PR #27358 完整报告

sgl-project/sglang

HiCache: Fix Flaky CI For 3FS Backend

合并时间：2026-06-05 22:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27358

执行摘要

- 一句话：修复 HiCache 3FS 测试的 CI 配置
- 推荐动作：建议回退该 PR 的变更，或同步更新 CI 套件配置（如 .github/workflows/pr-test.yml 等），确保 base-b-test-4-gpu-h100 套件存在。同时，应审查测试是否确实需要 4 GPU 资源，或者是否可以通过其他方式（如增加超时、重试机制）来解决不稳定问题。

功能与动机

该 PR 旨在修复 HiCache 3FS 后端测试的不稳定 CI 问题。PR body 提到 'Fix Flaky CI For 3FS Backend'，目的是通过调整测试运行的 GPU 配置来解决测试失败。

实现拆解

1. 修改测试注册配置：在文件 test/registered/hicache/test_hicache_storage_3fs_backend.py 中，将 register_cuda_ci 调用的 runner_config 参数从 2-gpu-large 改为 4-gpu-h100，其他参数保持不变。
2. 保留 AMD 配置：AMD CI 的注册行 register_amd_ci(est_time=300, suite="base-b-test-2-gpu-large") 未作更改。
3. 提交说明：单个 commit，提交信息为 "Change machine"。

关键文件：

- test/registered/hicache/test_hicache_storage_3fs_backend.py（模块 测试配置；类别 test；类型 test-coverage）：唯一变更文件，将 CUDA CI 注册的 runner_config 从 '2-gpu-large' 改为 '4-gpu-h100'，以修复测试不稳定问题。

关键符号：未识别

关键源码片段

test/registered/hicache/test_hicache_storage_3fs_backend.py

唯一变更文件，将 CUDA CI 注册的 runner_config 从 '2-gpu-large' 改为 '4-gpu-h100'，以修复测试不稳定问题。

```
"""
# 文件: test/registered/hicache/test_hicache_storage_3fs_backend.py
```

```
# 变更说明：将 CUDA CI 注册从 2-gpu-large 改为 4-gpu-h100，
# 期望提升测试稳定性。注意：同时需要确保 CI 套件配置中存在
# base-b-test-4-gpu-h100 套件，否则会导致 CI 套件验证失败。
"""

import json
import os
import unittest

from test_hicache_storage_file_backend import HiCacheStorageBaseMixin

from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.test_utils import CustomTestCase

# 变更行：runner_config 从 "2-gpu-large" 改为 "4-gpu-h100"
register_cuda_ci(est_time=150, stage="base-b", runner_config="4-gpu-h100")
register_amd_ci(est_time=300, suite="base-b-test-2-gpu-large")


class HiCacheStorage3FSBackendBaseMixin(HiCacheStorageBaseMixin):
    """Base mixin class with common setup and utilities"""

    @classmethod
    def _get_additional_server_args_and_env(cls):
        # ... 剩余代码保持不变
```

评论区精华

该 PR 的讨论主要围绕 CI 验证。作者和合并者分别触发了 /rerun-test 命令，重新运行了 test_hicache_storage_3fs_backend.py 测试，两次都通过了。然而，合并后其他 PR 的 CI 运行出现了套件验证错误，提示该测试文件被注册到无效套件 base-b-test-4-gpu-h100，原因是 register_cuda_ci(est_time=150, stage="base-b", runner_config="4-gpu-h100") 中的 stage 和 runner_config 组合未在 CI 套件配置中定义。其他贡献者随后报告了该问题。

- CI 套件验证失败 (bug): 该问题由 PR 变更引起，需修复 CI 配置或回退变更。

风险与影响

- 风险：主要风险是 CI 配置错误：将 CUDA 测试的 runner_config 改为 4-gpu-h100 但未更新相应的 CI 套件配置，导致 CI 在套件验证阶段失败。该问题已在实际 CI 运行中被确认。修复后需确保 base-b-test-4-gpu-h100 套件已正确配置。
- 影响：直接影响是 HiCache 3FS 测试的 CUDA CI 运行环境从 2 GPU 变为 4 GPU H100，预期能提升测试稳定性。负面影响是当前配置导致了 CI 套件验证失败，影响了所有 PR 的 CI 运行。需紧急修复套件配置或回退变更。
- 风险标记：CI 配置错误 ，影响其他 PR CI，缺少配套配置更新

关联脉络

- PR #27366 [BugFix]: Fix HiMamba HiCache prefetch hang after L3 sidecar transfer: 相关 HiCache 修复 PR，体现了 HiCache 组件近期有多个 bugfix，本 PR 也是同一主题。