PR #25965 完整报告

sgl-project/sglang

cap API quota for runner-utilization / amd-ci-job-monitor

合并时间：2026-05-21 16:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25965

执行摘要

- 一句话：限制 CI 工作流 API 配额消耗
- 推荐动作：值得合并，以保护共享 API 配额。建议在后续 PR 中补充更新内部函数的类型提示以保持一致性。

功能与动机

PR body 指出长期存在的沙箱 PR（如 #25656）绕过 PR-trigger paths-filter，产生多个并发 24 小时 API 扫描，耗尽了共享的 15k/hr 安装令牌配额。

实现拆解

1. 为两个工作流添加并发控制：在 .github/workflows/runner-utilization.yml 和 .github/workflows/amd-ci-job-monitor.yml 中添加 concurrency 块，设置 group 为工作流名加 ref，启用 cancel-in-progress: true，确保同一 ref 同时只有一个运行实例。
2. 限制 PR 触发条件：在两个工作流的 job 定义中添加 if 条件，跳过 fork PR，且要求同仓库 PR 必须带有 run-ci 标签才能触发；schedule 和 workflow_dispatch 事件不受影响。
3. 缩短 PR 触发的扫描时间窗口：通过条件表达式 (github.event_name == 'pull_request' && '0.34') || inputs.hours || '24'，将 PR 触发的 --hours 参数设为 0.34 小时（约 20 分钟），而 schedule/dispatch 仍使用默认的 24 小时。
4. 调整脚本参数类型：将 scripts/ci/utils/query_job_status.py 和 scripts/ci/utils/runner_utilization_report.py 中的 --hours 参数类型从 int 改为 float，以支持小数小时数，并更新帮助文本。

关键文件：

- .github/workflows/runner-utilization.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：核心变更文件之一：添加并发控制、限制 PR 触发条件并缩短 PR 扫描窗口
- .github/workflows/amd-ci-job-monitor.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：与 runner-utilization 相同的变更模式，保护 AMD CI 作业监控工作流
- scripts/ci/utils/query_job_status.py（模块 CI 脚本；类别 infra；类型 infrastructure）：参数类型从 int 改为 float 以支持小数小时数

- scripts/ci/utils/runner_utilization_report.py（模块 CI 脚本；类别 infra；类型 infrastructure）：参数类型从 int 改为 float 以支持小数小时数

关键符号：未识别

关键源码片段

.github/workflows/runner-utilization.yml

核心变更文件之一：添加并发控制、限制 PR 触发条件并缩短 PR 扫描窗口

```
# .github/workflows/runner-utilization.yml
# 并发控制：同一 ref 同时只允许一个运行实例，取消正在进行的运行
concurrency:
  group: runner-utilization-${{ github.ref }}
  cancel-in-progress: true

jobs:
  report:
    name: Generate Report
    # 跳过 fork PR，同仓库 PR 需要 run-ci 标签；schedule/dispatch 始终运行
    if: >-
      github.event_name != 'pull_request' ||
      (github.event.pull_request.head.repo.full_name == github.repository &&
       contains(github.event.pull_request.labels.*.name, 'run-ci'))
    runs-on: ubuntu-latest
    steps:
      - name: Run report
        env:
          GH_TOKEN: ${{ secrets.GITHUB_TOKEN }}
        run: |
          # PR 触发仅做脚本冒烟检查，扫描 20 分钟窗口（0.34h）以控制 API 成本
          # schedule/dispatch 运行完整报告（默认 24h）
          python scripts/ci/utils/runner_utilization_report.py \
            --repo ${{ github.repository }} \
            --hours ${{ (github.event_name == 'pull_request' && '0.34') || inputs.hours || '24' }} \
            ${{ inputs.filter && format('--filter {0}', inputs.filter) || '' }}
```

.github/workflows/amd-ci-job-monitor.yml

与 runner-utilization 相同的变更模式，保护 AMD CI 作业监控工作流

```
# .github/workflows/amd-ci-job-monitor.yml
# 并发控制：防止同一 ref 多次触发导致 API 配额耗尽
concurrency:
  group: amd-ci-job-monitor-${{ github.ref }}
  cancel-in-progress: true

jobs:
  fetch-actions-data:
    name: Fetch Actions Snapshot
```

```
    # 跳过 fork PR，同仓库 PR 需要 run-ci 标签；schedule/dispatch 始终运行
    if: >-
      github.event_name != 'pull_request' ||
      (github.event.pull_request.head.repo.full_name == github.repository &&
       contains(github.event.pull_request.labels.*.name, 'run-ci'))
    runs-on: ubuntu-latest
    env:
      GH_TOKEN: ${{ secrets.GITHUB_TOKEN }}
    steps:
      - name: Fetch Actions data snapshot
        timeout-minutes: 30
        run: |
          # PR 触发仅做冒烟检查，扫描 20 分钟窗口
          python scripts/ci/utils/query_job_status.py \
            --repo ${{ github.repository }} \
            --workflow "${{ steps.select-workflows.outputs.workflows }}" \
            --hours ${{ (github.event_name == 'pull_request' && '0.34') || inputs.hours || '24' }} \
            --dump-data-file actions-job-snapshot.json
```

评论区精华

review 中 gemini-code-assist[bot] 指出 --hours 参数改为 float 后，内部函数（如 get_workflow_runs、fetch_all_jobs_snapshot、format_markdown 等）的类型提示仍是 int，建议同步更新为 float 以避免静态分析警告。该建议未在评论中得到回复或解决。

- 类型提示未同步更新为 float (style): 未在评论中得到答复或解决。

风险与影响

- 风险：低风险。变更仅限于 CI 工作流配置和两个脚本的参数解析，不涉及核心服务代码。但未按 review 建议更新内部函数类型提示，可能导致类型检查器产生警告，不过不影响运行时行为。
- 影响：影响面局限于 CI 基础设施。长期存在的沙箱 PR 将不再无限制消耗 API 配额，fork PR 完全跳过这两个工作流，PR 触发的扫描时间窗口从 24 小时缩短至 20 分钟，显著降低 API 调用量。schedule 和 workflow_dispatch 触发的完整报告不受影响。
- 风险标记：类型提示不一致

关联脉络

- PR #25656 [WIP] kv_canary: PR body 中提及此长期存在的沙箱 PR 是导致 API 配额耗尽的原因之一