PR #46548 完整报告

vllm-project/vllm

[ROCm] Fix OOB During Model Warmup With `ROCM_ATTN` and MRV2

合并时间：2026-06-24 23:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46548

执行摘要

- 一句话：修复 ROCm paged attention 内核 OOB 访问
- 推荐动作：建议合并。这是一个精准的 bugfix，修复根因而非绕行，且经过性能验证。测试扩展有效防止回归。

功能与动机

修复 issue #46179 以及大量 AMD CI 测试失败。在 MRV2 默认启用 Llama（PR #43458）后，ROCm paged attention 内核在 head_size=64 时存在越界读取，当 num_seqs=1024 进行预热时触发 segfault。

实现拆解

1. 修复内核越界（csrc/rocm/attention.cu）：在 paged_attention_ll4mi_QKV_mfma4_kernel 中，将 qhead_elemh8 的计算从 laneid / 4 改为 MIN(laneid / 4, HEAD_SIZE / 8 - 1)，确保多余的 lane 不会越界读取 query 张量。
2. 还原权宜回退（tests/kernels/quantization/test_triton_scaled_mm.py）：移除之前对 test_rocm_compressed_tensors_w8a8 测试强制使用 TRITON_ATTN 的限制，因为底层 bug 已修复。
3. 扩展测试覆盖（tests/kernels/attention/test_attention.py）：在 NUM_HEADS 中新增 (32, 8) 组合，在 HEAD_SIZES 中新增 64，并修改 opcheck 条件使其在 head_size==64 时执行，确保该场景被验证。同时更新 opcheck 调用参数以匹配最新接口。

关键文件：

- csrc/rocm/attention.cu（模块 ROCm 内核；类别 other；类型 core-logic；符号 paged_attention_ll4mi_QKV_mfma4_kernel）：核心修复文件，修改了 ROCm paged attention 内核中 query 索引的计算，防止越界访问
- tests/kernels/attention/test_attention.py（模块 注意力测试；类别 test；类型 test-coverage）：扩展测试矩阵，新增 head_size=64 覆盖，调整 opcheck 条件确保该场景被验证
- tests/kernels/quantization/test_triton_scaled_mm.py（模块 量化测试；类别 test；类型 test-coverage）：移除了之前为绕过此 bug 而对 TRITON_ATTN 的强制回退

关键符号：paged_attention_ll4mi_QKV_mfma4_kernel

关键源码片段

csrc/rocm/attention.cu

核心修复文件，修改了 ROCm paged attention 内核中 query 索引的计算，防止越界访问

```
// csrc/rocm/attention.cu
__launch_bounds__(NUM_THREADS) void paged_attention_ll4mi_QKV_mfma4_kernel(
    ...)
{
  // ...
  const scalar_t* q_ptr = q + query_start_off * q_stride + wg_start_head_idx * HEAD_SIZE;
  const _B16x8* q_ptrh8 = reinterpret_cast<const _B16x8*>(q_ptr);
  // 修复：将 qhead_elemh8 限制在有效范围内，防止当 laneid/4 等于 HEAD_SIZE/8 时越界
  // 多余的 lane 会读取最后一个有效元素，但这些值不会被用于最终输出
  const int qhead_elemh8 = MIN(laneid / 4, HEAD_SIZE / 8 - 1);
  // ...
}
```

tests/kernels/attention/test_attention.py

扩展测试矩阵，新增 head_size=64 覆盖，调整 opcheck 条件确保该场景被验证

```
# tests/kernels/attention/test_attention.py
# 扩展测试参数组合
NUM_HEADS = [(32, 8), (40, 40), (64, 8)] # 新增 (32,8)
HEAD_SIZES = [32, 64, 80, 128, 256] # 新增 64

# 修改 opcheck 条件，确保 head_size==64 时也运行
opcheck(
    torch.ops._rocm_C.paged_attention,
    (...),
    cond=(head_size == 64 and block_size == BLOCK_SIZES[0]),
)
```

tests/kernels/quantization/test_triton_scaled_mm.py

移除了之前为绕过此 bug 而对 TRITON_ATTN 的强制回退

```
# tests/kernels/quantization/test_triton_scaled_mm.py
@pytest.mark.skipif(not current_platform.is_rocm(), reason="Should only run on ROCm")
def test_rocm_compressed_tensors_w8a8(
    vllm_runner, example_prompts, model_path, max_tokens, num_logprobs
):
    dtype = "bfloat16"
    # 不再需要强制使用 TRITON_ATTN 来回避 ROCm paged attention 的 bug
    with vllm_runner(model_path, dtype=dtype) as vllm_model:
        vllm_model.generate_greedy_logprobs(example_prompts, max_tokens, num_logprobs)
```

评论区精华

无正式 review 评论，仅审核人 AndreasKaratzas 批准并感谢贡献。Author 和 co-author 在 issue 评论中详细分析了 bug 根因：内核在 head_size=64 时，laneid/4 可能计算出等于 HEAD_SIZE/8 的索引，导致读取 Q 元素时越界。修复仅增加一条 v_min_u32 指令（clamp），性能无影响，已通过性能扫描验证。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。修复仅在内核中增加一个 clamp 操作，且已确认性能无退化。测试扩展确保了 head_size=64 场景的回归覆盖。唯一潜在风险是 opcheck 条件从通用条件改为硬编码 head_size==64，可能漏检其他 head_size 的 opcheck，但原条件 head_size==HEAD_SIZES[0] 也仅覆盖 32，且 test_paged_attention 本身对所有 HEAD_SIZES 都会运行，opcheck 只是附加检查。
- 影响：直接影响 ROCm 平台上使用 MRV2 且 head_size=64 的模型（如 Llama-3.2-1B），修复了预热时的 segfault。间接修复了大量 AMD CI 测试失败，并移除了 TRITON_ATTN 的临时回退。不影响 CUDA 平台。
- 风险标记：ROCM 特定修复，GPU 内核修改

关联脉络

- PR #43458 [MRV2] Add Llama to default architectures: 该 PR 将 Llama 加入 MRV2 默认架构，暴露了此 ROCm 内核 bug
- PR #46180 [ROCm][CI] Pin test_rocm_compressed_tensors_w8a8 to TRITON_ATTN: 临时回退，本 PR 将其还原