PR #34019 完整报告

sgl-project/sglang

[SM12x] Default the fused MHC post+pre path on

合并时间：2026-08-15 03:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34019

执行摘要

- 一句话：SM12x 默认启用 fused MHC post+pre，decode 提速约一成
- 推荐动作：值得精读。虽然只有 2 行改动，但它是"小改动 + 重验证"的范例：作者用 kernel 级 profile 把混测收益按 kernel 族拆分归属，用 GSM8K 双样本量说明准确率可比性陷阱，并主动文档化 split-K 归约带来的位级不可复现问题及退出开关。关注点：(1) is_set() 守卫 + 架构默认值的组合模式，可复用到其他硬件特判；(2) 默认行为变更与位级可复现性之间的取舍文档；(3) 架构门控路径的测试盲区如何用硬件验证 + 外部 benchmark 补位。对在 SM12x 上部署 DeepSeek-V4 的团队，建议升级后先用 SGLANG_OPT_FUSE_MHC_POST_PRE=0 对照跑一遍业务评测，确认可复现性要求是否被满足。

功能与动机

在 sm120/sm121 上，server_args.py 的 SM120 块把 SGLANG_OPT_USE_TILELANG_MHC_PRE 设为 False，DeepSeek-V4 的 hc_pre 因此落到 hc_pre_torch_impl——一个形状 [M, 16384] × [16384, 24] 的 fp32 F.linear，cuBLAS 用 cutlass_80_simt_sgemm 服务，纯 CUDA core、完全不用 tensor core。在 2× DGX Spark（GB10 / sm_121, TP=2）跑 deepseek-ai/DeepSeek-V4-Flash-0731 + DSPARK 时，该回退每 decode 步 93 次调用、9.54 ms，约占步长的 14%，而投影权重只有 1.5 MB。mhc_fused_post_pre 覆盖相同计算且 _is_fused_mhc_post_pre_enabled() 已对该架构特判，但 SGLANG_OPT_FUSE_MHC_POST_PRE 全局默认 False，导致注释里预想的组合永远不会开箱即达——"the SIMT fallback is what everyone gets"。

实现拆解

1. 变更入口：python/sglang/srt/server_args.py 的 _handle_model_specific_adjustments()，位于 DeepseekV4ForCausalLM 分支的 is_sm120_supported() 块内。该块原本已集中处理 SM120 硬件限制（无 tcgen05/TMEM、SMEM 上限约 99KB），并关闭独立 TileLang pre 路径。
2. 核心改动：在 SGLANG_OPT_USE_TILELANG_MHC_PRE.set(False) 之后追加两行：if not envs.SGLANG_OPT_FUSE_MHC_POST_PRE.is_set(): envs.SGLANG_OPT_FUSE_MHC_POST_PRE.set(True)。is_set() 保证用户显式设置（无论 True/False）优先于架构默认值；全局默认值不变，非 SM12x 架构不受影响。

3. 生效链路：该 flag 决定 hc_pre 是否由 mhc_fused_post_pre 接管。_is_fused_mhc_post_pre_enabled() 本就独立于 USE_TILELANG_MHC_PRE 按 batch 大小分发，本 PR 只是让默认值与既有特判对齐，从根上消除 hc_pre_torch_impl 的 fp32 cuBLAS SIMT sgemm 回退。
4. 验证与配套：无新增单测（架构门控 + CI 无 runner），改为硬件验证——GSM8K 准确率（n=200 与 n=1000 两个样本量均无回归）+ bench_one_batch_server 吞吐 + 20 步 kernel 级 profile；benchmark 与 FP8 wo_a（#34018）混测，用 kernel 族（cuBLAS SIMT / TileLang / SM80-WMMA / deep_gemm）分离归属。
5. 兼容性说明：合并后 SM12x 用户默认获得新数值路径；追求位级可复现需显式 SGLANG_OPT_FUSE_MHC_POST_PRE=0。

关键文件：

- python/sglang/srt/server_args.py（模块 启动配置；类别 source；类型 core-logic；符号 _handle_model_specific_adjustments）：唯一变更文件，负责 DeepSeek-V4 在 SM120/SM121 上的启动参数适配；本次新增的 2 行让 fused MHC post+pre 成为该架构的开箱默认，直接决定 hc_pre 是走 TileLang fused 内核还是 fp32 SIMT 回退，是全部性能收益与数值语义变化的来源。

关键符号：_handle_model_specific_adjustments

关键源码片段

python/sglang/srt/server_args.py

唯一变更文件，负责 DeepSeek-V4 在 SM120/SM121 上的启动参数适配；本次新增的 2 行让 fused MHC post+pre 成为该架构的开箱默认，直接决定 hc_pre 是走 TileLang fused 内核还是 fp32 SIMT 回退，是全部性能收益与数值语义变化的来源。

```
# python/sglang/srt/server_args.py —— DeepSeek-V4 模型参数调整片段
def _handle_model_specific_adjustments(self):
    # ... 前面的模型分支省略 ...
    elif model_arch in ["DeepseekV4ForCausalLM"]:
        from sglang.srt.arg_groups.deepseek_v4_hook import (
            validate_deepseek_v4_cp,
            validate_deepseek_v4_mega_moe_token_budget,
        )

        validate_deepseek_v4_cp(self)
        validate_deepseek_v4_mega_moe_token_budget(self)

        # SM120 的 marlin fallback 已迁到 overrides.py 的
        # _deepseek_v4_sm120_moe 中，这里保留旧调用槽位。
        from sglang.srt.arg_groups.overrides import (
            _deepseek_v4_sm120_moe,
            run_post_process_pass,
        )

        run_post_process_pass(self, _deepseek_v4_sm120_moe)
```

```
if is_sm120_supported():
    # SM120 没有 tcgen05 与 TMEM：关闭依赖 DeepGEMM 或大于
    # 99KB SMEM 的特性（如 topk_v2）。
    envs.SGLANG_OPT_FP8_WO_A_GEMM.set(False)
    envs.SGLANG_OPT_USE_TOPK_V2.set(False)
    envs.SGLANG_OPT_USE_TILELANG_MHC_PRE.set(False)

    # 【本 PR 新增】fused post+pre 不读独立 pre 路径的标志，
    # 自己按 batch 大小分发；默认开启后 hc_pre 不再落入
    # fp32 cuBLAS SIMT sgemm。is_set() 保证用户显式设置优先。
    if not envs.SGLANG_OPT_FUSE_MHC_POST_PRE.is_set():
        envs.SGLANG_OPT_FUSE_MHC_POST_PRE.set(True)

    envs.SGLANG_OPT_DEEPGEMM_HC_PRENORM.set(False)
    envs.SGLANG_FP8_PAGED_MQA_LOGITS_TORCH.set(True)
    # 优先用 TileLang indexer，而不是 Torch 回退。
    envs.SGLANG_OPT_USE_TILELANG_INDEXER.set(True)
elif is_hip():
    envs.SGLANG_OPT_DEEPGEMM_HC_PRENORM.set(False)
    envs.SGLANG_OPT_USE_FUSED_COMPRESS.set(True)
    envs.SGLANG_OPT_FP8_WO_A_GEMM.set(False)
    envs.SGLANG_OPT_USE_JIT_INDEXER_METADATA.set(False)
    envs.SGLANG_OPT_USE_TOPK_V2.set(False)
    envs.SGLANG_OPT_USE_AITER_INDEXER.set(True)
    envs.SGLANG_OPT_USE_TILELANG_MHC_PRE.set(False)
    envs.SGLANG_OPT_USE_TILELANG_MHC_POST.set(False)
    envs.SGLANG_FP8_PAGED_MQA_LOGITS_TORCH.set(True)
    envs.SGLANG_OPT_USE_MULTI_STREAM_OVERLAP.set(False)
    envs.SGLANG_EAGER_INPUT_NO_COPY.set(True)
# ... 其余模型分支省略 ...
```

评论区精华

核心讨论集中在三处：(1) ormandj 的独立复测给出与作者一致的 kernel 级数据——“Target verifier graph span: 14.872 ms → 13.403 ms (-9.88%)”、“Component-median complete step: 16.694 ms → 15.237 ms (-8.73%)”，并观察到 FP32 SIMT MHC-pre GEMM 等从 86 次 launch 降到 1-4 次，说明收益可跨硬件复现；(2) 作者主动声明 “This change is not bit-identical”，split-K 归约改变累加顺序，最后一位差异可能翻转 MHC 的离散 cluster 指派并让 temperature 0 的 greedy 发散，需要用 SGLANG_OPT_FUSE_MHC_POST_PRE=0 显式退出；(3) 无单测问题——作者解释架构门控 + CI 无 runner 使测试无法在行为差异处执行，表态愿意按社区偏好的 arch-gated 测试模式补齐。审阅人 b8zhong 直接 APPROVED，无 review comment。

- 非位级一致：split-K 归约改变 greedy 可复现性 (correctness): 接受该数值语义变化，提供 SGLANG_OPT_FUSE_MHC_POST_PRE=0 显式退出；合并时未要求位级对齐。
- 独立复测：2× RTX PRO 6000 上的 kernel 级 profile (performance): 与作者 DGX Spark 数据互相印证，性能收益确认。

- 是否补单测（架构门控路径） (testing): 审阅者接受硬件验证方式，PR 未加单测即合并；作者表示如有 arch-gated 测试模式可补。
- DSPARK 在 sm12x 无法用 stock flashinfer 启动 (other): 外部阻塞，依赖 flashinfer#4309 或 sglang#33407 合入；本 PR 不解决该问题。

风险与影响

- 风险：风险集中在四点：(1) 数值语义变更——fused kernel 全程 fp32 但采用 split-K 归约，累加顺序与回退不同，hc_pre 结果最后一位可能翻转，MHC 的离散 cluster 指派随之翻转，temperature 0 的 greedy 输出可能与旧默认不可复现；升级用户若依赖位级一致会静默拿到不同结果。(2) 缺少自动化测试——改动只在 sm120/sm121 上生效，CI 无该 runner，后续重构可能悄悄破坏该路径而无人察觉；当前仅靠硬件上的 GSM8K 与 profile 背书。(3) 外部依赖阻塞——在 stock flashinfer（0.6.15.post1）下 DSPARK 在 sm12x 根本无法启动（topk=192 缺失），本 PR 的基准收益需要 flashinfer#4309 或 sglang#33407 合入后才对最终用户可见。(4) 收益结构依赖——profile 显示 fused TileLang 调用从 92 次涨到 180 次（0.36 ms → 1.75 ms），净收益来自消除 9.54 ms SIMT GEMM；若未来改动让小 batch 下 fused kernel 变慢，收益可能被侵蚀，需要保留 kernel 级回归观测。
- 影响：对用户：SM120/SM121 上的 DeepSeek-V4（尤其 DeepSeek-V4-Flash-0731 + DSPARK）开箱即获得 decode 步长约 7-10% 的加速（作者 DGX Spark 单步 -10.1%、净 -7.3 ms；ormandj 复测 step -8.73%），代价是输出不再位级可复现；其他架构（HIP、非 SM120 的 CUDA）不受影响。对系统：server_args.py 启动参数路径增加一个架构默认值，is_set() 守卫保证显式配置优先，全局 flag 默认值不变；由于改动在 DeepseekV4ForCausalLM 分支内，仅在该模型 + 该架构组合下触发。对团队：需要继续跟进 flashinfer#4309 / sglang#33407 才能让 DSPARK + SM12x 在 stock 依赖下跑通；同时应考虑为架构门控路径补充 CI 或硬件测试模式，避免该默认值后续被无感回退。
- 风险标记：数值语义变更（greedy 不可复现），缺少单测覆盖，依赖外部 flashinfer 修复，默认行为变更影响升级用户

关联脉络

- PR #34018 （PR body 引用的 FP8 wo_a 相关改动，标题未提供）: PR body 的 benchmark 将本改动与 FP8 wo_a（#34018）混测，并用 kernel 族 profile 把收益拆分归属；两者同属 SM12x + DeepSeek-V4 的 decode 性能优化。
- PR #33407 Fix DSPARK SM120 decode dispatch for non-instantiated topk widths: PR body 的测试环境 caveat 列出该 open PR 是 stock flashinfer 下 DSPARK 启动崩溃的已知修复；与本文共享 sm12x + DeepSeek-V4 + DSPARK 场景。
- PR #4309 feat(sm120): support DeepSeek-V4 top-k 192: flashinfer 侧补上 topk=192 的 decode/prefill 实例化；本 PR 的 DGX Spark 基准依赖其本地补丁才能启动 DSPARK，是同一 topk=192 缺口的外部修复。
- PR #34816 [Perf] Publish the WAR read-done event at DSPARK verify: 同为 DSPARK decode 步长性能优化（在 verify 阶段发布 WAR read-done 以改善重叠调度），与本 PR 同处 DeepSeek-V4 + DSPARK 解码路径优化线。

- PR #34788 [Fix] Restore layer-level DSV4 RoPE policy: 同为 DeepSeek-V4 解码路径修复（恢复层级 RoPE 策略），反映该模型在维护期的迭代；与本 PR 无直接依赖但同属模型功能线。