PR #36543 完整报告

sgl-project/sglang

[kernel] Tune LingBot-Video MoE TMA configs for H100

合并时间：2026-08-27 17:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36543

执行摘要

- 一句话：H100 MoE 新增 TMA 配置，LingBot-Video 推理提速约 2%
- 推荐动作：值得精读。虽然只是两个 JSON 文件 + 一个测试函数，但它展示了：以真实负载 shape（5034/5049 token）为基准做内核配置调优、用 ABBA 交叉验证和 SHA256 断言保证无数值回归、以及将调优结果固化为可被测试保护的配置资产。对于要在新 Triton 版本或新硬件上做 MoE 性能调优的团队，这是很好的参考样例。

功能与动机

PR body 指出，真实 SGLang 负载 robbyant/lingbot-video-moe-30b-a3b 在 384x640x17、12 步、quality=high 下使用 4096-token 配置处理 5034/5049-token 的 denoise shape；该 PR 希望通过为 Triton 3.7.1 增加 H100 BF16 配置并在 4096 桶上启用 TMA，压缩 MoE up/down 投影内核的耗时，从而端到端降低视频生成时延。

实现拆解

1. 新增 up 投影配置：python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_7_1/E=128,N=768,device_name=NVIDIA_H100_80GB_HBM3.json，内含 1~4096 共 18 个 token 桶的 BLOCK_SIZE_M/N/K、GROUP_SIZE_M、num_warps、num_stages 组合。
2. 新增 down 投影配置：同名后缀 _down.json，内容与 up 版一致，供 down projection 单独读取；若不存在该文件，已有运行时逻辑会回退到 up 配置。
3. 关键变化集中在 4096 桶："USE_TMA": true；其余桶的取值与 Triton 3.2 fallback 完全一致（PR 验证了删除 USE_TMA 后两份文件与旧配置逐字节一致），将行为变化面收缩到单个形状。
4. 测试配套：在 test/registered/unit/layers/moe/test_fused_moe_triton_config.py 中新增 test_h100_lingbot_video_configs_enable_tma_only_for_the_tuned_shape，断言 4096 桶带 USE_TMA 且其余桶不带，防止后续调参把 TMA 扩散到未验证形状。
5. 验证方法：ABBA 顺序各跑 4 次 baseline/candidate，记录 denoise 与 E2E 时延；并用 8 个输出 MP4 的 SHA256 一致性证明无数值回归；微基准显示 up 投影从 625.611 us 降到 530.563 us，down 投影从 316.475 us 降到 283.427 us。

关键文件：

- python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_7_1/E=128,N=768,device_name=NVIDIA_H100_80GB_HBM3.json（模块 MoE 配置；类别 config；类型 configuration）：新增 Triton 3.7.1 H100 BF16 MoE 主（up）投影调优配置，覆盖 1~4096 token 桶，并在 4096 桶启用 USE_TMA，这是本次性能提升的核心数据输入。
- python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_7_1/E=128,N=768,device_name=NVIDIA_H100_80GB_HBM3_down.json（模块 MoE 配置；类别 config；类型 configuration）：新增 down 投影的独立配置，与 up 版保持一致，确保 down projection 在 4096 桶同样启用 TMA；若缺失该文件时运行时回退到 up 配置。
- test/registered/unit/layers/moe/test_fused_moe_triton_config.py（模块 配置测试；类别 test；类型 test-coverage；符号 test_h100_lingbot_video_configs_enable_tma_only_for_the_tuned_shape）：新增测试函数验证 TMA 只出现在 4096 桶，保护配置边界，防止未来误开 TMA 到未验证 shape。

关键符号：test_h100_lingbot_video_configs_enable_tma_only_for_the_tuned_shape

评论区精华

本 PR 没有公开的 review 评论（review_comments_count 为 0）。实际上，作者在 PR body 中用“ABBA 对照实验 + SHA256 一致性”替代了文字讨论：四个 baseline 与四个 candidate 交替运行，denoise 从 3045.514 ms 降到 2980.128 ms、E2E 从 3720.085 ms 降到 3657.357 ms，且 8 个 MP4 哈希完全一致。这说明对于纯配置型变更，数字自证比口头讨论更有说服力。

需要留意的是 CI 三个运行（PR Test、Extra、AMD ROCm 7.2）都标记为失败，但 PR 仍被合并，未在仓库内看到失败说明；不排除是环境或无关 flaky 问题，但合并前未留下显式说明。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，但需注意： 1) 配置仅覆盖 Triton 3.7.1 + H100 80GB + BF16 + E=128/N=768 的窄场景，其他硬件或 Triton 版本会自动回退，行为不变； 2) 只在 4096 token bin 启用 TMA，若用户 shape 不落在该桶（如非 5034/5049 token），性能可能回落，但不会出错； 3) 新增 JSON 文件仅被 fused_moe_triton_config 读取，无源码路径改动，回归面小； 4) CI 三个 run 失败未见说明，需关注是否有未覆盖的验证缺口。数值一致性已通过输出 MP4 SHA256 全同验证。
  - 影响：影响范围：仅使用 Triton 3.7.1 + H100 80GB + LingBot-Video-MoE-30B（E=128, N=768）的用户；MoE up/down 投影微基准分别提速约 15.2% 和 10.4%，端到端 denoise 提升 2.15%、E2E 提升 1.69%；模型输出不变（8 个 MP4 的 SHA256 完全一致）。对团队而言，该 PR 为 Triton 3.7.1 配置族补上了 H100 场景，并固化了“只对实测桶启用 TMA”的保守策略。
  - 风险标记：配置数据变更，仅 H100 单点验证，非 4096 shape 未验证，CI 失败未说明

关联脉络

- PR #35634 [Feature] Add DeepEPv2 (ElasticBuffer) MoE A2A backend: 同属 MoE 内核性能优化线，TMA 与 A2A 后端共同压缩 MoE 耗时。
- PR #33871 [Performance] Reduce idle DP work in breakable prefill CUDA graphs: 同为 MoE 执行效率优化，说明仓库持续投入 MoE 性能调优。