PR #38043 完整报告

vllm-project/vllm

{ROCm]: gpt-oss fusion/padding fixes

合并时间：2026-03-28 00:19

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38043

执行摘要

本 PR 修复了 ROCm 平台上 gpt-oss 模型的两个问题：将 MI300 的 padding 对齐从 128 调整为 256，并重新启用 RMSNorm+padding fusion，旨在提升性能和正确性。变更涉及配置和量化模块，风险较低但需注意相关潜在错误。

功能与动机

PR 目的是跟进 #37787，解决 gpt-oss RMSNorm+padding fusion 和 padding 本身的 minor issues。具体动机来自历史讨论：MI300 padding 应切换为 256（hidden_size 3072 替代 2944），参考了 #32307 和 #34285；同时，由于 #34304 已落地，重新启用 rmsnorm+pad fusion 以支持 Triton 和 wvsplitKrc GEMMs。

实现拆解

实现主要修改三个文件：

- vllm/config/vllm.py：更新 *enable_norm_pad_fusion* 函数，移除 not rocm_aiter_ops.is_triton_gemm_enabled() 检查，简化条件仅基于 hidden size 是否为 2880。python def enable_norm_pad_fusion(cfg: "VllmConfig") -> bool: """Enable if using AITER RMSNorm and hidden size is 2880 i.e. gpt-oss.""" from vllm._aiter_ops import rocm_aiter_ops return (rocm_aiter_ops.is_rmsnorm_enabled() and cfg.model_config is not None and cfg.model_config.get_hidden_size() == 2880)
- vllm/model_executor/layers/fused_moe/oracle/mxfp4.py：在 mxfp4_round_up_hidden_size_and_intermediate_size 函数中，将 ROCm 平台的 padding 对齐硬编码为 256。python elif current_platform.is_rocm(): intermediate_size = round_up(intermediate_size, 256) hidden_size = round_up(hidden_size, 256)
- vllm/model_executor/layers/quantization/utils/mxfp4_utils.py：删除 get_padding_alignment 函数，因为不再需要动态获取对齐值。

评论区精华

review 中，gemini-code-assist[bot] 评论指出另一个文件中的 hidden_pad 计算错误：

> "The calculation for hidden_pad appears to be incorrect. It's currently being set by getting hidden_pad from extra_weight_attrs, but this key is not present, so it defaults to 0. This will be incorrect when the hidden size is padded, potentially leading to memory access errors."

此问题未在本 PR 解决，但审核者均批准，表明变更被接受，但需关注潜在影响。

风险与影响

- 风险：padding 对齐变更可能影响依赖旧值的其他配置；移除 get_padding_alignment 函数需确保无残留引用；bot 提到的 hidden_pad 错误可能导致内存访问问题，需后续处理。
- 影响：主要影响 ROCm 平台上的 gpt-oss 模型，提升性能并修复 padding 错误；对系统优化了计算逻辑；对团队增强了代码稳定性。

关联脉络

与本 PR 最相关的是历史 PR #34285，它重构了 FusedMoE 的 padding 逻辑，为本 PR 的调整提供了上下文。其他关联 PR 如 #37787、#32307、#34304 在动机中提到，但不在近期分析列表中，显示了跨 PR 的功能演进方向：逐步优化 ROCm 平台的量化性能和融合支持。