PR #40754 完整报告

vllm-project/vllm

[Bugfix][ROCm] Fix gemm_a4w4 call to use updated AITER API signature

合并时间：2026-04-29 08:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40754

执行摘要

- 一句话：修复 AITER gemm_a4w4 API 变更导致的 MXFP4 GEMM bug
- 推荐动作：值得集成，变更集中且验证充分；对于维护 ROCm 推理管线的工程师，建议关注 AITER 的 API 变化及时跟进类似调整。

功能与动机

AITER 库的 gemm_a4w4 API 签名为避免多余显存分配，移除了预分配输出张量参数并改为直接返回输出结果。旧调用方式将 out 张量作为参数传递，在新 API 下被错误地当作 bias 参数，导致 MXFP4 GEMM 计算错误，量化模型无法正常运行。

实现拆解

1. 移除手动输出张量分配：删除对 torch.empty 的调用，该调用原用于预分配维度对齐后的输出张量，以匹配 gemm_a4w4 旧 API 的需求。
2. 更新 gemm_a4w4 调用签名：将 y（预分配张量）从参数中移除，改用关键字参数 dtype=out_dtype 指定输出数据类型，并以函数返回值的形式获取计算结果。
3. 保持后续行为不变：调用返回后，仍执行 return y[:M] 截断输出行数，与之前逻辑一致。

关键文件：

- vllm/model_executor/layers/quantization/quark/schemes/quark_ocp_mx.py（模块 量化层；类别 source；类型 data-contract）：包含对 AITER gemm_a4w4 调用的修复，移除已废弃的预分配输出张量参数，改用新返回值 API。

关键符号：未识别

关键源码片段

vllm/model_executor/layers/quantization/quark/schemes/quark_ocp_mx.py

包含对 AITER gemm_a4w4 调用的修复，移除已废弃的预分配输出张量参数，改用新返回值 API。

```
# vllm/model_executor/layers/quantization/quark/schemes/quark_ocp_mx.py

if rocm_use_aiter_fp4_asm_gemm:
    # ... 前面的条件分支，另一个 GEMM 路径不变 ...
```

```
else:
    if x_scales is None:
        x_q, x_s = per_1x32_f4_quant_hip(x, shuffle=True)
    else:
        x_q = x
        x_s = x_scales

    # AITER 新 API：不再需要调用方预分配输出张量，
    # gemm_a4w4 直接返回结果，dtype 通过关键字参数传递。
    y = gemm_a4w4(
        x_q,
        weight.view(x_q.dtype),
        x_s,
        weight_scale.view(x_s.dtype),
        dtype=out_dtype, # 新增参数， 替代旧式的预分配 out 张量
        bpreshuffle=True,
    )
# 截断输出行数，与之前逻辑保持一致
return y[:M]
```

评论区精华

无有意义的审查讨论；机器人审查员和两位维护者（Rohan138, tjtanaa）均批准了变更。Rohan138 指出该修复由 AITER PR#1679 引起，已在当前版本 AITER 0.1.10.post3 中生效。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. API 兼容性风险：若部署环境中的 AITER 版本低于 0.1.10.post3（即不支持新 API），调用会失败。但 AITER 已发布包含此变更的版本，且修复随 vLLM 升级即可同步。
  2. 回归风险：移除预分配和填充逻辑可能影响其他使用 gemm_a4w4 的路径。但当前 quark_ocp_mx.py 中只有这一处调用，且验证通过。
  3. 测试覆盖缺失：无对应单元测试变更，依赖手动验证（Llama-3.1-405B-MXFP4 模型），回归测试未自动化。 - 影响：直接影响使用 MXFP4 量化模型的 ROCm 用户，修复后模型可正确推理；间接影响依赖 AITER GEMM 的其他量化方案，但本变更仅修改 quark_ocp_mx.py，范围有限。影响程度：中（功能性修复，仅限特定硬件 + 量化组合）。 - 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- 暂无明显关联 PR