

PR #38770 完整报告

sgl-project/sglang

[CI] Temporarily disable GB300 tests

合并时间: 2026-09-10 09:31

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/38770>

执行摘要

由于 GB300 GPU 运行器集群完全宕机，所有派发到该硬件的 CI 测试和构建任务无限期排队，阻塞了全仓库的 PR 合并流程（`pr-test-finish` 无法完成）。本 PR 是一个临时的运维规避措施，通过注释掉三个 GitHub Actions 工作流文件中与 `4-gpu-gb300` 标签相关的所有任务定义和依赖，立即解除了对 CI 管线的阻塞。测试本身在 `run_suite.py` 中保持注册，恢复时只需 revert 本提交。

功能与动机

根据 PR 描述，问题根源是“Both GB300 runners went down tonight, independently... every GB300 job sits queued indefinitely and holds up `pr-test-finish` on every PR.”这导致了两个集群的运行器同时不可用，一个集群的节点注册完全丢失，另一个运行器的 Pod 被移除且未恢复。这直接造成了所有需要 GB300 硬件的测试任务（PR 提交级、每夜定期、以及轮子构建）在队列中无限等待，而 `pr-test-finish` 作业依赖于这些任务的完成，因此阻塞了所有 PR 的合并状态检查。

实现拆解

- 禁用 PR 测试阶段：在 `.github/workflows/pr-test.yml` 中，注释掉了 `base-c-test-4-gpu-gb300` 作业的完整定义，同时注释掉了 `pr-test-finish` 作业 `needs` 列表中对该作业的引用。这阻止了每次提交触发的测试阶段向失效的运行器派发任务。
- 禁用每夜测试阶段：在 `.github/workflows/nightly-test-nvidia.yml` 中，注释掉了 `nightly-4-gpu-gb300` 作业的完整定义，以及其在 `nightly-test-finish` 和 `nightly-test-finish-extra` 这两个最终聚合作业 `needs` 列表中的两个引用。这防止了每夜测试套件尝试使用 GB300 运行器。
- 禁用轮子构建：在 `.github/workflows/release-whl-deepgemm.yml` 中，注释掉了矩阵条目中指定 `runner: 4-gpu-gb300` 的 `sm100-aarch64` 轮子构建配置。这停止了失败的构建任务进入队列。
- 为恢复保留配置：PR 明确说明测试仍在 `run_suite.py` 中注册，且 `4-gpu-gb300` 条目保留在 `runner_configs.yml` 中。这意味着重新启用只需 revert 本次提交。同时，`rust-ext-build-aarch64` 作业被刻意保留，以为未来恢复时预热 `aarch64` 缓存。

本次变更是纯 YAML 配置注释，没有引入新的逻辑或函数，因此无需展示关键源码片段。

评论区精华

没有技术讨论。PR 由 [mmangkad](#) 以空消息批准 (APPROVED)，表明这是一个被快速认可的运维修复。

风险与影响

风险：主要风险是运维性的而非技术性的。如果 GB300 运行器恢复上线后，团队忘记 revert 本提交，相关测试将保持禁用状态，可能允许针对该特定硬件的回归问题被遗漏。另一个较小的风险是保留的 [rust-ext-build-aarch64](#) 作业是否可能依赖于已禁用的 GB300 阶段，但 PR 作者注明这是有意为之。

影响：

- 对 CI 管线：此变更立即解除了所有 PR 和每夜运行的 CI 阻塞。pr-test-finish 作业将不再被挂起的 GB300 作业阻塞，允许合并操作继续。
- 对测试覆盖：所有专门针对 GB300 硬件的测试被临时禁用，这减少了对该特定 GPU 平台的测试覆盖，直到运行器恢复并 revert 本更改。
- 对发布：sm100-aarch64 目标架构的轮子构建暂停，这意味着在该架构的新轮子产出之前，相关发布将受影响。

关联脉络

这是一个独立的运维修复 PR，与近期历史中的其他 PR 没有直接的代码逻辑关联。它临时中断了涉及 GB300 硬件的 CI 信号流，其恢复 (revert) 将是未来一个必要的后续操作。