PR #46401 完整报告

vllm-project/vllm

[CI][ROCm] Restrict MLA cross-layer KV cache test to supported backends on ROCm

合并时间：2026-06-23 06:05

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46401

执行摘要

- 一句话：限制 ROCm MLA 跨层 KV 缓存测试后端
- 推荐动作：该 PR 为简单的测试适配，值得快速合并。关注关联 issue #46411 以了解未来在其他 ROCm 后端的支持扩展。

功能与动机

PR body 明确指出修复 test_verified_mla_backends_support_cross_layer_kv_cache[FlashAttnMLABackend] 在 ROCm (MI300/MI325) 上的失败。测试结果显示，在 ROCm 上该测试原本 1 failed, 3 passed, 1 skipped，修改后 3 passed。

实现拆解

1. 添加平台判断导入：在 tests/v1/kv_connector/unit/test_kv_cache_layout.py 中增加 from vllm.platforms import current_platform 导入，用于条件判断当前平台是否为 ROCm。
2. 修改参数化列表：将 test_verified_mla_backends_support_cross_layer_kv_cache 的 backend_path 参数化列表改为条件表达式：
   - 在 ROCm 上，列表仅包含 TritonMLABackend（已验证支持）。
   - 在其他平台上，保持原有完整后端列表（TritonMLABackend、CutlassMLABackend、FlashAttnMLABackend、FlashMLABackend、FlashInferMLABackend）。
3. 添加 Issue 引用：在代码注释中关联了上游 issue #46411，用于追踪后续在其他后端的支持扩展。

关键文件：

- tests/v1/kv_connector/unit/test_kv_cache_layout.py（模块 缓存布局；类别 test；类型 test-coverage）：唯一的变更文件，通过条件参数化修复 ROCm 上的测试失败。

关键符号：test_verified_mla_backends_support_cross_layer_kv_cache

关键源码片段

tests/v1/kv_connector/unit/test_kv_cache_layout.py

唯一的变更文件，通过条件参数化修复 ROCm 上的测试失败。

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

```
import pytest

from vllm.platforms import current_platform


# ... 其他测试保持不变 ...

@pytest.mark.parametrize(
    "backend_path",
    # See: https://github.com/vllm-project/vllm/issues/46411
    [
        "vllm.v1.attention.backends.mla.triton_mla.TritonMLABackend",
    ]
    if current_platform.is_rocm()
    else [
        "vllm.v1.attention.backends.mla.triton_mla.TritonMLABackend",
        "vllm.v1.attention.backends.mla.cutlass_mla.CutlassMLABackend",
        "vllm.v1.attention.backends.mla.flashattn_mla.FlashAttnMLABackend",
        "vllm.v1.attention.backends.mla.flashmla.FlashMLABackend",
        "vllm.v1.attention.backends.mla.flashinfer_mla.FlashInferMLABackend",
    ],
)
def test_verified_mla_backends_support_cross_layer_kv_cache(backend_path):
    """Backends whose decode kernels honor the cache's block-dim stride opt
    in to the cross-layer layout with a non-identity permutation placing
    num_blocks first in physical layout."""
    module_path, name = backend_path.rsplit(".", 1)
    backend = getattr(
        pytest.importorskip(module_path, reason="backend deps unavailable"), name
    )

    stride_order = backend.get_kv_cache_stride_order(include_num_layers_dimension=True)
    assert stride_order == (1, 0, 2, 3)
    assert stride_order[0] != 0 # num_blocks first => cross-layer supported
    assert backend.get_kv_cache_stride_order(include_num_layers_dimension=False) == (
        0,
        1,
        2,
    )
```

评论区精华

无实质性 review 讨论。审核者 AndreasKaratzas 批准并表示感谢。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅限于测试文件，且通过条件判断确保不影响非 ROCm 平台。ROCm 上仅保留 TritonMLABackend 可能暂时降低测试覆盖率，但避免了因不受支持的后端导致的测试失败。
- 影响：直接修复 ROCm 平台（MI300/MI325）上一个特定测试失败。影响范围仅限于测试文件，不涉及任何生产代码。对其他平台无影响。
- 风险标记：仅测试变更

关联脉络

- PR #45111 [Attention] Re-enable cross-layer KV cache layout for MLA via stride-aware kernels: 该 PR 引入了跨层 KV 缓存布局支持和相关测试，本 PR 是其后续 ROCm 适配修复。