PR #34503 完整报告

sgl-project/sglang

[Diffusion] Tune QK head LayerNorm for SM103

合并时间：2026-08-12 16:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34503

执行摘要

- 一句话：SM103 调优 QK head LayerNorm 启动，B300 提速 1.46x
- 推荐动作：值得快速阅读。适合关注 diffusion 内核优化或 CUDA 架构差异的工程师，尤其可以学习“按架构分派 launch 配置”这一低成本高性能调优模式；对不涉及该内核的读者价值有限。

功能与动机

PR #34349 已对 SM120 调优并保留 H100/H200 配置，但 B300（即 SM103）在 SM90 的 ROWS=64, num_warps=2 配置下同样受寄存器压力影响，且其最优 launch 参数与 SM120 不同，因此需要独立的架构分支。PR body 用 B300 全 row/warp 扫描和 Nsight Compute 数据佐证 16 行 / 1 warp 更优。

实现拆解

1. 重构架构判断：删除 _is_sm120_or_newer() 布尔函数，新增 _qk_head_launch_config()，返回当前架构对应的 (rows, num_warps) 元组；分支顺序为先判 SM103（16 / 1），再判 SM120+（8 / 4），其余架构回退到原 SM90 配置（64 / 2）。
2. 更新调用点：在 fused_qk_head_layernorm() 中改为解包 rows, num_warps = _qk_head_launch_config()，并将两个变量分别传入 ROWS 与 num_warps，一处代码同时驱动所有架构。
3. 验证配套：B300 上运行 test_glm_image_ln_modulate.py，6 项全部通过；生产形状 (1, 4360, 32, 128) 的 Q/K 输出与 *eager* LayerNorm 保持 bit-exact；Nsight Compute 显示寄存器 / thread 从 168 降到 80，*achieved occupancy* 从 17.5% 提升到 33.4%。未新增测试文件，改动仅限既有内核文件。

关键文件：

- python/sglang/kernels/ops/diffusion/triton/layernorm_modulate.py（模块 内核调优；类别 source；类型 performance-tuning；符号 _qk_head_launch_config, _is_sm120_or_newer, fused_qk_head_layernorm）：唯一变更文件：将架构布尔判断重构为返回 (rows, num_warps) 的分派函数，为 SM103 增加 16 / 1 配置，并在 fused_qk_head_layernorm 调用点注入参数。

关键符号：_qk_head_launch_config, fused_qk_head_layernorm

关键源码片段

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

唯一变更文件：将架构布尔判断重构为返回 (rows, num_warps) 的分派函数，为 SM103 增加 16 / 1 配置，并在 fused_qk_head_layernorm 调用点注入参数。

```
# 返回当前 JIT 目标架构下 QK head LayerNorm 的 (rows, num_warps)。
# 每个架构族均基于生产形状完整扫描选择最优配置。
def _qk_head_launch_config() -> tuple[int, int]:
    arch = get_jit_cuda_arch()

    # SM103（即 B300）：SM90 的 64 行 / 2 warp 配置受寄存器压力拖累，
    # 生产形状扫描显示 16 行 / 1 warp 最优（约 1.46 倍加速）。
    if arch.major == 10 and arch.minor == 3:
        return 16, 1

    # SM120+（即 RTX 5090）：沿用 PR #34349 的 8 行 / 4 warp 配置。
    if arch.major * 10 + arch.minor >= 120:
        return 8, 4

    # 其他架构（H100 / H200 等）：保持原 64 行 / 2 warp 字节级不变。
    return 64, 2
```

评论区精华

本 PR 没有实质性的 review 讨论，唯一评论是作者 BBuf 触发的 /tag-and-rerun-ci 指令，属于 CI 重跑操作，不涉及设计争议。设计决策与数据全部来自 PR body 的自述（B300 全 row/warp 扫描 + Nsight Compute 佐证）。

- 暂无高价值评论线程

风险与影响

- 风险：精度风险低：B300 生产形状与 *eager* LayerNorm 保持 bit-exact，SM120+ 与 SM90 配置未动。兼容性风险较低：_is_sm120_or_newer 被移除，若有其他调用方依赖该符号会编译失败，从改动面看该符号仅在 layernorm_modulate.py 内部使用。架构风险：SM103 分支使用硬编码 arch.major == 10 and arch.minor == 3，未来新架构若沿用 SM103 设备号可能继承该配置。性能回归风险：改动仅影响 launch 参数，若其他模型输入形状与生产形状差异大，16 / 1 可能不是最优，但不会导致正确性问题。
- 影响：对用户：B300/GB300 上的 diffusion（尤其 GLM 系列）推理中，QK layernorm 单点延迟下降约 1.46 倍，寄存器压力和 *occupancy* 显著改善。对系统：改动限定在单内核 launch 参数，不改变数值路径，无精度风险。对团队：形成按架构族分派 launch 配置的复用模式，为后续 Blackwell Ultra 调优提供模板。影响范围小，风险低。
- 风险标记：SM103 硬编码分支，仅 B300 验证，端到端收益未量化

关联脉络

- PR #34349 [Diffusion] Tune QK head LayerNorm for SM120: 本 PR 的直接前作，SM120+ 分支沿用其 8 / 4 配置；PR body 明确引用其结论。
- PR #34507 [Diffusion][Z-Image] Tune native QK RMSNorm launch for SM103: 同期对 SM103 的 diffusion 内核启动调优，与本次改动属于同一系列架构专门化工作。
- PR #34505 [Diffusion][MiniMax H3] Extend exact QKNorm+RoPE rounding to SM103: 同期将 SM103 精确舍入修复扩展到扩散内核，共同构成 Blackwell Ultra 内核适配脉络。