

PR #22830 完整报告

sgl-project/sglang

ci: enable /rerun-test for nightly test suites

合并时间: 2026-04-22 09:28

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/22830>

执行摘要

- 一句话: 为夜间测试套件添加 /rerun-test 命令支持, 扩展 CI 工具能力。
- 推荐动作: 此 PR 值得快速浏览以了解 CI 工具的扩展点, 但不需深入分析。关注点在于如何将 CI 工作流配置映射到命令行工具, 这对于维护类似基础设施的团队有参考价值。

功能与动机

根据 PR 描述, 当前 /rerun-test 命令对任何注册在 nightly-* 套件中的测试文件都会失败, 因为 CUDA_SUITE_TO_RUNNER 字典只包含 PR 测试套件映射 (stage-a、stage-b、stage-c)。此变更添加了缺失的夜间和每周套件条目, 使开发者能够从 PR 评论中重新运行单个夜间测试。

实现拆解

1. 扩展 CI 配置映射: 在 scripts/ci/utils/slash_command_handler.py 文件的 CUDA_SUITE_TO_RUNNER 字典中添加了 17 个新的键值对, 包括 nightly-1-gpu、nightly-4-gpu、nightly-8-gpu-h200 等夜间测试套件, 以及 weekly-8-gpu-h200 每周测试套件, 每个套件映射到对应的 GPU runner (如 1-gpu-h100、4-gpu-h100)。
2. 映射来源: 这些 runner 分配源自 .github/workflows/nightly-test-nvidia.yml 作业定义, 确保与现有夜间 CI 流程保持一致。
3. 无其他配套改动: 此 PR 仅修改了这一个配置文件, 没有涉及测试代码、文档或部署脚本的变更。

关键文件:

- scripts/ci/utils/slash_command_handler.py (模块 CI 工具; 类别 infra; 类型 configuration; 符号 CUDA_SUITE_TO_RUNNER): 这是唯一被修改的文件, 包含了 /rerun-test 命令的核心配置映射。

关键符号: handle_rerun_stage

关键源码片段

scripts/ci/utils/slash_command_handler.py

这是唯一被修改的文件, 包含了 /rerun-test 命令的核心配置映射。

```
# 在 handle_rerun_stage 函数中定义的 CUDA_SUITE_TO_RUNNER 字典
CUDA_SUITE_TO_RUNNER = {
```

```

# PR 测试套件映射（原有部分保持不变）
"stage-a-test-1-gpu-small": "1-gpu-5090",
"stage-a-test-cpu": "ubuntu-latest",
# ... 其他 PR 套件映射

# 新增：夜间测试套件（NVIDIA）
"nightly-1-gpu": "1-gpu-h100", # 映射到 1 个 H100 GPU 的 runner
"nightly-4-gpu": "4-gpu-h100", # 映射到 4 个 H100 GPU 的 runner
"nightly-4-gpu-b200": "4-gpu-b200", # 映射到 4 个 B200 GPU 的 runner
"nightly-8-gpu-common": "8-gpu-h200", # 映射到 8 个 H200 GPU 的 runner（通用套件）
"nightly-8-gpu-h200": "8-gpu-h200", # 映射到 8 个 H200 GPU 的 runner
"nightly-8-gpu-h20": "8-gpu-h20", # 映射到 8 个 H20 GPU 的 runner
"nightly-8-gpu-b200": "8-gpu-b200", # 映射到 8 个 B200 GPU 的 runner
"nightly-eval-text-2-gpu": "2-gpu-h100", # 文本评估套件
"nightly-eval-trlm-2-gpu": "2-gpu-h100", # 视觉语言模型评估套件
"nightly-perf-text-2-gpu": "2-gpu-h100", # 文本性能测试套件
"nightly-perf-trlm-2-gpu": "2-gpu-h100", # 视觉语言模型性能测试套件
"nightly-kernel-1-gpu": "1-gpu-h100", # 内核测试套件（1 GPU）
"nightly-kernel-8-gpu-h200": "8-gpu-h200", # 内核测试套件（8 GPU）

# 新增：每周测试套件
"weekly-8-gpu-h200": "8-gpu-h200", # 映射到 8 个 H200 GPU 的 runner
}

```

评论区精华

此 PR 的 review 过程非常简单，只有 Kangyan-Zhou 的批准评论，没有其他讨论或争议点。这表明变更被团队认为是直接且必要的改进。

- 暂无高价值评论线程

风险与影响

- 风险：低风险：
 - 回归风险：变更只添加新映射，不修改现有 PR 测试套件的映射，因此不会影响现有的 `/rerun-test` 功能。
 - 配置错误风险：如果映射的 runner 名称与 CI 系统中实际可用的 runner 不匹配，可能导致 `/rerun-test` 命令执行失败。但根据 PR 描述，这些映射源自现有的 `nightly-test-nvidia.yml`，降低了配置错误的风险。
 - 性能与安全风险：无直接影响。
- 影响：影响范围：
 - 对开发者：显著提升了开发体验，现在可以通过 `/rerun-test` 命令方便地重新运行夜间测试，无需手动触发完整 CI 流程。
 - 对系统：扩展了 CI 工具链的功能完整性，使 `/rerun-test` 命令支持更广泛的测试套件类型。
 - 对团队：减少了夜间测试调试的摩擦，可能加快问题排查和验证速度。影响程度：中等，主要影响使用夜间测试的开发者工作流程，不改变核心业务逻辑。

- 风险标记：配置映射风险

关联脉络

- PR #23417 [ci] split stage-c-test-4-gpu-b200 to enable a low-disk runner pool: 同样修改了 slash_command_handler.py 文件，涉及 CI 测试套件和 runner 的配置调整。
- PR #23413 Fix docker build error: 同属基础设施（infra）类别的 PR，关注 CI/ 部署工具的改进。