PR #22489 完整报告

sgl-project/sglang

[AMD] Replace push trigger with scheduled runs and enable parallel stage execution

合并时间：2026-04-14 13:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22489

执行摘要

本 PR 将 AMD CI 工作流的触发方式从 push 改为每 6 小时调度运行，以减少 main 合并时的冗余运行压力，并引入等待作业实现调度运行中的阶段并行执行。变更优化了资源使用，但需关注并发设置和作业依赖风险，已在 review 中修复关键 bug。

功能与动机

为什么做：根据 PR body，主要动机是“移除 push 触发器以避免每次合并到 main 时的冗余 CI 运行，减少 AMD runner 压力”，同时“添加 schedule 触发器（每 6 小时）以定期验证 main 分支”。这旨在平衡 CI 资源消耗和 main 分支稳定性验证需求。

实现拆解

关键改动点（基于文件 .github/workflows/pr-test-amd.yml）：

1. 触发器变更：将 on: push 替换为 on: schedule，使用 cron 表达式 '0 */6 * * *' 每 6 小时运行。
2. 并发控制：调整 concurrency 组逻辑，使调度运行和 run_all_tests 运行使用唯一组（基于 run_id），避免相互取消；PR 运行共享组以支持新推送取消旧运行。
3. 作业逻辑：
   - 引入 wait-for-stage-a-amd 和 wait-for-stage-b-amd 作业，通过 GitHub API 轮询控制阶段执行顺序。
   - 在 check-changes 作业中添加 continue_on_error 输出，自动为调度运行启用错误继续。
   - 添加条件 if: github.event_name != 'schedule' 使调度运行跳过 call-gate 作业。
4. 运行模式判断：在 run-mode 步骤中，根据 inputs.run_all_tests 或 github.event_name == 'schedule' 设置 run_all_tests=true，确保调度运行执行全量测试。

评论区精华

核心讨论：来自 amd-bot 的 review 评论指出关键 bug：

> “Bug 1 — check-changes will be skipped on schedule runs (Critical)”

该问题源于 check-changes 作业依赖 call-gate，而调度运行中 call-gate 被跳过，导致 check-changes 可能无法执行。讨论结论是通过后续 commit 修复，调整了作业依赖和条件逻辑，确保 check-changes 在调度运行中正常进行。

风险与影响

风险：

- check-changes 作业在调度运行中被跳过的风险（已修复），否则可能导致变更检查遗漏和测试覆盖不足。
- concurrency 设置复杂，错误配置可能引发 PR 运行与调度运行冲突，或调度运行相互取消。
- 并行阶段执行在调度运行中可能增加资源竞争，影响其他 CI 作业性能。

影响：

- 系统：减少 AMD runner 压力，提高资源利用率；调度运行定期验证 main 分支，增强稳定性。
- 团队：CI 运行频率变化，需适应新的调度周期；并行执行加速测试但增加调试复杂度。
- 用户：无直接影响，属内部优化。

关联脉络

与历史 PR 的关系：

- PR 22534（CI 重跑优化）和 PR 22733（GB200 流水线门控）均为 CI 基础设施变更，与本 PR 共同体现团队对 CI 资源管理和触发机制的持续改进趋势。
- PR 21097（AMD MoE 权重填充）涉及 AMD 平台支持，与本 PR 共同影响 AMD CI 的测试覆盖和运行效率。演进方向：显示仓库在 AMD 硬件和 CI 基础设施方面的投入增加，通过调度触发和并行执行优化资源分配，平衡测试覆盖与运行压力。