PR #22305 完整报告

sgl-project/sglang

[CI] Update est_time for 64 tests based on actual elapsed times

合并时间：2026-04-10 11:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22305

执行摘要

- 一句话：更新 64 个 CI 测试的估计时间，基于实际耗时优化分区平衡。
- 推荐动作：对于一般工程师，此 PR 无需精读，除非关注 CI 优化方法。值得注意的决策是使用严格标准（≥2 数据点且≥50% 差异 >60s）来确保更新可靠性，可借鉴于类似估计调整场景。

功能与动机

从 PR body 引用：'Drifted estimates cause poor CI partition balancing and misleading timeout alerts.' 即估计漂移导致 CI 分区不平衡和误导性超时警报。

实现拆解

实现方案是机械更新：遍历 test/registered/ 目录下的 64 个测试文件，修改每个文件中 register_cuda_ci 或 register_amd_ci 调用的 est_time 参数。更改仅基于实际耗时数据，不涉及测试逻辑或代码功能。关键步骤包括数据收集（从 2463 个计时记录）、应用更新标准（≥2 数据点且≥50% 差异 >60 秒），以及四舍五入到 10 秒。

关键文件：

- test/registered/spec/eagle/test_deepseek_v3_fp4_mtp_small.py（模块 speculative-decoding）：估计时间从 900 秒大幅减少到 240 秒，显示推测解码测试性能优化显著，影响 CI 分区。
- test/registered/quant/test_w4a8_deepseek_v3.py（模块 quant）：估计时间从 520 秒增加到 700 秒，表明量化测试变慢，需关注潜在性能回归。
- test/registered/perf/test_bench_serving_1gpu_part1.py（模块 performance）：估计时间从 1000 秒增加到 1140 秒，影响 CI 大测试套件分区，反映实际耗时增长。
- test/registered/vlm/test_vlm_input_format.py（模块 multimodal）：估计时间从 447 秒增加到 620 秒，多模态测试耗时增加，可能涉及模型加载或处理逻辑变化。

关键符号：register_cuda_ci, register_amd_ci

评论区精华

review 评论为空，无讨论或争议点。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：变更仅影响 CI 估计时间，可能调整测试分区，但基于实际数据减少了不确定性。潜在风险包括数据采样不足（如 <2 个点）导致更新不准确，但标准已规避；或极端耗时波动未被捕捉，影响 CI 调度。
- 影响：改善 CI 系统的资源调度效率，减少分区不平衡和虚假超时警报，提升团队开发体验和 CI 可靠性。对用户和系统无直接影响，属于后台优化。
- 风险标记：估计数据采样不足，CI 调度依赖变更

关联脉络

- PR #22483 [CI] Remove Slack notification from ci-auto-bisect workflow: 同为 CI 工作流优化，涉及 run-ci 标签，展示团队持续改进 CI 基础设施。
- PR #22478 [Docker] Fix CI docker target after Dockerfile restructure: 修复 CI 相关基础设施错误，与本 PR 共同维护 CI 可靠性。
- PR #22160 [Docker] Optimize Dockerfile for BuildKit layer caching: 优化 CI 构建过程性能，体现跨 PR 的 CI 效率提升趋势。