PR #21332 完整报告

sgl-project/sglang

[GLM-5] Apply trtllm MHA kernel for GLM-5 on Blackwell

合并时间：2026-06-04 08:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21332

执行摘要

- 一句话：GLM-5 在 Blackwell 上改用 trtllm MHA 内核，移除 FA4 workaround
- 推荐动作：建议阅读该 PR 以理解 GLM-5 在 Blackwell 的注意力内核选型历史。核心设计决策是：在外部内核（FA4）不稳定的情况下，等待上游修复后改用更稳定的 trtllm 实现，而非长期维护 workaround。这一决策思路值得在依赖外部库时参考。

功能与动机

PR 正文指明该变更是被 flashinfer #3064 和后续 flashinfer v0.6.12 升级阻塞的。作者在 Issue 评论中解释："The FA4 kernel will crash under large concurrency. Let's wait for the merge of flashinfer #3064 and use trtllm ragged kernel instead"。即原 FA4 内核在高并发场景下崩溃，需要迁移至更稳定的 trtllm MHA 内核。

实现拆解

1. 删除 Blackwell 特殊处理：在 python/sglang/srt/server_args.py 的 _handle_model_specific_adjustments 方法中，移除了 if model_arch == "GlmMoeDsaForCausalLM" and is_blackwell_supported(): envs.SGLANG_DSA_PREFILL_DENSE_ATTN_KV_LEN_THRESHOLD.set(0) 这一特殊分支。该分支原本在 Blackwell 上强制将所有预填充注意力切换为稀疏 MLA（即禁用 MHA_ONE_SHOT），以避免 FA4 崩溃。
2. 统一 DSA 阈值逻辑：将 GLM-5 纳入与其他 DSA 模型（如 DeepSeek V3/V3.2）相同的处理流程：若 SGLANG_DSA_PREFILL_DENSE_ATTN_KV_LEN_THRESHOLD 已通过环境变量手动设定，则发出警告；若未设定，则自动从模型配置中获取 index_topk 作为阈值。
3. 依赖升级：此变更依赖于外部依赖升级——flashinfer 库需包含 PR #3064 的改动，且已通过之前的版本升级（如 #26854）生效。
4. 无测试文件变更：未新增或修改测试。已有的 GLM-5 端到端测试（test_dsa_glm5_tp_mtp.py、test_dsa_glm5_dp_mtp.py）和 piecewise cuda graph 测试（test_pcg_glm5_fp4.py）被用于验证正确性，根据 PR 评论中的 /rerun-test 结果，这些测试均通过。

关键文件：

- python/sglang/srt/server_args.py（模块 配置；类别 source；类型 dependency-wiring；符号 _handle_model_specific_adjustments）：唯一变更文件；删除了 GLM-5 在

Blackwell 上的特殊阈值逻辑，统一了 DSA 模型的 attention 配置路径。

关键符号：server_args.ServerArgs._handle_model_specific_adjustments

关键源码片段

python/sglang/srt/server_args.py

唯一变更文件；删除了 GLM-5 在 Blackwell 上的特殊阈值逻辑，统一了 DSA 模型的 attention 配置路径。

```
# python/sglang/srt/server_args.py ( 片段 )
# 修改前：GLM-5 on Blackwell 强制设置阈值 0，禁用 MHA_ONE_SHOT
# 修改后：与其他 DSA 模型一致，仅在阈值未设定时自动获取 index_topk

if model_arch in [
    "DeepseekV3ForCausalLM",
    "DeepseekV32ForCausalLM",
    "KimiK25ForConditionalGeneration",
    "MistralLarge3ForCausalLM",
    "PixtralForConditionalGeneration",
    "GlmMoeDsaForCausalLM", # 新增此行，使 GLM-5 进入统一分支
]:
    if is_deepseek_dsa(hf_config): # DeepSeek 3.2 / GLM 5
        # 移除了 Blackwell 特殊 if-else，直接判断阈值是否已被手动设置
        if envs.SGLANG_DSA_PREFILL_DENSE_ATTN_KV_LEN_THRESHOLD.is_set():
            logger.warning(
                f"Dense attention kv len threshold is manually set to "
                f"{envs.SGLANG_DSA_PREFILL_DENSE_ATTN_KV_LEN_THRESHOLD.get()} for DSA.
                "
                f"Caution: This may cause performance regression if the threshold "
                f"is larger than the index topk of model."
            )
        else:
            # 当阈值未手动设定时，自动使用模型的 index_topk
            from sglang.srt.configs.model_config import get_dsa_index_topk
            envs.SGLANG_DSA_PREFILL_DENSE_ATTN_KV_LEN_THRESHOLD.set(
                get_dsa_index_topk(hf_config)
            )
            logger.warning(
                f"Set dense attention kv len threshold to model index_topk="
                f"{envs.SGLANG_DSA_PREFILL_DENSE_ATTN_KV_LEN_THRESHOLD.get()} for
                DeepSeek with DSA."
            )
```

评论区精华

唯一的实质性讨论来自作者 Fridge003 在 Issue 中的评论：“The FA4 kernel will crash under large concurrency. Let's wait for the merge of flashinfer #3064 and use trtllm ragged kernel instead。”这解释了从 FA4 迁移到 trtllm MHA 的原因，且表明该决策是等待

上游修复后的主动切换。无 review 评论或争议。

- FA4 内核崩溃与迁移到 trtllm MHA (correctness): 采用 trtllm MHA 内核替代 FA4，删除 workaround。

风险与影响

- 风险:
  1. 回归风险（低）：改动仅针对 GLM-5 在 Blackwell 上的配置路径，其他模型或平台不受影响。但需确认 trtllm MHA 内核在所有并发场景下均稳定运行，避免类似 FA4 的崩溃。
  2. 性能风险（低）：原先强制阈值 0 会使预填充始终使用稀疏 MLA，可能对长序列有性能影响。切换到自适应阈值后理论上更优，但需基准测试验证。
  3. 依赖风险（低）：正确性依赖于 flashinfer 版本是否正确升级。由于 #26854 已完成升级，此风险已解除。
- 影响:
  - 用户影响：仅影响在 Blackwell GPU 上使用 GLM-5 模型的用户。修复了高并发下推理崩溃的问题，并可能提升性能（因不再强制使用稀疏 MLA）。
  - 系统影响：无架构性改动；移除了一段特殊处理，使 GLM-5 的配置路径与其他 DSA 模型保持一致，降低了维护复杂度。
  - 团队影响：较小的清理改动，便于未来对 DSA 模型进行统一调整。
  - 风险标记：依赖外部库，仅 Blackwell 平台影响

关联脉络

- PR #26854 [Deps] Bump FI to 0.6.12 and cutedsl to 4.5.2: 此 PR 将 flashinfer 升级到 0.6.12，包含了所需的 flashinfer #3064 修改，使 trtllm MHA 内核可用。