PR #45404 完整报告

vllm-project/vllm

fix(moe_wna16): access tp_size via moe_config for RoutedExperts compatibility

合并时间：2026-06-23 23:46

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45404

执行摘要

- 一句话：修复 MoE WNA16 加载时 tp_size 属性缺失
- 推荐动作：此 PR 修复清晰且必要，无复杂设计决策。建议精读以了解 RoutedExperts 和 MoE 量化加载的数据结构差异，对其他 MoE 加载器具有参考价值。

功能与动机

Issue #45403 报告了在使用 TP=2 加载 AWQ MoE 模型（Qwen3-Coder-30B-A3B 等）时出现 AttributeError: 'RoutedExperts' object has no attribute 'tp_size'。根本原因是 moe_wna16_weight_loader 直接访问 layer.tp_size，但 RoutedExperts 类从未设置 self.tp_size，该值仅通过 layer.moe_config.tp_size 获取。

实现拆解

1. 删除局部变量 tp_size：在 moe_wna16_weight_loader 函数开头移除 tp_size = layer.moe_config.moe_parallel_config.tp_size 这一行，消除未使用变量（F841）。
2. 修改 w13_qzeros 分支：将 loaded_weight.view(tp_size, -1, loaded_weight.size(1)) 改为 loaded_weight.view(layer.moe_config.tp_size, -1, loaded_weight.size(1))。
3. 修改 w2_qzeros 分支：将 loaded_weight.size(0), tp_size, -1 改为 loaded_weight.size(0), layer.moe_config.tp_size, -1。
4. 代码质量修复：通过多个提交修复 pre-commit 检查（ruff 格式化、导入排序）和 DCO 签名，最终通过所有 CI 检查后合并。

关键文件：

- vllm/model_executor/layers/quantization/moe_wna16.py（模块 量化层；类别 source；类型 data-contract；符号 moe_wna16_weight_loader）：唯一的变更文件，修改了 moe_wna16_weight_loader 函数中两处 tp_size 的访问方式，是修复的核心。

关键符号：moe_wna16_weight_loader

关键源码片段

vllm/model_executor/layers/quantization/moe_wna16.py

唯一的变更文件，修改了 moe_wna16_weight_loader 函数中两处 tp_size 的访问方式，是修复的核心。

```
# 修改后的 moe_wna16_weight_loader 函数相关片段
# 之前通过局部变量 tp_size 访问 layer.tp_size,
# 但 RoutedExperts 将 tp_size 存储在 moe_config 中

def moe_wna16_weight_loader(
    param: torch.nn.Parameter,
    loaded_weight: torch.Tensor,
    weight_name: str,
    shard_id: str,
    expert_id: int,
    return_success: bool = False,
):
    # ...
    device = get_tp_group().device
    tp_rank = get_tensor_model_parallel_rank()
    # 删除上一行 : tp_size = layer.moe_config.moe_parallel_config.tp_size
    loaded_weight = loaded_weight.to(device)
    shard_size = layer.intermediate_size_per_partition

    # ... 权重格式转换 ...

    if "w13_qzeros" in weight_name:
        # 使用 layer.moe_config.tp_size 替换之前的 tp_size
        tensor = loaded_weight.view(
            layer.moe_config.tp_size, -1, loaded_weight.size(1)
        )[tp_rank]
        if shard_id == "w1":
            param.data[expert_id, : shard_size // 2] = tensor
        else:
            param.data[expert_id, shard_size // 2 :] = tensor
        return True if return_success else None
    elif "w2_qzeros" in weight_name:
        param.data[expert_id] = loaded_weight.view(
            loaded_weight.size(0), layer.moe_config.tp_size, -1
        )[:, tp_rank]
        return True if return_success else None
    else:
        # ... 委托给原始加载器 ...
        return weight_loader(...)
```

评论区精华

讨论主要集中在 pipeline 的自动检查上：Mergify 多次提示 pre-commit 失败和合并冲突，作者 Oxygen56 逐一修复（rebase、解决冲突、通过 DCO）。审核人 yewentao256 和 bnellnm 最终批准，未出现实质性技术争议。

- 暂无高价值评论线程

风险与影响

- 风险：此变更仅修改一个文件中的两处属性访问路径，且与已有模式（layer.moe_config.tp_size）保持一致，回归风险低。但代码未包含新增测试来验证 RoutedExperts 下的加载行为，可能遗漏未来重构导致的回归。
- 影响：修复了 AWQ MoE 模型在 TP≥2 时加载失败的直接影响。受影响用户为使用 AWQ MoE 模型（如 Qwen3 系列）且启用张量并行的用户。变更范围极小，不涉及 API 变化或性能改变。
- 风险标记：缺少测试覆盖，核心路径变更

关联脉络

- PR #45403 AttributeError: 'RoutedExperts' object has no attribute 'tp_size' when loading AWQ MoE models with TP≥2: 关联 Issue，报告了此 PR 修复的 bug。