PR #34795 完整报告

sgl-project/sglang

[MoE] Add H20 fp8_w8a8 tuned configs for Qwen3.8 (triton 3.7.1) + fix Qwen3_5MoeForCausalLM tuning

合并时间：2026-08-17 10:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34795

执行摘要

- 一句话：新增 H20 Qwen3.8 MoE 调优配置与工具修复
- 推荐动作：值得快速阅读：若团队在 H20 或同类 GPU 上做 MoE 推理调优，这份 PR 演示了 tuned config 从生成（tuning_fused_moe_triton.py）、加载匹配到验证（micro benchmark + e2e + 数值一致性）的完整闭环。重点看配置文件的 M 分桶规律及其与默认启发式的差异、common_utils.py 架构分支组织方式，以及作者对精度噪声的论证。若没有 Qwen3.8/H20 场景，可只关注调优工具的架构识别修复一处。

功能与动机

Qwen3.8 使用 Qwen3_5MoeForCausalLM 架构，在 8×H20（tp=8）上服务时缺少 triton 3.7.1 的 tuned MoE 配置——上游经 torch 2.13.0 在 Linux 固定 triton==3.7.1，当前加载器对所有 shape 都回退到旧版本目录，无法发挥 kernel 性能。同时 benchmark/kernels/fused_moe_triton/common_utils.py 的 get_model_config 未识别该架构，落入 Mixtral 默认分支读取不存在的 num_local_experts，导致 tuning_fused_moe_triton.py 以 AttributeError 崩溃，阻塞后续配置生成。

实现拆解

1. 新增 triton 3.7.1 调优配置目录与 Qwen3.8 配置：在 python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_7_1/ 下新增 E=512,N=256,device_name=NVIDIA_H20,dtype=fp8_w8a8,block_shape=[128, 128].json，以 M=1 到 4096 共 18 个桶给出 BLOCK_SIZE_M/N/K、GROUP_SIZE_M、num_warps、num_stages。该文件由 benchmark/kernels/fused_moe_triton/tuning_fused_moe_triton.py 在 8×NVIDIA H20（tp=8，模型 Qwen3.8-2.4T-A95B-FP8）上生成。运行时 MoE runner 按文件名（E、N、设备名、dtype、block_shape）匹配并在 M 桶内查表，未命中时回退旧版本目录；配置说明中也注明 up-projection 配置在无 _down 文件时复用于 down-projection。
2. 修复调优工具架构识别：common_utils.py 的 get_model_config 将 Qwen3_5MoeForCausalLM 加入 Qwen 系列架构分支（与 Qwen2MoeForCausalLM、Qwen3MoeForCausalLM、Qwen3NextForCausalLM、Qwen3_5MoeForConditionalGeneration 等并列），使 E=num_experts/ep_size、topk=num_experts_per_tok、intermediate_size=moe_intermediate_size 被正确读取；此

前该架构落入 Mixtral 默认分支读取 num_local_experts 而崩溃。

3. 性能与精度验证：micro benchmark（CUDA-graph 回放、100 轮）显示 M=1..256 相对默认启发式加速 1.59x-1.88x；端到端 bench_one_batch_server（in=1024/out=1024）batch 1 从 52.38 提升到 60.79 tok/s（+16.1%）、batch 16 从 366.28 提升到 513.30 tok/s（+40.1%）。GSM8K 8-shot 500 题精度 0.884 vs 0.874（标准误约 1.5pp），30 个固定 prompt 中 29/30 生成 token 一致。测试配套：本 PR 未新增自动化测试文件，验证主要依赖 micro/e2e benchmark 与人工一致性检查；triton_3_7_1 目录为首次引入，后续 triton 3.7.1 的配置将沿用该目录组织。

关键文件：

- python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_7_1/E=512,N=256,device_name=NVIDIA_H20,dtype=fp8_w8a8,block_shape=[128, 128].json（模块 MoE 调优；类别 config；类型 configuration）：核心产物：首份 triton 3.7.1 MoE tuned config，18 个 M 桶的 kernel 启动参数带来 1.6x-1.9x kernel 加速与最高约 40% 端到端吞吐提升，并确立 triton_3_7_1 配置目录。
- benchmark/kernels/fused_moe_triton/common_utils.py（模块 调优工具；类别 source；类型 core-logic；符号 get_model_config）：修复 tuning 工具：get_model_config 未识别 Qwen3_5MoeForCausalLM，导致 Qwen3.8 调参脚本以 AttributeError 崩溃；加入 Qwen3 分支后调优 harness 可用于该架构，是配置生成的前置条件。

关键符号：get_model_config

关键源码片段

python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_7_1/E=512,N=256,device_name=NVIDIA_H20,dtype=fp8_w8a8,block_shape=[128, 128].json

核心产物：首份 triton 3.7.1 MoE tuned config，18 个 M 桶的 kernel 启动参数带来 1.6x-1.9x kernel 加速与最高约 40% 端到端吞吐提升，并确立 triton_3_7_1 配置目录。

```
// Qwen3.8 在 8×H20、tp=8 下 MoE 上投影（运行时复用于下投影）的 tuned 启动参数；
// 键为一次前向的 token 数 M，每个 M 桶独立选择 kernel tile 尺寸与流水参数。
{
    "1": {
        "BLOCK_SIZE_M": 16,   // 小 M 用小块，避免线程空转
        "BLOCK_SIZE_N": 64,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 16,   // 组大小影响 L2 数据复用
        "num_warps": 4,
        "num_stages": 4       // 小 M 下多级流水隐藏访存延迟
    },
    "8": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,  // N 取 128 提升连续访存效率
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 1,
```

```
        "num_warps": 8,
        "num_stages": 3
    },
    "256": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 2
    },
    "2048": {
        "BLOCK_SIZE_M": 64,  // 大 M 放大 BLOCK_SIZE_M，降低调度开销
        "BLOCK_SIZE_N": 64,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 16,
        "num_warps": 4,
        "num_stages": 3
    }
    // 其余 M 桶（2/4/16/24/32/48/64/96/128/512/1024/1536/3072/4096）结构相同，
    // 全部由 benchmark/kernels/fused_moe_triton/tuning_fused_moe_triton.py 在 8×H20
    上扫描生成
}
```

benchmark/kernels/fused_moe_triton/common_utils.py

修复 tuning 工具：get_model_config 未识别 Qwen3_5MoeForCausalLM，导致 Qwen3.8 调参脚本以 AttributeError 崩溃；加入 Qwen3 分支后调优 harness 可用于该架构，是配置生成的前置条件。

```
# get_model_config 中的架构分支：Qwen 系列统一读取三类 MoE 维度。
# 本 PR 新增 Qwen3_5MoeForCausalLM，避免 Qwen3.8 落入 Mixtral 默认分支，
# 进而读取不存在的 num_local_experts 属性导致 AttributeError。
elif architecture in [
    "Qwen2MoeForCausalLM",
    "Qwen3MoeForCausalLM",
    "Qwen3NextForCausalLM",
    "Qwen3VLMoeForConditionalGeneration",
    "Qwen3_5MoeForCausalLM", # 本 PR 新增：Qwen3.8 系列（如 Qwen3.8-2.4T-A95B-FP8）
    "Qwen3_5MoeForConditionalGeneration",
    "InternS2PreviewForConditionalGeneration",
    "MellumForCausalLM",
]:
    # Qwen3 分支字段在 Qwen3_5MoeTextConfig 中均存在
    E = config.num_experts // ep_size
    topk = config.num_experts_per_tok
    intermediate_size = config.moe_intermediate_size
```

评论区精华

b8zhong 在评论中要求补充端到端证据："is there any e2e acc? and e2e perf gain numbers?"。TobyMint 随后给出 8×H20（TP=8、PP=4、4 节点）实测：GSM8K 8-shot 500 题精度 0.884（默认） vs 0.874（tuned），1 个百分点差异在约 1.5pp 标准误内；30 个固定 prompt 中 29/30 生成 token 一致，唯一差异出现在 max_new_tokens 截断边界，符合 fp8 累加顺序噪声；端到端吞吐 batch 1 +16.1%、batch 16 +40.1%。回复后 b8zhong 给予 APPROVED，无未解决疑虑。

- 端到端精度与性能证据 (question): TobyMint 补充 8×H20（TP=8、PP=4）实测：GSM8K 8-shot 500 题 0.884 vs 0.874（标准误约 1.5pp，属统计噪声）；固定 30 个 prompt 中 29/30 token 一致；batch 1 吞吐 +16.1%、batch 16 +40.1%。

风险与影响

- 风险：数值精度：改动只调整 kernel 启动参数（block size、warps、stages）而非计算逻辑，作者用 29/30 token 一致性论证数值基本不变，但 GSM8K 上仍观测到 1pp 差异，建议更大样本确认。回归范围：配置文件按（E、N、device_name、dtype、block_shape、M）精确匹配，仅影响 8×H20 + tp=8 且 shard_intermediate_size=512/N=256 的 Qwen3.8 服务场景；若设备命名或中间尺寸变化，该文件不会被加载，副作用面很小。测试覆盖：本 PR 无自动化测试，架构分支修复缺少回归用例，未来新增 MoE 架构时仍可能静默落入错误分支。回退机制：triton_3_7_1 目录存在后 loader 按新目录优先查找，个别未覆盖 shape 回退到旧版本目录，行为与现状一致。
- 影响：直接受益方为在 8×NVIDIA H20（tp=8）上服务 Qwen3.8-2.4T-A95B-FP8 的用户：kernel 级 1.6x-1.9x、端到端 batch 16 约 40% 的吞吐提升（在调度、显存不成为瓶颈时）会直接反映到线上吞吐。triton_3_7_1 配置目录的建立为后续 triton/torch 升级场景确立了配置组织先例。tuning 工具链对 Qwen3.8 系列可用后，社区可继续为更多形状与设备生成配置。影响范围局部且明确，不涉及核心调度、KV cache、接口行为或其它硬件平台。
- 风险标记：端到端精度差异需更大样本确认，配置变更缺少自动化测试覆盖

关联脉络

- PR #34744 （前身 PR，标题未在材料中提供）：TobyMint 在评论中说明，Qwen3_5MoeForCausalLM 的 tuning 工具修复最初在 #34744 中，按建议整合进本 PR。
- PR #23682 Qwen3.5-397B MoE tuned configs: 评论中作为同类 MoE tuned config PR 被引用，属于同一调优配置系列。
- PR #13815 GLM-4.6 fp8_w8a8 MoE tuned configs: 评论中作为同类 fp8_w8a8 tuned config PR 被引用，属于同一调优配置系列。