PR #47164 完整报告

vllm-project/vllm

fix: skip cooperative top-K on SM120

合并时间：2026-07-01 12:32

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47164

执行摘要

- 一句话：跳过 SM120 上的 cooperative top-K
- 推荐动作：可以安全合并。变更简单、目标明确，已由作者手工验证。建议在 SM120 上增加自动化测试覆盖，但非阻塞。

功能与动机

SM120 GPU 上 clustered kernel 启动会失败，返回 cudaErrorInvalidValue，导致合作 top-K 无法正常工作。需要跳过 cooperative top-K 以避免错误，并利用现有的 persistent top-K 实现作为 fallback。

实现拆解

在 vllm/model_executor/layers/sparse_attn_indexer.py 文件中，修改 use_cooperative_topk 的判断条件：

1. 在原有条件（平台为 CUDA、topk_tokens 为 512/1024/2048、num_rows <= 32、stride 对齐 4、计算能力 >= 90）的基础上，增加 not current_platform.is_device_capability_family(120)。
2. 当 SM120 时，use_cooperative_topk 为 False，代码会判断 use_persistent_topk（仅依赖 CUDA 平台和 topk_tokens 值）为 True，从而执行 torch.ops._C.persistent_topk 作为 fallback。 整修变更仅一行新增代码，不涉及测试或其他文件修改。

关键文件：

- vllm/model_executor/layers/sparse_attn_indexer.py（模块 核函数；类别 source；类型 data-contract）：修改了 cooperative top-K 的判断条件，增加了 SM120 排除逻辑。

关键符号：sparse_attn_indexer

关键源码片段

vllm/model_executor/layers/sparse_attn_indexer.py

修改了 cooperative top-K 的判断条件，增加了 SM120 排除逻辑。

```
# file: vllm/model_executor/layers/sparse_attn_indexer.py
# 决定是否使用 cooperative top-K 的条件判断
# 新增一行排除 SM120，因为该架构上 clustered kernel 会失败
```

```
use_cooperative_topk = (
    current_platform.is_cuda()
    and topk_tokens in (512, 1024, 2048)
    and num_rows <= 32
    and logits.stride(0) % 4 == 0 # TMA 16-byte alignment
    and current_platform.has_device_capability(90)
    and not current_platform.is_device_capability_family(120) # 新增：跳过 SM120
)
```

评论区精华

仅有自动化审查评论（由 claude[bot] 提及来自 fork 的 PR 自动审查被禁用）和 zyongye 的批准，没有实质性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅增加一个对 SM120 的排除检查，不影响其他 GPU 的行为。SM120 上原本 cooperative top-K 会失败，现在使用 persistent top-K 正常 fallback，不会引入新的回归。但若 future SM 系列也遇到类似问题，可能需要更通用的架构范围判断而非硬编码。
- 影响：仅影响 SM120 计算能力的 NVIDIA GPU。在这些 GPU 上，合作 top-K 被禁用，但 persistent top-K 作为 fallback 继续工作。对其他所有 GPU 无影响。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR