PR #26281 完整报告

sgl-project/sglang

[CI] Enable EPD CI for EPD architecture enhancements

合并时间：2026-05-25 23:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26281

执行摘要

- 一句话：将 EPD CI 测试移至 base-c stage
- 推荐动作：

功能与动机

PR 说明 'We are enhancing the EPD architecture, a nightly CI test is not enough to guard the correctness of the incoming following PRs, so I added it back to the base-c test.' 关联 Issue #24945 记录 EPD 性能 / 架构 / 可观测性增强的完整 RFC。

实现拆解

1. 测试注册调整（test/registered/disaggregation/test_epd_disaggregation.py）：将 register_cuda_ci 调用从 suite="nightly-4-gpu", nightly=True 改为 stage="base-c", runner_config="4-gpu-h100"，使 EPD 测试在常规 CI base-c 阶段以 4-GPU H100 runner 执行。
2. 源码适配（python/sglang/srt/disaggregation/encode_receiver.py）：在 WaitingImageRequest.__init__ 和 _process_waiting_requests 中新增 model_type 参数的传递与存储，为后续 EPD 架构改进准备接口。
3. 配套修复：第二个 commit 修复了首次 CI 运行中的问题。

关键文件：

- test/registered/disaggregation/test_epd_disaggregation.py（模块 测试 CI；类别 test；类型 test-coverage）：EPD 测试注册配置变更，将测试从 nightly 移至 base-c CI stage
- python/sglang/srt/disaggregation/encode_receiver.py（模块 EPD 编码接收；类别 source；类型 core-logic）：新增 model_type 参数传递，为 EPD 架构增强准备

关键符号：未识别

关键源码片段

python/sglang/srt/disaggregation/encode_receiver.py

新增 model_type 参数传递，为 EPD 架构增强准备

```
# WaitingImageRequest.__init__ 新增 model_type 参数
class WaitingImageRequest:
```

```
    def __init__(
        self,
        rid: str,
        recv_req: TokenizedGenerateReqInput,
        mm_processor,
        encoder_urls,
        model_type, # 新增：记录模型类型，供后续 EPD 逻辑使用
        host_name,
        receive_count,
    ):
        # ... 原有初始化 ...
        self.model_type = model_type # 存储模型类型
        # ...

# _process_waiting_requests 中传递 model_type
waiting_req = waiting_cls(
    rid=recv_req.rid,
    recv_req=recv_req,
    mm_processor=self.mm_processor,
    encoder_urls=self.encode_urls,
    model_type=self.model_type, # 新增传递
    host_name=self.hostname,
    receive_count=self.tp_size,
)
```

评论区精华

Review 评论指出 PR 描述称 change 'lightweight'，但实际启用了 4-GPU 测试而跳过了 3-GPU 测试，存在逻辑不一致。建议更新估计执行时间以确保准确性。作者未进一步回应。

- CI 测试配置逻辑一致性 (design): 作者未回应，CI 已成功运行，问题未完全解决。

风险与影响

- 风险：
- 影响：
  - 风险标记：CI 配置不一致，仅 2 个文件变更

关联脉络

- PR #24945 [RFC] SGLang EPD Performance / Architecture / Observability Enhancements: 关联 Issue，记录 EPD 架构增强的完整 RFC，本 PR 是其 CI 配套。
- PR #26295 Refactor HiCache stack dispatch into strategies: 同为 CI/ 测试注册调整类 PR，涉及测试配置变更。