PR #47154 完整报告

vllm-project/vllm

[Bugfix] compressed-tensors: allow int8 grouped WNA16 MoE on Marlin

合并时间：2026-07-01 03:50

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47154

执行摘要

- 一句话：移除 int8 分组量化断言限制
- 推荐动作：变更简单明确，值得所有使用 compressed-tensors + Marlin MoE 的团队立即合入。建议补充一个针对 group_size=128 的 int8 MoE 正确性测试，防止未来重构时再次引入类似约束。

功能与动机

PR body 指出：num_bits==8 分支断言 group_size==-1，拒绝了使用 group_size（如 128）的 int8 专家权重，但 Marlin 内核实际上支持 uint8b128 量化类型，且 MARLIN_SUPPORTED_GROUP_SIZES 包含 128。该断言冗余且阻碍了混合精度模型（如 poolside/Laguna-XS.2-INT4，同时包含 INT4 和 INT8 层）的正常使用。

实现拆解

仅有一处变更：

1. 在 vllm/model_executor/layers/quantization/compressed_tensors/compressed_tensors_moe/compressed_tensors_moe_wna16_marlin.py 的 __init__ 方法中，num_bits==8 分支下删除了 assert self.group_size == -1 这一行。
2. 删除后，scale 直接赋值为 kInt8StaticGroupScale（原本只有在通过断言后才会执行），这与 Marlin 内核实际支持的 group_size 范围（-1，32，64，128）一致。

关键文件：

- vllm/model_executor/layers/quantization/compressed_tensors/compressed_tensors_moe/compressed_tensors_moe_wna16_marlin.py（模块 量化层；类别 source；类型 data-contract；符号 CompressedTensorsWNA16MarlinMoEMethod.init）：删除导致 int8 分组量化被拒绝的断言，是变更的唯一文件

关键符号：CompressedTensorsWNA16MarlinMoEMethod.init

关键源码片段

vllm/model_executor/layers/quantization/compressed_tensors/compressed_tensors_moe/compressed_tensors_moe_wna16_marlin.py

删除导致 int8 分组量化被拒绝的断言，是变更的唯一文件

```
# compressed_tensors_moe_wna16_marlin.py（修改后）
# 原 __init__ 方法中 num_bits==8 分支的 assert 被移除
if self.num_bits == 4:
    if self.group_size == 32:
        scale = kInt4Static32GroupScale
    else:
        scale = kInt4StaticGroupScale
elif self.num_bits == 8:
    # 移除了 assert self.group_size == -1
    # Marlin 内核实际支持 group_size = -1, 32, 64, 128
    scale = kInt8StaticGroupScale
else:
    raise ValueError(
        "CompressedTensorsWNA16MarlinMoEMethod only supports int4 and int8 now."
    )
```

评论区精华

讨论较少：mgoin 和 yewentao256 直接批准，claude[bot] 因来自 fork 未执行自动审查。Mergify 提醒了 pre-commit 检查失败，但后续提交已修复。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅删除一个冗余断言，不影响其他分支逻辑。原有断言拒绝的合法配置（如 group_size=128）现在会被正确接受，不会造成回归。但需注意缺失对应测试覆盖：本次变更没有新增或修改测试文件来验证分组 int8 MoE 场景。
- 影响：影响范围有限，但对使用 compressed-tensors 量化且包含 int8 分组 MoE 层的用户是直接阻塞修复。特别是 poolside/Laguna-XS.2-INT4 等混合精度模型现在可以正常加载。不影响 int4 或其他量化格式。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR