PR #22545 完整报告

sgl-project/sglang

feat: add weekly workflow to update CI test est_time values

合并时间：2026-04-11 06:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22545

执行摘要

本次 PR 引入每周自动化工作流和脚本，基于历史 CI 日志更新测试估计时间值，优化负载均衡以减少资源浪费，是一个有意义的 CI 基础设施改进。

功能与动机

为什么做：CI 测试注册调用中的 est_time 值随时间漂移，这些值驱动 LPT 负载均衡算法分区并行测试作业。不准确的估计导致分区不平衡，浪费 CI 资源。引用 PR body："Inaccurate estimates lead to unbalanced partitions, wasting CI resources."

实现拆解

新增文件：

- .github/workflows/weekly-update-est-time.yml：每周六 00:00 UTC 触发的工作流，运行更新脚本，并在检测到变更时自动创建 PR。
- scripts/ci/update_est_time.py：Python 脚本，核心功能包括：
  - 通过 GitHub API 获取最近 20 个调度的 PR Test 运行日志。
  - 使用正则表达式解析日志中的文件执行时间（例如 filename='.../test_x.py', elapsed=120）。
  - 计算每个测试文件最后 10 次成功执行的中位数。
  - 通过正则表达式匹配并更新源代码中的 est_time 字面量。

关键逻辑代码块示例（来自脚本）：

```
LOG_PATTERN = re.compile(
    r"filename='[^']*?/sglang/((?:test|python)/[^']+\.py)', elapsed=(\d+),"
)
def determine_backend(job_name):
    name = job_name.lower()
    for backend in ["cpu", "amd", "npu"]:
        if backend in name:
            return backend
    return "cuda"
```

评论区精华

review 中 gemini-code-assist[bot] 提出两个关键讨论：

- 后端检测扩展：

"The current implementation only recognizes 'cpu' and defaults everything else to 'cuda'... should be updated to correctly identify them from job names." 结论：作者更新 determine_backend 函数以支持 'amd' 和 'npu'，确保时间数据按硬件平台正确分离。

- 更新逻辑稳健性：

"The regex and update logic have several issues:

1. Strictness... 2. Type Handling... 3. Logic Flaw..." 结论：作者改进了逻辑，解决空格、浮点数处理等问题，确保更新准确。

风险与影响

风险：

1. 依赖外部 API：脚本依赖 GitHub API 获取日志，若 API 变更或故障，更新流程可能中断。
2. 正则表达式脆弱：匹配逻辑可能无法覆盖所有代码格式（如多行调用），导致部分 est_time 值未被更新。
3. 数据采样偏差：仅基于最后 10 次执行计算中位数，在测试频繁变更时可能产生不准确估计。

影响：

- 积极影响：提升 CI 负载均衡准确性，减少资源浪费，估计可改善测试分区效率（例如案例中 test_deepseek_v3_fp4_4gpu.py 从 450 秒更新至 1146 秒）。
- 范围：仅影响 CI 测试配置，不改变模型推理逻辑或用户功能。

关联脉络

与历史 PR 的关系：

- PR #22461：添加 GB200 夜间性能回归管道，同为 CI 基础设施改进，显示团队持续优化测试监控。
- PR #22465：更新 CI 权限，与本 PR 的自动化 PR 创建权限管理相关。演进趋势：这表明 sglang 项目正加强 CI 自动化和资源管理，以减少手动维护成本并提升测试效率。