PR #23130 完整报告

sgl-project/sglang

[AMD]Fix AMD multimodal-gen-test-2-gpu timeout by adding partition for standalone test

合并时间：2026-04-19 23:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23130

执行摘要

- 一句话：为 AMD 多模态 2-GPU 测试增加分区，解决因单分区运行过多参数化测试导致的超时问题。
- 推荐动作：该 PR 变更简单直接，主要涉及 CI 配置调整，无需深入阅读源码。对于关注 CI 基础设施或 AMD 平台测试稳定性的工程师，可快速浏览以了解分区策略的优化方法。

功能与动机

根据 PR body 的描述，动机是修复 multimodal-gen-test-2-gpu-amd 在分区 0 上持续超时的问题。根本原因在于分区逻辑：总分区数（total_partitions）为 2，其中有一个独立测试文件（test_disagg_server.py），根据公式 parametrized_partitions = total_partitions - len(standalone_files)，只有 1 个分区用于运行所有 22 个参数化测试，导致该分区超时；而另一个分区仅运行独立测试文件（约 11 分钟）。

实现拆解

1. 修改 CI 配置文件：在 .github/workflows/pr-test-amd.yml 中，将 multimodal-gen-test-2-gpu 作业的 part 矩阵从 [0, 1] 改为 [0, 1, 2]，并将 --total-partitions 参数从 2 改为 3。
2. 影响：此变更确保总分区数增加后，参数化测试能被分配到 2 个分区（通过 LPT 调度算法），同时独立测试文件 test_disagg_server.py 仍拥有自己的专属分区，从而平衡负载，避免单个分区因运行过多测试而超时。
3. 配套改动：无其他测试、配置或部署配套改动，仅调整 CI 工作流配置。

关键文件：

- .github/workflows/pr-test-amd.yml（模块 CI 配置；类别 infra；类型 configuration）：这是唯一被修改的文件，直接调整了 AMD CI 测试的分区配置，是解决超时问题的核心。

关键符号：未识别

关键源码片段

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

这是唯一被修改的文件，直接调整了 AMD CI 测试的分区配置，是解决超时问题的核心。

```
# 在 pr-test-amd.yml 的 multimodal-gen-test-2-gpu 作业中
```

```
jobs:
  multimodal-gen-test-2-gpu:
    max-parallel: 1 # 一次运行一个，避免资源耗尽
    matrix:
      runner: [linux-mi325-2gpu-sglang]
      part: [0, 1, 2] # 从 [0, 1] 改为 [0, 1, 2]，总分区数增加到 3
    runs-on: ${{matrix.runner}}
    steps:
      - name: Run multimodal 2-gpu test suite
        run: |
          python3 sglang/multimodal_gen/test/run_suite.py \
            --suite 2-gpu \
            --partition-id ${{ matrix.part }} \
            --total-partitions 3  # 从 2 改为 3，确保参数化测试分配到 2
            个分区，独立测试文件有专属分区
```

评论区精华

Review 中仅有一次批准（由 bingxche 执行），无具体评论或争议点。这表明变更直接明了，团队对修复方案达成共识。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅涉及 CI 工作流配置，不触及核心业务逻辑、性能或安全。潜在风险包括：
  - 配置错误：如果分区数调整不当（例如，增加过多可能导致资源浪费），但当前从 2 到 3 的调整基于明确的计算逻辑（22 个参数化测试需至少 2 个分区），风险可控。
  - CI 稳定性：修改可能影响其他 CI 作业的调度，但仅限于 AMD 2-GPU 测试，影响范围小。
  - 影响：影响范围有限，但直接解决 CI 阻塞问题。
  - 对用户：无直接影响，属于内部 CI 优化。
  - 对系统：修复 AMD 多模态 2-GPU 测试的超时问题，提升 CI 稳定性和运行效率。
  - 对团队：减少因 CI 超时导致的开发中断，加速 PR 合并流程。
  - 风险标记：配置调整

关联脉络

- PR #23045 [AMD] Fix AMD Multimodal Test - skip nvfp4 tests: 同属 AMD 多模态测试 CI 修复，涉及类似配置调整（跳过特定测试），但本 PR 专注于分区优化以解决超时。
- PR #23119 [CI] Add per-job uv venv isolation and upgrade CI version to Cuda 13: 同属 CI 基础设施改进，优化依赖管理和环境一致性，而本 PR 聚焦于测试分区调度。