PR #24614 完整报告

sgl-project/sglang

[AMD] Route PR multimodal tests to MI325

合并时间：2026-05-07 23:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24614

执行摘要

- 一句话：AMD 多模态测试路由到 MI325 并开启并发
- 推荐动作：该 PR 属于基础设施调整，技术深度较低。建议快速合入，但可提醒后续关注 MI325 资源使用率和并发稳定性的监控。

功能与动机

PR body 明确指出 'Route the AMD PR-triggered multimodal jobs to MI325 by default and let their partitions run concurrently'，目的是利用 MI325 更充足的资源，避免 AITER kernel JIT 编译导致的资源耗尽（原注释），并通过并发执行缩短 PR 测试流水线耗时。

实现拆解

1. 修改作业命名模板：在 pr-test-amd.yml 中，针对 multimodal-gen-test-1-gpu-amd 和 multimodal-gen-test-2-gpu-amd 两个作业，将 name 字段中的 runner 选择逻辑从条件判断 (inputs.runner_arch || (github.event_name == 'pull_request' && 'mi300' || 'mi325')) 简化为 inputs.runner_arch || 'mi325'，确保 PR 事件默认使用 mi325 runner。
2. 修改运行器标签：同样在两个作业的 runs-on 字段中，将 runner 选择逻辑替换为固定 fallback 到 'mi325'，使得 PR 触发的 multimodal 测试默认运行在 MI325 而非 MI300。
3. 移除并发限制：删除 max-parallel: 1 配置行，使 matrix 中的多个 partition 可以并行执行，加速整体测试流程。
4. 准确性测试验证：PR 提供了 CI 运行链接（https://github.com/sgl-project/sglang/actions/runs/25490675144），标注测试通过。

关键文件：

- .github/workflows/pr-test-amd.yml（模块 CI 配置；类别 infra；类型 infrastructure）：唯一变更文件，修改了 runner 选择逻辑和并发参数，直接影响 AMD 多模态测试的执行环境与并行度。

关键符号：未识别

关键源码片段

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

唯一变更文件，修改了 runner 选择逻辑和并发参数，直接影响 AMD 多模态测试的执行环境与并行度。

```
# .github/workflows/pr-test-amd.yml 片段
  multimodal-gen-test-1-gpu-amd:
    # 将默认 runner 从 PR 条件判断 (mi300) 改为固定 mi325
    name: ${{ format('multimodal-gen-test-1-gpu-amd (linux-{0}-1gpu-sglang, {1})', inputs.runner_
    arch || 'mi325', matrix.part) }}
    needs: [check-changes, call-gate]
    if: ...
    strategy:
      fail-fast: false
      # 移除 max-parallel: 1，允许 partition 并发执行
      matrix:
        part: [0, 1, 2, 3]
    runs-on: ${{ format('linux-{0}-1gpu-sglang', inputs.runner_arch || 'mi325') }}
    ...

  multimodal-gen-test-2-gpu-amd:
    # 同样修改 name 和 runs-on 的默认值
    name: ${{ format('multimodal-gen-test-2-gpu-amd (linux-{0}-2gpu-sglang, {1})', inputs.runner_
    arch || 'mi325', matrix.part) }}
    needs: [check-changes, call-gate]
    strategy:
      fail-fast: false
      # 移除 max-parallel: 1
      matrix:
        part: [0, 1, 2]
    runs-on: ${{ format('linux-{0}-2gpu-sglang', inputs.runner_arch || 'mi325') }}
```

评论区精华

本 PR 无 review 评论，仅由 bingxche 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。迁移到 MI325 后测试环境可能略有差异，但 PR 已通过关联的 CI 运行验证。并发执行可能暴露资源竞争（如 GPU 显存争用），但 multimodal-gen-test-1-gpu-amd 和 2-gpu-amd 各自使用独立 GPU。若 MI325 资源不足，并发可能导致偶发超时。
- 影响：仅影响 AMD 相关的 PR 触发 CI，尤其是 multimodal 测试。对非 AMD 用户无影响。团队可更早获得测试结果，提升迭代效率。
- 风险标记：并发执行可能引发资源竞争

关联脉络

- 暂无明显关联 PR