PR #24713 完整报告

sgl-project/sglang

[HiCache] ci: lower est_time for test_hicache_spec_file_storage

合并时间：2026-05-09 15:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24713

执行摘要

- 一句话：降低 HiCache 测试预估时间以平衡 CI 分区
- 推荐动作：可快速合入，无需深入精读。但值得关注 run_suite.py 分区器如何利用 est_time，可为其他测试的类似优化提供参考。

功能与动机

PR body 指出 test_hicache_spec_file_storage 注册的 est_time=600s 远超实际运行时间（~240s），导致 LPT 分区器将该测试单独分配一个分区，而其他分区过载，延长了阶段关键路径。

实现拆解

1. 在 test/registered/hicache/test_hicache_spec_file_storage.py 中将 register_cuda_ci(est_time=600, ...) 改为 register_cuda_ci(est_time=200, ...)。
2. 该值仅影响分区规划，不涉及超时或门控逻辑，无其他行为变更。

关键文件：

- test/registered/hicache/test_hicache_spec_file_storage.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，修改了 register_cuda_ci 的预估时间参数。

关键符号：未识别

关键源码片段

test/registered/hicache/test_hicache_spec_file_storage.py

唯一变更文件，修改了 register_cuda_ci 的预估时间参数。

```
# 在文件顶部，注册 CI 测试并设置预估时间（单位：秒）
# 基于实际观察到的约 240s 墙钟时间，将过高的 600s 降至 200s
# 以改善 LPT 分区器的负载均衡
register_cuda_ci(est_time=200, suite="stage-b-test-1-gpu-large")

@unittest.skipIf(is_hip(), "HiCache + EAGLE3 file-storage loadback e2e is CUDA-only.")
class TestHiCacheSpecFileStorage(CustomTestCase):
    model = DEFAULT_TARGET_MODEL_EAGLE3
    draft_model = DEFAULT_DRAFT_MODEL_EAGLE3
```

```
    # ... 其余代码不变
```

评论区精华

无 review 讨论。审核者 Kangyan-Zhou 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改一个元数据字段，不影响测试逻辑或系统行为。若实际运行时间再次变化，可能导致分区不平衡，但可通过再次调整缓解。
- 影响：影响局限于 CI 阶段 stage-b-test-1-gpu-large 内该测试所在的分区负载，预期减少分区空闲时间，提升整体 CI 效率。
- 风险标记：极低风险

关联脉络

- 暂无明显关联 PR