PR #46163 完整报告

vllm-project/vllm

[CI] Fix missing `tp_size` attribute on `RoutedExperts`

合并时间：2026-06-22 08:46

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46163

执行摘要

- 一句话：修复 RoutedExperts 缺少 tp_size 属性导致的崩溃
- 推荐动作：建议快速合并。这是一个典型的属性名兼容性修复，修改直接、风险低。对于关注 MoE 量化模型加载的开发者，该 PR 可作为了解 RoutedExperts 配置结构调整的参考。

功能与动机

修复 AMD CI 流水线中 Gemma-4 模型加载时的崩溃：AttributeError: 'RoutedExperts' object has no attribute 'tp_size'。根本原因是 layer.tp_size 已被移除，但 moe_wna16.py 中的权重加载器未同步更新。PR body 附带了完整的堆栈跟踪，显示崩溃发生在 gemma4.py 的 load_weights 中，最终归结到 moe_wna16_weight_loader。

实现拆解

1. 定位问题：在 vllm/model_executor/layers/quantization/moe_wna16.py 的 moe_wna16_weight_loader 函数中，第 467 行和 476-477 行使用了 layer.tp_size，但 RoutedExperts 层不再直接暴露该属性。
2. 新增局部变量：在第 429 行（原 428 行之后）插入 tp_size = layer.moe_config.moe_parallel_config.tp_size，从 MoE 并行配置对象中读取正确的 TP 大小。
3. 替换引用：将第 467-468 行的 loaded_weight.view(layer.tp_size, -1, ...) 替换为 loaded_weight.view(tp_size, -1, ...)，并将第 476-477 行的 loaded_weight.size(0), layer.tp_size, -1 替换为 loaded_weight.size(0), tp_size, -1。
4. 整理代码格式：合并了 w13_qzeros 分支的 view 调用，使代码更紧凑。整个修改仅涉及一个文件，共 7 行变更（+3/-4），无新增测试或配置变化。

关键文件：

- vllm/model_executor/layers/quantization/moe_wna16.py（模块 量化层；类别 source；类型 data-contract）：唯一修改的文件。修复了 MoE WNA16 权重加载器中对 tp_size 的错误引用，确保从正确的配置链中获取张量并行大小。

关键符号：moe_wna16_weight_loader

关键源码片段

vllm/model_executor/layers/quantization/moe_wna16.py

唯一修改的文件。修复了 MoE WNA16 权重加载器中对 tp_size 的错误引用，确保从正确的配置链中获取张量并行大小。

```
# vllm/model_executor/layers/quantization/moe_wna16.py
# 关键变更：新增局部变量 tp_size 以替换已移除的 layer.tp_size

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,
):
    # ... 前置转换逻辑 ...

    device = get_tp_group().device
    tp_rank = get_tensor_model_parallel_rank()
    # 修复：从 moe_config 中获取 tp_size，而非直接访问 layer.tp_size（已被移除）
    tp_size = layer.moe_config.moe_parallel_config.tp_size
    loaded_weight = loaded_weight.to(device)
    shard_size = layer.intermediate_size_per_partition

    # ... 重复 qzeros/scales 逻辑 ...

    if "w13_qzeros" in weight_name:
        # 使用 tp_size 变量来分割张量
        tensor = loaded_weight.view(tp_size, -1, loaded_weight.size(1))[tp_rank]
        # ... 后续赋值 ...
    elif "w2_qzeros" in weight_name:
        param.data[expert_id] = loaded_weight.view(
            loaded_weight.size(0), tp_size, -1
        )[:, tp_rank]
        # ... 后续 ...
```

评论区精华

PR 获得 3 个 Approved（BowenBao、yewentao256、mgoin），无 review 评论或争议。Issue 评论中 AndreasKaratzas 指出 AMD CI 测试将在其他 PR 中解决，本 PR 仅聚焦于属性兼容性修复。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：修改范围局限在单一函数 (moe_wna16_weight_loader)，仅将硬编码的属性访问改为从 moe_config 间接获取。由于 layer.moe_config.moe_parallel_config.tp_siz

e 是 MoE 层的标准配置路径，该修改不会引入回归，且不影响非 MoE 场景或无 TP 的量化模型。

- 影响：影响范围窄：仅影响使用 MoE + WNA16 量化（如 AWQ/GPTQ 量化 + MoE 模型）的模型加载路径，具体为 Gemma-4 等模型。修复后，这些模型可以在启用了张量并行的环境下正常加载权重。对不涉及 MoE 或 WNA16 量化的场景无影响。
- 风险标记：低影响

关联脉络

- PR #44935 [MoE] Refactor MoE parallel config (remove tp_size from RoutedExperts)：本 PR 修复的 bug 正是由于该重构移除了 layer.tp_size 但未更新 moe_wna16.py 中的引用所致。