PR #47314 完整报告

vllm-project/vllm

[BugFix] Fix packed HND KV cache reshape for FlashAttention

合并时间：2026-07-11 11:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47314

执行摘要

- 一句话：修复 packed HND KV cache 形状重塑错误
- 推荐动作：值得合并，修复明确且风险低。建议补充针对 packed HND KV cache reshape 的单元测试，以覆盖该路径。

功能与动机

Issue #47054 报告了在使用 CPUOffloadingConnector 和 enable_cross_layers_blocks 配置时，vLLM 在处理 GPT-OSS-120B 模型时崩溃。根本原因是 packed KV cache 在 reshape 时未正确处理物理布局到逻辑布局的转换，导致形状不匹配。

实现拆解

变更仅涉及 vllm/v1/worker/gpu/attn_utils.py 中的 _reshape_attention_kv_cache 函数，共一行代码改动。

1. 在函数中，当 packing is not None 时，代码从原始张量中提取 packed 数据并尝试 reshape。
2. 原代码使用 .view(kv_cache_shape) 应用物理布局（即 unpacked 前的形状），但此时数据已经被 unpacked，应使用逻辑布局 permuted_kv_cache_shape（即按 stride_order 排列后的形状）。
3. 修正为 .view(permuted_kv_cache_shape)，确保结果张量形状与后端期望的逻辑布局一致。

关键文件：

- vllm/v1/worker/gpu/attn_utils.py（模块 KV 缓存；类别 source；类型 core-logic；符号 _reshape_attention_kv_cache）：核心修复文件，_reshape_attention_kv_cache 函数中一行改动。

关键符号：_reshape_attention_kv_cache

关键源码片段

vllm/v1/worker/gpu/attn_utils.py

核心修复文件，_reshape_attention_kv_cache 函数中一行改动。

```
# vllm/v1/worker/gpu/attn_utils.py
```

```
def _reshape_attention_kv_cache(
    kv_raw_tensor: torch.Tensor,
    kv_cache_spec: AttentionSpec,
    kv_cache_shape: tuple[int, ...],
    kv_cache_stride_order: tuple[int, ...],
    num_blocks: int,
    packing: tuple[int, int] | None,
) -> torch.Tensor:
    # 根据 stride_order 计算逻辑布局的形状
    permuted_kv_cache_shape = tuple(kv_cache_shape[i] for i in kv_cache_stride_order)
    inv_order = [
        kv_cache_stride_order.index(i) for i in range(len(kv_cache_stride_order))
    ]
    dtype = kv_cache_spec.dtype

    if packing is not None:
        offset, block_stride = packing
        assert inv_order[0] == 0
        page_bytes = prod(kv_cache_shape[1:]) * get_dtype_size(dtype)
        kv_cache = (
            kv_raw_tensor.view(-1, block_stride)[:, offset : offset + page_bytes]
            .view(dtype)
            .view(permuted_kv_cache_shape) # 修复：使用逻辑布局而非物理布局
        )
    elif kv_cache_spec.page_size_padded is not None:
        # ... padding handling ...
    else:
        kv_cache = kv_raw_tensor.view(dtype).view(permuted_kv_cache_shape)
    return kv_cache
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：变更影响范围极小（仅一行），仅影响 packed KV cache 场景（packing is not None）。风险较低，但缺少直接针对此场景的回归测试。
- 影响：仅影响使用 packed KV cache 且 backend 为 FlashAttention 的 HND 格式用户，修复了 CPUOffloading 等跨层 KV 传输场景下的崩溃。
- 风险标记：缺少测试覆盖

关联脉络

- PR #47316 [Misc] Use meta tensor for KV cache stride calculation: 同一文件 vllm/v1/worker/gpu/attn_utils.py 的后续优化，改进 KV cache stride 计算。