PR #39904 完整报告

vllm-project/vllm

Add tuned triton fused_moe configs on H100 for gpt-oss

合并时间：2026-04-28 18:38

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39904

执行摘要

该 PR 为 gpt-oss 120b 模型在 H100 GPU 上新增了一个 Triton fused MoE 调优配置文件，平均降低内核时间约 3.35%，最大降低 8.09%。变更仅涉及一个 JSON 文件，无代码逻辑修改，安全且收益明确。

功能与动机

根据 PR body，目的是“Optimize gpt-oss 120b Fused MoE kernels Optimization configs”。默认的 fused MoE 配置并非针对该模型和 GPU 调优，通过提供手调的参数组合，可以显著减少内核执行时间，从而提升整体推理性能。

实现拆解

1. 调优参数搜索：使用 Triton 自动化工具对 16 种典型 batch size（1 到 4096）分别搜索最优的 BLOCK_SIZE_M、BLOCK_SIZE_N、BLOCK_SIZE_K、GROUP_SIZE_M、num_warps、num_stages 组合。
2. 生成配置文件：将搜索结果写入 E=128,N=2880,device_name=NVIDIA_H100_80GB_HBM3.json，文件命名遵循 数据集 约定，便于自动加载。
3. 性能验证：在 gpt-oss 120b 模型上对比默认配置，报告显示几何平均加速比约 1.035x，即内核时间减少约 3.35%，其中 batch size=8 时加速比最大（1.088x）。
4. 集成与兼容：vllm 已有配置加载机制，新增文件后无需额外代码，框架会自动选择该配置。
5. 版本同步：Review 中 reviewer 要求使用最新 Triton 版本（3.6.0）调优，作者将 triton_version 从 3.5.1 更新为 3.6.0 后获批准。

vllm/model_executor/layers/fused_moe/configs/E=128,N=2880,device_name=NVIDIA_H100_80GB_HBM3.json

新增的 Triton 调优配置文件，直接影响 fused MoE 内核性能，是 PR 的唯一变更。

```
// Triton fused MoE 调优配置，适用于 NVIDIA H100 80GB HBM3 GPU
// 模型：gpt-oss 120b，128 专家（E=128），中间维度（N=2880）
// 每个 entry 对应一个 batch size 的最优参数组合
{
    "triton_version": "3.6.0", // 调优所用 Triton 版本
    "1": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 32,
```

```
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 4,
        "num_stages": 4
    },
    "2": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 32,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 4,
        "num_stages": 3
    },
    // ... 中间配置略（完整见文件）...
    // 大 batch size 使用更大块和更多 warps 以充分利用 HBM 带宽
    "4096": {
        "BLOCK_SIZE_M": 128,
        "BLOCK_SIZE_N": 256,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 8,
        "num_stages": 4
    }
}
```

评论区精华

- ZJY0516评论道："I think vllm is using triton 3.6.0. Could you please use latest version to tune the config? " 要求使用当前项目所用的 Triton 版本（3.6.0）进行调优，以确保兼容性和最佳性能。
- zhangxin81回应："Updated already, please review again" 并更新了配置文件中的版本号。
- 最终 ZJY0516 批准："Thanks for contribution"。

风险与影响

- 风险：极低。仅为配置文件变更，无逻辑修改。如果未来 Triton 版本升级，该配置可能不再最优，但框架会回退到默认配置，不会引发错误。文件名包含 GPU 型号（H100），确保其他硬件不会误用。
- 影响：仅影响在 H100 上运行 gpt-oss 120b 的用户，带来约 3-4% 的内核性能提升，无负面兼容性影响。

关联脉络

该 PR 与历史 PR #39141（更新 TRTLLM MoE 路由方法）在 fused MoE 性能优化方向上一致，但后者侧重路由算法，本 PR 侧重内核级调优。当前 PR 展示了通过简单配置即可实现性能提升的模式，未来可为其他模型和 GPU 贡献类似的优化文件。