PR #24080 完整报告

sgl-project/sglang

[CI] Broaden stage-b-test-1-gpu-large runner pool to H100 + H200

合并时间：2026-04-30 03:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24080

执行摘要

本次 PR 将 CI 中 stage-b-test-1-gpu-large job 的 runner 标签从 1-gpu-h100 切换为共享标签 1-gpu-h100-h200，使其能被 H100 或 H200 1-GPU runner 承接，从而扩大 runner 池，减少排队等待。同时更新斜杠命令处理器中的对应映射。Multimodal_gen 相关 job 保持 1-gpu-h100 标签不变，以避免硬件差异带来的不稳定。

功能与动机

PR 目标是扩大 stage-b-test-1-gpu-large 的 runner 池，使其在 H100 和 H200 上均可运行，以减少排队等待时间并提高 CI 资源利用率。这一模式沿用了 #23505 为 4-gpu-b200-low-disk 采用的共享标签方式。Multimodal_gen 因依赖 H100 特有的硬件行为，仍限制在 H100 上。

实现拆解

1. 修改 pr-test.yml 中的 runner 标签：在 .github/workflows/pr-test.yml 文件中，将 stage-b-test-1-gpu-large job 的 runs-on 字段从 1-gpu-h100 改为 1-gpu-h100-h200。该变更仅影响一个 job，其他 job 标签不变。
2. 更新斜杠命令处理器的映射：在 scripts/ci/utils/slash_command_handler.py 中，将 CUDA_SUITE_TO_RUNNER 字典中 stage-b-test-1-gpu-large 对应的标签值也改为 1-gpu-h100-h200。这样可以确保 /rerun-stage stage-b-test-1-gpu-large 命令的 runner 健康检查与新的标签一致。

无可用源码片段。

评论区精华

本 PR 没有 review 讨论。但 PR 描述中提到，合并前需要运维人员手动将 1-gpu-h100-h200 标签添加到所有 1-GPU H100 和 H200 runner 上，否则 job 会无限排队。

风险与影响

风险：

- 若 runner 端未提前添加共享标签，stage-b-test-1-gpu-large 会永久排队，但不影响其他 job。
- H100 和 H200 在性能上存在差异，可能导致测试时间波动，但这是一个已知且可接受的风险。

影响：

- 范围：仅影响 stage-b-test-1-gpu-large 一个 CI job。
- 程度：正面影响是减少排队等待，提高 CI 吞吐量；负面影响是测试时间基线可能因 runner 而异。
  - Multimodal_gen 相关 job 不受影响。

关联脉络

本 PR 遵循了 #23505 引入的共享标签模式。在 #23505 中，4-gpu-b200-low-disk 标签被同时应用到旧和新 b200 runner 上，实现了类似的扩池效果。