PR #26838 完整报告

sgl-project/sglang

Skip flaky mamba extra_buffer disagg test

合并时间：2026-05-31 15:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26838

执行摘要

- 一句话：跳过 CI 中 flaky 的 Mamba extra_buffer 测试
- 推荐动作：该 PR 为临时权宜之计，变更微小且合理。建议阅读者关注关联问题（PR#15829）的修复进展，并在修复后及时恢复测试。

功能与动机

测试 TestDisaggregationHybridAttentionMambaExtraBuffer.test_gsm8k 在 CI 中不稳定，分数在重试时从 0.82 降至 0.70（隔离运行正常 0.90+），阻塞了无关 PR。跟踪上游回归问题（关联 PR#15829 comment）。采用与已有 TestDisaggregationHybridAttentionGDN 相同的跳过策略，待底层修复后再恢复。

实现拆解

1. 定位 flaky 测试类：在 test/registered/disaggregation/test_disaggregation_hybrid_attention.py 中，TestDisaggregationHybridAttentionMambaExtraBuffer 类定义了 test_gsm8k 测试，使用 nvidia/NVIDIA-Nemotron-Nano-9B-v2 模型和 extra_buffer Mamba 调度策略启动 PD 拆分服务器，并评估 GSM8K 准确率。
2. 添加跳过装饰器：在类定义前添加 @unittest.skipIf(is_in_ci(), "Temporarily disable the flaky test.")，当 is_in_ci() 返回 True 时跳过整个测试类。该模式与同一文件中已有 TestDisaggregationHybridAttentionGDN 的跳过方式一致。
3. 无其他配套变更：仅测试文件修改，无配置或部署调整。

关键文件：

- test/registered/disaggregation/test_disaggregation_hybrid_attention.py（模块 PD 拆分测试；类别 test；类型 test-coverage）：唯一修改的文件，添加了 @unittest.skipIf 装饰器跳过 flaky 测试类。

关键符号：未识别

关键源码片段

test/registered/disaggregation/test_disaggregation_hybrid_attention.py

唯一修改的文件，添加了 @unittest.skipIf 装饰器跳过 flaky 测试类。

```
# 跳过 flaky 测试类，避免 CI 阻塞。上游回归跟踪：https://github.com/sgl-project/sglang/pull/
```

```
15829#issuecomment-4586094495
@unittest.skipIf(is_in_ci(), "Temporarily disable the flaky test.")
class TestDisaggregationHybridAttentionMambaExtraBuffer(PDDisaggregationServerBase):
    @classmethod
    def setUpClass(cls):
        super().setUpClass()
        cls.model = "nvidia/NVIDIA-Nemotron-Nano-9B-v2"
        # ... 服务器启动逻辑保持不变 ...

    def test_gsm8k(self): # 该测试在 CI 中不再执行
        args = SimpleNamespace(
            base_url=self.base_url,
            model=self.model,
            eval_name="gsm8k",
            api="completion",
            max_tokens=512,
            num_examples=200,
            num_threads=128,
        )
        metrics = run_eval(args)
        print(f"Evaluation metrics: {metrics}")
        self.assertGreater(metrics["score"], 0.87)
```

评论区精华

无 reviewer 讨论（仅自动 bot 评论）。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险：低。仅跳过 CI 中的 flaky 测试，不影响任何生产代码或其他测试。
  - 功能覆盖丢失：低。该测试因上游回归已不可靠，跳过不会掩盖新引入的 bug，且隔离运行仍可手动验证。
  - 长期风险：需确保上游修复后及时移除跳过装饰器，否则可能遗漏回归。
- 影响：
  - 用户影响：无。
  - 系统影响：CI 流水线不再因该 flaky 测试失败而阻塞，提升其他 PR 的合并效率。
  - 团队影响：减少 CI 噪音，避免开发者反复重试。
  - 风险标记：测试覆盖缺失

关联脉络

- PR #15829 [Mamba] Fix extra_buffer disagg test flakiness: PR body 中提及该上游回归问题正在跟踪，是本 PR 跳过测试的根本原因。