PR #30207 完整报告

sgl-project/sglang

[AMD] Register 2 hardware-agnostic 1-GPU PR tests for AMD CI

合并时间：2026-07-08 05:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30207

执行摘要

- 一句话：AMD CI 新增两个硬件无关测试注册
- 推荐动作：建议精读本 PR 作为 CI 扩覆盖的参考流程：选型策略（硬件无关测试优先）、验证步骤（在两个 AMD 硬件版本上验证）、以及失败处理（明确记录并排除，不阻塞合并）。

功能与动机

缩小 AMD vs NVIDIA per-commit（PR 级）覆盖差距，参考 ROCm CI 仪表盘。两个测试已是 NVIDIA per-commit CI 的一部分，且硬件无关（mock 模型引擎测试 / 纯 Triton 内核），在 ROCm 上无需修改即可通过。

实现拆解

1. 在两个测试文件中添加 register_amd_ci(...) 调用：在 test/registered/lora/test_moe_lora_info.py 和 test/registered/unit/managers/test_customized_info_streaming.py 中，从 sglang.test.ci.ci_register 导入 register_amd_ci，并在已有的 register_cuda_ci(...) 之后添加对应 AMD CI 注册行。
2. 调整 AMD CI stage 配置：将两个测试注册到 stage-b-test-1-gpu-small-amd 阶段，预估时间分别为 5 秒和 120 秒，与 NVIDIA 侧保持一致。
3. 移除验证失败的 FP8 测试：初始提交包含 4 个测试，但 test_triton_moe_channel_fp8_kernel.py 和 test_fp8_kernel.py 在 ROCm 上存在数值精度差异或超阈值问题，最终提交将其排除，仅保留两个已确认通过的测试。

关键文件：

- test/registered/lora/test_moe_lora_info.py（模块 测试注册；类别 test；类型 test-coverage）：添加了 register_amd_ci(...) 调用，将纯 Triton 内核测试注册到 AMD CI。
- test/registered/unit/managers/test_customized_info_streaming.py（模块 测试注册；类别 test；类型 test-coverage）：添加了 register_amd_ci(...) 调用，将 mock 模型引擎端到端测试注册到 AMD CI。

关键符号：未识别

关键源码片段

test/registered/lora/test_moe_lora_info.py

添加了 register_amd_ci(...) 调用，将纯 Triton 内核测试注册到 AMD CI。

```
import sys

import pytest
import torch

from sglang.srt.lora.backend.base_backend import _compute_moe_lora_info
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci # 新增导入 register_amd_
ci

register_cuda_ci(est_time=5, stage="base-b", runner_config="1-gpu-small")
register_amd_ci(est_time=5, stage="stage-b", runner_config="1-gpu-small-amd") # 新增 AMD CI
注册

def _expected_adapter_enabled(
    lora_ranks: torch.Tensor,
    weight_indices: torch.Tensor,
) -> torch.Tensor:
    expected = torch.zeros_like(lora_ranks)
    expected.scatter_(
        0,
        weight_indices.long(),
        (lora_ranks[weight_indices.long()] > 0).to(torch.int32),
    )
    return expected
```

test/registered/unit/managers/test_customized_info_streaming.py

添加了 register_amd_ci(...) 调用，将 mock 模型引擎端到端测试注册到 AMD CI。

```
from __future__ import annotations

import unittest
from typing import TYPE_CHECKING, List

import torch

from sglang.srt.entrypoints.engine import Engine
from sglang.srt.layers.sampler import Sampler, register_sampler_backend
from sglang.srt.managers.scheduler import run_scheduler_process
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci # 新增导入 register_amd_
ci
from sglang.test.mock_model.utils import MOCK_MODEL_PATH
from sglang.test.test_utils import CustomTestCase

if TYPE_CHECKING:
    from sglang.srt.layers.logits_processor import LogitsProcessorOutput
    from sglang.srt.sampling.sampling_batch_info import SamplingBatchInfo
```

```
register_cuda_ci(est_time=120, stage="base-b", runner_config="1-gpu-small")
register_amd_ci(est_time=120, stage="stage-b", runner_config="1-gpu-small-amd") # 新增 AMD CI 注册
```

评论区精华

无 review 评论。PR 作者在 body 中明确说明了两个 FP8 测试被移除的原因：AMD 验证失败，已记录在 ROCm/sglang-ci#311。PR 获得 HaiShaw 批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅在两处测试导入和注册行新增内容，不修改任何业务逻辑、内核代码或 CI 配置模板。被注册的测试已在 NVIDIA CI 上稳定运行，且在 AMD 上预先验证通过。移除 FP8 测试不会影响 NVIDIA 侧覆盖。
- 影响：对用户无感知。对开发团队：缩小了 AMD 与 NVIDIA 之间的 PR 级测试覆盖差距，减少了 AMD 特有问题被疏忽的风险。对 CI 系统：AMD CI 新增约 2 分钟测试时长，但属于合理开销。
- 风险标记：仅测试注册变更，无风险

关联脉络

- PR #30309 [AMD] ci: run multimodal_gen unit suite on AMD: 同样是 AMD CI 扩覆盖的 PR，属于同一改进方向。