

PR #23417 完整报告

sgl-project/sglang

[ci] split stage-c-test-4-gpu-b200 to enable a low-disk runner pool

合并时间: 2026-04-22 09:33

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23417>

执行摘要

- 一句话: 分割 B200 CI 测试套件以支持低磁盘容量 runner 池, 优化资源分配。
- 推荐动作: 对于 CI 维护者和基础设施工程师, 值得精读以了解测试套件分割策略、runner 标签设计和负载均衡调整; 一般工程师可关注变更对测试运行时间和资源分配的影响, 无需深入代码细节。

功能与动机

PR body 明确指出: 'Splits the per-commit B200 stage-c suite into two so a new B200 runner with limited disk capacity can serve a subset of tests', 目的是优化 CI 资源分配, 避免磁盘瓶颈影响测试效率。

实现拆解

- 修改 GitHub Actions 配置: 在 .github/workflows/pr-test.yml 中, 添加新 job stage-c-test-4-gpu-b200-small, 设置 runs-on 为低磁盘 runner 标签, 并调整矩阵分区大小从 6 减为 3 以平衡负载; 同时扩展 set-runner 步骤输出 b200_low_disk_runner 标签。
- 更新测试文件套件标签: 修改 9 个测试文件 (如 test/registered/4-gpu-models/test_qwen35_fp4_mtp_v2.py) 中的 register_cuda_ci 调用, 将 suite 参数从 "stage-c-test-4-gpu-b200" 改为 "stage-c-test-4-gpu-b200-small"; 3 个低优先级测试文件 (如 test/registered/4-gpu-models/test_nvidia_nemotron_3_super_nvfp4.py) 改为 "nightly-4-gpu-b200" 并设置 nightly=True。
- 注册新套件: 在 scripts/ci/utis/slash_command_handler.py 中添加 stage-c-test-4-gpu-b200-small 到 nvidia_stages 和 CUDA_SUITE_TO_RUNNER 映射, 支持命令行重定向; 在 test/run_suite.py 的 PER_COMMIT_SUITES[HWBackend.CUDA] 中注册新套件名称, 确保测试运行器识别。
- 配套调整: 在 PR body 中详细规划了 runner 标签策略 (如 4-gpu-b200-low-disk), 但未在代码中直接修改, 依赖外部基础设施配置; 变更后需验证套件分割不影响现有 CI 流程。

关键文件:

- .github/workflows/pr-test.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure): 定义了新的 CI job 'stage-c-test-4-gpu-b200-small' 并调整现有 job, 是分割测试套件的核心配置入口。

- scripts/ci/utlis/slash_command_handler.py (模块 CI 工具; 类别 infra; 类型 configuration; 符号 nvidia_stages, CUDA_SUITE_TO_RUNNER) : 更新命令行处理逻辑, 添加新套件到映射表, 确保 /rerun-stage 命令能正确重定向测试。
- test/run_suite.py (模块 测试运行; 类别 test; 类型 configuration; 符号 PER_COMMIT_SUITES) : 注册新测试套件名称到全局套件映射, 确保测试运行器能识别和执行新套件。
- test/registered/4-gpu-models/test_qwen35_fp4_mtp_v2.py (模块 测试覆盖; 类别 test; 类型 test-coverage; 符号 register_cuda_ci) : 示例测试文件, 套件标签从 'stage-c-test-4-gpu-b200' 改为 'stage-c-test-4-gpu-b200-small', 反映分割策略。

关键符号: register_cuda_ci

关键源码片段

test/registered/4-gpu-models/test_qwen35_fp4_mtp_v2.py

示例测试文件, 套件标签从 'stage-c-test-4-gpu-b200' 改为 'stage-c-test-4-gpu-b200-small', 反映分割策略。

```
from sglang.test.ci.ci_register import register_cuda_ci

# 变更前: suite="stage-c-test-4-gpu-b200"
# 变更后: suite="stage-c-test-4-gpu-b200-small", 表示此测试现在属于低磁盘容量 runner 池
register_cuda_ci(est_time=540, suite="stage-c-test-4-gpu-b200-small")

# 其他测试逻辑保持不变, 仅套件标签更新以适配 runner 分割
class TestQwen35FP4MTPV2(ReasoningTokenUsageMixin, CustomTestCase):
    # ... 测试方法实现未改动
```

评论区精华

无 review 评论, 变更由作者 Kangyan-Zhou 直接提交并通过 /rerun-stage 命令验证执行成功, 表明无重大争议。

- 暂无高价值评论线程

风险与影响

- 风险:
 1. 配置错误风险: 新 job stage-c-test-4-gpu-b200-small 在 .github/workflows/pr-test.yml 中的条件逻辑 (if 语句) 或依赖关系若设置不当, 可能导致测试不运行或重复执行。
 2. 测试覆盖失衡: 测试分割可能影响负载均衡, 新套件分区大小调整为 3, 若时间估算不准确, 可能导致部分测试超时或资源浪费。
 3. runner 标签依赖: 变更依赖于外部 runner 池正确配置 4-gpu-b200-low-disk 等标签, 若标签未及时应用, 新 job 可能无法分配到 runner。

4. 回归风险：修改多个测试文件的 `register_cuda_ci` 调用，若套件名称拼写错误或未在 `run_suite.py` 中注册，可能导致测试被忽略。

- 影响：

1. 对用户：无直接影响，不涉及功能或性能变更。
2. 对开发团队：CI 测试更灵活，低磁盘 runner 可参与部分测试，可能减少队列等待时间，提高开发效率；但增加了 CI 配置复杂度，需团队熟悉新的套件分割策略。
3. 对系统：优化硬件资源使用，提升测试吞吐量；长期可能降低基础设施成本，但需维护额外的 runner 池和标签逻辑。 - 风险标记：配置依赖变更，测试覆盖调整，CI 流程复杂度增加

关联脉络

- PR #23413 Fix docker build error: 同为基础设施变更，涉及 CI/CD 配置修复，展示团队对基础设施优化的持续关注。
- PR #23409 feat: enable SGLANG_PATCH_TOKENIZER by default: 涉及环境变量默认值调整，虽非直接相关，但体现 CI 环境配置的演进脉络。