PR #22081 完整报告

sgl-project/sglang

[CI] Support CPU stage and auto-batch same-stage files in `/rerun-test`

合并时间：2026-04-04 06:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22081

执行摘要

本 PR 扩展了 CI 命令 /rerun-test，支持 CPU-only 测试并引入自动批处理机制，通过分组相同 runner 配置的文件来减少工作流运行次数，提升测试效率和资源利用率。

功能与动机

PR 的主要动机是增强 /rerun-test 命令的灵活性，支持使用 register_cpu_ci() 注册的 CPU-only 测试文件，并优化多个测试文件的分派逻辑。PR body 中说明："/rerun-test now supports CPU-only tests (files with register_cpu_ci()) by dispatching to ubuntu-latest runner" 和 "When multiple test files are specified, files targeting the same (runner_label, use_deepep, is_cpu) are batched into a single workflow run instead of one run per file"。

实现拆解

关键改动集中在两个文件：

1. scripts/ci/utils/slash_command_handler.py：
   - 将 detect_cuda_suite() 重命名为 detect_suite()，使其支持检测 register_cuda_ci 和 register_cpu_ci，返回 is_cpu 标志。
   - 拆分 _resolve_and_dispatch_ut() 为 _resolve_test_spec()（解析测试 spec）和 _dispatch_batch()（分派批处理组）。
   - 在 handle_rerun_test() 中，解析所有测试 spec 后按 (runner_label, use_deepep, is_cpu) 分组，每个组触发一个工作流运行。
2. .github/workflows/rerun-test.yml：
   - 添加 is_cpu 输入选项和 ubuntu-latest runner 选择。
   - 新增 rerun-test-cpu job（当 is_cpu == 'true'），模仿现有 CI 的 CPU 阶段设置（释放磁盘空间、安装 Python、使用 uv pip 安装依赖）。
   - 修改 rerun-test-cuda job 支持多行 test_command，通过 shell 循环逐行执行命令。

评论区精华

无正式的 review 讨论，但 issue 评论中展示了功能的测试验证：

- 作者执行 /rerun-test test_runai_utils.py，成功触发 CPU 测试并使用 ubuntu-latest runner。
- 执行多个文件如 /rerun-test test_srt_endpoint.py test_openai_server.py，正确批处理相同 GPU runner 的测试。
- 这验证了 CPU 测试支持和批处理逻辑的有效性。

风险与影响

- 技术风险：批处理中单个测试失败可能导致整个工作流停止，增加调试难度；CPU job 的依赖安装可能不稳定；detect_suite() 的正则匹配可能误判非标准测试文件。
- 影响分析：对开发者而言，提供了更便捷的测试重跑功能；对 CI 系统，减少了工作流数量，节约资源；但对团队需确保批处理稳定性，避免因失败传播影响测试反馈。

关联脉络

从近期历史 PR 看，本 PR 是 CI 基础设施持续优化的一部分：

- PR 22045 调整 CI 超时参数，解决测试配置问题。
- PR 22036 添加内核发布提示，完善工作流文档。
- PR 22018 修复构建失败，提升 CI 可靠性。 这些 PR 共同体现了团队对测试流程和资源效率的关注，本 PR 的批处理设计可能为未来更大规模的测试优化奠定基础。