

PR #36981 完整报告

sgl-project/sglang

[CI] Temporarily disable GB300 jobs

合并时间: 2026-08-29 14:56

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36981>

执行摘要

- 一句话: 临时禁用不可用的 GB300 CI 运行器
- 推荐动作: 值得精读, 作为临时禁用 CI 运行器的标准做法, 特别是保留注册以便回滚的设计。

功能与动机

4-gpu-gb300 运行器当前不可用, 需防止 CI 和 release 工作流等待或分配该失效运行器, 直至其恢复。

实现拆解

1. 在 `.github/workflows/_pr-test-stage.yml` 的共享 CUDA 阶段, 在运行器分配前增加守卫, 跳过 `inputs.runner_config == '4-gpu-gb300'`, 覆盖 PR 与 nightly 套件。
2. 在 `.github/workflows/rerun-test.yml` 中, 为 CUDA 和 multimodal 的 rerun 增加守卫, 拒绝针对 4-gpu-gb300 的重跑请求。
3. 在 `.github/workflows/release-whl-deepgemm.yml` 中, 临时注释掉 GB300 aarch64 的矩阵条目, 避免 release 构建尝试使用该运行器。
4. 保留 GB300 套件和测试注册, 以便运行器恢复后快速安全回滚。

关键文件:

- `.github/workflows/_pr-test-stage.yml` (模块 CI 工作流; 类别 infra; 类型 infrastructure) : 共享 CUDA 阶段增加运行器守卫, 覆盖 PR 和 nightly 套件, 是核心变更。
- `.github/workflows/rerun-test.yml` (模块 CI 工作流; 类别 infra; 类型 infrastructure) : 拒绝针对 GB300 的 CUDA 和 multimodal rerun, 避免手动重跑等待失效运行器。
- `.github/workflows/release-whl-deepgemm.yml` (模块 CI 工作流; 类别 infra; 类型 infrastructure) : 临时注释掉 GB300 aarch64 构建矩阵条目, 避免 release 构建使用失效运行器。

关键符号: 未识别

关键源码片段

`.github/workflows/_pr-test-stage.yml`

共享 CUDA 阶段增加运行器守卫，覆盖 PR 和 nightly 套件，是核心变更。

```
# .github/workflows/_pr-test-stage.yml
# 临时跳过失效的 GB300 运行器，避免 CI 等待或分配。
if: |
  always() &&
  inputs.runner_config != '4-gpu-gb300' &&
  ((github.event_name == 'schedule' || fromJson(inputs.caller_inputs).test_parallel_dispatch ==
  true) ||
  (!failure() && !cancelled())) &&
  (fromJson(inputs.check_changes).main_package == 'true' || fromJson(inputs.check_changes).
  sgl_kernel == 'true')
# runs-on 从 runs_on_map 解析；check-changes 已替换。
```

[.github/workflows/rerun-test.yml](#)

拒绝针对 GB300 的 CUDA 和 multimodal rerun，避免手动重跑等待失效运行器。

```
# .github/workflows/rerun-test.yml
# 临时拒绝针对失效 GB300 运行器的 rerun 请求。
rerun-test-cuda:
  if: inputs.mode == 'cuda' && inputs.runs_on != '4-gpu-gb300'
  runs-on: ${{ inputs.runs_on }}
  # ...

rerun-test-multimodal-gen:
  if: inputs.mode == 'multimodal_gen' && inputs.runs_on != '4-gpu-gb300'
  runs-on: ${{ inputs.runs_on }}
  # ...
```

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，仅涉及工作流调度，不影响模型输出。但若 GB300 运行器恢复后未及时回滚，可能导致相关测试和构建被永久跳过，需要持续跟踪运行器状态。
- 影响：影响 CI 与 release 工作流，GB300 相关的测试和构建暂时不执行，可能延迟相关硬件的回归测试。对用户无直接功能影响。
- 风险标记：临时变更需及时回滚，影响 CI 覆盖

关联脉络

- 暂无明显关联 PR