# PR #22910 完整报告

sgl-project/sglang

ci: re-enable fp8 nightly benchmark configs

合并时间：2026-04-16 06:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22910

## PR 22910 分析报告

### 执行摘要

本 PR 恢复了先前为测试 AI 日志分析器而临时禁用的 GB200 FP8 夜间基准测试配置，确保 CI 系统恢复完整的测试覆盖。这是一个基础设施维护操作，直接影响夜间基准测试的配置范围，风险较低但需要关注配置正确性。

### 功能与动机

根据 PR 正文描述，这些 FP8 配置在 PR #22899 中被临时注释掉，目的是为了隔离已知失败的 FP4 任务以测试新引入的 AI 日志分析器。现在日志分析器已经通过端到端验证（能够自动提 issue 并生成报告），因此需要恢复完整的夜间测试覆盖，确保 FP8 推理路径在持续集成中得到持续监控。

### 实现拆解

#### 1. 配置恢复

唯一的变更文件是 scripts/ci/slurm/nightly-configs.yaml，该文件定义了夜间基准测试的各种配置。本 PR 将先前被注释掉的 dsr1-fp8-gb200-dynamo-sglang 配置块取消注释，使其重新生效。

#### 2. 配置内容分析

恢复的配置定义了 DeepSeek-R1-0528 模型在 GB200 runner 上的 FP8 基准测试：

```
dsr1-fp8-gb200-dynamo-sglang:
  model: deepseek-ai/DeepSeek-R1-0528
  model-prefix: dsr1
  runner: gb200
  precision: fp8 # 使用FP8精度
  framework: dynamo-sglang # 使用dynamo-sglang框架
  multinode: true # 多节点测试
  disagg: true # 分离式架构支持
  seq-len-configs:
    - isl: 1024
      osl: 1024
      search-space:
        - conc-list: [1024, 2048, 4096, 6144] # 测试不同并发数
```

```
      config_file: recipes/gb200-fp8/1k1k/max-tpt.yaml # 最大吞吐量配置
  - conc-list: [4096]
    config_file: recipes/gb200-fp8/1k1k/ultra-tpt.yaml # 超高吞吐量配置
```

### 3. 基础设施影响

此变更直接影响 CI/CD 流水线，确保夜间基准测试包含 FP8 配置。配置中引用了外部 *recipe* 文件（位于 srt-slurm 仓库），这些文件定义了具体的测试参数和性能目标。

## 评论区精华

该 PR 没有 review 评论，表明这是一个相对简单且无争议的基础设施恢复操作，符合团队对 CI 配置变更的常规处理流程。

## 风险与影响

### 风险

1. 配置错误：如果恢复的配置中存在路径错误或参数不匹配，可能导致夜间测试失败。
2. 外部依赖：配置引用的外部 *recipe* 文件如果发生变化，可能影响测试结果的一致性。
3. 资源消耗：重新启用 FP8 测试会增加 CI 运行时间和计算资源消耗。

### 影响

1. CI 系统：夜间基准测试将重新包含 FP8 配置，提供更全面的性能数据，有助于检测 FP8 相关的性能回归。
2. 开发团队：获得更完整的基准测试报告，但需要关注可能增加的 CI 失败率。
3. 系统稳定性：间接提高 FP8 推理路径的持续验证，增强系统整体稳定性。

## 关联脉络

本 PR 与 PR #22899 直接相关，后者在测试 AI 日志分析器时临时禁用了这些 FP8 配置。从近期历史 PR 看，该仓库持续优化 CI 基础设施，包括日志分析器（#22899、#22903、#22859）、基准测试配置（#22854）和代码质量工具（#22912）。本 PR 是这一系列基础设施改进的延续，体现了团队对 CI 可靠性和测试覆盖的重视。