

PR #25939 完整报告

sgl-project/sglang

[AMD] ci: register 8 framework / unit tests to run on AMD CI

合并时间: 2026-06-10 23:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/25939>

执行摘要

- 一句话: 将 8 个测试注册到 AMD CI, 填补 NV→AMD 覆盖差距
- 推荐动作: 本次变更是典型的 CI 覆盖增强, 值得其他硬件平台 (如 NPU、XPU) 参考其筛选和验证流程。设计决策上, 选择“不注册并禁用”而不是“注册但标记为 `always skip`”的策略, 确保了 CI 清单的简洁性。建议关注后续是否会有类似批次将更多测试引入 AMD CI。

功能与动机

根据 PR body, 这是“Second batch of the NV→AMD coverage gap audit (follow-up to #25208)”, 目的是将不依赖 NVIDIA 特定内核的测试注册到 AMD CI 上执行, 确保 AMD 平台也能获得同等覆盖。PR 指出“All target `stage-b-test-1-gpu-small-amd`, the existing per-commit AMD CI suite — no workflow changes needed。”

实现拆解

1. 导入 `register_amd_ci`: 在 8 个测试文件的 import 语句中, 从 `sglang.test.ci.ci_register` 同时导入 `register_amd_ci` 和已有的 `register_cuda_ci` (或 `register_cpu_ci`)。
2. 添加注册调用: 在每个文件中紧挨着 `register_cuda_ci(...)` 之后, 调用 `register_amd_ci(est_time=..., suite="stage-b-test-1-gpu-small-amd")`, 其中 `est_time` 与 CUDA 估算时间相同。
3. 处理 AMD 不兼容测试: 在 `test_dsa_pool_host_unit.py` 中, 为 `test_device_to_host_indexer_kernel` 方法添加 `@unittest.skipIf(is_hip(), ...)` 装饰器, 当运行在 AMD (ROCm) 时跳过该测试, 因为其 kernel IO 后端仅支持 CUDA。
4. lint 修复: 在一次合并后, 将 `test_transformers_backend_eval.py` 中的 import 改为括号换行, 以符合 black 的 88 字符限制。
5. 删除候选中的 4 个文件: 在首个 AMD CI 运行后发现 4 个文件存在真正的 ROCm 不兼容问题 (如 `deep_gemm` 依赖、Triton NV-only 操作等), 因此从 PR 中删除其 AMD 注册, 保持 CUDA-only。

关键文件:

- `test/registered/unit/mem_cache/test_dsa_pool_host_unit.py` (模块 内存池; 类别 test; 类型 test-coverage; 符号 `test_device_to_host_indexer_kernel`): 该文件除了注册 AMD CI 外, 还添加了 `@unittest.skipIf(is_hip(), ...)` 装饰器, 跳过不兼容的 kernel IO 后端测试, 是唯一包含逻辑调整的文件。

- test/registered/models/test_transformers_backend_eval.py (模块 评估; 类别 test; 类型 test-coverage) : 该文件是 transformers 后端评估测试, 已注册 CUDA 和 CPU CI, 新增 AMD CI 注册, 且 import 行因代码风格被调整为多行格式。
- test/registered/disaggregation/test_specv2_kvcache_offloading.py (模块 分解; 类别 test; 类型 test-coverage) : specv2 KV 缓存卸载测试, 仅添加 AMD 注册, 无其他修改, 代表大多数文件的典型变更。

关键符号: 未识别

关键源码片段

test/registered/unit/mem_cache/test_dsa_pool_host_unit.py

该文件除了注册 AMD CI 外, 还添加了 `@unittest.skipIf(is_hip(), ...)` 装饰器, 跳过不兼容的 `kernel` IO 后端测试, 是唯一包含逻辑调整的文件。

```
import inspect
import unittest

import torch

from sglang.srt.mem_cache.memory_pool import DSATokenToKVPool
from sglang.srt.mem_cache.memory_pool_host import (
    ALLOC_MEMORY_FUNCS,
    DSAIndexerPoolHost,
    MLATokenToKVPoolHost,
    alloc_with_pin_memory,
)
from sglang.srt.utils import is_cuda, is_hip, is_npu, is_xpu
# 同时导入 CUDA 和 AMD 注册函数
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

register_cuda_ci(est_time=9, stage="base-b", runner_config="1-gpu-small")
# 新增 AMD 注册, est_time 与 CUDA 一致
register_amd_ci(est_time=9, suite="stage-b-test-1-gpu-small-amd")

...

# 跳过 kernel IO 后端: 该路径在 AMD 上不受支持 (ValueError)
@unittest.skipIf(
    is_hip(),
    '`io_backend="kernel"` path in memory_pool_host.backup_from_device_all_layer '
    "raises ValueError on AMD (only the `direct` IO backend is wired for ROCm). "
    "The other 62 tests in this file pass on AMD.",
)
def test_device_to_host_indexer_kernel(self):
    self._run_device_to_host_indexer_copy(io_backend="kernel")
```

test/registered/models/test_transformers_backend_eval.py

该文件是 transformers 后端评估测试，已注册 CUDA 和 CPU CI，新增 AMD CI 注册，且 import 行因代码风格被调整为多行格式。

```
"""A small end-to-end eval coverage for the transformers modeling backend."""

import unittest
from types import SimpleNamespace

# 同时导入 CUDA、AMD、CPU 注册函数
from sglang.test.ci.ci_register import (
    register_amd_ci,
    register_cpu_ci,
    register_cuda_ci,
)
from sglang.test.few_shot_gsm8k import run_eval
from sglang.test.server_fixtures.default_fixture import DefaultServerBase

register_cuda_ci(est_time=48, stage="base-b", runner_config="1-gpu-small")
# 新增 AMD 注册，est_time 与 CUDA 相同
register_amd_ci(est_time=48, suite="stage-b-test-1-gpu-small-amd")
register_cpu_ci(est_time=320, suite="base-b-test-cpu")
```

test/registered/disaggregation/test_specv2_kvcache_offloading.py

specv2 KV 缓存卸载测试，仅添加 AMD 注册，无其他修改，代表大多数文件的典型变更。

```
from sglang.srt.disaggregation.decode_kvcache_offload_manager import (
    DecodeKVCacheOffloadManager,
)
from sglang.srt.disaggregation.kv_events import OffloadedState
# 新增 AMD 注册导入
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

register_cuda_ci(est_time=8, stage="base-b", runner_config="1-gpu-small")
# 注册 AMD CI
register_amd_ci(est_time=8, suite="stage-b-test-1-gpu-small-amd")
```

评论区精华

该 PR 没有 review 评论。在 Issue 中，作者请求 amd-bot 提供 CI 状态，bot 确认所有 8 个注册测试在 AMD CI 上通过，并指出 PR 的 NVIDIA 测试失败是由于不相关的 PD 分解 KV 传输问题。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于新增的 AMD CI 注册可能引入误报或漏报。由于 PR 仅添加注册行和一处 skip 装饰器，不修改测试逻辑，因此回归风险较低。但是，est_time 参数是基于 CUDA 估算的，可能不准确，但 PR 提到后续可通过 update-est-time bot 调优。另外，

test_dsa_pool_host_unit.py 中的 skip 装饰器仅跳过单个方法，其他 62 个测试仍会运行，不存在覆盖丢失。

- 影响：影响范围限于 CI 基础设施：AMD CI 套件 stage-b-test-1-gpu-small-amd 将新增 8 个测试，总估计时间约 237 秒。测试涵盖 transformers 后端评估、DSA 池主机单元、specv2 KV 缓存卸载、受限推理 e2e、优先级调度、SWA 锁释放、统一 radix 缓存等模块。这些测试之前仅在 CUDA 上运行，现在 AMD 上也获得了持续验证，有助于发现跨平台兼容性问题。
- 风险标记：CI 配置与编排变更，硬件平台兼容风险（已处理跳过）

关联脉络

- PR #25208 [AMD] ci: register first batch of tests to run on AMD CI: 该 PR 是 #25208 的延续，采用相同的模式将更多测试注册到 AMD CI。