PR #23785 完整报告

sgl-project/sglang

chore: update CI test est_time values

合并时间：2026-04-27 11:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23785

执行摘要

- 一句话：更新 CI 测试预估时长以优化并行调度
- 推荐动作：该 PR 属于纯运维类更新，无设计决策需要关注。但对 CI 调度策略感兴趣的读者可以留意其测量和更新流程，以及 est_time=0 的遗留问题。建议后续补充校验逻辑，避免零值。

功能与动机

为了保持 LPT 负载均衡算法对测试分区调度的准确性，需要根据实际执行时间更新 est_time。PR body 说明数据来源为最近 15 次成功执行（来自 main 分支上的定时 PR Test 运行）的 90 百分位数。

实现拆解

该 PR 共修改 268 个文件，每个文件仅改动一行：调用 register_cuda_ci 或 register_cpu_ci 时传递的 est_time 参数。在自动汇总运行数据后，将每个测试的预估时长更新为 90 百分位实测值。典型变化如 test_awq.py 从 950 秒降至 226 秒，test_disaggregation_hybrid_attention.py 从 317 秒升至 695 秒。因未改动测试逻辑本身，不影响测试正确性。

关键文件：

- test/registered/unit/tools/test_get_version_tag.py（模块 单元测试；类别 test；类型 test-coverage）：该文件收到 review 评论，指出 est_time=0 可能导致的调度问题，是讨论焦点。
- test/registered/4-gpu-models/test_gpt_oss_4gpu.py（模块 多 GPU 测试；类别 test；类型 test-coverage）：展示了典型的 est_time 更新，且变化较为显著（H100 从 328 升至 392，B200 从 740 降至 584）。
- test/registered/4-gpu-models/test_qwen35_fp4_mtp_v2.py（模块 Qwen3.5 测试；类别 test；类型 test-coverage）：Qwen3.5 FP4 MTP 测试的 est_time 从 540 降至 422，降幅约 22%，反映实测性能改善。
- test/registered/4-gpu-models/test_qwen35_fp4_triton.py（模块 Qwen3.5 测试；类别 test；类型 test-coverage）：Qwen3.5 FP4 Triton 测试的 est_time 从 720 降至 563，降幅约 22%。
- test/registered/8-gpu-models/test_deepseek_v32_indexcache.py（模块 DeepSeek 测试；类别 test；类型 test-coverage）：DeepSeek V3.2 indexcache 测试的 est_time 从

354 升至 492，增幅约 39%，反映实际运行耗时增加。

关键符号：未识别

关键源码片段

test/registered/unit/tools/test_get_version_tag.py

该文件收到 review 评论，指出 est_time=0 可能导致的调度问题，是讨论焦点。

```
# test/registered/unit/tools/test_get_version_tag.py
import os
import sys
import tempfile

from sglang.test.ci.ci_register import register_cpu_ci

# 注意：est_time 设为 0 可能导致 LPT 调度忽略该测试
register_cpu_ci(est_time=0, suite="stage-a-test-cpu")

class TestGetVersionTag(unittest.TestCase):
    # 测试内容未变动
    pass
```

test/registered/4-gpu-models/test_gpt_oss_4gpu.py

展示了典型的 est_time 更新，且变化较为显著（H100 从 328 升至 392，B200 从 740 降至 584）。

```
# test/registered/4-gpu-models/test_gpt_oss_4gpu.py
import unittest
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.gpt_oss_common import BaseTestGptOss

# 基于最近 15 次运行的 90 百分位更新预估时间
register_cuda_ci(est_time=392, suite="stage-c-test-4-gpu-h100") # 旧值 328
register_cuda_ci(est_time=584, suite="stage-c-test-4-gpu-b200-small") # 旧值 740

class TestGptOss4Gpu(BaseTestGptOss):
    def test_bf16_120b(self):
        self.run_test(model_variant="120b", quantization="bf16",
                      expected_score_of_reasoning_effort={"low": 0.58},
                      other_args=["--tp", "4", "--cuda-graph-max-bs", "200"])

    def test_mxfp4_120b(self):
        self.run_test(model_variant="120b", quantization="mxfp4",
                      expected_score_of_reasoning_effort={"low": 0.58},
                      other_args=["--tp", "4", "--cuda-graph-max-bs", "200"])

if __name__ == "__main__":
    unittest.main()
```

评论区精华

1 条来自 gemini-code-assist[bot] 的评论指出，test_get_version_tag.py 的 est_time 被设为 0 可能导致 LPT 调度算法将其视为无耗时测试，产生分区扭曲，建议取最小值为 1。该 PR 未采纳该建议，最终合入版本仍为 0。

- est_time=0 可能导致 LPT 调度问题 (correctness): 建议未采纳，PR 合入版本仍为 0。

风险与影响

- 风险：主要风险在于 est_time 数据本身可能存在噪声。若某次运行异常慢或快，90 百分位数可能不能代表正常耗时，导致调度偏差。此外 est_time=0 的测试会被调度器忽略，可能与其他任务冲突。但因改动仅为量值调整，不涉逻辑，风险较低。
- 影响：直接影响 CI 作业的划分均衡性，可能降低总体排队时间或减少个别作业超时。对用户功能无影响。对团队而言，需持续关注调度效果以确保调整有效。
- 风险标记：est_time 零值风险，数据噪声可能影响调度

关联脉络

- 暂无明显关联 PR