PR #25188 完整报告

sgl-project/sglang

[SMG] Fix matrix-sibling concurrency collision in PR Test (SMG)

合并时间：2026-05-14 04:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25188

执行摘要

- 一句话：修复 PR 测试中矩阵兄弟并发冲突
- 推荐动作：本 PR 是一次精巧的 CI 配置 bugfix，值得所有涉及矩阵并发场景的仓库参考。建议精读 PR body 中对 concurrency group 行为的解释，以及如何通过追加 matrix.name 解决兄弟冲突。对于维护高并发 CI 的团队，此模式可作为最佳实践。

功能与动机

PR body 指出：两个矩阵条目共享 runner（如 benchmarks 和 chat-completions-4gpu 均使用 4-gpu-h100），它们映射到同一个 concurrency group，而 GitHub Actions 每个 group 只允许一个 running 和一个 pending 任务，新进入的任务会驱逐之前的 pending 任务（即使 cancel-in-progress: false）。这导致 benchmarks 任务在 chat-completions-4gpu 成为 eligible 时被取消，造成整体 workflow 结论为 cancelled。

实现拆解

1. 定位冲突来源：在 .github/workflows/pr-test-rust.yml 中，gateway-e2e 矩阵作业的 concurrency.group 被设为 pr-test-rust-${{ matrix.runner }}，其中 runner 为硬件标签（如 4-gpu-h100），导致同一 runner 的不同矩阵条目（由 matrix.name 区分）共享同一 group。
2. 修改 concurrency group 定义：在 group 值后追加 ${{ matrix.name }}，将其变为 pr-test-rust-${{ matrix.runner }}-${{ matrix.name }}，使得每个矩阵条目拥有独立的 group。
3. 保留跨 PR 串行化：不同 PR 中相同 runner+name 的作业仍共享同一 group（如两个 PR 的 benchmarks on 4-gpu-h100 均使用 group pr-test-rust-4-gpu-h100-benchmarks），从而维持 runner 稀缺性下的硬件级串行化，避免超载。
4. 无其他文件修改：变更仅涉及一个 CI 配置文件的 8 行改动（+6/-2）。

关键文件：

- .github/workflows/pr-test-rust.yml（模块 CI 配置；类别 infra；类型 infrastructure）：唯一修改的文件，调整 concurrency group 定义以修复矩阵兄弟冲突。

关键符号：未识别

关键源码片段

.github/workflows/pr-test-rust.yml

唯一修改的文件，调整 concurrency group 定义以修复矩阵兄弟冲突。

```
# .github/workflows/pr-test-rust.yml ( 片段 )
# Self-hosted GPU runners are scarce; serialize per hardware type so
# 2-gpu-h100 and 4-gpu-h100 each run one job at a time across all
# in-flight PRs. Queue rather than cancel — different refs shouldn't
# interrupt each other. Include matrix.name in the group so siblings
# within the same run don't collide: GitHub keeps only one pending job
# per concurrency group, so matrix entries sharing a group (e.g.
# benchmarks + chat-completions-4gpu on 4-gpu-h100) would evict each
# other regardless of cancel-in-progress: false.
concurrency:
  group: pr-test-rust-${{ matrix.runner }}-${{ matrix.name }}
  cancel-in-progress: false
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及 GitHub Actions 的 concurrency group 字符串拼接，作用域局部且逻辑直观。潜在风险是拼写错误导致 group 不匹配（如 name 变量未定义），但 CI 运行后即可验证。由于 matrix.name 已在其他类似 CI 作业中使用，该风险几乎不存在。
- 影响：影响范围：仅影响 pr-test-rust.yml 中 gateway-e2e 矩阵作业的并发行为。对用户无影响；对开发者而言，同一个 PR 内所有矩阵条目现在都能独立排队，不再因兄弟条目的出现而被取消，从而避免 CI 误报为 cancelled。跨 PR 的硬件串行化行为保持不变。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR