PR #44936 完整报告

vllm-project/vllm

[ROCm][V2] Fix failed assertion in Llama models when using EAGLE with `ROCM_AITER_FA`

合并时间：2026-06-10 02:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44936

执行摘要

- 一句话：修复 ROCm EAGLE 推断时断言失败
- 推荐动作：建议快速合入并 backport 到 ROCm 相关分支。此 PR 虽小，但体现了对硬件特定路径中控制流顺序的严谨思考，值得相关开发者了解。

功能与动机

PR#43458 为 Llama 模型启用了 MRV2，导致测试 tests/v1/e2e/spec_decode/test_spec_decode.py::test_eagle_correctness_heavy[ROCM_AITER_FA-llama3_eagle] 在 ROCm 上断言失败。断言 assert common_attn_metadata.seq_lens_cpu_upper_bound is not None 在纯解码批次中不适用，因为该字段仅用于 prefill 场景，且为其赋值需要无意义的 D→H 拷贝。

实现拆解

将 split_decodes_prefills_and_extends 函数中针对纯解码批次的提前返回逻辑（if max_query_len <= decode_threshold）移动到 seq_lens_cpu_upper_bound 断言之前。这样在纯解码批次中函数会在访问该字段前直接返回，避免断言失败和不必要的 Host 端数据拷贝。

关键文件：

- vllm/v1/attention/backends/utils.py（模块 注意力后端；类别 source；类型 core-logic；符号 split_decodes_prefills_and_extends）：核心变更文件，调整了 split_decodes_prefills_and_extends 函数中的控制流顺序，将纯解码提前返回移至断言之前。

关键符号：split_decodes_prefills_and_extends

关键源码片段

vllm/v1/attention/backends/utils.py

核心变更文件，调整了 split_decodes_prefills_and_extends 函数中的控制流顺序，将纯解码提前返回移至断言之前。

```
# vllm/v1/attention/backends/utils.py

def split_decodes_prefills_and_extends(
    common_attn_metadata: CommonAttentionMetadata,
    decode_threshold: int = 1,
```

```
) -> tuple[int, int, int, int, int, int]:
    """
    假设批次已重排序，找到 prefill 与 decode 请求的分界。
    """
    max_query_len = common_attn_metadata.max_query_len
    num_reqs = common_attn_metadata.num_reqs
    num_tokens = common_attn_metadata.num_actual_tokens
    query_start_loc = common_attn_metadata.query_start_loc_cpu

    # 【修改】先处理纯 decode 批次，避免无谓的断言和 D→H 拷贝
    if max_query_len <= decode_threshold:
        return num_reqs, 0, 0, num_tokens, 0, 0

    # seq_lens_cpu_upper_bound 只在有 prefill 时才有意义
    assert common_attn_metadata.seq_lens_cpu_upper_bound is not None
    seq_lens = common_attn_metadata.seq_lens_cpu_upper_bound

    query_lens = query_start_loc[1:] - query_start_loc[:-1]
    is_prefill_or_extend = query_lens > decode_threshold
    is_prefill = (seq_lens == query_lens) & is_prefill_or_extend
    # ... 后续逻辑不变 ...
```

评论区精华

该 PR 无 review 评论。审阅者 AndreasKaratzas 批准并指出此路径仅被 ROCm AITER Flash Attention 使用，不影响 CUDA。

- 暂无高价值评论线程

风险与影响

- 风险：变更极小且经过测试验证，风险很低。仅影响 ROCm AITER FA 后端的纯解码批次分支，不改变其他分支逻辑。但需注意若未来有代码依赖该断言提前触发（例如确保 seq_lens_cpu_upper_bound 被正确填充），此改动可能掩盖未初始化的数据。
- 影响：影响范围仅限于 ROCm 平台使用 EAGLE 投机解码且启用 ROCM_AITER_FA 后端的场景。修复后相应测试通过，Llama 模型在该配置下恢复正常运行。对其他平台、后端或模型无影响。
- 风险标记：最小变更，仅影响 ROCm AITER FA 后端

关联脉络

- PR #43458 Enable MRV2 for Llama models: 该 PR 为 Llama 启用了 MRV2，导致当前 PR 修复的断言失败。