PR #36329 完整报告

vllm-project/vllm

[Bugfix] Fix Qwen3.5 GatedDeltaNet in_proj_ba Marlin failure at TP>=2

合并时间：2026-05-21 12:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/36329

执行摘要

- 一句话：修复 Qwen3.5 GDN 在 TP>=2 时 Marlin 量化失败
- 推荐动作：值得所有 Qwen3.5 用户和关注量化兼容性的工程师阅读。本 PR 展示了如何在不改动权重加载和模块映射的前提下，通过 disable_tp 优雅地规避量化内核的分片限制，并利用 split_ba 保持下游 kernel 的兼容性。review 中的设计权衡讨论也很有启发性。

功能与动机

Issue #35924 报告：Qwen3.5-27B/397B 使用 GPTQ/AWQ Marlin 量化且在 TP>=2 时，GatedDeltaNet 的 in_proj_ba 因 MergedColumnParallelLinear 输出分片小于 64 而抛出 ValueError。用户 naroam1 确认该问题阻塞其生产部署的迁移。PR 作者 sonusflow 经分析确定根因为 Marlin 内核的 MIN_THREAD_N=64 限制。

实现拆解

1. 新增量化配置导入：在 gdn_linear_attn.py 中导入 AutoGPTQConfig, AWQMarlinConfig, INCConfig，用于后续类型判断。
2. 添加 maybe_disable_tp 方法：判断当前是否使用了 Marlin 系列量化（AWQ/GPTQ/INC）且平台为 CUDA、布局非 interleaved，若是则返回 True，表示应对 ba_proj 禁用 TP 分片。
3. 修改 create_ba_proj：在调用 MergedColumnParallelLinear 时传入 disable_tp=self.maybe_disable_tp(quant_config)，使得在需要时每个 rank 持有完整权重而非分片。
4. 新增 split_ba 方法：对 ba 输出执行 chunk(2, dim=-1) 后，若已禁用 TP 且 TP>1，则根据 tp_rank 和 tp_size 切片到本地输出，保证下游 kernel 形状正确。
5. 适配 forward 路径：在 forward_cuda 和 forward_xpu 中用 split_ba 替代直接的 ba.chunk(2, dim=-1)，同时为 XPU 路径添加了条件切片逻辑。

关键文件：

- vllm/model_executor/layers/mamba/gdn_linear_attn.py（模块 模型层；类别 source；类型 core-logic；符号 maybe_disable_tp, split_ba, create_ba_proj）：核心修改文件，新增 maybe_disable_tp 和 split_ba 方法，修改 create_ba_proj 传入 disable_tp 参数，并在 forward 路径中适配。

关键符号：maybe_disable_tp, split_ba, create_ba_proj

关键源码片段

vllm/model_executor/layers/mamba/gdn_linear_attn.py

核心修改文件，新增 maybe_disable_tp 和 split_ba 方法，修改 create_ba_proj 传入 disable_tp 参数，并在 forward 路径中适配。

```
# 在 __init__ 中，创建 ba_proj 并记录是否禁用 TP
self.in_proj_ba = self.create_ba_proj(
    hidden_size=self.hidden_size,
    num_v_heads=self.num_v_heads,
    quant_config=quant_config,
    prefix=f"{prefix}.in_proj_ba",
)
# 记录是否需要禁用 TP 分片，用于 split_ba 中判断
self.disable_tp_for_ba_proj = self.maybe_disable_tp(quant_config)

def create_ba_proj(self, hidden_size, num_v_heads, quant_config, prefix):
    # ... 原有实现
    return MergedColumnParallelLinear(
        input_size=hidden_size,
        output_sizes=[num_v_heads] * 2,
        bias=False,
        quant_config=quant_config,
        prefix=prefix,
        # 当使用 Marlin 系列量化且平台为 CUDA 时禁用 TP 分片
        disable_tp=self.maybe_disable_tp(quant_config),
    )

def maybe_disable_tp(self, quant_config: QuantizationConfig | None) -> bool:
    """判断是否应对 ba_proj 禁用 TP 分片。
    Marlin 内核要求输出维数 >= MIN_THREAD_N=64,
    Qwen3.5 非 interleaved 布局的 ba_proj 输出为 [num_v_heads]*2,
    TP 分片后可能低于阈值（如 64/4=16），因此需要禁止 TP 分片,
    使每个 rank 持有完整权重，并在 forward 中手动切片。
    Qwen3-Next interleaved 布局不受影响。
    """
    return (
        current_platform.is_cuda() # Marlin 仅 CUDA
        and not self.gqa_interleaved_layout # 仅非 interleaved
        and isinstance(quant_config, (
            AWQMarlinConfig,
            AutoGPTQConfig,
            INCConfig,
        )) # 仅 Marlin 类量化
    )

def split_ba(self, ba: torch.Tensor) -> tuple[torch.Tensor, torch.Tensor]:
    """将 ba_proj 输出按最后一个维度 split 为 b 和 a,
    若已禁用 TP 则需要根据 tp_rank 和 tp_size 切片到本地输出。"""
```

```
b, a = ba.chunk(2, dim=-1)
if self.disable_tp_for_ba_proj and self.tp_size > 1:
    chunk_size = self.num_v_heads // self.tp_size
    start = self.tp_rank * chunk_size
    b = b[:, start : start + chunk_size]
    a = a[:, start : start + chunk_size]
return b, a
```

评论区精华

审查者 Isotr0py 最初反对直接替换为 ReplicatedLinear，认为会引起性能退化，建议改为使用 MergedColumnParallelLinear 的 `disable_tp=True` 参数。作者采纳该建议并进行了重构，新增 `maybe_disable_tp` 和 `split_ba` 方法。此外，Isotr0py 指出 Marlin 内核仅支持 CUDA，需要在条件中加入 `current_platform.is_cuda()` 检查，作者相应调整。

- 使用 ReplicatedLinear 替换可能导致性能退化 (design): 作者采纳建议，改用 disable_tp=True 方案，并新增 maybe_disable_tp 和 split_ba 方法。
- Marlin 内核仅支持 CUDA 平台 (correctness): 作者添加了 is_cuda() 条件，并相应调整了 forward_xpu 中的切片逻辑。

风险与影响

- 风险:
  1. 性能风险：ReplicatedLinear 的复制计算可能引入额外开销，但实测 37 tok/s 匹配 FP16 基线。瓶颈在于 Marlin 已可高效运行，整体性能优化。
  2. 兼容性风险：仅对 Marlin 系列量化（AWQ/GPTQ/INC）生效，其他量化方式或非 CUDA 平台不受影响。XPU 路径也同步添加了条件切片，兼容性有保证。
  3. 量化精度：quant_config 被正确传递给 MergedColumnParallelLinear，量化精度完全保持。
  4. 回归风险：仅修改 ba_proj 一个模块，且通过 maybe_disable_tp 条件控制，默认行为不变。 - 影响：影响所有使用 Qwen3.5 GatedDeltaNet 并在 TP>=2 时启用 Marlin 量化（AWQ/GPTQ/INC）的用户。此前模型加载直接崩溃，修复后可以正常运行，实测单用户吞吐 37 tok/s。Qwen3-Next（interleaved 布局）不受影响。社区用户 naroam1 已确认该 PR 是 Qwen3.5-27B-AWQ 在 TP=2 下迁移的关键阻塞。 - 风险标记：量化精度保持，性能无退化，CUDA 平台依赖，非 interleaved 布局限制

关联脉络

- 暂无明显关联 PR