PR #23719 完整报告

sgl-project/sglang

add H100 configs for GLM-4.7-Flash

合并时间：2026-04-27 15:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23719

PR 分析报告：为 GLM-4.7-Flash 添加 H100 Triton MoE 配置

执行摘要

本 PR 为 GLM-4.7-Flash 在 H100 GPU 上补充缺失的 Triton fused MoE 内核配置，避免 fallback 带来的性能损失。仅新增两个 JSON 配置文件，实测 TTFT 降低 16%~24%。需关注配置文件目录版本注册问题。

功能与动机

在 H100 上运行 GLM-4.7-Flash 时，SGLang 因缺少 E=65,N=1536 形状的配置而回退到默认 Triton MoE 配置，导致显著延迟。PR 作者 BBuf 在 profiling 中捕捉到该回退，并通过 A/B 测试验证了优化收益。

实现拆解

1. 主配置：新增 E=65,N=1536,device_name=NVIDIA_H100_80GB_HBM3.json，包含 M=1 到 8192 共 18 个条目的 BLOCK_SIZE、num_warps、num_stages 参数。M 值覆盖 [1,2,4,8,16,24,32,48,64,96,128,256,512,1024,1536,2048,3072,4096,8192]。
2. Down 配置：新增同名带 _down 后缀的文件，内容与主配置一致，用于 down-projection 的 fused MoE。
3. 验证：配置基于 vLLM 的 E=64 配置初始生成，并在真实模型 zai-org/GLM-4.7-Flash 上通过两轮 A/B 测试确认性能提升。

以下为新增配置文件的完整内容，展示了不同 M 值下的 Triton 内核启动参数：


```
{
    "1": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 256,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 16,
        "num_warps": 4,
        "num_stages": 3
    },
    "2": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 64,
        "BLOCK_SIZE_K": 128,
```

```
    "GROUP_SIZE_M": 16,
    "num_warps": 4,
    "num_stages": 4
},
"4": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 64,
    "BLOCK_SIZE_K": 256,
    "GROUP_SIZE_M": 16,
    "num_warps": 4,
    "num_stages": 3
},
"8": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 32,
    "BLOCK_SIZE_K": 256,
    "GROUP_SIZE_M": 1,
    "num_warps": 4,
    "num_stages": 2
},
"16": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 32,
    "BLOCK_SIZE_K": 128,
    "GROUP_SIZE_M": 16,
    "num_warps": 4,
    "num_stages": 5
},
"24": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 128,
    "BLOCK_SIZE_K": 256,
    "GROUP_SIZE_M": 32,
    "num_warps": 4,
    "num_stages": 2
},
"32": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 256,
    "BLOCK_SIZE_K": 128,
    "GROUP_SIZE_M": 1,
    "num_warps": 4,
    "num_stages": 3
},
"48": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 256,
    "BLOCK_SIZE_K": 128,
    "GROUP_SIZE_M": 1,
```

```
    "num_warps": 4,
    "num_stages": 3
},
"64": {
    "BLOCK_SIZE_M": 16,
    "BLOCK_SIZE_N": 256,
    "BLOCK_SIZE_K": 128,
    "GROUP_SIZE_M": 1,
    "num_warps": 4,
    "num_stages": 3
},
"96": {
    "BLOCK_SIZE_M": 32,
    "BLOCK_SIZE_N": 256,
    "BLOCK_SIZE_K": 128,
    "GROUP_SIZE_M": 1,
    "num_warps": 4,
    "num_stages": 3
},
"128": {
    "BLOCK_SIZE_M": 32,
    "BLOCK_SIZE_N": 128,
    "BLOCK_SIZE_K": 128,
    "GROUP_SIZE_M": 1,
    "num_warps": 4,
    "num_stages": 3
},
"256": {
    "BLOCK_SIZE_M": 64,
    "BLOCK_SIZE_N": 64,
    "BLOCK_SIZE_K": 64,
    "GROUP_SIZE_M": 1,
    "num_warps": 4,
    "num_stages": 3
},
"512": {
    "BLOCK_SIZE_M": 128,
    "BLOCK_SIZE_N": 128,
    "BLOCK_SIZE_K": 64,
    "GROUP_SIZE_M": 1,
    "num_warps": 8,
    "num_stages": 3
},
"1024": {
    "BLOCK_SIZE_M": 128,
    "BLOCK_SIZE_N": 256,
    "BLOCK_SIZE_K": 64,
    "GROUP_SIZE_M": 1,
    "num_warps": 8,
```

```
        "num_stages": 4
    },
    "1536": {
        "BLOCK_SIZE_M": 128,
        "BLOCK_SIZE_N": 256,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 8,
        "num_stages": 4
    },
    "2048": {
        "BLOCK_SIZE_M": 128,
        "BLOCK_SIZE_N": 256,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 8,
        "num_stages": 4
    },
    "3072": {
        "BLOCK_SIZE_M": 128,
        "BLOCK_SIZE_N": 256,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 32,
        "num_warps": 8,
        "num_stages": 4
    },
    "4096": {
        "BLOCK_SIZE_M": 128,
        "BLOCK_SIZE_N": 256,
        "BLOCK_SIZE_K": 64,
        "GROUP_SIZE_M": 1,
        "num_warps": 8,
        "num_stages": 4
    },
    "8192": {
        "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

- gemini-code-assist[bot]指出配置放在 triton_3_5_1 目录但该版本未在 supported_triton_versions 中注册，导致配置不会被 fallback 发现。建议要么移动目录要么注册版本。该问题未被解决。

风险与影响

- 风险极低：纯配置文件，不涉及代码逻辑。
- 主要风险：若目录未注册，配置可能不生效，用户需确保 Triton 版本为 3.5.1 或手动更新支持列表。
- 影响范围限于 GLM-4.7-Flash 在 H100 上的 MoE 性能，TTFT 收益显著。

关联脉络

本 PR 是 GLM-4.7-Flash 模型 H100 优化链的一环，与近期 MoE 重构 PR（如 #23707 废弃 act_and_mul_triton）无直接依赖，但共享同一 MoE 配置框架。未来类似模型可直接参考此配置模板。