PR #46148 完整报告

vllm-project/vllm

[ROCm][CI] Only require q_scale==1.0 for fp8 query in RocmAttention

合并时间：2026-06-23 07:25

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46148

执行摘要

- 一句话：修复 ROCm 注意力层 q_scale 误报问题
- 推荐动作：值得合并，修复明确且影响范围有限。推荐关注后续注释精简的清理 PR。

功能与动机

修复 ROCm 上 test_kv_scale_reload 等测试失败：当 KV cache 为 FP8 且模型携带非 1.0 的 q_scale（如 nm-testing/Llama-3.2-1B-Instruct-FP8-KV）时，引擎初始化期间因断言 assert layer._q_scale_float == 1.0 失败。

实现拆解

1. 在 vllm/v1/attention/backends/rocm_attn.py 第 421-435 行，将原本的硬断言替换为带条件检查：当 is_quantized_kv_cache(self.kv_cache_dtype) 为真时，不再直接断言 q_scale == 1.0，而是先判断 query.dtype == self.fp8_dtype。
2. 仅在查询为 FP8 且 q_scale != 1.0 时，抛出 NotImplementedError，表明该情况暂不支持。
3. 当查询为非 FP8（如 bf16）时，忽略 q_scale，因为 chunked_prefill_paged_decode 不量化查询，q_scale 不参与计算，行为与 CUDA 后端一致。

关键文件：

- vllm/v1/attention/backends/rocm_attn.py（模块 注意力层；类别 source；类型 core-logic）：核心修复文件。修改了 forward() 方法中对 q_scale 的断言逻辑，将无条件断言改为条件检查（仅当查询为 FP8 时检查）。

关键符号：forward

关键源码片段

vllm/v1/attention/backends/rocm_attn.py

核心修复文件。修改了 forward() 方法中对 q_scale 的断言逻辑，将无条件断言改为条件检查（仅当查询为 FP8 时检查）。

```
# 文件：vllm/v1/attention/backends/rocm_attn.py（修改后）
if is_quantized_kv_cache(self.kv_cache_dtype):
    key_cache = key_cache.view(self.fp8_dtype)
```

```
value_cache = value_cache.view(self.fp8_dtype)
# chunked_prefill_paged_decode 使用全精度查询，
# 不量化 Q 也不消费 q_scale，
# 因此 q_scale 只在查询本身为 fp8 时有意义。
# 对于非 fp8 查询，q_scale 不适用且被忽略。
# 这避免了加载带有非 1.0 q_scale 的检查点时报错。
if query.dtype == self.fp8_dtype and layer._q_scale_float != 1.0:
    raise NotImplementedError(
        "A non 1.0 q_scale with an fp8 query is not currently "
        "supported by RocmAttentionImpl."
    )
```

评论区精华

Rohan138 建议精简代码中的注释（'Can we minimize/remove this comment?'），但原工程师休假，AndreasKaratzas 同意在后续 PR 中处理，先合并此修复。

- 注释长度 (style): AndreasKaratzas 同意在后续 PR 中处理注释精简，先合并修复。

风险与影响

- 风险：风险低。改动仅调整了断言条件，不影响 FP8 查询下的行为（仍会正确报错），且 TritonAttentionImpl 已有类似逻辑，对齐后数值正确性已通过测试验证（test_kv_scale_reload 中重载后困惑度断言通过）。
- 影响：直接影响 ROCm 平台下使用 FP8-KV 量化且附带非 1.0 q_scale 的模型加载路径。修复了引擎初始化崩溃，提升了 ROCm 上混合精度检查点的兼容性。对 CUDA 和其他后端无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR