(est_time=30, suite="nightly-amd-kernel-1-gpu", nightly=True)
```

评论区精华

PR 仅有一个来自 HaiShaw 的审核, 状态为 APPROVED, 无 review 评论。无实质讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。变更纯为测试注册, 不涉及任何生产代码逻辑修改。每个文件仅修改 2 行 (导入 + 注册), 且已在两个 ROCm 版本上验证通过。存在极小的风险: 如果 AMD CI 环境的配置与验证环境有差异, 可能导致测试失败, 但 PR body 已通过 dispatched CI 验证

确认通过。

- 影响：
 - 用户：无直接影响。
 - 系统 /CI: AMD 夜间 CI 将新增 5 个 JIT 内核测试，增加约 45-120 秒的运行时间，确保 AMD 平台对这些内核的持续兼容性。
 - 团队：AMD 相关开发者将获得更及时的回归反馈。
 - 风险标记：纯测试基础设施变更

关联脉络

- PR #28967 [AMD] Register 7 JIT kernel unit tests for AMD nightly CI: 本 PR 是 #28967 的延续，同为注册 JIT 内核测试到 AMD 夜间 CI，且 PR body 中明确提及这是继 #28967 后的新一批。