PR #48825 完整报告

vllm-project/vllm

Perf/h20 moe config e256 n512

合并时间：2026-08-04 08:47

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/48825

执行摘要

本 PR 为 NVIDIA H20 GPU 上 E=256/N=512 的 MoE 形状新增了 Triton 内核调优配置，解决 Qwen3.5-35B-A3B 在 TP=1 时缺失设备专属配置、回落默认配置的问题。调优配置覆盖 18 个 batch size，端到端测试显示 decode 吞吐提升约 15.4%、P99 TPOT 下降约 17%。变更仅新增一个 JSON 配置，无源码改动，风险集中在配置适用范围和测试环境差异上。

功能与动机

Qwen3.5-35B-A3B 的 MoE 层形状为 E=256、N=512，在 H20 上 TP=1 部署时，vLLM 没有 E=256,N=512,device_name=NVIDIA_H20.json，会回落到默认 MoE 配置，kernel 无法利用 H20 的访存与计算特性。PR body 明确指出 "Without this file, vLLM reports that the device-specific configuration is missing and falls back to the default MoE configuration."。作者用 vLLM 官方 benchmark_moe.py tuner 在 Triton 3.6.0 下生成该配置，并给出了三组端到端 A/B 数据作为支撑。

实现拆解

1. 环境准备与数据采集：在 H20 GPU、Triton 3.6.0、vLLM 0.23.1rc1.dev714 下，用官方 benchmarks/kernels/benchmark_moe.py 的 --tune 模式分两批跑 18 个 batch size（1、2、4、8、16、24、32、48、64、96、128、256、512、1024、1536、2048、3072、4096），分别存到 part0/part1 目录。
2. 配置生成与入库：将 tuner 输出整理为一个 JSON 文件，包含 triton_version 字段与每个 batch size 的 BLOCK_SIZE_M/N/K、GROUP_SIZE_M、num_warps、num_stages；文件名按 vLLM 约定的 E=...,N=...,device_name=... 命名，确保运行时按形状与设备命中。
3. 验证环节：用 python -m json.tool 校验 JSON 合法性，pre-commit 通过；再用官方 kernel benchmark 对 batch size 64 做非调优模式复测，选中 BLOCK_SIZE_M=16、BLOCK_SIZE_N=64、BLOCK_SIZE_K=128、GROUP_SIZE_M=64、num_warps=4、num_stages=2，测量 423.80 us。
4. 端到端 A/B 测试：对 Qwen3.5-35B-A3B 启动 vllm serve，有 / 无该配置各跑三轮、每轮 50 个请求；默认配置回落默认 MoE 配置，新配置自动加载，得到吞吐与延迟对比数据。
5. 配套改动：本 PR 仅新增配置文件，不涉及源码、测试或文档改动。

vllm/model_executor/layers/fused_moe/configs/E=256,N=512,device_name=NVIDIA_H20.json

这是本 PR 唯一变更文件，为 NVIDIA H20 上 E=256/N=512 的 MoE 形状新增 Triton 内核调优配置，包含 18 档 batch size 的 launch 参数，直接影响 H20 上 Qwen3.5-35B-A3B 的性能。

该文件是 tuner 输出的 JSON 配置，以下片段截取几个有代表性的 batch size 档位，展示调优参数随 batch 规模的变化趋势：

```
{
    "triton_version": "3.6.0", // 用于匹配 Triton 版本，避免其他版本误用本配置
    "1": { // 小 batch 场景：BLOCK_SIZE_K 取 64，减少寄存器与共享内存占用
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 32,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 4,
        "num_stages": 4
    },
    "24": { // 中等 batch：BLOCK_SIZE_N 与 BLOCK_SIZE_K 放大到 128，num_warps 提升到 8
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 1,
        "num_warps": 8,
        "num_stages": 3
    },
    "512": { // 大 batch：BLOCK_SIZE_N 拉满到 256，减少 K 方向分块次数
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 256,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 16,
        "num_warps": 8,
        "num_stages": 3
    },
    "4096": { // 超大 batch：BLOCK_SIZE_M 放大到 64，并降低 num_stages 控制共享内存
        "BLOCK_SIZE_M": 64,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 16,
        "num_warps": 4,
        "num_stages": 2
    }
}
```

评论区精华

> 作者 @zzt93：“@ZJY0516 hello, do you have time to review this pr?”
>
> claude[bot]：“This pull request is from a fork — automated review is disabled. A repository maintainer can comment @claude review to run a one-time review.”

整个 PR 没有出现技术性 review 分歧，也没有未解决的疑虑；最终由 ywang96 合入。

风险与影响

- 覆盖范围有限：配置只覆盖 18 档 batch size，超出范围的 batch size 会回落到默认 MoE 配置，性能收益无法保证。
- 环境依赖：配置基于 Triton 3.6.0 与 H20 硬件生成，其他 Triton 版本下 launch 参数可能不是最优，且若加载逻辑按 triton_version 匹配，旧版本可能直接跳过该配置。
- 验证场景单一：测试环境为 cu129 nightly、TP=1，未见 TP>1 或其他模型验证。
- 缺少自动化回归测试：没有 CI 测试保护，未来若 GPU 名称或 Triton 默认参数变化，该配置可能失效且无人察觉。

影响范围：H20 上 Qwen3.5-35B-A3B 用户获得约 15% 的 decode 吞吐提升与约 17% 的 TPOT 改善；其他形状与 GPU 不受影响。整体扩散面很小，风险可控。

关联脉络

该 PR 是 vLLM fused_moe 内核调优领域的增量补充。仓库中已有 H20 的 E=256,N=256 配置，本 PR 补齐 E=256,N=512 形状；与近期 [MRV2] Enable routed-experts capture、[ROCm][Kimi-K3] aiter moe environment variable cleanup 等 PR 共同体现了 vLLM 在 MoE 内核执行路径上的持续调优与整理趋势。这类配置文件的积累方式也说明：vLLM 依赖社区 + 官方 tuner 的方式逐步覆盖更多 GPU 形状。