

PR #32719 完整报告

sgl-project/sglang

[CI] Re-enable GB300 CI jobs

合并时间: 2026-08-01 05:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32719>

执行摘要

- 一句话: 恢复 GB300 CI 任务, nightly、PR 与 DeepGEMM 测试回线
- 推荐动作: 值得快速浏览, 作为“runner 恢复后如何回滚禁用 CI 的模板”; 对源码阅读没有价值。重点关注两点: restartPolicy 对 runner 可用性的保障方法, 以及评论中对 EADDRINUSE 的系统性排查 (ephemeral 范围、外部占用排除)。不建议精读。

功能与动机

GB300 是 Blackwell 系列重要推理平台, 此前因 runner 不可用 (个别 pod 丢失、宿主重启后无法恢复) 在 #31764 中禁用相关 CI。作者在 PR body 中说明: 两个 GB300 runner 已重新上线并携带 4-gpu-gb300 标签, 重新创建的 gb300-4-gpu-03 这次使用了 restartPolicy: Always, 宿主重启也不会再让它永久离线; 并在真实安装路径 (GRACE_BLACKWELL=1 ci_install_deepep.sh) 上 smoke test 通过, 因此可以恢复被禁用的任务。

实现拆解

1. 恢复夜间性能套件: 在 .github/workflows/nightly-test-nvidia.yml 中取消 nightly-test-perf-4-gpu-gb300 job 的注释, 恢复 4 个 model 矩阵 (glm5-nvfp4、qwen35-fp8、deepseek-v4-pro-fp4、kimi-k25-nvfp4), 并在 workflow 调度列表 (on: 事件与 consolidate-metrics 的依赖项) 中重新加入该 job。该 job 使用 GRACE_BLACKWELL=1 ci_install_deepep.sh 安装依赖, 通过 run_suite.py --hw cuda --suite ... --nightly --continue-on-error 运行, 并保留 trace 发布与 coredump 上传步骤。
2. 恢复 PR 门禁: 在 .github/workflows/pr-test.yml 中恢复 base-c-test-4-gpu-gb300, 复用 _pr-test-stage.yml 模板, 设置 runner_config: 4-gpu-gb300、单文件超时 1800 秒, 并把该 job 加回 pr-test-finish 的 needs 列表, 使 GB300 测试重新成为每次 commit 合并前的必过门禁。
3. 恢复 wheel 构建: 在 .github/workflows/release-whl-deepgemm.yml 中恢复 sm100-aarch64 矩阵项 (runner: 4-gpu-gb300), 保留 sm120 的注释状态 (#32193 的 skip 与本次恢复无关)。
4. 配套说明: 三个改动均为 workflow 配置, 无产品代码、测试或 schema 变更; 实际验证依赖 GitHub Actions 的 rerun (/rerun-test) 与两次真实 job 运行。

关键文件:

- `.github/workflows/nightly-test-nvidia.yml` (模块 夜间测试; 类别 `infra`; 类型 `infrastructure`; 符号 `nightly-test-perf-4-gpu-gb300`) : 恢复了整个 GB300 夜间性能套件的 `job` 定义与调度入口, 是本 PR 最大的配置恢复, 影响 4 个模型矩阵的 `nightly` 基准。
- `.github/workflows/pr-test.yml` (模块 PR 测试; 类别 `infra`; 类型 `infrastructure`; 符号 `base-c-test-4-gpu-gb300`, `pr-test-finish`) : 恢复 `per-commit` GB300 测试并把其加入 `pr-test-finish` 门禁, 直接影响每个 PR 的合并质量门槛。
- `.github/workflows/release-whl-deepgemma.yml` (模块 发版构建; 类别 `infra`; 类型 `infrastructure`; 符号 `sm100-aarch64`) : 恢复 `aarch64` DeepGEMM wheel 的构建矩阵项, 低频但影响发版物覆盖。

关键符号: 未识别

评论区精华

1. `alisonshao` 第一次 rerun 后, bot 回报 4-gpu-gb300 上 4 个测试中 `test_deepseek_v3_cutedsl_4gpu` 失败 (两次在 `torch-dist` init 因 `EADDRINUSE` 退出, 端口 39539、39537, exit code -3), 其余通过; 4-gpu-b200 上 `test_numa_utils` 通过。
 2. `alisonshao` 的归因分析: 失败端口落在宿主 `ephemeral` 范围 32768-60999 内, 属于服务器自己端口规划冲突; 主机上该范围 0 监听、0 socket 绑定, 排除外部占用; `test_flashinfer_a2a` 通过说明 `NVSHMEM` bootstrap 正常。她明确指出这是 `main` 上 #31731 修复的竞态, 与本 PR 无关 (本 PR 只取消注释)。
 3. 第二次 rerun 后, 4-gpu-gb300 4 个测试与 4-gpu-b200 1 个测试全部通过, 恢复的 `job` 在真实 runner 上稳定; reviewer `Fridge003` 直接 approve。
- 首次 rerun 结果: GB300 测试失败报告 (testing): 失败仅限单个 DeepSeek 4-GPU 测试, 其余 GB300 测试 (`test_numa_utils`、`test_flashinfer_a2a`、`test_disaggregation_aarch64`) 均通过。
 - `EADDRINUSE` 失败归因讨论 (correctness): 确认失败非本 PR 引入, 不阻塞恢复 GB300 任务。
 - 二次 rerun 验证 (testing): 恢复的 GB300 CI 任务在真实 runner 上稳定通过, `Fridge003` approve。

风险与影响

- 风险:
 1. `nightly-test-nvidia.yml` 的 `job` 使用 `runs-on: 4-gpu-gb300-nightly`, `body` 明确当前没有任何 runner 携带该标签, 恢复的 `nightly` perf suite 会一直排队, 直到运维添加标签——这是最直接的配置与资源不匹配风险。
 2. `pr-test.yml` 将 `base-c-test-4-gpu-gb300` 重新纳入 `pr-test-finish` 门禁, 若 GB300 runner 再次离线 (尽管有 `restartPolicy: Always`, 仍可能因其他原因下线), PR 合并会被 `pending` 阻塞。
 3. GB300 上曾出现的 `EADDRINUSE` 失败被定位为 `main` 上已存在的端口竞态 (`get_free_port()` 与 `ephemeral` 端口范围冲突), 本 PR 不包含修复, 依赖 #31731/#33044 等后续 PR; 合入时 `main` 已含修复, 残余风险低。

4. 三个 workflow 文件都是纯注释解除与矩阵恢复，没有本地可执行验证手段，只能依赖 Actions 实际运行；一旦 YAML 缩进或 key 错误会静默失效，但已通过 rerun 实测。 - 影响：对用户无运行时影响，整个 PR 仅涉及 CI/CD 配置。对开发者：GB300/Blackwell 相关改动会重新触发 4-gpu-gb300 的 PR 测试；nightly 恢复 4 个模型的 GB300 性能基准（glm5-nvfp4、qwen35-fp8、deepseek-v4-pro-fp4、kimi-k25-nvfp4）。对发布流程：sm100-aarch64 DeepGEMM wheel 重新进入构建矩阵。对团队：恢复 Grace-Blackwell 平台的回归防线，同时需要运维补充 4-gpu-gb300-nightly runner 标签避免夜间任务积压。 - 风险标记：runner 标签不匹配，夜间任务可能排队，CI 配置无本地验证

关联脉络

- PR #33044 [CI] Pin NCCL ports for GB300 PR tests: 同是 GB300 CI 恢复线上的后续修复复合入；本 PR 恢复到 GB300 PR 测试后暴露了 EADDRINUSE 端口冲突，该 PR 通过固定 NCCL 端口消除同类竞争失败，属于同一 CI 功能线的收尾。