PR #24495 完整报告

sgl-project/sglang

ci: drop 1-gpu-h100-h200 shared label

合并时间：2026-05-06 16:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24495

执行摘要

此 PR 清理了已退役的 1-GPU H200 runner 池的共享标签 1-gpu-h100-h200，将 stage-b-test-1-gpu-large 作业重新指向完整的 1-GPU H100 runner 集群，避免 CI 任务因 runner 数量不足而排队。变更仅涉及 3 个 CI 配置文件，风险极低。

功能与动机

此前，stage-b-test-1-gpu-large 使用共享标签 1-gpu-h100-h200，该标签同时被 1-GPU H100 和 1-GPU H200 runner 池宣传。由于 1-GPU H200 池已退役，共享标签仅剩下 8 个仍带该标签的 runner，导致作业无法利用完整的约 32 个 H100 runner，造成 CI 瓶颈。PR 描述原文："Without this change, stage-b-test-1-gpu-large was bottlenecked on the 8 runners that still carried the legacy shared label"。

实现拆解

1. 更新主 CI 工作流(.github/workflows/pr-test.yml): 将 stage-b-test-1-gpu-large 的 runs-on 从 1-gpu-h100-h200 改为 1-gpu-h100。
2. 更新重跑命令映射(scripts/ci/utils/slash_command_handler.py): 在 CUDA_SUITE_TO_RUNNER 字典中将对应的 runner 值从 1-gpu-h100-h200 改为 1-gpu-h100，确保 /rerun-test 命令使用正确的标签。
3. 清理重跑工作流选项(.github/workflows/rerun-test.yml): 从 runner-label 的 options 中删除 1-gpu-h100-h200，防止用户通过 UI 选择已退役的标签。

无需展示，变更均为简单的字符串替换和选项移除。

评论区精华

无实质讨论。机器人审查无反馈，Fridge003 直接批准。

风险与影响

- 风险：极低。仅修改 CI 配置字符串，不涉及任何运行时逻辑。若未来 H100 池也退役，需要类似清理。
- 影响：stage-b-test-1-gpu-large 将调度到完整的 1-GPU H100 集群，预期可减少 CI 排队时间。对产品功能无影响。

关联脉络

此 PR 是基础设施清理的一部分，与 H200 runner 退役的运营决策相关。无关联的代码变更。