PR #23505 完整报告

sgl-project/sglang

[CI] Broaden stage-b-test-4-gpu-b200 runner pool to low-disk label

合并时间：2026-04-23 08:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23505

执行摘要

- 一句话：扩展 CI 中 B200 测试任务的 runner 选择范围，支持低磁盘标签以提升调度弹性。
- 推荐动作：该 PR 变更简单直接，主要涉及 CI 配置调整。对于关注 CI 基础设施或 B200 测试环境的工程师，可以快速浏览以了解 runner 标签的使用策略；对于其他开发者，无需深入阅读。

功能与动机

根据 PR 描述，目的是将 stage-b-test-4-gpu-b200 任务的 runner 选择范围从特定的 b200_runner（对应 4-gpu-b200/4-gpu-b200-kernel 标签）扩展到 b200_low_disk_runner（对应 4-gpu-b200-low-disk/4-gpu-b200-kernel-low-disk 标签）。*-low-disk 标签同时被现有的大磁盘 B200 runner 和新引入的低磁盘 runner 所支持，因此该任务现在可以落在任一资源池中，这与同一工作流中已分区的 B200 任务模式保持一致，旨在提高 CI 资源的利用率和任务调度弹性。

实现拆解

1. 核心配置变更：修改 .github/workflows/pr-test.yml 文件中 stage-b-test-4-gpu-b200 任务的 runs-on 字段，从引用 needs.check-changes.outputs.b200_runner 变量改为引用 needs.check-changes.outputs.b200_low_disk_runner 变量。
2. 测试与验证：提交历史显示，作者先临时硬编码 runner 标签为 4-gpu-b200-low-disk-test 进行测试，验证新低磁盘 runner 池的功能，随后恢复为使用变量引用，确保任务能正常回退到内核 / 非内核的常规拆分逻辑。
3. 配套改动：无其他源码、测试或文档改动，变更仅限于 CI 工作流配置。

关键文件：

- .github/workflows/pr-test.yml（模块 CI 配置；类别 infra；类型 configuration）：这是唯一被修改的文件，直接定义了 CI 工作流中 B200 测试任务的 runner 选择逻辑。

关键符号：未识别

关键源码片段

.github/workflows/pr-test.yml

这是唯一被修改的文件，直接定义了 CI 工作流中 B200 测试任务的 runner 选择逻辑。

```
# 在 pr-test.yml 的 jobs 部分，stage-b-test-4-gpu-b200 任务配置：
stage-b-test-4-gpu-b200:
  # ... 其他配置 ...
  # 条件判断：仅当主包或 sgl-kernel 有变更时才运行此任务
  if: ((needs.check-changes.outputs.main_package == 'true') || (needs.check-changes.outputs.
  sgl_kernel == 'true'))
  # 关键变更：将 runs-on 从 b200_runner 改为 b200_low_disk_runner
  # b200_low_disk_runner 变量解析为 `4-gpu-b200-low-disk` 或 `4-gpu-b200-kernel-low-disk`
  标签
  # 这些标签同时被现有的大磁盘 B200 runner 和新引入的低磁盘 runner
  支持，因此任务可以落在任一资源池
  runs-on: ${{ needs.check-changes.outputs.b200_low_disk_runner }}
  timeout-minutes: 240
  strategy:
    fail-fast: false
  # ... 后续步骤 ...
```

评论区精华

PR 的评论中没有 review 讨论，只有作者执行 /rerun-stage 命令触发 CI 运行和 bot 的响应。这表明变更可能较为直接，或已在团队内部达成共识。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅影响 CI runner 的标签选择，不涉及业务逻辑。潜在风险包括： 1) 如果低磁盘 runner 的磁盘空间不足，可能导致构建或测试失败（但 PR 描述指出 *-low-disk 标签也被大磁盘 runner 支持，因此有回退机制）。 2) 变量 b200_low_disk_runner 必须在 CI 上下文中正确定义，否则任务可能无法找到合适的 runner。
  - 影响：影响范围有限。直接影响 CI 系统中的 stage-b-test-4-gpu-b200 任务，使其能够使用更广泛的 runner 资源池，可能提高任务调度成功率和资源利用率。对用户、系统核心功能或团队开发流程无直接影响。
  - 风险标记：配置依赖风险

关联脉络

- PR #23510 [CI] /rerun-stage: fix workflow-run URL lookup for sgl-kernel PRs: 同属 CI 基础设施改进，涉及 /rerun-stage 命令和 CI 工作流配置。
- PR #23492 [CI] /rerun-stage: auto-include wheel build when PR modifies sgl-kernel/: 同属 CI 基础设施改进，涉及 /rerun-stage 命令和 CI 工作流配置，且都修改了 .github/workflows/pr-test.yml 文件。
- PR #23086 [CI] GB200 nightly: on-demand PR/branch image build and config filter: 同属 CI 基础设施改进，涉及特定硬件（GB200/B200）的 CI 配置和资源调度。