

PR #30496 完整报告

sgl-project/sglang

[AMD] Gate stage-c on stage-b-test-1-gpu-large-amd

合并时间: 2026-07-08 15:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30496>

执行摘要

- 一句话: 修复 AMD CI 流水线 stage-c 门控缺失
- 推荐动作: 值得合并, 逻辑清晰, 改动小且经过 NVIDIA 工作流验证。可快速合入。

功能与动机

PR body 明确指出: `wait-for-stage-b-amd` 只监控了 `stage-b-test-1-gpu-small-amd` 和 `stage-b-test-2-gpu-large-amd`, 而 `stage-b-test-1-gpu-large-amd` 缺失, 导致该任务失败时 gate 不会快速失败, stage-c 仍然调度, 浪费多 GPU 运行器。

实现拆解

在 `.github/workflows/pr-test-amd.yml` 文件中, `wait-for-stage-b-amd` 任务的 `jobs` 列表中添加了一项:

1. 新增 `{"prefix": "stage-b-test-1-gpu-large-amd", "expected_count": 3}`, 对应矩阵 part: `[0,1,2]`, 与其他 stage-b 任务格式一致。

关键文件:

- `.github/workflows/pr-test-amd.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : 唯一修改的文件, 在 `wait-for-stage-b-amd` 中新增 `stage-b-test-1-gpu-large-amd` 任务监控。

关键符号: 未识别

关键源码片段

`.github/workflows/pr-test-amd.yml`

唯一修改的文件, 在 `wait-for-stage-b-amd` 中新增 `stage-b-test-1-gpu-large-amd` 任务监控。

```
# .github/workflows/pr-test-amd.yml 中的 wait-for-stage-b-amd 任务
jobs: |
[
  {"prefix": "stage-b-test-1-gpu-small-amd", "expected_count": 14},
  # 新增: 监控 1-gpu-large 任务, 预期 3 个 (对应 part: [0,1,2])
  {"prefix": "stage-b-test-1-gpu-large-amd", "expected_count": 3},
  {"prefix": "stage-b-test-2-gpu-large-amd", "expected_count": 2}
]
```

max-wait-minutes: '480'

评论区精华

无 review 讨论。作者在 PR body 中说明了设计选择：本次仅添加 `stage-b-test-1-gpu-large-amd`，其他 stage-b 任务（如 `-nondeterministic`、`-mi35x` 等）有意留作后续处理。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅涉及 CI workflow 配置，不影响任何模型、内核或业务逻辑。若 `expected_count` 配置错误（如与实际矩阵维度不匹配），可能导致门控等待超时或误判，但 3 与现有矩阵 `part: [0,1,2]` 一致，风险可接受。
- 影响：直接影响 AMD PR 测试流水线的 `stage-c` 触发时机，避免因 `stage-b-test-1-gpu-large-amd` 失败导致的资源浪费。对用户无影响，对开发团队而言，提高了 CI 资源利用率和反馈速度。
- 风险标记：CI 配置变更，低风险

关联脉络

- PR #27436 [diffusion] Enable breakable CUDA graph (BCG) for diffusion DiTs: 同仓库 CI workflow 相关 PR，但与本 PR 无直接关联。