PR #24987 完整报告

sgl-project/sglang

[AMD] Run jit kernel PR test through run_suite.py register mechanism

合并时间：2026-05-13 17:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24987

执行摘要

- 一句话：AMD JIT kernel 测试接入 run_suite.py 注册机制
- 推荐动作：建议合并。这是一次良好的基础设施统一化变更，提高了可维护性和可扩展性。

功能与动机

AMD JIT kernel 测试目前使用硬编码的 pytest 命令，无法通过注册机制扩展。PR body 指出「makes it inconsistent with NVIDIA's equivalent job and hard to extend — adding a new jit kernel test to AMD CI requires editing the workflow file rather than just adding a register_amd_ci(...) call to the test」。

实现拆解

1. 注册测试：在 python/sglang/jit_kernel/tests/test_store_cache.py 中添加 register_amd_ci(est_time=55, suite="jit-kernel-unit-test-amd")，与已有的 register_cuda_ci 并列。
2. 添加 suite 到配置：在 test/run_suite.py 的 PER_COMMIT_SUITES[HWBackend.AMD] 列表中插入 "jit-kernel-unit-test-amd"，使其被 run_suite.py 识别。
3. 更新工作流：修改 .github/workflows/pr-test-amd.yml 和 .github/workflows/pr-test-amd-rocm720.yml，将原来的 pytest -q python/sglang/jit_kernel/tests/test_store_cache.py 替换为 run_suite.py --hw amd --suite jit-kernel-unit-test-amd，并调整工作目录到 test 目录。行为不变，但未来添加新测试只需注册即可。

关键文件：

- python/sglang/jit_kernel/tests/test_store_cache.py（模块 测试注册；类别 test；类型 test-coverage）：添加了 register_amd_ci 调用，将测试注册到 AMD CI suite 中。
- test/run_suite.py（模块 测试编排；类别 test；类型 test-coverage）：在 AMD 的 PER_COMMIT_SUITES 列表中添加了 'jit-kernel-unit-test-amd' suite。
- .github/workflows/pr-test-amd.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：将硬编码的 pytest 命令替换为 run_suite.py 调用。
- .github/workflows/pr-test-amd-rocm720.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：同样替换硬编码的 pytest 命令。

关键符号：未识别

关键源码片段

python/sglang/jit_kernel/tests/test_store_cache.py

添加了 register_amd_ci 调用，将测试注册到 AMD CI suite 中。

```
import itertools
import sys
import pytest
import torch
from sglang.jit_kernel.kvcache import can_use_store_cache, store_cache
from sglang.jit_kernel.utils import get_ci_test_range
# 同时导入 register_amd_ci 和 register_cuda_ci
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 已有的 CUDA 注册
register_cuda_ci(est_time=28, suite="stage-b-kernel-unit-1-gpu-large")
register_cuda_ci(est_time=120, suite="nightly-kernel-1-gpu", nightly=True)
# 新增的 AMD 注册，suite 名为 "jit-kernel-unit-test-amd"
register_amd_ci(est_time=55, suite="jit-kernel-unit-test-amd")

BS_LIST = [2**n for n in range(0, 15)]
BS_LIST += [x + 1 + i for i, x in enumerate(BS_LIST)]
BS_LIST = get_ci_test_range(BS_LIST, [1, 9, 256, 16399])
HIDDEN_DIMS = get_ci_test_range([64, 128, 256, 512, 1024, 96, 98, 100], [64, 512, 1024, 98]
)
CACHE_SIZE = 1024 * 1024
DTYPE = torch.bfloat16
DEVICE = "cuda"

@pytest.mark.parametrize("batch_size,element_dim", list(itertools.product(BS_LIST, HIDDEN_
DIMS)))
def test_store_cache(batch_size: int, element_dim: int) -> None:
    # 测试 store_cache 的正确性
    k = torch.randn((batch_size, element_dim), dtype=DTYPE, device=DEVICE)
    v = torch.randn((batch_size, element_dim), dtype=DTYPE, device=DEVICE)
    k_cache = torch.randn((CACHE_SIZE, element_dim), dtype=DTYPE, device=DEVICE)
    v_cache = torch.randn((CACHE_SIZE, element_dim), dtype=DTYPE, device=DEVICE)
    indices = torch.randperm(CACHE_SIZE, device=DEVICE)[:batch_size]
    store_cache(k, v, k_cache, v_cache, indices)
    assert torch.all(k_cache[indices] == k)
    assert torch.all(v_cache[indices] == v)
```

test/run_suite.py

在 AMD 的 PER_COMMIT_SUITES 列表中添加了 'jit-kernel-unit-test-amd' suite。

```
# ... 其他 backend 的 suite 列表 ...
HWBackend.AMD: [
```

```
    "stage-a-test-1-gpu-small-amd",
    "stage-b-test-1-gpu-small-amd",
    "stage-b-test-1-gpu-small-amd-nondeterministic",
    "stage-b-test-1-gpu-small-amd-mi35x",
    "stage-b-test-large-8-gpu-35x-disaggregation-amd",
    "stage-b-test-1-gpu-large-amd",
    "stage-b-test-2-gpu-large-amd",
    "jit-kernel-unit-test-amd", # 新增的 suite，对应注册的 AMD JIT kernel 测试
    "stage-c-test-4-gpu-amd",
    "stage-c-test-large-8-gpu-amd",
    "stage-c-test-large-8-gpu-amd-mi35x",
],
# ... 其他 backend ...
```

评论区精华

无 review 评论。两位 reviewer（bingxche, HaiShaw）均直接 approve，无需修改。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅涉及测试注册和工作流命令替换，行为不变（仍只运行同一个测试文件）。但需注意：
  - 工作流中工作目录从 /sglang-checkout 改为 /sglang-checkout/test，可能影响其他路径引用，但测试表明正确。
  - run_suite.py 需要能正确发现并执行注册的测试，已通过 CI 验证。
  - 影响：影响范围：仅限于 AMD CI 中的 JIT kernel 单元测试 job。对用户无影响。对团队：后续添加 AMD JIT kernel 测试只需在测试文件中加入 register_amd_ci() 调用，无需修改工作流，降低了维护成本。
  - 风险标记：低风险，基础设施变更

关联脉络

- PR #24572 [AMD] Register 5 server-style 1-GPU tests for AMD PR CI: 同样是 AMD CI 测试注册相关工作，使用了相同的注册机制。