PR #23447 完整报告

sgl-project/sglang

[CI] Move disaggregation basic CI back to 2-gpu suite

合并时间：2026-04-22 17:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23447

执行摘要

- 一句话：将解聚基本 CI 测试从 4 卡迁移回 2 卡套件
- 推荐动作：该 PR 为纯粹的 CI 配置调整，无核心逻辑变更，不值得精读。但可作为 CI 资源优化的参考案例。

功能与动机

PR body 明确指出："TCP fallback is compatible with CUDA 13 now. Move disaggregation basic CI back to 2-gpu suite"，目的是在 TCP fallback 兼容后，将测试回归到更经济的 2 卡环境，节省 CI 资源。

实现拆解

1. 修改 CI 注册套件：在 test/registered/disaggregation/test_disaggregation_basic.py 中，将 register_cuda_ci 的 suite 参数从 "stage-c-test-4-gpu-h100" 改为 "stage-b-test-2-gpu-large"，使该测试在 2-GPU CI 套件中运行。
2. 启用 TCP 连接池环境变量：在 python/sglang/test/server_fixtures/disaggregation_fixture.py 的 PDDisaggregationServerBase.setUpClass() 中新增 os.environ["MC_TCP_ENABLE_CONNECTION_POOL"] = "true"，确保 TCP fallback 模式下使用连接池。
3. 清理环境变量：在 tearDownClass() 中新增 os.environ.pop("MC_TCP_ENABLE_CONNECTION_POOL")，避免影响后续测试。

整体变更仅涉及两处测试配置和夹具的微小调整，无核心逻辑改动。

关键文件：

- test/registered/disaggregation/test_disaggregation_basic.py（模块 测试套件；类别 test；类型 configuration）：将测试套件从 4-GPU 迁移到 2-GPU，是本次 CI 调整的核心变更。
- python/sglang/test/server_fixtures/disaggregation_fixture.py（模块 测试夹具；类别 test；类型 test-coverage）：新增 TCP 连接池环境变量设置，确保 TCP fallback 模式下正确运行。

关键符号：未识别

关键源码片段

python/sglang/test/server_fixtures/disaggregation_fixture.py

新增 TCP 连接池环境变量设置，确保 TCP fallback 模式下正确运行。

```
class PDDisaggregationServerBase(CustomTestCase):
    @classmethod
    def setUpClass(cls):
        # 启用 TCP 连接池以支持 Mooncake TCP fallback 场景
        os.environ["MC_TCP_ENABLE_CONNECTION_POOL"] = "true"
        # ... 其余初始化代码

    @classmethod
    def tearDownClass(cls):
        # 清理环境变量，避免影响后续测试
        os.environ.pop("MC_TCP_ENABLE_CONNECTION_POOL")
        # ... 其余清理代码
```

评论区精华

无人工审核评论，仅自动化机器人评论确认变更内容。PR 作者通过多次 /rerun-test 命令手动触发 CI 验证，测试均在 2-GPU H100 上通过，并在最后触发 8-GPU H20 CI 套件验证无回归。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅涉及测试套件归属和环境变量设置，不修改任何生产代码。主要风险是 TCP 连接池环境变量是否在所有 2 卡环境中正确生效，但已在 CI 中成功验证。
- 影响：用户影响：无。系统影响：解聚基础 CI 测试从 4 卡迁移至 2 卡，释放 4 卡 CI 资源，降低 CI 成本。团队影响：CI 配置更合理，减少不必要的资源占用。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR