PR #42736 完整报告

vllm-project/vllm

[Kernel][Test] Make kernel tests for mamba dual-HW (CUDA + XPU)

合并时间：2026-06-08 08:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42736

执行摘要

- 一句话：使 Mamba 混合内核测试可同时在 CUDA 和 XPU 上运行
- 推荐动作：值得快速浏览，展示了 vLLM 平台层进行跨平台测试适配的推荐模式。如果负责 XPU 或多平台测试，可以精读以复用该模式。PR 作者工作规范，测试结果详尽，提交历史清晰，适合作为测试适配的参考范例。

功能与动机

使 MambaMixer Triton kernel 测试在 Intel XPU 上可运行。底层内核已经是纯 Triton 实现，只在 XPU 上工作，但测试框架硬编码了 'cuda' 设备，导致测试无法在 XPU 上执行。PR body 原文：'Make the MambaMixer Triton kernel tests under tests/kernels/mamba/ runnable on Intel XPU in addition to CUDA/ROCm. The underlying kernels ... are pure Triton and already work on XPU - only the test harness was pinning device="cuda".'

实现拆解

1. 在每个测试文件中导入 from vllm.platforms import current_platform 并定义 DEVICE = current_platform.device_type。
2. 添加模块级 pytestmark = pytest.mark.skipif(...)，确保文件仅在 CUDA-alike 或 XPU 平台上运行，否则跳过并给出明确原因。
3. 将文件中所有 device="cuda" 的硬编码字面量替换为 device=DEVICE，包括测试辅助函数 generate_random_inputs、generate_continuous_batched_examples 以及各个测试函数中的 device 变量。
4. 在 test_mamba_ssm.py 中，为依赖 CUDA-only C++ op ops.selective_scan_fwd 的测试（test_selective_scan 和 test_selective_scan_varlen）添加 @skip_unless_cuda_alike 装饰器，使其在非 CUDA 平台明确跳过而非报错。
5. 修复 test_causal_conv1d.py 中 test_causal_conv1d_varlen 函数里遗漏的 cumsum.cuda() 调用，改为 cumsum.to(device)，解决 XPU 上“Torch not compiled with CUDA enabled”断言失败。
6. 验证结果：在 Intel Battlemage B70（XPU）上完整运行，506 passed，56 skipped，exit code 0。

关键文件：

- tests/kernels/mamba/test_mamba_ssm_ssd.py（模块 测试；类别 test；类型 test-coverage；符号 generate_random_inputs, generate_continuous_batched_examples, test_mamba_chunk_scan_single_example）：核心变更，展示了如何为纯 Triton 测试添加 XPU 支持，包括平台导入、DEVICE 定义和模块跳过条件。
- tests/kernels/mamba/test_mamba_ssm.py（模块 测试；类别 test；类型 test-coverage；符号 test_selective_scan, test_selective_state_update, test_selective_state_update_varlen, test_selective_scan_varlen）：除通用适配外，额外处理了 CUDA-only 的 selective_scan 测试，通过 skip_unless_cuda_alike 装饰器正确跳过。
- tests/kernels/mamba/test_causal_conv1d.py（模块 测试；类别 test；类型 test-coverage；符号 test_causal_conv1d_update, test_causal_conv1d_update_with_batch_gather, test_causal_conv1d_varlen）：包含一个修复：将遗漏的 .cuda() 调用替换为 .to(device)，确保 XPU 上 varlen 测试通过。

关键符号：generate_random_inputs, generate_continuous_batched_examples, test_selective_scan, test_selective_state_update, test_selective_scan_varlen, test_causal_conv1d_update, test_causal_conv1d_varlen

关键源码片段

tests/kernels/mamba/test_mamba_ssm_ssd.py

核心变更，展示了如何为纯 Triton 测试添加 XPU 支持，包括平台导入、DEVICE 定义和模块跳过条件。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project

import pytest
import torch
import torch.nn.functional as F
from einops import rearrange, repeat

from vllm.model_executor.layers.mamba.ops.ssd_combined import (
    mamba_chunk_scan_combined_varlen,
)
from vllm.platforms import current_platform # 新增导入，用于获取当前平台类型
from vllm.utils.torch_utils import set_random_seed
from vllm.v1.attention.backends.mamba2_attn import compute_varlen_chunk_metadata

# 所有在此文件测试的内核都是纯 Triton 实现，因此可以在任何 vLLM 平台层
# 视为 CUDA-alike 或 XPU 的设备上运行。
DEVICE = current_platform.device_type # 动态获取设备类型，替代硬编码的 "cuda"

# 模块级跳过条件：仅当平台是 CUDA-alike 或 XPU 时才运行测试，
# 否则跳过并给出清晰原因。
pytestmark = pytest.mark.skipif(
```

```
    not (current_platform.is_cuda_alike() or current_platform.is_xpu()),
    reason="Mamba2 SSD Triton kernels require a CUDA-alike or XPU device.",
)

# ...（其余代码，包括 ssd_minimal_discrete 等，未改动）...

def generate_random_inputs(batch_size, seqlen, n_heads, d_head, itype, device=DEVICE):
    """生成随机输入，device 参数默认使用全局 DEVICE。"""
    set_random_seed(0)
    A = -torch.exp(torch.rand(n_heads, dtype=itype, device=device))
    dt = F.softplus(
        torch.randn(batch_size, seqlen, n_heads, dtype=itype, device=device) - 4
    )
    X = torch.randn((batch_size, seqlen, n_heads, d_head), dtype=itype, device=device)
    B = torch.randn((batch_size, seqlen, n_heads, d_head), dtype=itype, device=device)
    C = torch.randn((batch_size, seqlen, n_heads, d_head), dtype=itype, device=device)
    return A, dt, X, B, C
```

评论区精华

本 PR 的 review 没有实质性讨论。AndreasKaratzas 和 jikunshang 均直接批准，gemini-code-assist 机器人评论无反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试代码中的设备选择逻辑，没有修改任何内核、模型或推理路径。CUDA 行为通过 CI 保持不变。XPU 运行结果已完整验证，506 个测试全部通过，56 个预期跳过。唯一可能的风险是若未来平台层 API 变更可能导致测试跳过条件失效，但此类 API（is_cuda_alike、is_xpu）相对稳定。
- 影响：对最终用户无影响。对开发者：XPU 平台开发者现在可以自动运行这些测试，确保 Mamba 内核正确性。对团队：建立了跨平台测试适配的通用模式，可复用于其他测试套件。影响范围限定在测试基础设施，程度中等。
- 风险标记：仅测试变更，XPU 已验证，无核心逻辑改动

关联脉络

- 暂无明显关联 PR