PR #23120 完整报告

sgl-project/sglang

ci: run weekly est_time update on Monday using p90 of last 15 runs

合并时间：2026-04-20 05:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23120

执行摘要

- 一句话：将每周 CI 测试 est_time 更新调整到周一执行，并改用 p90 百分位数和更大样本窗口优化负载均衡。
- 推荐动作：此 PR 值得负责 CI 基础设施的工程师精读，特别是关注 scripts/ci/update_est_time.py 中的统计学方法和阈值设计，这些决策直接影响测试调度的准确性和效率。对于其他开发者，了解此变更有助于理解 CI 测试时间估计的更新机制。

功能与动机

PR body 中说明了三个优化动机：

1) 调度时间从周六 00:00 UTC 移至周一 00:00 UTC，使得刷新 PR 在工作周开始时落地，便于在工作时间审查而非闲置周末； 2) 估计方法从 median-of-last-10 改为 p90-of-last-15，因为中位数低估了偶尔峰值的测试，导致 LPT 下延迟任务主导总时长，p90 偏向于慢速但合理的尾部，且更大的窗口减少了噪声； 3) 自动生成的 PR body 现在包含显著 est_time 变更表，让审查者无需 diff 文件即可查看差异。

实现拆解

1. 调整调度时间：在 .github/workflows/weekly-update-est-time.yml 中，将 cron 表达式从 '0 0 * * 6' 改为 '0 0 * * 1'，从周六移到周一 UTC 零点。
2. 修改时间估计逻辑：在 scripts/ci/update_est_time.py 中，将 TARGET_DATA_POINTS 从 10 增加到 15，MAX_RUNS 从 20 增加到 25；将 compute_medians 函数重命名为 compute_p90，使用 statistics.quantiles 计算 90% 百分位数。
3. 添加变更总结功能：在脚本中添加新的辅助函数 is_significant 和 write_summary，基于绝对和相对阈值（$|\Delta| \geq 30$s 且 $|\Delta|/\text{old} \geq 30\%$）识别显著变更，并生成 Markdown 表格。
4. 更新工作流以生成总结文件：在工作流文件中，修改运行命令以传递 --summary-file /tmp/est_time_summary.md 参数，并在创建 PR 时使用 --body-file 包含总结内容。
5. 配套调整：脚本中的 update_est_times 函数现在返回变更列表，用于生成总结；工作流中 PR body 的生成逻辑改为从文件中读取。

关键文件：

- scripts/ci/update_est_time.py（模块 CI 脚本；类别 infra；类型 infrastructure；符号 compute_medians, compute_p90, update_est_times, is_significant）：这是核心逻辑文

件，负责解析 CI 运行日志、计算时间估计并更新测试注册文件中的 est_time 值，变更直接影响负载均衡算法的基础数据。

- .github/workflows/weekly-update-est-time.yml（模块 工作流配置；类别 infra；类型 infrastructure）：这是 GitHub Actions 工作流配置文件，定义了每周自动更新 est_time 的触发条件和执行步骤，变更涉及调度时间和 PR 生成逻辑。

关键符号：compute_p90, update_est_times, is_significant, write_summary

关键源码片段

scripts/ci/update_est_time.py

这是核心逻辑文件，负责解析 CI 运行日志、计算时间估计并更新测试注册文件中的 est_time 值，变更直接影响负载均衡算法的基础数据。

```
def compute_p90(timings):
    """Compute 90th percentile of last TARGET_DATA_POINTS timings for each entry.

    Returns dict mapping (rel_path, suite, backend) -> p90 (int).
    Only includes entries with >= MIN_DATA_POINTS data points.
    """
    p90s = {}
    for key, values in timings.items():
        recent = values[:TARGET_DATA_POINTS] # 取最近 TARGET_DATA_POINTS 个数据点
        if len(recent) < MIN_DATA_POINTS:
            continue # 数据点不足，跳过此项
        # 使用 statistics.quantiles 计算 90% 百分位数（索引 8 对应 n=10）
        p90s[key] = round(statistics.quantiles(recent, n=10, method="inclusive")[8])
    return p90s
```

评论区精华

此 PR 没有 review 评论，表明变更较为直接或已通过内部讨论。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于估计方法的改变可能影响 CI 测试的调度效率。从 median 改为 p90 可能导致 est_time 值偏高，增加 LPT 负载均衡中分配给测试的预算，从而可能减少并行度或增加总运行时间。此外，更大的样本窗口（15 runs）依赖于足够的成功运行数据，如果 CI 运行不稳定，可能导致估计不准确。脚本中的显著性阈值（30s 和 30%）是硬编码的，可能不适用于所有测试场景。但总体风险较低，因为这只是自动化工具，不直接影响生产代码。
- 影响：对用户无直接影响，因为这是内部 CI 流程。对系统影响：优化了 CI 测试的负载均衡，减少了因低估测试时间导致的延迟，从而可能提高 CI 运行效率和稳定性。对团队影响：周一生成 PR 便于及时审查，自动总结表格减少了审查工作量，提升了开发体验。影响范围限于 CI 基础设施，程度为中等。
- 风险标记：估计方法变更可能影响调度，依赖足够 CI 运行数据，硬编码阈值

关联脉络

- PR #23119 [CI] Add per-job uv venv isolation and upgrade CI version to Cuda 13: 同为 CI 基础设施优化，展示了团队在提升 CI 环境一致性和依赖管理方面的持续努力。
- PR #23108 Update CI_PERMISSIONS: 涉及 CI 权限配置，反映团队在 CI 自动化流程上的维护和调整。