PR #22395 完整报告

sgl-project/sglang

[CI] Increase stage-c-test-4-gpu-b200 partitions from 4 to 5

合并时间：2026-04-09 07:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22395

执行摘要

本次 PR 将 GitHub Actions 中 stage-c-test-4-gpu-b200 测试套件的分区数量从 4 个增加到 5 个，以解决因测试时间增长导致的 CI 步骤超时问题。变更仅涉及 CI 配置文件，通过降低每个分区的平均执行时间（从 29.2 分钟降至 23.4 分钟），为设置开销提供约 6 分钟缓冲，从而提升 Blackwell B200 GPU 上 4-GPU 测试的稳定性。

功能与动机

为什么做：近期新增的 3 个 LoRA 测试（PR #21466、#21469、#21570）使 stage-c-test-4-gpu-b200 套件总预估时间达到 7010 秒（116.8 分钟）。在原有 4 个分区下，平均每个分区耗时 29.2 分钟，接近 30 分钟步骤超时限制，无法覆盖约 2 分钟的设置开销（依赖安装、验证），导致 分区 2 超时中断。

关键数据：

- 新增 LoRA 测试贡献 620 秒（10.3 分钟）。
- 4 分区时平均耗时 29.2 分钟 / 分区，缓冲不足 1 分钟。
- 5 分区时平均耗时 23.4 分钟 / 分区，缓冲约 6 分钟。

实现拆解

仅修改一个文件：.github/workflows/pr-test.yml。

变更位置	原值	新值	作用
矩阵策略 part	[0, 1, 2, 3]	[0, 1, 2, 3, 4]	增加一个分区索引
运行命令 --auto-partition-size	4	5	匹配新分区数

变更后，测试套件将被均匀分配到 5 个并行作业中执行，每个作业时间压力显著降低。

评论区精华

Review 过程简单直接：

- 审核者 hnyls2002 直接批准，未留下评论。

- 表明变更逻辑清晰，无技术争议，属于常规 CI 优化。

风险与影响

风险分析：

1. 低风险：仅修改 CI 配置，不触及生产代码，回归风险可忽略。
2. 资源消耗：增加一个并行作业可能略微提升 CI 资源使用，但通过避免超时重试，整体效率可能提升。
3. 兼容性：分区逻辑基于现有 run_suite.py 脚本，扩展分区数属于支持范围内操作。

影响评估：

- 对用户：无感知，不影响系统功能或性能。
- 对团队：减少 CI 超时失败，提高测试可靠性，尤其保障 Blackwell B200 GPU 上关键测试的连续执行。
- 对系统：无直接影响，仅优化测试执行策略。

关联脉络

相关 PR：

1. PR #21466、#21469、#21570：新增 LoRA 测试，导致测试时间增长，是本 PR 的直接诱因。
2. PR #22346：通过设置内存限制解决测试超时，与本 PR 同属 CI 优化范畴，展示不同维度的调优手段。
3. PR #22237：通过降低准确度阈值减少 CI 误报，与本 PR 共同体现团队对 CI 稳定性的持续改进。

演进趋势：

- 随着模型测试复杂度增加（如 LoRA、多 GPU、Blackwell 支持），测试时间压力上升，CI 配置需动态调整。
- 本 PR 是典型的“响应式”优化，通过增加分区应对时间增长，未来可能需更系统化的测试时间监控与自动分区策略。