

PR #28736 完整报告

sgl-project/sglang

[AMD] register 3 tests to stage-b-test-1-gpu-large-amd (batch-6)

合并时间: 2026-06-20 06:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28736>

执行摘要

- 一句话: AMD CI 新增 3 个测试注册
- 推荐动作: 该 PR 是纯粹的 CI 配置变更, 无源码逻辑改动, 适合快速合入。

功能与动机

作为 NV→AMD CI 覆盖审计的批次 6, 将经过验证的、后端无关或 Triton 路径的 CUDA 测试注册到 AMD CI 中, 以提升 AMD 平台测试覆盖。PR body 详细说明了已通过和未通过的测试及原因。

实现拆解

1. 调整导入: 在每个受影响的测试文件中, 将 `from sglang.test.ci.ci_register import register_cuda_ci` 修改为 `from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci`, 以导入 `register_amd_ci` 函数。
2. 添加 AMD 注册: 在每个文件中原有 `register_cuda_ci` 调用之后, 新增 `register_amd_ci` 调用, 指定预估执行时间和目标套件名称 (`stage-b-test-1-gpu-large-amd`)。
3. 文件清单:
 - `test/registered/lora/test_lora_update.py`: 注册 LoRA 更新测试, AMD 预估时间 730s。
 - `test/registered/attention/test_normal_decode_set_metadata.py`: 注册 `decode metadata` 测试, AMD 预估时间 17s。
 - `test/registered/unit/spec/test_resolve_swa_kv_pool.py`: 注册 SWA KV 池解析测试, AMD 预估时间 8s。

关键文件:

- `test/registered/lora/test_lora_update.py` (模块 LoRA; 类别 test; 类型 test-coverage) : 注册 LoRA 更新测试到 AMD CI, 该测试真实运行 Llama-3.1-8B 模型, 是 3 个测试中最重要的一個。
- `test/registered/attention/test_normal_decode_set_metadata.py` (模块 注意力机制; 类别 test; 类型 test-coverage) : 注册 fused Triton decode metadata kernel 测试到 AMD CI, 验证 page size 1/16/64 及 SWA 场景。
- `test/registered/unit/spec/test_resolve_swa_kv_pool.py` (模块 KV 缓存; 类别 test; 类型 test-coverage) : 注册 SWA KV pool 解析单元测试到 AMD CI, 属于纯逻辑测试 (

mocked) 。

关键符号：未识别

关键源码片段

[test/registered/lora/test_lora_update.py](#)

注册 LoRA 更新测试到 AMD CI，该测试真实运行 Llama-3.1-8B 模型，是 3 个测试中最重要中的一个。

```
# 注册 CUDA CI
register_cuda_ci(
    est_time=487,
    stage="base-b",
    runner_config="1-gpu-large",
)
# 注册 AMD CI
register_amd_ci(
    est_time=730, # AMD 预估执行时间，约 CUDA 1.5 倍
    suite="stage-b-test-1-gpu-large-amd",
)
```

评论区精华

PR 仅有一条 bot 评论和一条 approve 评论，没有实质性讨论。

- AMD bot CI status (question): amd-bot 回复指出 PR 缺少 run-ci-extra 标签导致额外工作流失败，但这些失败不相关，且 AMD 测试尚未在 PR CI 中运行。

风险与影响

- 风险：风险极低。变更仅限于 CI 注册调用，不修改任何业务逻辑。所有测试已在两条 AMD 流水线上通过，回退成本很小。
- 影响：对用户和系统无直接影响。对团队而言，AMD CI 覆盖范围扩大，有助于及早发现 AMD 平台上的回归问题。
- 风险标记：暂无

关联脉络

- PR #28739 refactor(runner): move kernel warmup into the shared runner lifecycle (warmup()): 同为 CI 基础设施改进，但无关。