PR #7363 完整报告

PaddlePaddle/FastDeploy

[CI] Modify 4-card container startup config and move test case

合并时间：2026-04-13 20:23

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7363

执行摘要

本次 PR 优化了 FastDeploy 仓库中 4 卡 CI 作业的容器启动配置，通过添加 RDMA 设备支持、内存锁定权限和环境变量，提升了在 RDMA 环境下的测试稳定性和兼容性，同时重组了 e2e 测试用例以改善代码组织。变更属于基础设施优化，对用户无直接影响，但有助于团队更可靠地执行多卡推理测试。

功能与动机

PR 的动机明确：提升 4 卡 CI 作业在 RDMA 环境下的稳定性和兼容性，并确保正确的设备挂载和资源限制。作者在 PR body 中写道："Improve stability and compatibility of 4-card CI jobs with RDMA-enabled environments" 和 "Ensure proper device mounting and resource limits for multi-card inference tests"。这反映了团队在分布式推理测试中遇到的环境配置问题，需要通过 CI 优化来保证测试的可靠性。

实现拆解

实现分为两个关键部分：

1. CI 工作流配置更新（.github/workflows/_gpu_4cards_case_test.yml）：
   - 添加 RDMA 设备自动检测与挂载：export RDMA_DEVICES=$(find /dev/infiniband/uverbs* -maxdepth 1 -not -type d | xargs -I{} echo '--device {}:{}')
   - 挂载 RDMA 通信设备：--device=/dev/infiniband/rdma_cm
   - 设置内存锁定无限制：--ulimit memlock=-1:-1
   - 添加必要权限：--cap-add=SYS_PTRACE --cap-add=IPC_LOCK
   - 调整共享内存大小：--shm-size=64G
   - 新增环境变量：FD_ROUTER_PORT、FD_CONNECTOR_PORT、FD_RDMA_PORT、CLEAN_CUDA=1
2. 测试用例重组（tests/e2e/4cards_cases/test_ernie_03b_pd_router_v1_rdma_tp2.py）：
   - 将文件从 tests/e2e/ 移动到 tests/e2e/4cards_cases/ 目录
   - 修正导入路径：从 from utils.serving_utils import 改为 from e2e.utils.serving_utils import
   - 调整 GPU 设备分配：CUDA_VISIBLE_DEVICES 从 "1" 改为 "2"

评论区精华

review 讨论较少，主要亮点来自 fastdeploy-bot 的 AI 代码审查：

> 虽然不在本次 PR 变更范围内，但注意到 .github/workflows/_gpu_4cards_case_test.yml:158 的 PORTS 数组未包含本次新增的 FD_ROUTER_PORT、FD_CONNECTOR_PORT、FD_RDMA_PORT，可能需要后续补充端口清理逻辑。

这指出了潜在的正确性问题，但被标记为非阻塞性建议。yuanlehome 直接批准了 PR，表明变更被认可。

风险与影响

风险分析：

- 配置变更可能引入不稳定性，但针对 RDMA 环境优化，风险可控。
- 测试文件移动可能导致导入错误，但 PR 中已修正路径。
- 新增权限（如 IPC_LOCK）在 CI 环境中安全风险有限。
- 端口清理逻辑缺失可能引发端口冲突，需后续关注。

影响分析：

- 对用户无直接影响，属于内部 CI 改进。
- 提升 4 卡 CI 作业在 RDMA 环境下的可靠性，支持更稳定的多卡推理测试。
- 测试重组使代码更清晰，便于团队维护。

关联脉络

从近期历史 PR 看，本 PR 与多个 CI 优化 PR 相关：

- PR #7335 和 #7315 都涉及容器清理和资源管理，与本 PR 的基础设施优化目标一致。
- 整体趋势显示团队持续改进 CI 流程，特别是在多卡和 RDMA 环境下的测试稳定性。

本 PR 是这一系列优化的一部分，专注于 4 卡 RDMA 配置，反映了 FastDeploy 在分布式推理测试基础设施上的演进。