PR #22565 完整报告

sgl-project/sglang

chore: update CI test est_time values

合并时间：2026-04-11 09:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22565

执行摘要

本 PR 由 sglang-bot 自动化生成，更新了 250 个 CI 测试文件的 est_time 估计时间值，基于最近 10 次成功执行的中位数，旨在优化 LPT 负载均衡算法，提升并行 CI 作业的测试分配效率。变更仅涉及数值参数调整，无代码逻辑修改，已由 ch-wan 直接合并。

功能与动机

更新动机源于保持 CI 测试负载均衡准确性的需求。根据 PR body 描述："This keeps the LPT load-balancing algorithm accurate for partitioning tests across parallel CI jobs." 通过基于 scheduled PR Test runs on main 的最近 10 次成功执行中位数更新 est_time 值，确保测试在多个 CI 作业间合理分配，避免资源浪费或超时。

实现拆解

实现方案聚焦于修改测试文件中的 register_cuda_ci 和 register_cpu_ci 调用参数。关键改动点按模块梳理：

- 4-GPU 模型测试：例如 test/registered/4-gpu-models/test_gpt_oss_4gpu.py，est_time 从 300 调整为 328（H100）和 312（B200）。
- 8-GPU 模型测试：例如 test/registered/8-gpu-models/test_deepseek_v3_basic.py，est_time 从 275 调整为 320。
- 注意力内核测试：例如 test/registered/attention/test_fa3.py，est_time 从 390 调整为 386。
- 其他模块：涉及量化、LoRA、推测解码等多个标签对应测试，数值更新幅度从 -50% 到 +50% 不等（如 test/registered/moe/test_cutedsl_moe.py 从 300 改为 13）。所有变更均通过脚本自动化执行，共修改 250 个文件。

评论区精华

无 review 讨论或争议点，表明变更被视为常规维护。提交历史中有一个修复脚本的提交（64dd689），修正了后端匹配问题，但未在 PR 讨论中展开。

风险与影响

- 技术风险：主要风险是更新后的 *est_time* 值可能不准确，若历史数据中位数无法反映最新测试性能，会导致 CI 调度不均衡。但由于基于统计方法，风险可控。变更不涉及代码逻辑，因此无回归或安全风险。
- 影响分析：对用户透明，无直接功能影响。系统层面，CI 测试调度更准确，可能提升团队开发效率，减少测试超时。影响范围限于 CI 基础设施，程度中等。

关联脉络

本 PR 是 CI 基础设施自动化更新流程的一部分，与近期多个 PR 关联：

- PR #22563和 #22557：修复 *est_time* 更新脚本的后端和套件匹配逻辑，为本 PR 的自动化执行奠定基础。
- PR #22545：添加每周工作流来自动化更新 *est_time* 值，本 PR 是该工作流的直接产出。

这些 PR 共同揭示了仓库在 CI 测试负载均衡和自动化维护方面的持续演进，旨在提升测试可靠性和资源利用率。