PR #32914 完整报告

vllm-project/vllm

[ROCm][perf] Shuffle KV cache to use paged_attention_common

合并时间：2026-04-01 11:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/32914

执行摘要

本 PR 通过引入 aiter 的 paged_attention_common 函数优化 ROCm 平台上的 shuffle KV 缓存性能，解决了 Qwen 模型在小并发下性能下降的问题。变更影响 ROCm 注意力后端，显著提升吞吐量，但需注意对特定模型（如 Qwen3.5）的正确性风险和兼容性限制。建议团队关注后续修复和动态路由设计。

功能与动机

为什么做：根据 PR body，对于 Qwen/Qwen3-235B-A22B-Instruct-2507-FP8 模型，当前启用 VLLM_ROCM_SHUFFLE_KV_CACHE_LAYOUT=1 时，在小并发场景下性能比 =0 更差。PR 引用 aiter 的 paged_attention_common（见 https://github.com/ROCm/aiter/pull/1821）来修复此问题，旨在提升 ROCm 平台的注意力计算效率。

实现拆解

关键改动点：

- vllm/_aiter_ops.py：新增 paged_attention_common 静态方法，封装 aiter 库的同名函数，跳过 @is_aiter_supported 装饰器以允许显式后端选择。
- vllm/v1/attention/backends/rocm_aiter_fa.py：在 forward 函数中，当 rocm_aiter_ops.is_shuffle_kv_cache_enabled() 为真时，改用 rocm_aiter_ops.paged_attention_common，并创建临时张量（tmp_out、exp_sums、max_logits）以适配新接口。关键参数调整包括 KV 缩放传递和缓存视图重构。

评论区精华

核心讨论线程：

1. 设计权衡：tjtanaa 询问是否移除 VLLM_ROCM_SHUFFLE_KV_CACHE_LAYOUT 标志，samutamm 回应“shuffle-path 在部分场景有收益”，最终团队决定保留标志，因为 paged_attention_common 不支持滑动窗口等特性。

“So far VLLM_ROCM_SHUFFLE_KV_CACHE_LAYOUT seems useful in cases we've seen.” – samutamm

2. 正确性问题：tjtanaa 报告 Qwen3.5-397B 模型输出错误，tuukkjs 指出 pa_fwd_asm 仅支持 head_dim=128，需在 aiter 中添加路由逻辑，后续 PR 39192 跟进修复。

“pa_fwd_asm supports only head_dim=128. This model has afaik head_dim=256.” – tuukkjs

3. 代码质量：gemini-code-assist[bot] 指出变量重定义和死代码，提交中修复以提升可维护性。

风险与影响

具体风险：

- 回归风险：对于 head_dim!=128 的模型（如 Qwen3.5），paged_attention_common 可能路由到不支持的内核，导致输出错误或崩溃。
- 兼容性限制：与滑动窗口注意力不兼容，需保留环境变量标志，可能影响其他优化（如 rope+kvcache 融合）的集成。
- 性能依赖：动态路由阈值基于硬件配置，需在不同平台上验证以优化性能。

影响评估：

- 用户：ROCm 用户启用 shuffle kv cache 后将看到性能提升（测试显示 0.5% 到 13.4% 改善），但需注意模型兼容性。
- 系统：改进注意力计算路径，减少延迟，但增加代码复杂性和外部依赖。
- 团队：需监控后续正确性修复，并可能调整其他 ROCm 相关优化策略。

关联脉络

与历史 PR 的关系：

- PR 39192：直接后续，修复本 PR 中 paged_attention_common 对 Qwen3.5 模型的正确性问题。
- 其他 ROCm 相关 PR：如 39088（XPU 修复）、39087（AMD 内核修复），显示团队在跨平台性能优化上的持续投入。

演进趋势：本 PR 是 ROCm 平台上注意力内核优化的一部分，通过引入统一接口 paged_attention_common，为动态内核选择和性能调优奠定基础，未来可能扩展支持更多模型和特性。