PR #34507 完整报告

sgl-project/sglang

[Diffusion][Z-Image] Tune native QK RMSNorm launch for SM103

合并时间：2026-08-12 16:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34507

执行摘要

- 一句话：SM103 调优 QK RMSNorm 启动，B300 提速 1.72x
- 推荐动作：值得快速阅读的低风险性能优化样本。核心价值不在改动本身，而在于其方法论：先用 Nsight Compute 定位 launch-bound，再用穷举 sweep 选定 row/warp 组合，最后以 bit-exact 与 torch.equal 作为验收闸门降低验证成本。对于计划做多架构内核调优的工程师，建议关注 get_jit_cuda_arch() 的跨平台行为，并考虑将此类按 SM 分流的 launch 配置抽象成统一配置表，避免每个内核各自 if-else。

功能与动机

PR body 明确指出原实现的问题："The bit-exact native Z-Image Q/K RMSNorm kernel used a single launch configuration on every GPU. At the production Z-Image shape, B300 is launch-bound with rows_per_prog=8, num_warps=8." 即同一 launch 配置无法适配所有 GPU，B300 上出现 launch-bound 瓶颈；本次改动目标是消除该瓶颈，同时保持 bit-exact 数值行为不变，降低验证成本与回归风险。

实现拆解

改动集中在 python/sglang/kernels/ops/diffusion/triton/zimage_native_norm.py，共 10 行新增、2 行删除，分四步完成：

1. 引入架构探测工具：文件顶部新增 from sglang.kernels.jit.utils import get_jit_cuda_arch()，为后续按 SM 架构分流提供能力，这是全仓库已有的 JIT utils 工具，无需新增依赖。
2. 按架构选择 launch 参数：在 zimage_qk_rmsnorm_native() 中，原先硬编码的 rows_per_prog = 8 改为先获取 arch 并判断 is_sm103（arch.major == 10 and arch.minor == 3）；SM103 使用 rows_per_prog = 16、num_warps = 4，其他架构保持 rows_per_prog = 8、num_warps = 8。选择依据是 B300 上对 row/warp 组合的穷举扫描。
3. 保持内核数学不变：_qk_rmsnorm_native_kernel 内部实现、bit-exact 舍入顺序完全未动，改动只影响 launch 网格形状（grid 由 rows_per_prog 决定）和 num_warps，因此位级一致性风险被限制在启动参数层面。
4. 测试与性能验证：B300 上运行 test/zimage_qknorm_fusion.py 两个用例全部通过；strided fused-QKV 路径与 eager ZImageRMSNorm 参考实现 torch.equal 逐位相等；性

能数据由 B300 穷举 sweep 与 Nsight Compute 双重复核。

无配置文件、部署脚本或公共 API 改动，属于纯内核层的定点调优。

关键文件：

- python/sglang/kernels/ops/diffusion/triton/zimage_native_norm.py（模块 内核算子；类别 source；类型 performance-tuning；符号 zimage_qk_rmsnorm_native）：本 PR 唯一改动文件。zimage_qk_rmsnorm_native() 从硬编码 rows_per_prog=8、num_warps=8 改为按 SM103 分流为 16/4，内核数学与 bit-exact 舍入顺序不变，是全部性能收益与风险的来源。

关键符号：zimage_qk_rmsnorm_native

关键源码片段

python/sglang/kernels/ops/diffusion/triton/zimage_native_norm.py

本 PR 唯一改动文件。zimage_qk_rmsnorm_native() 从硬编码 rows_per_prog=8、num_warps=8 改为按 SM103 分流为 16/4，内核数学与 bit-exact 舍入顺序不变，是全部性能收益与风险的来源。

```
def zimage_qk_rmsnorm_native(x, ...):
    # 前面省略：空输入 / 不支持类型的守卫判断
    nheads = x.shape[2]
    n_rows = x.shape[0] * x.shape[1] * nheads

    # 按目标架构选择启动配置。生产 Z-Image shape（1, 4096, 24, 128）
    # 在 B300（SM103）上是 launch-bound，改用 16 rows/prog + 4 warps
    # 可把内核延迟从 24.58 us 降到 14.33 us（约 1.72x）。
    arch = get_jit_cuda_arch()
    is_sm103 = arch.major == 10 and arch.minor == 3
    rows_per_prog = 16 if is_sm103 else 8
    num_warps = 4 if is_sm103 else 8

    y = torch.empty(x.shape, dtype=x.dtype, device=x.device)
    grid = (triton.cdiv(n_rows, rows_per_prog),)
    with torch.get_device_module().device(x.device):
        # 内核数学与 bit-exact 舍入顺序保持原样，只调整 launch 形状
        _qk_rmsnorm_native_kernel[grid](
            x,
            y,
            n_rows,
            nheads,
            head_dim,
            eps,
            rows_per_prog=rows_per_prog,
            num_warps=num_warps,
        )
    return y
```

评论区精华

该 PR 没有任何 review 评论，3 条 issue 评论均为作者 BBuf 触发的 CI 自动化指令（/tag-and-rerun-ci、/rerun-failed-ci 及一条 CI 运行链接），不存在设计争议或未解决疑虑。设计权衡在 PR body 中由作者自行说明：B300 上穷举 row/warp 组合后选定 16/4，且明确要求在内核数学与 bit-exact 舍入顺序不变的前提下验证，以最小化回归面。merged_by 为作者本人，属于小而明确的自合改动。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 架构特定代码路径：仅 SM103（arch.major==10 && arch.minor==3）分支变化，非 SM103 完全走原 8/8 路径，回归面非常小；但 SM103 分支缺少在其他 shape 下的性能验证，16/4 是生产 shape（1, 4096, 24, 128）的穷举结果，非常规 shape 下未必最优。
  2. 依赖 get_jit_cuda_arch() 行为：新增的架构探测若在非 NVIDIA 平台（AMD、XPU、MLX 等）返回异常结构或抛异常，会导致内核启动失败；虽然默认 else 分支保持了原配置，但该 util 的跨平台行为未在本 PR 中覆盖验证。
  3. 测试覆盖范围：测试仅在 B300 上执行并通过，非 SM103 架构没有针对该改动的回归测试，不过该路径逻辑上未变化，风险可接受。
  4. 安全：不涉及网络、输入校验或权限逻辑，无安全风险。 - 影响：对用户：B300（SM103）上运行 Z-Image 生产 shape 的用户可直接获得 QK RMSNorm 内核约 1.72x 的延迟下降（24.58 us -> 14.33 us），denoise 整体收益取决于该内核在整条流水线中的占比；strided fused-QKV 路径数值与 eager 参考实现逐位相等，无精度回退。对系统：无公共 API、无配置项、无依赖变更，非 SM103 用户完全不受影响。对团队：确立了 "get_jit_cuda_arch() 按 SM 分流 + 穷举 sweep 选 launch 参数 + 保持内核数学不变 " 的调优范式，可复用于后续 diffusion Triton 内核的架构特调。 - 风险标记：架构特定代码路径，依赖 JIT arch 探测，测试仅覆盖 B300

关联脉络

- PR #34349 [Diffusion] Tune QK head LayerNorm for SM120: 同属 "diffusion QK norm 内核按 SM 架构特调 " 系列，SM120 与 SM103（B300）两处调优可相互对照，形成同类问题的处理范式。
- PR #34347 [Diffusion][MiniMax H3] Fix SM120 QKNorm+RoPE rounding: 同族 QKNorm 内核在 SM120 上的数值修复，说明 QK norm 内核正处于多架构适配与精度优化的活跃迭代期。
- PR #34508 [Diffusion][LTX-2] Allocate AdaLN outputs from one contiguous slab: 同为 diffusion Triton 内核性能优化，反映 sglang 对 diffusion 内核 launch 与访存模式的持续调优方向。