PR #44830 完整报告

vllm-project/vllm

[Kernel][Perf] Tune fused_moe FP8 config for Qwen3-Next-80B tp=4 on H100 (+25% at batch 96-512)

合并时间：2026-06-09 19:15

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44830

执行摘要

本 PR 为 Qwen3-Next-80B-A3B-Instruct-FP8 模型在 H100 GPU、tp=4 场景下，添加了一组调优后的 fused_moe Triton 内核配置。通过将 GROUP_SIZE_M 从默认的 32 调整为 64，在主流服务批量大小（96~512）下实现了 20%~25% 的加速，同时不影响其他批量大小的性能。

功能与动机

Qwen3-Next-80B 具有 512 专家、每个 token 激活 10 个专家，MoE 中间层大小为 512。在 tp=4 时，每个 GPU 处理的内核形状为 E=512、N=128。默认配置在该批量范围下比调优配置慢约 25%，直接影响生产服务的吞吐和延迟。

实现拆解

1. 调优过程：使用 benchmark_moe.py --tune 在 H100 80GB HBM3 上对 18 个批量大小（1~4096）进行调优，共评估 10,368 种配置，耗时 77 分钟。
2. 配置筛选：筛选出 5 个批量大小（96, 128, 256, 512, 1024）的调优配置，核心调整是将 GROUP_SIZE_M 从 32 提升至 64，其他参数如 BLOCK_SIZE_M、BLOCK_SIZE_N、num_warps、num_stages 保持默认或微调。
3. 集成方式：将配置写入 JSON 文件放入 configs/ 目录，利用 vLLM 现有的 fallback 机制，使未配置的批量大小自动使用 get_default_config 的默认配置。

vllm/model_executor/layers/fused_moe/configs/E=512,N=128,device_name=NVIDIA_H100_80GB_HBM3,dtype=fp8_w8a8,block_shape=[128,128].json

新增的调优配置文件，包含 5 个批量大小的最优参数，是本次 PR 的核心变更。

```
{
    "triton_version": "3.6.0",
    "96": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 64, // 核心变动：从默认 32 提升至 64，优化 L2 缓存 tile 遍历顺序
        "num_warps": 4,
        "num_stages": 2
    },
    "128": { /* 同 96 配置 */
```

```
        "BLOCK_SIZE_M": 16, "BLOCK_SIZE_N": 128, "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 64, "num_warps": 4, "num_stages": 2
    },
    "256": { /* 同 96 配置 */
        "BLOCK_SIZE_M": 16, "BLOCK_SIZE_N": 128, "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 64, "num_warps": 4, "num_stages": 2
    },
    "512": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 256, // 批量较大时增加 K 分块大小以提高局部性
        "GROUP_SIZE_M": 64,
        "num_warps": 4,
        "num_stages": 2
    },
    "1024": {
        "BLOCK_SIZE_M": 32, // 最大批量增加 M 分块以容纳更多行
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 64,
        "num_warps": 4,
        "num_stages": 2
    }
}
```

评论区精华

本 PR 无 review 评论，仅由 claude[bot] 自动评论，且因为来自 fork 被阻止了自动审查。最后由 ZJY0516 直接批准合并。

风险与影响

- 风险极低：仅新增一个 JSON 配置文件，不涉及任何代码逻辑改动；调优配置只影响指定 GPU 型号和内核形状，其他场景自动 fallback 到默认配置；调优结果已通过 benchmark 验证，无回归。
- 影响范围明确：对使用 Qwen3-Next-80B-FP8 模型、H100 GPU、tp=4 的用户，在批量 96~512 时获得 20%~25% 性能提升；对其他批量大小和模型无影响。

关联脉络

本 PR 与 #35808（E=256,N=512 调优）属于同一类内核配置调优，但针对不同模型形状；与 #44273、#44152、#44553 等 H20 调优 PR 类似，但目标硬件为 H100。这些 PR 共同体现了 vLLM 社区通过逐步积累专家调优配置来优化特定模型 - 硬件组合吞吐的工程实践。