PR #39177 完整报告

vllm-project/vllm

[ROCm][Perf] Expose AITER MoE sorting dispatch policy via env var

合并时间：2026-05-27 13:11

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39177

执行摘要

- 一句话：暴露 AITER MoE 调度策略环境变量，支持 ROCm 性能调优
- 推荐动作：推荐合并。改动简洁、默认向后兼容，且提供了明确性能收益。建议后续引入自动化测试，并在文档中记录该环境变量。

功能与动机

ROCm AITER 的 fused MoE 内核支持 moe_sorting_dispatch_policy 参数，但 vLLM 中无法配置，且上游 AITER 的 bool 类型注释错误（ROCm/aiter#2576）导致无法传递 >1 的值。修复后（ROCm/aiter#2639），需在 vLLM 侧暴露该参数，允许用户根据模型和工作负载选择最优策略，提升 MoE 模型性能。

实现拆解

1. 新增环境变量：在 vllm/envs.py 的 EnvVars 类中添加 VLLM_ROCM_AITER_MOE_DISPATCH_POLICY: int = 0，并在 environments 字典中定义读取逻辑，附带策略注释。
2. 缓存层封装：在 vllm/_aiter_ops.py 的 rocm_aiter_ops 类中添加类方法 get_moe_dispatch_policy()，使用 @if_aiter_supported 装饰并在类变量 _MOE_DISPATCH_POLICY 中缓存 env 值，避免热路径反复调用 envs.XXX。
3. 调用链串联：修改 vllm/model_executor/layers/fused_moe/experts/rocm_aiter_moe.py 中的 rocm_aiter_fused_experts 函数签名，增加 moe_sorting_dispatch_policy 参数；在 AiterExperts.apply() 中调用缓存方法获取值并传递，最终经 _aiter_ops.py 的 fused_moe 静态方法传递到 torch.ops.vllm.rocm_aiter_fused_moe。

关键文件：

- vllm/envs.py（模块 环境变量；类别 source；类型 configuration）：定义 VLLM_ROCM_AITER_MOE_DISPATCH_POLICY 环境变量的类型、默认值和读取逻辑，并附带策略说明注释，是用户入口。
- vllm/_aiter_ops.py（模块 算子层；类别 source；类型 core-logic；符号 get_moe_dispatch_policy）：添加 get_moe_dispatch_policy 缓存类方法，修改 fused_moe 静态方法签名以透传策略参数，是核心逻辑。
- vllm/model_executor/layers/fused_moe/experts/rocm_aiter_moe.py（模块 MoE 专家；类别 source；类型 data-contract）：修改 rocm_aiter_fused_experts 函数签名增加策略

参数，并在 AiterExperts.apply 热路径中调用缓存获取策略值，完成链路打通。

关键符号：get_moe_dispatch_policy, rocm_aiter_fused_experts, AiterExperts.apply

评论区精华

- gshtras 提出是否应采用 #41159 的通用 env 注册方案代替逐个添加变量，tpopp 认为方向更好但不应该阻塞 PR，最终保持现状。
- tjtanaa 指出不应在热路径中直接调用 envs.XXX，因为开销大，要求缓存。作者在 _aiter_ops.py 中添加了缓存类方法。
- AndreasKaratzas 建议为每个策略值添加说明和性能数据，tjtanaa 同意，最终在 env 注释中补充了详细描述。
- tjtanaa 对 or 0 提出疑问，作者修复为直接传递缓存值。
- 关于增加更多 aiter 环境变量的替代方案 (design): PR 保持原有方式，后续可参考 #41159 重构。
- 避免热路径直接读取 env 变量 (performance): 作者添加了 _MOE_DISPATCH_POLICY 类变量和 get_moe_dispatch_policy 方法缓存值。
- 为策略值添加描述性注释 (documentation): 作者在 env 配置注释中添加了详细策略描述。

风险与影响

- 风险：变更范围小（+32/-0），默认值 0 无行为变化，风险低。但缺少测试覆盖，且新环境变量依赖 AITER 版本（需 ≥ 修复后的版本）。热路径中增加了条件判断，但因缓存几乎无影响。
- 影响：对用户：ROCm 用户可以设置环境变量调优 MoE 性能，不影响其他平台。对系统：无性能回退风险。对团队：增加了一个持久化的环境变量，未来可能需跟随 AITER 扩展策略值。
- 风险标记：新环境变量，缺少测试覆盖，依赖上游 aiter 修复

关联脉络

- PR #2639 Annotate moe_sorting_dispatch_policy as int for fused_moe: 上游 aiter 修复，使本 PR 生效的前提条件
- PR #2576 [BUG] fused_moe moe_sorting_dispatch_policy wrong type: 触发本 PR 的 bug 报告