PR #42749 完整报告

vllm-project/vllm

[Model][Hardware][AMD]: Part 1/2 -> Enable e2e QK Norm + RoPE + KV Cache runtime fusion for Qwen3-30B-A3B on ROCM_AITER_FA, and ROCM_AITER_UNIFIED_ATTN

合并时间：2026-07-17 06:39

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42749

执行摘要

- 一句话：新增 ROCm AITER QK Norm+RoPE+KV Cache 融合 pass
- 推荐动作：值得精读，特别是对 ROCm 平台推理优化和 TorchInductor pattern matcher 感兴趣的开发者。该 PR 展示了如何安全地将多次 kernel launch 融合为单个自定义 op，并处理了不同 attention 后端的布局差异。

功能与动机

从原始大 PR #39527 拆分而来，旨在优化 Qwen3-30B-A3B 模型在 ROCm 上的推理性能。QK Norm+RoPE+KV Cache 序列存在多次 kernel launch 和显存读写，融合为一个 kernel 可显著降低开销。

实现拆解

1. 自定义 op 与模式匹配：在 vllm/compilation/passes/fusion/qk_norm_rope_kvcache_fusion.py 中注册 Torch custom op fused_qk_norm_rope_and_unified_kv_cache_update，并实现 QkNormRopeKvCachePattern 类，使用 inductor pattern matcher 匹配 unfused 子图（split→RMSNorm→RoPE→unified_kv_cache_update）并替换为 fused op。
2. AITER kernel 封装：在 vllm/_aiter_ops.py 中新增 fused_qk_norm_rope_and_cache 静态方法，调用 aiter HIP kernel fused_qk_norm_rope_cache_pts_quant_shuffle；再提供 do_qk_norm_rope_kvcache_update 高层封装，处理 fp8 视图转换和部分旋转维度逻辑。
3. 后端抽象接口：在 vllm/v1/attention/backend.py 的 AttentionImplBase 中添加 fused_qk_norm_rope_kvcache_supported 和 do_qk_norm_rope_kvcache_update 抽象方法，默认返回 False 和 raise NotImplementedError。
4. 具体后端实现：在 rocm_aiter_fa.py 和 rocm_aiter_unified_attn.py 中分别实现上述方法，根据各自 KV cache layout（unbind 维度不同）拆分 cache 并调用 _aiter_ops.do_qk_norm_rope_kvcache_update。UNIFIED_ATTN 路径设置 use_shuffle_layout=False。
5. 配置与测试：在 vllm/config/vllm.py 和 compilation.py 中添加 fuse_qk_norm_rope_kvcache 配置项及 rope_kvcache_fusion_max_token_num 限制，自动启用逻辑。新增单元测试文件 test_rocm_aiter_qk_norm_rope_kvcache_fusion.py，验证 fusion 前后输出一致性和 K cache 正确性。

关键文件：

- vllm/compilation/passes/fusion/qk_norm_rope_kvcache_fusion.py（模块 编译融合；类别 source；类型 core-logic；符号 fused_qk_norm_rope_and_unified_kv_cache_update_impl, fused_qk_norm_rope_and_unified_kv_cache_update_fake, QkNormRopeKvCachePattern, init）：新增的核心融合 pass 和 custom op，是整个 PR 的核心逻辑所在。
- tests/compile/passes/test_rocm_aiter_qk_norm_rope_kvcache_fusion.py（模块 测试；类别 test；类型 test-coverage；符号 QKNormRoPEKVCacheTestModel, init, build_attn_metadata, forward）：新增的单元测试，验证融合 pass 的正确性和数值精度。
- vllm/_aiter_ops.py（模块 底层算子；类别 source；类型 core-logic；符号 fused_qk_norm_rope_and_cache, do_qk_norm_rope_kvcache_update）：封装了 AITER 融合内核的调用，是后端实现与底层 kernel 的桥梁。
- vllm/v1/attention/backend.py（模块 attention 后端；类别 source；类型 core-logic；符号 fused_qk_norm_rope_kvcache_supported, do_qk_norm_rope_kvcache_update）：定义了融合后端的抽象接口，是整个融合扩展的基础。
- vllm/v1/attention/backends/rocm_aiter_unified_attn.py（模块 attention 后端；类别 source；类型 core-logic；符号 fused_qk_norm_rope_kvcache_supported, do_qk_norm_rope_kvcache_update）：实现了 UNIFIED_ATTN 后端的融合方法，处理 K/V-first layout。
- vllm/v1/attention/backends/rocm_aiter_fa.py（模块 attention 后端；类别 source；类型 core-logic；符号 fused_qk_norm_rope_kvcache_supported, do_qk_norm_rope_kvcache_update）：实现了 FA 后端的融合方法，使用 unbind(1) 分割 KV cache。
- vllm/config/vllm.py（模块 配置；类别 source；类型 config-change；符号 enable_qk_norm_rope_kvcache）：配置自动启用逻辑和开关默认值。
- vllm/config/compilation.py（模块 配置；类别 source；类型 config-change）：添加 fusion 开关的编译配置。

关键符号：fused_qk_norm_rope_and_unified_kv_cache_update_impl, fused_qk_norm_rope_and_unified_kv_cache_update_fake, QkNormRopeKvCachePattern, fused_qk_norm_rope_kvcache_supported, do_qk_norm_rope_kvcache_update, fused_qk_norm_rope_and_cache, do_qk_norm_rope_kvcache_update

关键源码片段

vllm/compilation/passes/fusion/qk_norm_rope_kvcache_fusion.py

新增的核心融合 pass 和 custom op，是整个 PR 的核心逻辑所在。

```
# vllm/compilation/passes/fusion/qk_norm_rope_kvcache_fusion.py
# 自定义 op 的实现：融合 QK Norm + RoPE + KV cache 更新

def fused_qk_norm_rope_and_unified_kv_cache_update_impl(
    q_out: torch.Tensor,
    k_out: torch.Tensor,
```

```
    qkv: torch.Tensor,
    positions: torch.Tensor,
    q_weight: torch.Tensor,
    k_weight: torch.Tensor,
    rms_norm_eps: float,
    cos_sin_cache: torch.Tensor,
    is_neox: bool,
    layer_name: str = "",
) -> torch.Tensor:
    """
    执行融合后的 QK Norm + RoPE + KV cache 更新。
    从全局上下文中获取当前 attention layer 和 kv cache,
    然后委托给 backend 的 do_qk_norm_rope_kvcache_update 方法。
    """
    _, attn_layer, kv_cache, layer_slot_mapping = get_attention_context(layer_name)
    if layer_slot_mapping is not None:
        # 正式运行时，调用具体后端实现的融合方法
        attn_layer.impl.do_qk_norm_rope_kvcache_update(
            attn_layer,
            qkv,
            q_out,
            k_out,
            positions,
            q_weight,
            k_weight,
            rms_norm_eps,
            cos_sin_cache,
            is_neox,
            kv_cache,
            layer_slot_mapping,
        )
    else:
        # profiling 阶段：仅填充预分配输出占位
        q_out.zero_()
        k_out.zero_()
    return torch.empty(0, device=qkv.device, dtype=qkv.dtype)

# fake 实现用于元数据传播
def fused_qk_norm_rope_and_unified_kv_cache_update_fake(
    q_out: torch.Tensor,
    k_out: torch.Tensor,
    qkv: torch.Tensor,
    positions: torch.Tensor,
    q_weight: torch.Tensor,
    k_weight: torch.Tensor,
    rms_norm_eps: float,
    cos_sin_cache: torch.Tensor,
    is_neox: bool,
    layer_name: str = "",
```

```
) -> torch.Tensor:
    return torch.empty(0, device=qkv.device, dtype=qkv.dtype)

# 注册自定义 op
direct_register_custom_op(
    op_name="fused_qk_norm_rope_and_unified_kv_cache_update",
    op_func=fused_qk_norm_rope_and_unified_kv_cache_update_impl,
    mutates_args=["q_out", "k_out"],
    fake_impl=fused_qk_norm_rope_and_unified_kv_cache_update_fake,
)
```

评论区精华

关键讨论

1. Head size 限制必要性：dllehr-amd 质疑在 qk_norm_rope_fusion.py 中添加 head dim 限制是否必要。jhu960213 解释底层 AITER kernel 仅支持 64/128/256 三种 head dim，限制可防止不支持的 dim 导致崩溃。双方同意保留。
2. 删除 ROCM_ATTN 空存根：Rohan138 建议不在当前 PR 中添加 ROCM_ATTN 后端的 noop 实现，留待 Part 2 处理。jhu960213 同意并移除了相关变更，减少了干扰。
3. 默认关闭融合：Rohan138 指出 O1 优化级别不应默认开启此融合（依赖 Inductor partition）。jhu960213 调整配置使其默认关闭，用户需显式设置 fuse_qk_norm_rope_kvcache=true。
4. SymInt workaround：Rohan138 建议将 SymInt 处理上移至 pattern 预注册阶段。jhu960213 通过自定义 search_pattern 内嵌 SymInt wildcarding 实现，避免了模式匹配时的符号冲突。

- Fusion head size guard 与后端兼容性 (design): 双方同意保留限制，作为防御性措施。
- 删除 ROCM_ATTN 中的 noop 代码 (design): ROCM_ATTN 后端的融合实现移至后续 PR。
- 默认关闭 fuse_qk_norm_rope_kvcache (performance): 融合 pass 默认关闭，用户需显式设置 fuse_qk_norm_rope_kvcache=true 启用。

风险与影响

- 风险：
  1. 回归风险：新 fusion pass 可能与其他编译 pass（如 RopeKVCacheFusionPass）冲突，导致意外模式被匹配或漏匹配。已在测试中覆盖非融合场景。
  2. 性能影响：融合 kernel 仅支持有限 head dim，不匹配时静默跳过，不会造成错误但无法获得加速。用户需注意 rope_kvcache_fusion_max_token_num 限制。
  3. 兼容性：该融合专用 ROCm AITER 后端，NVIDIA 或其他平台上不会开启。但不影响其他后端的正常路径。
  4. 配置复杂性：用户需通过 --compilation-config 传递 JSON 配置开启，使用门槛较高。- 影响：用户影响：Qwen3-30B-A3B 模型在 ROCm 平台上的推理延迟和吞吐量有显著改善（具体数据参考 PR 截图）。用户需手动设置编译配置以启用融合。系统影响：新增编译 pass 和 custom op，但不影响未使用融合的模型。团队影响：PR 被拆分为 Part

1/2，Part 1 聚焦 AITER 后端，Part 2 将添加 ROCM_ATTN 支持。降低了单次 review 负担。

- 风险标记：ROCm 专用，新 fusion pass，依赖 Inductor 分区，需要手动启用

关联脉络

- PR #39527 [Optimize] Optimize Qwen3-30B with ROCm AITER fusions: 此 PR 是 #39527 的拆分 Part 1，原始 PR 包含更广泛的 ROCm 融合优化。