PR #39387 完整报告

vllm-project/vllm

[ROCm] Disable fused_silu_mul_block_quant on ROCm

合并时间：2026-04-10 01:59

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39387

执行摘要

- 一句话：临时禁用 ROCm 平台的特定量化融合，避免模型启动失败。
- 推荐动作：此 PR 变更简单但涉及平台兼容性设计，值得 ROCm 用户或关注量化编译的开发者精读，重点关注如何通过平台检查实现优雅降级，以及 review 中讨论的一致性考量。

功能与动机

PR body 中明确指出：'After https://github.com/vllm-project/vllm/pull/38817, some models are failing on ROCm.' 错误源于 QuantKey(f8e4m3fnuz,scale(f32,dynamic,GroupShape(row=1, col=128)),symmetric) 不被支持，因为 'torch.ops._C.per_token_group_fp8_quant' 在 ROCm 上尚未启用，需要临时禁用该路径以避免崩溃。

实现拆解

仅修改了 vllm/compilation/passes/fusion/act_quant_fusion.py 文件的一行代码：将 `if current_platform.is_cuda_alike():` 改为 `if current_platform.is_cuda():`。这限制了 SiluMulBlockQuantPattern 模式仅对 CUDA 平台注册，从而在 ROCm 上跳过相关融合，避免因缺失量化操作而导致的断言错误。

关键文件：

- vllm/compilation/passes/fusion/act_quant_fusion.py（模块 compilation/fusion）：包含平台检查逻辑，直接决定 SiluMulBlockQuantPattern 是否在 ROCm 上注册，是修复的核心文件。

关键符号：SiluMulBlockQuantPattern.init, current_platform.is_cuda

评论区精华

review 中 gemini-code-assist[bot] 指出使用 hasattr 与 QUANT_OPS 填充逻辑不一致，可能导致未来 ROCm 支持时仍出错，建议使用 current_platform.is_cuda()；tjtanaa 同意此建议。最终结论是采纳 is_cuda 作为临时修复，并计划后续通过 vLLM IR Ops 启用内核支持。

- 平台检查方法的选择 (design): 采纳建议，改为使用 current_platform.is_cuda() 作为临时修复，避免未来 ROCm 支持时的潜在断言错误。

风险与影响

- 风险：风险较低：变更仅影响条件检查，不会引入回归错误。但可能导致 ROCm 用户在特定 FP8 量化场景下性能下降，因为融合被禁用。兼容性依赖于平台检测的正确性，若 is_cuda() 在混合环境中误判，可能影响其他平台。
- 影响：对 ROCm 用户：解决了 Llama-3.1-70B-Instruct-FP8-KV 等模型的启动失败问题，提升稳定性。系统层面：在 ROCm 上禁用了 fused_silu_mul_block_quant 融合，可能轻微影响 FP8 量化推理性能。团队影响：需跟踪 ROCm 内核开发进度，以在未来重新启用融合优化。
- 风险标记：ROCm 功能降级，平台检测依赖

关联脉络

- PR #38817 未知：从 PR body 提及，引入了 fused_silu_mul_block_quant 融合，导致 ROCm 失败，是本 PR 的直接触发原因。
- PR #39122 [ROCm] Remove unnecessary fp8 roundtrip in gather cache NHD dequant: 同是 ROCm 平台上的 FP8 量化修复，涉及类似平台特定优化问题，可参考跨平台支持策略。