PR #37159 完整报告

sgl-project/sglang

[Kernel] Add GB300 Triton MoE configs for GLM-4.5 FP8

合并时间：2026-08-31 21:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/37159

执行摘要

本 PR 为 GLM-4.5-FP8 在 NVIDIA GB300（SM103）+ Triton 3.7.1 上的 per-channel FP8 Triton MoE lane 新增一组 tuned 配置，覆盖 M=1 到 16384 共 7 个 token 档位。变更仅为一个 JSON 配置文件（58 行新增），利用文件名选择器精确命中 E=161、N=192 的几何，其它硬件、Triton 版本、量化模式与 MoE 形状全部保持原 fallback。实测端到端 decode 吞吐 +21.88%、混合负载 +21.48%，TPOT 降低约 19%，最大数值偏差 2.4e-7，属于低成本、高收益的纯配置性能优化。

功能与动机

GLM-4.5-FP8 使用 per-channel FP8 Triton MoE，per-rank TP8 几何为 E=161、H=5120、I=192、topk=9。SGLang 的 Triton 3.7.1 GB300 配置字典中没有这条精确 lane，因此此前一直走通用启发式调度，存在明显性能缺口。缺口是在验证 KDA-Pilot PR #197 时定位的；作者同时评估了更大的自定义 kernel，但 tuned generic kernel 更快，因此本 PR 只做小而准的配置补充。

实现拆解

1. 定位配置入口：Triton MoE runner 按 device_name、Triton 版本、dtype、量化 scale 模式、E/N 形状从 triton_utils/configs// 读取 JSON 配置，缺少匹配文件时回落到启发式。新增文件即可精准覆盖目标 lane。
2. 新增配置表：文件 .../triton_3_7_1/E=161,N=192,device_name=NVIDIA_GB300,dtype=fp8_w8a8,per_channel_quant=True.json 提供 1/16/256/512/1024/4096/16384 七档。小 batch 用 BLOCK_SIZE_M=16 与较少 stages 控延迟；M=16 将 BLOCK_SIZE_N/K 提到 128 并用 4 级流水；M>=256 用 BLOCK_SIZE_M=64、GROUP_SIZE_M=32 提升 SM 利用率与 L2 复用。
3. 精度验证：默认 vs tuned 调度对比使用真实 [E,N] per-channel scale 布局，M=1/16/256/2048 的最大绝对差为 2.4e-7、最大平均相对差为 1.67e-8；真实 TP8 服务完成全部 74 次 prefill 与 52 次 decode CUDA Graph 捕获。
4. 性能验证：fused-experts 全调用（up GEMM + SiLU-mul + down GEMM + 路由加权 + combine）在 M=1/16/256/512 分别提速 1.067x、1.440x、1.727x、1.522x；两节点 TP8 GLM-4.5-FP8 服务端到端 decode +21.88%、混合 +21.48%，三个随机种子下结果稳定。
5. 测试配套：第一版曾包含一个 config availability CPU 测试，在第二个 commit 中被移除，最终合入只有配置文件本身，未引入自动化测试。

python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_7_1/E=161,N=192,device_name=NVIDIA_GB300,dtype=fp8_w8a8,per_channel_quant=True.json

唯一变更文件，为 GLM-4.5-FP8 在 GB300 + Triton 3.7.1 的 per-channel FP8 Triton MoE lane 提供 7 档 tuned 配置；文件名选择器确保只影响该精确组合，其余硬件与形状保持原 fallback。

```
// SGLang Triton MoE tuned config: E=161, N=192, GB300 (SM103), Triton 3.7.1
// dtype=fp8_w8a8, per_channel_quant=True. Key is token count M.
// 只有文件名精确匹配此组合才命中，其他硬件 / 版本 / 量化 / 形状走原 fallback。
{
    "1": {   // decode 单 token：小 N 切块 32，3 级流水压低延迟
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 32,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 4,
        "num_stages": 3
    },
    "16": {  // 小 batch：N 与 K 放大到 128，4 级流水提升吞吐
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 1,
        "num_warps": 4,
        "num_stages": 4
    },
    "256": { // 中等 batch：GROUP_SIZE_M=32 改善 L2 复用
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 3
    },
    "512": { // 大 batch：BLOCK_SIZE_M 升到 64，充分利用 SM103 并行度
        "BLOCK_SIZE_M": 64,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 3
    },
    // 1024/4096/16384 沿用大 batch 最优组合，覆盖 chunked prefill 至 16K
    "1024": {
        "BLOCK_SIZE_M": 64,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 64,
```

```
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 3
    },
    "4096": {
        "BLOCK_SIZE_M": 64,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 3
    },
    "16384": {
        "BLOCK_SIZE_M": 64,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 3
    }
}
```

评论区精华

本 PR 没有 review 评论，唯一的 PR 评论是 /tag-and-rerun-ci 重跑 CI 指令。设计权衡记录在 PR body 中：

> 作者评估了更大的自定义 kernel，但 tuned generic kernel 更快，因此只提交更小的配置变更。

commit 历史还揭示测试被有意移除：

> Remove GLM-4.5 config availability test

说明作者最终选择不把配置可用性测试纳入合入集，依赖配置目录结构和 fallback 机制保证隔离。

风险与影响

- 回归面：文件名强绑定 NVIDIA_GB300 + Triton 3.7.1 + fp8_w8a8 + per_channel + E=161/N=192，其它组合不受影响，回归面极小。
- 测试缺失：配置没有自动化测试保护，未来配置查找逻辑重构或目录整理可能使这条 lane 静默丢失。
- Triton 版本耦合：用户使用 Triton 3.8+ 时文件名不匹配会走 fallback，行为安全但性能退回启发式。
- 数值一致性：FP8 调度顺序变化带来 2.4e-7 级输出差异，对 bitexact 场景需要留意。
- 覆盖上限：tuned 配置最到 16384 token，更大 prefill 依赖 fallback。

影响范围：仅 GLM-4.5-FP8 在 GB300 上的 Triton MoE 路径；用户侧吞吐提升约 21-22%、TPOT 下降约 19%，无需改动代码或启动参数。

关联脉络

本 PR 与 #36985（恢复 FlashInfer per-token NVFP4 覆盖）同属 GB300/Blackwell 上 FP8 量化 MoE 的精度与性能验证脉络，说明该平台 FP8 MoE 路径正在被系统性补齐。它源于 KDA-Pilot 项目的调优实践（外部 PR #197），也是 SGLang 配置文件即 kernel 调优接口这一设计的具体应用。