PR #46122 完整报告

vllm-project/vllm

[ROCm] [Performance] Optimize aiter moe for DeepSeekV4

合并时间：2026-06-26 21:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46122

执行摘要

- 一句话：优化 ROCm AITER MoE 权重预处理，提升 DeepSeekV4 性能约 9%
- 推荐动作：该 PR 是典型的小范围性能优化，改动集中但收益明确。建议 ROCm/DeepSeekV4 用户必须同步升级 AITER 版本；对于追求极致 MoE 推理性能的开发者，值得阅读 mxfp4.py 中的权重预处理逻辑以复用模式；注意 TODO 后续应跟踪 AITER 修复以移除临时环境变量。

功能与动机

针对 DeepSeekV4 模型，通过更优的权重预处理方案（参考 ROCm ATOM 仓库的'speed of light'参考实现）来提升 MoE 前向计算的性能。PR body 明确指出'This feature is validated with aiter v0.1.15.post1'。基准测试显示在并发 64 场景下，输出 tok/s 提升约 8.81%，总 tok/s 提升约 8.81%。

实现拆解

1. 移除旧有导入与重写 shuffle 逻辑：在 vllm/model_executor/layers/fused_moe/oracle/mxfp4.py 的 convert_weight_to_mxfp4_moe_kernel_format 函数中，AITER_MXFP4_BF16 分支原本通过 vllm._aiter_ops 和 aiter.utility.fp4_utils.e8m0_shuffle 执行权重 shuffle。现在改为从 aiter.ops.shuffle 直接导入 shuffle_weight 和 shuffle_scale，并移除无用的 e, n, k = w13_weight.shape 计算。
2. 新增环境变量以避免 AITER 崩溃：插入 os.environ["AITER_BF16_FP8_MOE_BOUND"] = "0" 调用，注释说明此乃临时方案（TODO），等待 AITER 侧修复后才能移除。该变量在 DeepSeekV4 权重预处理期间设置，避免因权重布局不匹配导致 AITER 内部崩溃。
3. 精细化 shuffle 参数传递：shuffle_weight 调用新增 is_guinterleave=True 和 gate_up 参数：w13 权重 gate_up=True，w2 权重 gate_up=False。shuffle_scale 调用也新增 num_experts 参数及两个布尔标志（表示使用 gu_interleave 和 gate_up），实现与 AITER 内核期望的精确布局匹配。
4. 更新注释与清理代码：移除 # No de-interleave: standard _load_w13 already produces [gate_all, up_all] layout 等过时注释，新增 # Initially introduced for DeepSeekV4 和 # TODO: Remove this once AITER is fixed 等明确注释。同时删除了对 w13_weight_scale.view 的调用替换为 reshape。

关键文件：

- vllm/model_executor/layers/fused_moe/oracle/mxfp4.py（模块 MoE 层；类别 source；类型 data-contract）：唯一变更文件，重写了 AITER_MXFP4_BF16 分支的权重 shuffle 逻辑并添加 AITER 临时规避的环境变量。

关键符号：convert_weight_to_mxfp4_moe_kernel_format

关键源码片段

vllm/model_executor/layers/fused_moe/oracle/mxfp4.py

唯一变更文件，重写了 AITER_MXFP4_BF16 分支的权重 shuffle 逻辑并添加 AITER 临时规避的环境变量。

```
# vllm/model_executor/layers/fused_moe/oracle/mxfp4.py
# 位于 convert_weight_to_mxfp4_moe_kernel_format 函数内，AITER_MXFP4_BF16 分支
elif mxfp4_backend == Mxfp4MoeBackend.AITER_MXFP4_BF16:
    # 该分支最初为 DeepSeekV4 引入
    if w13_bias is not None:
        w13_bias = w13_bias.data.to(torch.float32)
    if w2_bias is not None:
        w2_bias = w2_bias.data.to(torch.float32)

    import os
    from aiter.ops.shuffle import shuffle_scale as _shuf_s
    from aiter.ops.shuffle import shuffle_weight as _shuf_w

    # TODO: Remove this once AITER is fixed
    # 临时环境变量，避免 AITER 因权重交错内部崩溃
    os.environ["AITER_BF16_FP8_MOE_BOUND"] = "0"

    # w13 (gate+up): 使用 AITER 官方 shuffle API，is_guinterleave=True 表示交错权重
    w13_weight = torch.nn.Parameter(
        _shuf_w(
            w13_weight.data.view(torch.float4_e2m1fn_x2),
            is_guinterleave=True,
            gate_up=True, # gate+up 投影
        ),
        requires_grad=False,
    )
    shuffled_w13_scale = _shuf_s(
        w13_weight_scale.reshape(-1, w13_weight_scale.shape[-1]),
        num_experts,
        True, # use_gu_interleave
        True, # gate_up
    )

    # w2 (down-proj): gate_up=False
    w2_weight = torch.nn.Parameter(
```

```
        _shuf_w(
            w2_weight.data.view(torch.float4_e2m1fn_x2),
            is_guinterleave=True,
            gate_up=False,
        ),
        requires_grad=False,
    )
    shuffled_w2_scale = _shuf_s(
        w2_weight_scale.reshape(-1, w2_weight_scale.shape[-1]),
        num_experts,
        True, # use_gu_interleave
        False, # gate_up
    )

    return (
        w13_weight,
        w2_weight,
        shuffled_w13_scale,
        shuffled_w2_scale,
        w13_bias,
        w2_bias,
    )
```

评论区精华

Reviewer Rohan138 指出，TODO 注释中提及的 PR#3741 可能造成误解——该 PR 并未真正修复 AITER 缺少针对 DeepSeekV4 交错权重的 abf16w4 MoE GEMM 的底层问题，设置环境变量仍是权宜之计。作者 tjtanaa 随即更新了注释，改为表述'Necessary for AITER side from crashing'。后续 Rohan138 验证了 gpt-oss-120b 模型不受影响（GptOssMxfp4MoEMethod 是独立类）。最终获得 dllehr-amd 批准。

- TODO 注释的误导性 (documentation): 作者 tjtanaa 更新了注释，去掉对 PR#3741 的引用，改为通用表述 'Necessary for AITER side from crashing'。

风险与影响

- 风险：
  1. AITER 版本依赖：变更依赖 AITER v0.1.15.post1 以上的特定 API（shuffle_weight 和 shuffle_scale 的新签名），若 AITER 版本不匹配会导致运行时导入错误。
  2. 环境变量副作用：AITER_BF16_FP8_MOE_BOUND=0 是进程级全局环境变量，可能影响同一进程中其他使用 AITER 的模块（如 gpt-oss-120b），但经过 reviewer 验证，GptOssMxfp4MoEMethod 与 Mxfp4MoEMethod 分离，且环境变量仅在 DeepSeekV4 权重预处理分支中设置，影响范围可控。
  3. 缺失测试覆盖：仅有一个文件变更（43 行），没有新增测试用例，回归风险依赖手动基准测试。 - 影响：影响范围：仅限 ROCm 平台上使用 Mxfp4MoeBackend.AITER_MXFP4_BF16 后端的 DeepSeekV4 模型推理路径。其他模型（如 gpt-oss-120b）不受影响。性能收益：输出 token 吞吐量提升 3%-9%，TPOT 改

善 3%-8%，TTFT 在高并发下有 6%-20% 的提升。开发影响：无 API 变更，对调用方完全透明。

- 风险标记：外部依赖版本敏感，环境变量副作用，缺少测试覆盖

关联脉络

- PR #46419 [ROCm]Enable AITER MoE backend for MiniMax-M3-MXFP4: 同样修改了同一文件（vllm/model_executor/layers/fused_moe/oracle/mxfp4.py），涉及 AITER MoE 后端，功能上有延续性（均为 ROCm MoE 性能优化）。