

PR #44692 完整报告

vllm-project/vllm

[Bugfix][Kernel] Fix mHC fused-RMSNorm big-fuse miscompile for hidden_size != 4096

合并时间: 2026-06-06 10:44

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44692>

执行摘要

- 一句话: 修复 mHC fused-RMSNorm 在 hidden_size != 4096 时编译错误
- 推荐动作: 建议合并。这是一行关键修复, 定位清晰、验证充分。

功能与动机

修复 fused-RMSNorm 大融合核函数在 non-4096 hidden_size 下的静默错误。详细错误表现: hidden_size 2048/3072 返回 NaN; 5120/6144/7168/8192 相对误差 0.17-0.55; 仅 4096 正确。该问题源于 TileLang 流水线编译器在结合循环携带的 sumsq 归约和持久化共享缓冲区时的误编译。

实现拆解

1. 定位问题: 在 vllm/model_executor/kernels/mhc/tilelang_kernels.py 的 mhc_pre_big_fuse_with_norm_tilelang 函数中, Pass 1 的 T.Pipelined 循环使用了 num_stages=3, 而对应的无 norm 版本 mhc_pre_big_fuse_tilelang 使用 num_stages=2 正确。
2. 分析根因: num_stages=3 本身并非错误, 但结合循环携带的 sumsq 归约和持久化的 output_shared 缓冲区时, TileLang 流水线编译器在序言 / 尾声生成中存在缺陷。
3. 修复方案: 将 Pass 1 的 num_stages=3 降为 num_stages=2, 与无 norm 版本保持一致。
4. 验证结果: 隐藏层大小 2048-8192 的 BFloat16 相对误差降至约 1.6e-3, 所有 token 数 (1-16384) 下均确定性正确。

关键文件:

- vllm/model_executor/kernels/mhc/tilelang_kernels.py (模块 kernel; 类别 source; 类型 core-logic; 符号 mhc_pre_big_fuse_with_norm_tilelang): 唯一变更文件, 修改了 mhc_pre_big_fuse_with_norm_tilelang 中 Pass 1 的流水线阶段数, 修复了 fused-RMSNorm 的误编译。

关键符号: mhc_pre_big_fuse_with_norm_tilelang

关键源码片段

[vllm/model_executor/kernels/mhc/tilelang_kernels.py](#)

唯一变更文件，修改了 `mhc_pre_big_fuse_with_norm_tilelang` 中 Pass 1 的流水线阶段数，修复了 fused-RMSNorm 的误编译。

```
# vllm/model_executor/kernels/mhc/tilelang_kernels.py (line 312)
# 修复前：num_stages=3 与循环携带的 sumsq 归约 + 持久化 shared 缓冲区组合时，
# TileLang 流水线编译器生成错误代码。
# 修复后：num_stages=2，与无 norm 版本的流水线深度一致，消除误编译。
for i0_h in T.Pipelined(hidden_size // hidden_block, num_stages=2):
    xs = T.alloc_shared((hc_mult, hidden_block), T.bfloat16)
    xl = T.alloc_fragment((hc_mult, hidden_block), T.float32)
    T.copy(residual[i, 0, i0_h * hidden_block], xs)
    T.copy(xs, xl)

    ol = T.alloc_fragment(hidden_block, T.float32)
    T.clear(ol)

    for i_hc in T.serial(hc_mult):
        pre = pre_mix_shared[i_hc]
        for i1_h in T.Parallel(hidden_block):
            ol[i1_h] += pre * xl[i_hc, i1_h]

    for i1_h in T.Parallel(hidden_block):
        sumsq_per_pos[i1_h] += ol[i1_h] * ol[i1_h]
        output_shared[i0_h * hidden_block + i1_h] = T.bfloat16(ol[i1_h])
```

评论区精华

本次 PR 未产生 review 讨论。Jeejeelee 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅涉及一行参数的修改，且与同一文件中已验证正确的无 norm 内核的流水线阶段数保持一致。非融合路径 (mhc_pre, mhc_post) 完全不受影响。从 patch 和验证结果看，回归可能性极低。
- 影响：直接影响：修复 DeepSeek-V4 等模型在非标准 hidden_size 下的 kernel 正确性。影响范围狭小，仅涉及 mhc_pre_big_fuse_with_norm 路径，用户无需额外配置。对于已经使用 hidden_size=4096 的用户，性能无退化。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #42735 mHC bf16 shared staging: PR body 提及此 PR 已使用 num_stages=2，与当前修复一致。