PR #22557 完整报告

sgl-project/sglang

fix: track est_time per suite instead of per backend

合并时间：2026-04-11 07:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22557

执行摘要

本 PR 修复了 CI 测试时间估算脚本中的关键缺陷：原脚本按后端硬件（如 H100/B200）而非测试套件统计时间，导致同一测试文件在不同硬件上获得相同错误时间值，影响 CI 负载均衡。通过新增套件名提取函数并将统计键从 (filename, backend) 改为 (filename, suite)，确保每个测试套件获得独立准确的时间估算。变更仅涉及基础设施脚本，风险低但对 CI 效率有重要改进。

功能与动机

PR #22545 引入的每周 update_est_time.py 脚本旨在从 CI 日志中自动更新测试文件的 est_time 值以优化负载均衡，但其按 (filename, backend) 键统计时间，忽略了同一文件可能在不同测试套件（对应不同硬件，如 H100 与 B200）上运行的事实。如 PR body 所述，这导致像 test_gpt_oss_4gpu.py 这样的文件在 stage-c-test-4-gpu-h100 和 stage-c-test-4-gpu-b200 套件上错误地获得相同 est_time=304，而实际执行时间因硬件差异应不同。本 PR 旨在纠正这一统计逻辑，确保时间估算准确反映各套件性能。

实现拆解

修改集中于 scripts/ci/update_est_time.py 文件：

1. 新增套件名提取函数：job_name_to_suite() 使用正则表达式 re.sub(r"\s*\\(\\d+\\)$", "", job_name) 从 CI 任务名（如 "stage-c-test-4-gpu-h100 (2)"）中去除分区后缀 " (N)"，得到套件名 "stage-c-test-4-gpu-h100"。
2. 更改时间统计键：在 collect_timings() 函数中，将时间数据字典的键从 (rel_path, backend) 改为 (rel_path, suite)，并更新所有相关注释和文档字符串以反映这一变更。
3. 确保套件匹配：update_est_times() 函数通过匹配函数名和 suite= 参数，为每个 register_cuda_ci 调用分配正确的套件特定时间中位数。

关键代码片段：

```
def job_name_to_suite(job_name):
    """Extract the suite name from a job name.
    Job names look like "stage-c-test-4-gpu-h100 (2)" or "stage-a-test-cpu".
    Strip the partition suffix " (N)" to get the suite name.
    """
    return re.sub(r"\s*\(\d+\)$", "", job_name)
```

评论区精华

无 review 评论，PR 由作者直接合并。从提交信息看，变更逻辑直接明了，旨在快速修复前序 PR 引入的问题。

风险与影响

风险：

- 若 CI 任务命名规范变更（如分区后缀格式），job_name_to_suite 函数可能提取错误套件名，导致时间统计不准确。
- 未添加单元测试，依赖现有 CI 流程验证，可能掩盖边缘情况。

影响：

- 正面：提升 CI 时间估算准确性，优化测试队列负载均衡，间接加快开发迭代。
- 范围：仅影响基础设施脚本，不涉及模型推理或核心测试逻辑，对终端用户无直接影响。

关联脉络

本 PR 是 CI 时间估算自动化流程的一部分：

- PR #22545：引入了每周更新 est_time 的工作流和初始脚本，但存在统计逻辑缺陷。
- PR #22550：展示了错误时间估算的实际案例（由工作流自动生成），凸显了修复必要性。

结合近期历史 PR，该仓库持续投入 CI 基础设施优化（如 PR #22461、#22545），本 PR 是这一趋势中的精细调整，体现了对 CI 资源调度效率的持续关注。