

PR #50930 完整报告

vllm-project/vllm

[Test] Add ROCm AITER MLA op registration and env gating tests

合并时间: 2026-08-08 03:07

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50930>

执行摘要

- 一句话: 新增 AITER MLA 自定义算子注册与 opcheck 契约测试
- 推荐动作: 值得 kernel/ 自定义算子开发者精读, 尤其是“用单次 opcheck 替代多条冗余断言、将 env 行为测试排除出内核测试域”的范围取舍; 该模式可复用于其他 torch.library 自定义算子 (如 mxfp4、fp8、flashattention 封装的注册验证)。

功能与动机

PR body 指出需要内核级测试验证 rocm_aiter_mla_decode_fwd 自定义算子正确注册、支持 fake tensors (供 torch.compile 追踪), 并遵守 VLLM_ROCM_USE_AITER / VLLM_ROCM_USE_AITER_MLA 环境变量门控。实际 review 后仅保留 opcheck 用例, 因为注册与别名校验都被 opcheck 单步覆盖, env 门控行为不属于内核测试范畴。

实现拆解

1. 新增 tests/kernels/attention/test_rocm_aiter_mla_op_registration.py (+79 行): 通过 pytestmark 将用例限定在 ROCm 且 MI300/MI350 (on_mi3xx) 平台; _require_aiter() 在 is_aiter_found_and_supported() 为假时跳过用例, 避免在未安装 AITER 的 CI 镜像上报错。
2. 核心用例 test_mla_decode_fwd_op_schema() 构造小形状推理张量 (q、kv_buffer、o、qo_indptr、kv_indptr、kv_indices、kv_last_page_lens 等), 显式 import vllm._aiter_ops 触发算子注册后调用 OPHECK, 一次覆盖注册、schema 一致性、fake-tensor 可追踪性与 mutates_args=["o"] 别名声明。
3. 按 review 删减: 初版约 181 行含注册存在性断言、手动 mutates_args 校验、4 组 env 门控组合与 refresh_env_variables 运行时刷新测试; 维护者指出这些都是冗余或越界, 合并版仅保留 79 行 opcheck 用例。
4. CI 接线: .buildkite/test-amd.yaml (+4 行) 在 Kernels MLA (MI300) 与 Kernels MLA (MI355) 两个 job 的 source_file_dependencies 追加新测试文件, 并在 commands 追加对应 pytest 命令, 使改动 rocm_aiter_mla.py、selector.py、_aiter_ops.py、rocm.py 的 PR 自动触发该测试。

关键文件:

- tests/kernels/attention/test_rocm_aiter_mla_op_registration.py (模块 内核测试; 类别 test; 类型 test-coverage; 符号 _require_aiter, test_mla_decode_fwd_op_schema): 本

PR 的核心产出：新增内核级 opcheck 契约测试，一次调用覆盖

rocm_aiter_mla_decode_fwd 的注册、schema、fake-tensor 支持与 mutates_args=["o"] 别名校验，是保护 ROCm MLA decode 路径 torch.compile 追踪能力的关键回归防线。

- .buildkite/test-amd.yaml（模块 CI 配置；类别 config；类型 configuration）：将新测试接入两个 ROCm 硬件线的 CI job，确保任何改动 AITER MLA 相关源码的 PR 都会自动运行该契约测试；缺失此接线则测试无法在上游 CI 生效。

关键符号：_require_aiter, test_mla_decode_fwd_op_schema

关键源码片段

tests/kernels/attention/test_rocm_aiter_mla_op_registration.py

本 PR 的核心产出：新增内核级 opcheck 契约测试，一次调用覆盖

rocm_aiter_mla_decode_fwd 的注册、schema、fake-tensor 支持与 mutates_args=["o"] 别名校验，是保护 ROCm MLA decode 路径 torch.compile 追踪能力的关键回归防线。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
"""ROCm 自定义算子模式测试：AITER MLA decode."""

import pytest
import torch

from tests.kernels.utils import opcheck
from vllm.platforms import current_platform

# 仅在 ROCm 平台且为 MI300 (gfx942) / MI350 (gfx950) 系列时执行
_SKIP_NON_MI3XX = True
if current_platform.is_rocm():
    from vllm.platforms.rocm import on_mi3xx

    _SKIP_NON_MI3XX = not on_mi3xx()

pytestmark = [
    pytest.mark.skipif(not current_platform.is_rocm(), reason="ROCm-specific tests"),
    pytest.mark.skipif(_SKIP_NON_MI3XX, reason="MI300/MI350 ROCm only"),
]

# MLA 的维度：Q 头维度 = kv_lora_rank + qk_rope_dim；V 头维度 = kv_lora_rank
Q_HEAD_DIM = 576
V_HEAD_DIM = 512

def _require_aiter():
    """AITER 库不可用时跳过用例，避免在未安装实验组件的镜像上报错。"""
    from vllm._aiter_ops import is_aiter_found_and_supported

    if not is_aiter_found_and_supported():
```

```
pytest.skip("aiter is required on supported ROCm hardware for this test")
```

```
@torch.inference_mode()
def test_mla_decode_fwd_op_schema() -> None:
    """一次 opcheck 同时覆盖：算子已注册、schema 与实现一致、fake-tensor 可追迹、
    `mutates_args=["o"]` 就地产出别名声明正确。"""
    _require_aiter()
    # 显式 import 触发自定义算子的注册
    from vllm._aiter_ops import rocm_aiter_ops # noqa: F401

    batch_size, nhead = 4, 128

    q = torch.randn(batch_size, nhead, Q_HEAD_DIM, dtype=torch.bfloat16, device="cuda")
    kv_buffer = torch.randn(64, 1, 1, Q_HEAD_DIM, dtype=torch.bfloat16, device="cuda")
    o = torch.zeros(batch_size, nhead, V_HEAD_DIM, dtype=torch.bfloat16, device="cuda")
    qo_indptr = torch.arange(0, batch_size + 1, dtype=torch.int32, device="cuda")
    kv_indptr = torch.arange(0, batch_size + 1, dtype=torch.int32, device="cuda") * 16
    kv_indices = torch.arange(0, 64, dtype=torch.int32, device="cuda")
    kv_last_page_lens = torch.ones(batch_size, dtype=torch.int32, device="cuda")

    # opcheck 会同时核对 torch.library 注册信息、composite/fake 实现与真实 kernel,
    # 并验证 `o` 被声明为 in-place 输出别名，任何不一致都会直接失败，
    # 因此无需再单独写注册存在性断言或手动 mutates_args 校验。
    opcheck(
        torch.ops.vllm.rocm_aiter_mla_decode_fwd,
        (q, kv_buffer, o, qo_indptr, 1),
        {
            "kv_indptr": kv_indptr,
            "kv_indices": kv_indices,
            "kv_last_page_lens": kv_last_page_lens,
            "sm_scale": Q_HEAD_DIM**-0.5,
            "logit_cap": 0.0,
            "q_scale": None,
            "kv_scale": None,
            "work_meta_data": None,
            "work_indptr": None,
            "work_info_set": None,
            "reduce_indptr": None,
            "reduce_final_map": None,
            "reduce_partial_map": None,
        },
    )
```

评论区精华

维护者 AndreasKaratzas 提出两处疑问：一是注册存在性检查是否有必要 ("Is this really needed?")，二是手动 mutates_args 断言是否与 opcheck 重复 ("We are testing it afterwards too right?")；作者回应这两项均由 opcheck 覆盖并删除。另一关键讨论是 env

门控测试的归属——"I dont think that this belongs here. This tries to check whether envvar follows design but this is the spot for kernels tests", 作者接受并移除全部 env 相关用例。最终 AndreasKaratzas 以 "LGTM" 批准合并。

- 注册存在性检查是否冗余 (testing): 作者回复 opcheck 在算子未注册时本身就会失败, 且 aliasing 检查已覆盖 mutates_args=['o'], 故删除冗余断言。
- 手动 mutates_args 断言与 opcheck 重复 (testing): 作者确认 opcheck 的 aliasing 检查已覆盖 mutates_args=['o'], 删除手动校验, 合并版只保留 opcheck 单步验证。
- env 门控测试是否属于内核测试范畴 (design): 作者接受意见, 移除全部 env 门控组合与 refresh_env_variables 测试, 最终文件仅保留针对算子 schema 的 opcheck 用例。

风险与影响

- 风险: 测试在 AITER 库缺失或非 MI300/MI350 平台时跳过而非失败, 可能掩盖 AITER 环境之外的注册回归, 但这是有意的平台门控, 常规 x86/ 非 MI3xx CI 不受阻塞。opcheck 仅验证注册与 schema/fake-tensor 一致性, 不覆盖数值正确性, 数值问题仍依赖 test_rocm_aiter_mla_decode.py 等既有用例。本 PR 无 runtime 代码改动, 对性能与安全无影响; 仅给两个 Buildkite job 增加一次轻量 pytest 执行, CI 耗时增幅可忽略。
- 影响: 影响范围限于 ROCm CI 与内核开发者: 为 MI300 (gfx942) / MI355 (gfx950) 上的 AITER MLA decode 算子建立 schema/fake-tensor 回归防线, 任何改动 rocm_aiter_mla.py、selector 或 _aiter_ops 的 PR 都会自动运行该测试。对最终用户无功能性影响, 属于纯质量保障改动。
- 风险标记: 仅在 MI300/MI355 硬件运行, AITER 缺失时测试跳过, 仅验证 schema 而非数值正确性, CI 作业耗时小幅增加

关联脉络

- PR #49373 [Bugfix][ROCM] Fix ROCM_AITER_FA & ROCM_AITER_UNIFIED_ATTENTION QK-Norm+RoPE+KVCache fusion for the packed KV-cache [BLOCKS, HEADS, BLOCK_SIZE, 2*HEAD_DIM] layout: 同属 ROCm AITER 注意力后端质量收敛线, 补强 kernel 级回归验证 (该 PR 修 fusion 并加编译 pass 测试, 本 PR 补自定义算子契约测试)。
- PR #51357 Fix ROCm architecture import on non-ROCM platforms: 与 _aiter_ops / rocm 平台导入门控逻辑相关, 同属 ROCm 专项代码的平台兼容与门控保障。