PR #48350 完整报告

vllm-project/vllm

[Model] Optimize Qwen3.5 on H20

合并时间：2026-07-13 11:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/48350

执行摘要

本 PR 为 Qwen3.5 模型在 NVIDIA H20 GPU 上添加了专用的 MoE 内核调优配置文件，通过自动调优器生成的参数显著提升了解码阶段吞吐量并降低延迟。变更仅包含一个 JSON 文件，无代码修改，风险极低。

功能与动机

Qwen3.5 在 H20 上的默认 MoE 内核未充分利用硬件特性，导致性能瓶颈。PR 作者使用 vLLM 调优器针对 H20 生成最优的 BLOCK_SIZE、GROUP_SIZE 和 warp 数目，旨在提升 Tensor-Parallel 推理效率。

实现拆解

1. 自动调优：以 Triton 3.6.0 为目标生成 18 组专家容量对应的内核参数组合。
2. 配置部署：将调优结果写入 vllm/model_executor/layers/fused_moe/configs/E=256,N=256,device_name=NVIDIA_H20.json，设备自动匹配加载。
3. 验证：在 2×H20、TP=2 环境下对 7 种输入场景进行基准测试，并对 GSM8K 数据集做准确率验证。

vllm/model_executor/layers/fused_moe/configs/E=256,N=256,device_name=NVIDIA_H20.json

这是本 PR 唯一的变更文件，新增针对 H20 的 MoE 内核调优配置，直接决定了 Qwen3.5 在该 GPU 上的推理性能。

```
{
  // 生成此配置使用的 Triton 编译器版本
  "triton_version": "3.6.0",
  // 不同激活专家数（M）对应的内核参数
  "16": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 256,  // N 维度较大，适合中等容量
    "BLOCK_SIZE_K": 64,
    "GROUP_SIZE_M": 64,   // 较大的 group size 提升并行度
    "num_warps": 8,       // 8 warp 充分利用 SM
    "num_stages": 3
  },
  "128": {
```

```
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 64,
    "BLOCK_SIZE_K": 64,
    "GROUP_SIZE_M": 16,
    "num_warps": 4,
    "num_stages": 3
  },
  "2048": {
    "BLOCK_SIZE_M": 64,  // 大容量使用更大的 M 块
    "BLOCK_SIZE_N": 128,
    "BLOCK_SIZE_K": 64,
    "GROUP_SIZE_M": 32,
    "num_warps": 8,
    "num_stages": 3
  }
  // 其余 15 个条目省略，结构类似
}
```

评论区精华

无讨论。

风险与影响

- 风险：配置仅针对 Triton 3.6.0 和 H20 设备，若环境不符可能性能不达预期；但不会导致错误。
- 影响：仅对使用 Qwen3.5 + H20 的用户有正面影响，吞吐量提升最高 12%、TPOT 延迟降低最多 15%。

关联脉络

与 PR #47006（将 MoE all-reduce 替换为 reduce-scatter）属于同一模型（Qwen3.5）的优化系列，但侧重点不同，本 PR 聚焦内核级调优，二者可叠加提升性能。