PR #39225 完整报告

vllm-project/vllm

[Bug] Fix rocm sparse attn indexer issue

合并时间：2026-04-13 22:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39225

执行摘要

- 一句话：修复 ROCm 稀疏注意力索引器在推测解码下因张量填充导致的越界读取问题。
- 推荐动作：该 PR 代码简洁，但涉及底层内核安全，建议 ROCm 用户关注。值得精读 review 讨论中关于张量填充处理的权衡，理解为何未采纳 num_actual_tokens 方案。

功能与动机

修复 PR #39219 中讨论的问题（链接：https://github.com/vllm-project/vllm/pull/39219#discussion_r3047394497）。在推测解码场景下，k 张量可能被填充到 CUDA 图批次大小，而 slot_mapping 仅覆盖实际令牌数，若不截断会导致 ROCm 稀疏注意力内核发生越界读取。

实现拆解

在 vllm/v1/attention/ops/rocm_aiter_mla_sparse.py 文件的 rocm_aiter_sparse_attn_indexer 函数中，添加 5 行代码：获取 slot_mapping 的形状长度作为实际令牌数，并截断 k 张量至该长度，然后传递给后续的量化缓存操作。

关键文件：

- vllm/v1/attention/ops/rocm_aiter_mla_sparse.py（模块 attention）：唯一修改文件，修复 ROCm 稀疏注意力索引器的越界读取问题。

关键符号：rocm_aiter_sparse_attn_indexer

评论区精华

review 中 gemini-code-assist[bot] 建议使用 attn_metadata.num_actual_tokens 而非 slot_mapping.shape[0]，因为 slot_mapping 在 CUDA 图执行时也可能被填充；同时建议同时截断 k 和 slot_mapping 以保证一致性。作者 yewentao256 回复“num_actual_tokens is not a good value to use”，最终维持原方案。tjtanaa 批准了该 PR。

- 使用 slot_mapping.shape[0] vs num_actual_tokens 进行截断 (correctness): 作者回复 num_actual_tokens 不可用，维持原方案。

风险与影响

- 风险：风险较低：1）变更仅涉及 ROCm 稀疏注意力路径，影响面有限；2）修复逻辑与已有 CUDA 路径（sparse_attn_indexer.py）保持一致，已有验证；3）但未采纳 review 中

关于使用 num_actual_tokens 和同时截断 slot_mapping 的建议，若 slot_mapping 确实被填充，则修复可能不彻底，仍存在潜在越界风险。

- 影响：影响范围：仅影响使用 ROCm 平台且启用稀疏注意力与推测解码的用户。修复后避免内核越界读取导致的未定义行为（如崩溃或错误结果），提升系统稳定性。对性能无显著影响，仅增加一次张量切片操作。
- 风险标记：潜在未彻底修复，内核安全风险

关联脉络

- PR #39219 未知：当前 PR 的动机源自该 PR 的讨论（#discussion_r3047394497），可能涉及相同问题或相关代码。
- PR #37376 fused qknorm+rope kernel optimization for SM9.0: 同属内核优化与 bugfix 领域，关注底层计算正确性。
- PR #39644 [Bugfix] [Tests] Enforce out tensor device in kernel/moe/test_cutedsl_moe.py: 同为设备相关 bugfix，涉及张量处理与内核安全。