

PR #27822 完整报告

sgl-project/sglang

[AMD] ci: add label-gated extra-a tier (kv_canary + mock_model unit tests)

合并时间: 2026-06-13 06:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27822>

执行摘要

- 一句话: 为 AMD CI 新增 label-gated extra-a 测试层
- 推荐动作: 值得关注的设计决策: 采用与 CUDA 模式一致的标签门控策略, 运行时检查标签以支持 rerun 时正确识别; 采用 single-job 而非分区以节省 AMD GPU 资源; 利用 workflow_call 将 extra-a 链入定时调度。这些模式可供后续 AMD 测试层扩展参考。

功能与动机

CUDA **extra-a** opt-in tier 由 PR #26648 引入, 但 AMD 端未有对应注册, 导致这些单元测试在 AMD 平台上缺乏 CI 覆盖。本 PR 弥补这一差距, 为 AMD 提供等效的测试套件, 确保 AMD 平台的质量保障。

实现拆解

1. 新增 workflow 文件 .github/workflows/pr-test-amd-extra.yml, 作为 AMD 的 extra-a 标签门控 CI 工作流, 镜像 CUDA 的 pr-test-extra.yml。
2. 修改 .github/workflows/pr-test-amd.yml 和 pr-test-amd-rocm720.yml, 通过 workflow_call 将 extra-a 工作流链入定时调度。
3. 修改 test/run_suite.py, 注册 AMD 的 extra-a 测试套件 extra-a-test-1-gpu-small-amd。
4. 修改 21 个测试文件 (kv_canary 和 mock_model), 在各文件的模块级加入 register_amd_ci() 调用, 并更新导入。
5. 经过验证, 所有 21 个测试在 mi325 上通过, 总执行时间约 233 秒, 分区后约 60-70 秒 /leg。

关键文件:

- .github/workflows/pr-test-amd-extra.yml (模块 CI 编排; 类别 infra; 类型 infrastructure): 新增的 AMD extra-a CI 工作流核心, 实现标签门控和套件调度。
- .github/workflows/pr-test-amd.yml (模块 CI 编排; 类别 infra; 类型 infrastructure): 将 extra-a 工作流链入默认 AMD 定时调度, 保证 main 分支覆盖。
- test/run_suite.py (模块 测试套件; 类别 test; 类型 test-coverage): 注册 AMD extra-a 测试套件标识, 使测试框架能够识别并调度该套件。
- test/registered/mock_model/test_self_unit_canary_mock_wiring.py (模块 单元测试; 类别 test; 类型 test-coverage): 作为测试文件添加 AMD CI 注册的示例, 展示

register_amd_ci 的使用。

关键符号：未识别

关键源码片段

[test/registered/mock_model/test_self_unit_canary_mock_wiring.py](#)

作为测试文件添加 AMD CI 注册的示例，展示 register_amd_ci 的使用。

```
from __future__ import annotations

import dataclasses
import unittest

import torch

from sglang.srt.kv_canary.expected_inputs import ExpectedInputs
from sglang.srt.kv_canary.token_oracle.oracle import HashOracle
from sglang.srt.kv_canary.token_oracle.sampler import install_oracle_sampler
from sglang.srt.model_executor.forward_batch_info import (
    ForwardMode,
    _stable_hash_str_to_i64,
)
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.mock_model.utils import mock_model_server_args, mock_model_server_env
from sglang.test.test_utils import CustomTestCase

# 模块级别注册 CI 触发：CUDA 在主测试阶段 extra-a，AMD 在对应套件
register_cuda_ci(est_time=60, stage="extra-a", runner_config="1-gpu-small")
register_amd_ci(est_time=60, suite="extra-a-test-1-gpu-small-amd")

# 测试辅助类（保持不变）
@dataclasses.dataclass
class _StubForwardBatch:
    input_ids: torch.Tensor
    # ... 其他字段
```

评论区精华

该 PR 由 HaiShaw 直接批准，无额外 comment 讨论。PR body 中详细说明了设计决策（如运行时标签检查、单 job 模式）和实施验证结果。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于 AMD 硬件上 KV-canary JIT kernel 未移植，导致 e2e 测试无法运行（已在 PR 中明确排除）。当前注册的单元测试通过 mock 模型避免了对 JIT kernel 的依赖，风险较低。另外，若 AMD CI 环境变化导致测试不稳定，可能会引起不必要的失败通知，

但标签门控限制了影响范围。

- 影响：对用户无直接影响。对系统：增加了 21 个测试在 AMD CI 的覆盖，提高 AMD 平台的质量保证。对 CI 基础设施：增加了新的工作流文件和定时任务，但通过标签门控和 `continue_on_error` 确保不影响基础 CI 门禁。
- 风险标记：AMD 硬件兼容，测试覆盖有限

关联脉络

- PR #26648 [CI] Split PP tests into base and extra suites: 引入了 extra-a tier 的概念, 本 PR 为其添加 AMD 镜像