PR #38501 完整报告

vllm-project/vllm

[ROCm][Quantization] Add asymmetric INT8 quantization support to TritonInt8ScaledMMLinearKernel

合并时间：2026-04-06 09:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38501

执行摘要

此 PR 为 ROCm 平台的 TritonInt8ScaledMMLinearKernel 添加非对称 INT8 量化支持，解决了非对称 INT8 模型在 AITER 未启用时无法运行的问题。实现沿用 Cutlass 内核的数学原理，涉及核心 kernel 修改和测试配置更新，是 ROCm 量化支持的重要扩展，建议 ROCm 开发者关注。

功能与动机

为什么做？根据 PR body，此前 TritonInt8ScaledMMLinearKernel 仅支持对称输入量化，导致 ROCm 平台上的非对称 INT8 模型（如 W8-Channel-A8-Dynamic-Asym-Per-Token）无法运行。当 AITER 未启用时，Triton kernel 是唯一的 INT8 回退方案，因此需扩展支持以解锁模型兼容性。

实现拆解

核心改动模块：

- 量化 kernel：vllm/model_executor/kernels/linear/scaled_mm/triton.py – 修改 process_weights_after_loading 和 apply_weights 方法：
  - 对于非对称量化，计算单 scale 和 azp，预计算 azp_adj（权重列和）。
  - 在 GEMM 后应用零点校正项 scale_a * scale_b * azp * azp_adj。
- 测试基础设施：.buildkite/lm-eval-harness/test_lm_eval_correctness.py – 新增 _check_rocm_gpu_arch_requirement 函数，根据 required_gpu_arch 配置跳过不兼容硬件的测试。
- CI 配置：.buildkite/test-amd.yaml – 添加新的测试步骤“LM Eval Small Models (MI325)”以验证非对称 INT8 模型。
- 配置文件：更新多个 YAML 配置文件，添加 required_gpu_arch 字段，并扩展模型列表。

评论区精华

关键讨论线程：

1. 除以零风险：gemini-code-assist[bot] 指出代码中潜在除以零错误，建议添加 epsilon；作者回应：“This is copied directly from CutlassInt8ScaledMMLinearKernel. The

identical code runs in production on CUDA today without the epsilon guard." 此风险未解决，但基于现有实现被视为可接受。

2. 测试命名与架构检查：tjtanaa 建议："In this case, a proper way to gate this should be in test_lm_eval_correctness.py instead of through naming." 作者随后更新了测试逻辑，体现了设计权衡。

风险与影响

技术风险：

- 除以零风险：如果输入张量恒定，可能导致崩溃，但作者认为模型级别已损坏。
- 兼容性：需确保新逻辑与对称量化模型无回归。
- 测试覆盖：新增测试依赖特定 ROCm 硬件（如 gfx942、gfx950），可能覆盖不全。

影响范围：

- 用户：ROCm 用户可运行更多非对称 INT8 量化模型，提升平台吸引力。
- 系统：量化 kernel 功能增强，轻微增加计算开销，但无重大性能影响。
- 团队：代码库增加非对称处理逻辑，维护复杂性上升；CI 管道扩展以支持新测试。

关联脉络

与历史 PR 的关联：

- PR #38993：同为量化优化，涉及 Trtllm fp8 MoE 的权重布局调整，可参考跨内核的量化实现模式。
- PR #38870：修复 DeepSeek 模型在 FP8 量化下的权重加载 bug，与本 PR 的量化支持扩展技术相关。演进趋势：近期多个 PR（如 #38993、#32694）聚焦量化性能和支持清理，表明 vLLM 在量化领域持续优化，本 PR 是 ROCm 平台量化支持的重要补充，反映了跨平台一致性设计。