

PR #28967 完整报告

sgl-project/sglang

[AMD] Register 7 JIT kernel unit tests for AMD nightly CI

合并时间: 2026-06-25 16:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28967>

执行摘要

- 一句话: 为 7 个 JIT 内核测试增加 AMD 夜间 CI 注册
- 推荐动作: 该 PR 是标准的测试基础设施扩展, 无技术复杂性, 不值得精读。但可作为参考案例: 如何在保持 CUDA 注册的同时, 为 AMD 平台添加并行测试覆盖, 且不破坏现有 CI。

功能与动机

在 AMD GPU (MI35x / gfx950) 上持续验证 JIT 内核的正确性, 防止退化。PR body 明确指出这些测试在 ROCm 7.0 和 7.2 上均通过, 并提供了详细的运行时数据表格和 CI 任务链接。

实现拆解

1. 导入调整: 在每个目标测试文件中, 将 `from sglang.test.ci.ci_register import register_cuda_ci` 改为 `from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci`。
2. 新增注册行: 在现有的 `register_cuda_ci(...)` 行之后, 添加一行:
`register_amd_ci(est_time=<具体预估时间>, suite="nightly-amd-kernel-1-gpu", nightly=True)`。
3. 验证: 通过 `run_suite.py --hw amd --suite nightly-amd-kernel-1-gpu --nightly` 在 ROCm 7.0 和 7.2 上实际运行整个套件, 所有 7 个测试均绿色通过。

关键文件:

- `test/registered/jit/minimax/test_minimax_decode_topk.py` (模块 测试注册; 类别 test; 类型 test-coverage) : 新增 `register_amd_ci` 导入和注册行, 将 MiniMax-M3 decode topk 测试纳入 AMD 夜间 CI。
- `test/registered/jit/minimax/test_minimax_store_kv_index.py` (模块 测试注册; 类别 test; 类型 test-coverage) : 同上, 为 MiniMax-M3 KV 索引存储测试添加 AMD 夜间注册。
- `test/registered/jit/test_deepseek_v4_compress_state_runtime_shapes.py` (模块 测试注册; 类别 test; 类型 test-coverage) : 为 DeepSeek-V4 压缩状态形状测试添加 AMD 夜间注册, 该测试参数规模较大 (90 个 case), `est_time` 设为 25 秒。
- `test/registered/jit/test_fused_store_index_cache.py` (模块 测试注册; 类别 test; 类型 test-coverage) : 为 fused store index cache 测试添加 AMD 夜间注册, 注意该文件原有 CUDA 注册包含两个 (一个标准、一个夜间), AMD 注册仅需一个。

- test/registered/jit/test_fused_verify_triton_gdn.py (模块 测试注册; 类别 test; 类型 test-coverage) : 为 fused verify triton GDN 测试添加 AMD 夜间注册。
- test/registered/jit/test_kvcacheio_asymmetric.py (模块 测试注册; 类别 test; 类型 test-coverage) : 为 KV cache IO 不对称测试添加 AMD 夜间注册。
- test/registered/jit/test_mla_kv_pack_quantize_fp8.py (模块 测试注册; 类别 test; 类型 test-coverage) : 为 MLA KV pack 量化 FP8 测试添加 AMD 夜间注册。

关键符号: 未识别

关键源码片段

test/registered/jit/minimax/test_minimax_decode_topk.py

新增 `register_amd_ci` 导入和注册行, 将 MiniMax-M3 decode topk 测试纳入 AMD 夜间 CI。

```
# 变更仅涉及导入和注册行, 其余代码不变
from sglang.jit_kernel.minimax_decode_topk import minimax_decode_topk
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 原有 CUDA 注册保持不变
register_cuda_ci(est_time=40, suite="base-b-kernel-unit-1-gpu-large")
register_cuda_ci(est_time=40, suite="base-b-kernel-unit-1-gpu-b200")
# 新增 AMD 夜间注册, est_time 基于实际运行耗时
register_amd_ci(est_time=15, suite="nightly-amd-kernel-1-gpu", nightly=True)
```

评论区精华

无实质性 review 讨论。审核人 HaiShaw 直接批准, 无任何评论。仅有 amd-bot 的 CI 状态评论确认所有 CI 失败与本 PR 无关。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。变更纯粹是测试注册的添加, 不涉及任何生产代码、工作流或依赖。唯一风险是这些测试在 AMD 环境下可能存在后续回归, 但已被夜间 CI 覆盖, 且 PR 已手动验证通过。
- 影响: 影响范围: 仅 AMD 夜间 CI。影响程度: 低。7 个 JIT 内核测试将自动在 AMD nightly 流水线中运行, 有助于及早发现 AMD 平台上的内核退化。对 CUDA、MUSA 等其他平台无影响。
- 风险标记: 仅测试注册变更, 无生产代码影响

关联脉络

- PR #27644 [AMD] initial JIT kernel migration for ROCm: 本 PR 中部分测试 (test_fused_store_index_cache, test_mla_kv_pack_quantize_fp8 等) 首次在 AMD 上验证通过, 是该迁移工作的后续完善。