

PR #23420 完整报告

sgl-project/sglang

Update pr-test-xeon.yml cancel-in-progress config

合并时间: 2026-04-22 10:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23420>

执行摘要

- 一句话: 调整 Xeon CI 并发策略, PR 提交时取消排队中的旧运行以加速反馈。
- 推荐动作: 该 PR 变更简单直接, 无需精读源码; 建议团队关注 CI 配置的此类优化模式, 可在其他类似工作流中考虑应用相同策略以提升整体 CI 效率。

功能与动机

PR body 指出: 当前 Xeon CI 使用 `cancel-in-progress: false`, 导致同一 PR 的新提交必须在单个 xeon-gnr runner 上等待旧运行完成 (通常耗时数小时)。为与现有 XPU 策略对齐, 改为 PR/push 时取消同一 ref 的陈旧运行, 以加速反馈并尽早释放 runner; workflow_call (用于发布 / 夜间构建) 保持不变, 永不取消。

实现拆解

- 修改 CI 工作流配置: 在 `.github/workflows/pr-test-xeon.yml` 中, 将 `concurrency` 下的 `cancel-in-progress` 从固定值 `false` 改为条件表达式 `${{ github.event_name != 'workflow_call' }}`。
- 条件逻辑说明: 当触发事件为 PR 推送 (`pull_request`) 或直接推送 (`push`) 时, `github.event_name` 不等于 `'workflow_call'`, 表达式结果为 `true`, 启用取消排队中运行的功能; 当通过 `workflow_call` 触发 (如发布流程、夜间测试) 时, 表达式结果为 `false`, 保持不取消, 确保这些关键流程的稳定性。
- 影响范围: 仅影响 Xeon 平台的 CI 测试工作流, 不涉及源码、测试或部署脚本的其他部分。

关键文件:

- `.github/workflows/pr-test-xeon.yml` (模块 CI 配置; 类别 `infra`; 类型 `configuration`): 唯一变更文件, 定义了 Xeon CI 测试工作流的并发行为, 直接影响 CI 执行效率和资源使用。

关键符号: 未识别

关键源码片段

`.github/workflows/pr-test-xeon.yml`

唯一变更文件, 定义了 Xeon CI 测试工作流的并发行为, 直接影响 CI 执行效率和资源使用。

```
# 在 concurrency 配置块中, 调整 cancel-in-progress 策略
concurrency:
```

```
group: pr-test-xeon-${{ inputs.ref || github.ref }}
# 原为 false, 导致同一 PR 的多次提交排队等待
# 现改为条件判断: 仅当非 workflow_call 事件 (如 PR/push) 时取消排队中的运行
cancel-in-progress: ${{ github.event_name != 'workflow_call' }}
# 这样 PR 提交能更快获得反馈, workflow_call (如发布流程) 保持稳定不取消
```

评论区精华

无 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 风险较低但需注意:
 - 工作流中断风险: 条件表达式 `${{ github.event_name != 'workflow_call' }}` 依赖于 GitHub Actions 的事件上下文, 若事件类型识别错误 (如罕见边缘情况), 可能导致意外取消或保留运行, 影响 CI 可靠性。
 - 资源争用变化: 启用取消后, 同一 PR 的并发运行可能被中断, 若测试本身有状态依赖 (如临时文件、外部服务), 可能引入 flaky 测试。
 - 配置一致性: 仅调整 Xeon 工作流, 需确保与其他平台 (如 XPU) 的策略一致, 避免团队认知负担。
- 影响: 影响范围: 仅限 Xeon CI 测试工作流的执行行为。影响程度:
 - 对开发者: PR 提交后获得 CI 反馈的速度显著提升, 减少等待时间, 提高开发效率。
 - 对系统: runner 资源利用率优化, 避免同一 PR 的陈旧运行阻塞新提交。
 - 对团队: 与 XPU 策略对齐, 降低维护复杂度, 但需确保所有成员理解此变更 (可能通过文档或公告)。
- 风险标记: 条件表达式依赖, CI 行为变更

关联脉络

- PR #23417 [ci] split stage-c-test-4-gpu-b200 to enable a low-disk runner pool: 同为 CI 基础设施优化, 通过分割测试套件提升资源利用率, 与本 PR 的并发策略调整目标一致。
- PR #22830 ci: enable /rerun-test for nightly test suites: 涉及 CI 工具能力扩展, 与本 PR 同属 infra 标签下的 CI 改进。