PR #21896 完整报告

sgl-project/sglang

fix(ci): update est_time for 57 tests based on runtime analysis

合并时间：2026-04-02 11:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21896

执行摘要

基于实际 CI 运行时数据，本 PR 更新了 57 个测试文件的预估时间 est_time，旨在优化 LPT 分区算法的负载平衡，减少 CI 作业完成时间差异和资源浪费。

功能与动机

PR body 指出，LPT 分区算法依赖准确的 est_time 来分配测试到 CI 作业，但当前估计偏差导致分区失衡（部分作业 11 分钟完成，其他 25 分钟）和预算浪费。通过分析 6 次近期 CI 运行数据，识别出 43 个下估计和 17 个过估计测试，驱动了此次更新。

实现拆解

变更涉及两个主要方面：一是批量更新 register_cuda_ci 和 register_cpu_ci 调用中的 est_time 参数，以匹配平均实际耗时；二是在 test_disaggregation_decode_offload.py 中将测试跳过方式从 @unittest.skipIf(is_in_ci()) 改为 disabled 参数，避免占用分区时间。示例更新如下表：

测试文件	旧 est_time	新 est_time	变化原因
test_hybrid_attn_backend.py	200 秒	350 秒	实际耗时更高，下估计
test_fp8kv_triton.py	520 秒	80 秒	实际耗时低，过估计

评论区精华

PR 未收到任何 review 评论，所有变更由作者直接提交和合并，表明团队对数据驱动更新的信任或缺乏深入讨论。

风险与影响

风险包括新估计可能不适应未来测试变化、人为错误导致分区问题，以及 disabled 参数变更可能影响非 CI 环境测试。影响限于 CI 系统：预期改善负载平衡，减少资源浪费，但对用户无直接影响；团队需监控后续运行验证效果。

关联脉络

与近期 CI 相关 PR 如 #21897（增加超时）、#21898（移除崩溃测试）形成序列，反映团队持续优化 CI 测试配置和效率的趋势。本 PR 的数据驱动方法为未来类似更新提供了基础。