PR #39705 完整报告

vllm-project/vllm

[Bugfix][Kernel][ROCm] Fix triton_w4a16 scales mismatch when BLOCK_K > group_size

合并时间：2026-04-14 02:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39705

执行摘要

本 PR 修复了 Triton W4A16 GEMM 内核在 BLOCK_K 大于量化组大小时，因单个计算瓦片跨越多个 scale 组却只使用第一个组的 scale，导致尾部行数据静默损坏的问题。修复方法是在内核启动器中强制将 BLOCK_K 限制为不超过 group_size。该问题在 ROCm 平台上使用特定量化模型进行长上下文工具调用时表现为模型行为异常，修复后模型运行正确且效率提升。

功能与动机

问题背景：作者在使用 ROCm RDNA3 平台运行 Qwen3.5-35B-A3B-GPTQ-W4A16-G32 模型时，发现模型在超过 10K 令牌的长上下文工具调用中表现异常：重复调用相同参数的工具、未完成任务、或幻觉指令。

根本原因：Triton W4A16 GEMM 内核（由 PR #37352 引入）在 triton_w4a16_gemm 函数中，当 BLOCK_K（默认 32）大于量化 group_size（如 32）时，单个计算瓦片会跨越多个量化 scale 组，但内核只加载第一个组的 scale 应用于整个瓦片，导致尾部行使用错误的 scale 进行反量化，静默损坏权重数据。

引用 PR body 关键表述：

> "When BLOCK_K exceeds group_size, a single tile spans multiple scale groups, but only the first group's scales are applied to all rows in the tile. This silently corrupts the dequantized weights in the tail rows."

实现拆解

仅修改一个文件：vllm/model_executor/kernels/linear/mixed_precision/triton_w4a16.py。

关键改动：在 triton_w4a16_gemm 函数中，在设置 BLOCK_M、BLOCK_N、BLOCK_K 默认值后，添加条件检查：

```
if group_size < BLOCK_K:
    BLOCK_K = group_size
```

逻辑说明：

- 默认 BLOCK_K = 32，但量化组大小 group_size 可能更小（如 32）。
- 当 group_size < BLOCK_K 时，将 BLOCK_K 限制为 group_size，确保每个计算瓦片不超过一个量化组，避免 scale 错配。
- 这保证了内核加载的 scale 与瓦片内所有行对应同一个量化组。

评论区精华

Review 讨论较少，但有两个关键评论：

1. *gemini-code-assist[bot]*：

"This pull request introduces a safety check in the triton_w4a16_gemm function to clamp the BLOCK_K parameter to the group_size. This change prevents potential data corruption that could occur if a processing tile spans multiple quantization groups, ensuring that the correct scales and zeros are applied during dequantization."

2. *yewentao256*：

"LGTM, thanks for the work!"

没有争议点，修复被迅速批准。

风险与影响

技术风险：

1. 静默数据损坏风险已修复：原问题导致权重反量化错误，输出不可预测，修复后确保正确性。
2. 性能潜在影响：当 group_size 较小时（如 32），BLOCK_K 被限制为较小值，可能降低计算效率，但这是正确性必需的权衡。
3. 回归风险低：仅影响使用 triton_w4a16_gemm 且 BLOCK_K > group_size 的场景，其他场景不受影响。
4. 缺少测试覆盖：PR 未添加测试用例，但基于问题描述，修复已通过实际模型（Qwen3.5-35B-A3B-GPTQ-W4A16-G32）验证。

影响评估：

- 用户影响：使用 Triton W4A16 量化内核（特别是 ROCm 平台）的用户在长上下文推理中将获得正确结果，避免模型行为异常。
- 系统影响：确保量化权重反量化正确性，提升模型输出质量和可靠性。
- 团队影响：揭示了内核实现中一个隐蔽的正确性问题，提醒在量化内核设计中需考虑 BLOCK_K 与 group_size 的匹配关系。

关联脉络

与历史 PR 的关联：

- PR #37352：引入了 Triton W4A16 GEMM 内核，本 PR 修复了该内核中的一个 bug。PR body 中明确提及："with the kernel introduced in PR #37352"。

功能演进方向：

- 近期多个 PR 涉及量化内核的 bugfix 和优化（如 PR #39717、#39604、#39418、#38707），显示团队在持续完善量化支持，特别是在多平台（ROCm、XPU）上的正确性和性能。

- 本 PR 是这一趋势的一部分，专注于 ROCm 平台上量化内核的正确性修复，确保长上下文推理的可靠性。