PR #45011 完整报告

vllm-project/vllm

[Refactor] Rename rocm_moe.py to rocm_moe_rdna.py

合并时间：2026-06-10 17:02

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45011

执行摘要

- 一句话：重命名 ROCm MoE 模块，明确 RDNA3 目标
- 推荐动作：本 PR 是一次安全的重命名操作，值得快速合并。它虽然没有功能改进，但提升了代码可维护性，并解决了潜在的命名隐患。对于关注 ROCm 量化路线图的工程师，建议留意后续可能新增的 rocm_moe_rdna4.py 等文件。

功能与动机

应评审人 @tjtanaa 要求重命名，解决原有 rocm_moe.py 名称过于通用，与仅支持 RDNA3 的实际情况不符的问题。新名称 rocm_moe_rdna.py 明确标识了目标架构，提升了代码可读性，并避免未来添加新架构实现时产生命名冲突。

实现拆解

1. 重命名模块文件：将 rocm_moe.py 重命名为 rocm_moe_rdna.py，文件内容不变。
2. 更新导入语句：在 compressed_tensors_moe.py 中，将 from . import rocm_moe 改为 from . import rocm_moe_rdna，同时更新后续引用到 rocm_moe 的两处调用（is_supported 和 make_method）。
3. 更新测试引用：在 test_rdna3_compile_guards.py 中，将所有测试函数和类中的 import rocm_moe 替换为 import rocm_moe_rdna，并更新 docstring 和断言消息中的模块名。涉及 10 处引用，包括 test_rocm_moe_not_supported_on_non_gfx1100 和 TestMoEDispatchMocked 类中的所有方法。

关键文件：

- vllm/model_executor/layers/quantization/compressed_tensors/compressed_tensors_moe/compressed_tensors_moe.py（模块 量化调度；类别 source；类型 data-contract；符号 get_moe_method）：核心调度入口，更新了导入和调用以引用新模块名。
- vllm/model_executor/layers/quantization/compressed_tensors/compressed_tensors_moe/rocm_moe_rdna.py（模块 量化调度；类别 source；类型 rename-or-move；符号 is_supported, make_method）：实际被重命名的模块，内容不变，但文件名明确表明针对 RDNA3。
- tests/kernels/quantization/test_rdna3_compile_guards.py（模块 编译守卫；类别 test；类型 test-coverage；符号 test_rocm_moe_not_supported_on_non_gfx1100, TestMoEDispatchMocked）：测试文件，所有模块引用均同步更新，确保重命名后测试通过。

关键符号：get_moe_method, is_supported, make_method

关键源码片段

vllm/model_executor/layers/quantization/compressed_tensors/compressed_tensors_moe/rocm_moe_rdna.py

实际被重命名的模块，内容不变，但文件名明确表明针对 RDNA3。

```
# 模块原本名为 rocm_moe.py，现改为 rocm_moe_rdna.py
# 以下为完整文件内容（未改动）
"""ROCm MoE kernel dispatcher.

Selects architecture-specific native HIP MoE kernels in priority order.
Falls back to the Triton WNA16 path when no native kernel is available.
"""

import torch
from vllm.logger import init_logger

logger = init_logger(__name__)

def is_supported(weight_quant) -> bool:
    """Check if a native ROCm MoE kernel is available for this config."""
    if weight_quant.num_bits != 4:
        return False
    from vllm.platforms.rocm import on_gfx1100
    # RDNA3 (gfx1100). Future: add RDNA4 (gfx12x), CDNA (gfx94x), etc.
    return (
        on_gfx1100()
        and hasattr(torch.ops, "_rocm_C")
        and hasattr(torch.ops._rocm_C, "moe_gptq_gemm_rdna3")
    )

def make_method(weight_quant, input_quant, moe_config):
    """Create the native ROCm MoE method. Call only after is_supported()."""
    from vllm.platforms.rocm import on_gfx1100
    if on_gfx1100():
        from .compressed_tensors_moe_wna16_rdna3 import
        CompressedTensorsWNA16RDNA3MoEMethod
        logger.info_once(
            "Using CompressedTensorsWNA16RDNA3MoEMethod (native RDNA3 HIP kernel)"
        )
        return CompressedTensorsWNA16RDNA3MoEMethod(
            weight_quant, input_quant, moe_config
        )
    # Future: RDNA4, CDNA, etc.
    raise RuntimeError("is_supported() returned True but no kernel matched")
```

评论区精华

无 review 评论。PR 由 @yewentao256 和 @tjtanaa 审批通过，@tjtanaa 是重命名的原始请求者，@yewentao256 表示“LGTM, thanks for the work!”。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。本 PR 仅涉及模块重命名与引用更新，补丁完全对称（19+ / 19-），无功能语义变化。主要风险在于外部依赖如 __init__.py 或未跟踪的导入可能遗漏更新，但已通过测试文件验证了所有内部引用。
- 影响：影响范围仅限于 compressed_tensors_moe 包内部和对应的测试文件。对外部用户完全透明，不影响任何 API、模型加载或推理行为。未来引入 RDNA4/CDNA MoE 实现时，命名空间更加清晰，可避免歧义。
- 风险标记：极低风险，纯重命名

关联脉络

- 暂无明显关联 PR