

PR #30356 完整报告

sgl-project/sglang

Disable FA3 sparse mask kernels by default

合并时间: 2026-07-07 16:46

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30356>

执行摘要

- 一句话: 默认禁用 FA3 sparse mask 内核以减小 wheel 体积
- 推荐动作: 变更安全且收益明确, 值得快速合并。适合作为了解 sgl-kernel 编译系统的示例。

功能与动机

如果不禁用该 flag, wheel 体积很大, 超过 500MB。禁用后可减小至 361MB, 减少分发和存储成本。

实现拆解

1. 新增编译选项 `SGL_KERNEL_ENABLE_FA3_SPARSE_MASK`: 在 `sgl-kernel/CMakeLists.txt` 中新增该选项, 默认值为 OFF。
2. 根据选项控制宏定义: 当选项为 OFF 时, 在 FA3 编译单元中添加宏定义 `FLASHATTENTION_DISABLE_SPARSE_MASK`, 跳过 sparse mask 内核的编译。

关键文件:

- `sgl-kernel/CMakeLists.txt` (模块 内核编译; 类别 infra; 类型 configuration): 唯一变更文件, 新增编译选项和条件宏定义, 控制 FA3 sparse mask 内核是否编译。

关键符号: 未识别

关键源码片段

`sgl-kernel/CMakeLists.txt`

唯一变更文件, 新增编译选项和条件宏定义, 控制 FA3 sparse mask 内核是否编译。

```
# sgl-kernel/CMakeLists.txt
# 新增编译选项, 默认关闭 sparse mask 内核以减小 wheel 体积
option(SGL_KERNEL_ENABLE_FA3_SPARSE_MASK "Enable FA3 sparse mask kernels" OFF)

# ... 在 FA3 编译配置中, 如果选项关闭, 则添加宏定义跳过 sparse mask 内核
if(SGL_KERNEL_ENABLE_FA3)
# 其他配置 ...
if(NOT SGL_KERNEL_ENABLE_FA3_SPARSE_MASK)
    list(APPEND FLASH_OPS_COMPILE_DEFS FLASHATTENTION_DISABLE_SPARSE_MASK)
endif()
```

```
target_compile_definitions(flash_ops PRIVATE ${FLASH_OPS_COMPILE_DEFS})
endif()
```

评论区精华

变更较简单，无实质性 review 讨论。PR CI 中有两条 bot 评论，一条提示配额限制，另一条是 `/tag-and-rerun-ci` 命令，无技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 若某些实验场景需要 FA3 sparse mask 内核，用户需显式开启该选项。
 2. 变更仅影响编译时宏定义，不涉及运行时逻辑，回归风险极低。
- 影响：
 1. 用户：默认情况下 sgl-kernel wheel 体积显著减小（从 500+MB 降至 361MB），下载和安装更快。需要 sparse mask 功能的用户可手动开启选项重新编译。
 2. 系统：无运行时影响。
 3. 团队：降低了 CI 产物分发成本和存储空间。 - 风险标记：低频路径：需手动开启

关联脉络

- 暂无明显关联 PR