

PR #27722 完整报告

sgl-project/sglang

[AMD] Migrate 2-GPU kernel allreduce tests into the registered system

合并时间: 2026-06-10 10:39

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27722>

执行摘要

- 一句话: AMD 2-GPU allreduce 测试迁移至注册系统
- 推荐动作: 低风险测试基础设施变更, 推荐快速合并。值得关注的点: 迁移模式可作为后续更多测试迁移的参考。

功能与动机

使 AMD 2-GPU kernel 测试变为声明式注册, 便于覆盖工具识别和 run_suite/LPT 自动分区, 同时为后续迁移更多测试铺垫基础。PR body 明确指出这是“安全的第一个切片”, 不会影响 NV 或 ROCm-7.2 行为。

实现拆解

1. 注册 2-GPU allreduce 测试: 将 sgl-kernel/tests/test_amd_nccl_allreduce_determinism.py 和 sgl-kernel/tests/test_amd_deterministic_custom_allreduce.py 移动至 test/registered/jit/, 并在每个文件顶部添加 register_amd_ci(est_time=120, suite="sgl-kernel-unit-test-2-gpu-amd")。
2. 更新工作流: 在 pr-test-amd.yml 和 pr-test-amd-rocm720.yml 中, 将 2-GPU kernel 任务的硬编码 pytest 命令替换为 bash scripts/ci/amd/amd_ci_exec.sh -w "/sglang-checkout/test" python3 run_suite.py --hw amd --suite sgl-kernel-unit-test-2-gpu-amd。
3. 新增 suite 注册: 在 test/run_suite.py 的 AMD 有效套件列表中添加 "sgl-kernel-unit-test-2-gpu-amd"。
4. 注册 1-GPU JIT 测试: 在 test/registered/jit/test_activation.py 和 test/registered/jit/test_per_token_group_quant_8bit.py 中添加 register_amd_ci(est_time=20/16, suite="jit-kernel-unit-test-amd"), 与现有 CUDA 注册并存。

关键文件:

- test/registered/jit/test_amd_nccl_allreduce_determinism.py (模块 AMD CI; 类别 test; 类型 rename-or-move) : 从 sgl-kernel/tests 搬迁并注册到 AMD 2-GPU 套件
- test/registered/jit/test_amd_deterministic_custom_allreduce.py (模块 AMD CI; 类别 test; 类型 rename-or-move) : 从 sgl-kernel/tests 搬迁并注册到 AMD 2-GPU 套件

- `.github/workflows/pr-test-amd.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : 将硬编码的 `pytest` 命令替换为 `run_suite.py` 调用
- `test/registered/jit/test_activation.py` (模块 JIT 内核; 类别 test; 类型 test-coverage) : 新增 AMD CI 注册, 与 CUDA 注册并存
- `test/registered/jit/test_per_token_group_quant_8bit.py` (模块 JIT 内核; 类别 test; 类型 test-coverage) : 新增 AMD CI 注册, 与 CUDA 注册并存
- `test/run_suite.py` (模块 测试框架; 类别 test; 类型 test-coverage) : 添加新套件到 AMD 有效套件列表
- `.github/workflows/pr-test-amd-rocm720.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : 同步 `pr-test-amd.yml` 的变更到 ROCm7.2 工作流

关键符号: 未识别

关键源码片段

`test/registered/jit/test_amd_nccl_allreduce_determinism.py`

从 `sgl-kernel/tests` 搬迁并注册到 AMD 2-GPU 套件

```
"""
Test to confirm non-determinism of default NCCL all-reduce with batch size invariance.
...
"""

import multiprocessing as mp
import socket

import pytest
import torch
import torch.distributed as dist

from sglang.test.ci.ci_register import register_amd_ci

# 注册到 AMD 2-GPU 套件, 预估耗时 120s
register_amd_ci(est_time=120, suite="sgl-kernel-unit-test-2-gpu-amd")

def get_open_port():
    with socket.socket(socket.AF_INET, socket.SOCK_STREAM) as s:
        s.bind(("127.0.0.1", 0))
        return s.getsockname()[1]

def worker(world_size, rank, port):
    # ... 测试逻辑保持不变 ...
```

评论区精华

该 PR 没有公开的 review 评论。HaiShaw 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅涉及测试注册和 CI 工作流，不修改任何生产代码。已通过真实 MI325 2-GPU 环境验证，两个 allreduce 测试均通过。但工作流 dispatch 到 main 分支会因文件路径不同而失败（PR body 已说明，可接受）。
- 影响：影响范围仅限于 AMD CI： 1) 2-GPU allreduce 测试从硬编码变为注册，对开发者无感知； 2) JIT kernel 测试新增 AMD CI 覆盖，扩大测试面； 3) 不影响 CUDA/ROCm-7.2 或其他后端。
- 风险标记：低风险基础设施变更

关联脉络

- PR #27644 Migrate JIT kernel tests into registered system: 本 PR 遵循 #27644 建立的模式，将更多 AMD 测试迁移到注册系统。