PR #22150 完整报告

sgl-project/sglang

Fix flaky test_load_weights_from_remote_instance in CI

合并时间：2026-04-05 22:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22150

执行摘要

本 PR 通过将 remote_instance_weight_loader_start_seed_via_transfer_engine 参数硬编码为 False，修复了 CI 中 test_load_weights_from_remote_instance 测试因 transfer_engine 后端导致的挂起问题，提升了测试稳定性和 CI 可靠性。变更简单直接，但揭示了后端重构后的配置不一致性。

功能与动机

CI 中分布式权重加载测试间歇性失败，原因是测试随机选择 transfer_engine 后端，而该后端在 Engine 模式下存在 bug，导致 rank 1 初始化时挂起。根据开发者 ShangmingCai 的评论："The root cause is that the engine server doesn't need to start the bootstrap server anymore after refactoring"，因此需要调整参数设置以避免不必要的服务器启动。

实现拆解

仅修改了 test/registered/distributed/test_load_weights_from_remote_instance.py 文件：

- 在 init_process_dst 函数中，将 remote_instance_weight_loader_start_seed_via_transfer_engine 从基于后端的条件判断改为固定 False：python remote_instance_weight_loader_start_seed_via_transfer_engine=False,
- 移除了原有的条件逻辑，简化了代码。
- 在测试函数中添加了 TODO 注释，计划未来重构以减少随机行为。

评论区精华

- debug 代码移除：gemini-code-assist[bot] 指出在 model_runner.py 中添加的 time.sleep(30) 必须移除，以免影响性能。最终提交中该代码已被 revert。
- 测试覆盖率：reviewer 建议恢复随机选择以保持覆盖，但 PR 中暂时 hardcode 参数，并添加 TODO 注释："refactor this test to have less random behavior"。

风险与影响

- 风险：硬编码参数可能降低测试覆盖率，掩盖 transfer_engine 后端的其他问题；如果 debug 代码未彻底移除，可能导致性能下降。
- 影响：对用户无直接影响，但修复了 CI 不稳定性，提升团队开发效率；测试资源浪费减少，CI 成功率提高。

关联脉络

本 PR 与近期多个 CI 测试修复 PR 相关，如 PR #22140（修复夜间测试）和 PR #22137（移除不稳定测试），共同反映了团队对提升 CI 稳定性的持续努力。这些变更凸显了在重构后保持配置一致性和测试可靠性的重要性。