

PR #29067 完整报告

sgl-project/sglang

build(sgl-kernel): bump FlashMLA pin + fix cccl include for CUDA 13

合并时间: 2026-06-26 11:00

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29067>

执行摘要

- 一句话: FlashMLA pin 升级 + CUDA 13 cccl 修复
- 推荐动作: 建议合并。变更范围小, 验证充分, 且解决了 CUDA 13 兼容性问题。值得关注的是 FlashMLA 的 pin 管理策略, 未来类似依赖升级可参考此 PR 的验证流程。

功能与动机

1. 支持 head_dim_k=512 的纯 latent K head (无 RoPE tail) 模型, 使其能使用 sm90 dense decode 内核。
2. 兼容 CUDA 13 编译器, 解决 flashmla_ops 在 CUDA 13 下编译失败的问题。

实现拆解

1. FlashMLA pin 升级: 在 sgl-kernel/cmake/flashmla.cmake 中将 FetchContent_Declare 的 URL 从 df022eba 更新为 05e26647, 对应哈希同步更新。新 pin 包含了 sgl-project/FlashMLA#9 的变更, 使 HEAD_DIM_K 成为编译期模板参数, 支持 512 和 576 两种配置。
2. CUDA 13 cccl 包含路径修复: 通过 CMAKE_CUDA_COMPILER_VERSION >= 13.0 条件判断, 使用 find_path 查找 cccl 包含目录 (内含 cuda/std/utility), 将其添加到 flashmla_ops 的 include directories 中。这样 host 端 C++ 翻译单元 (由 g++ 编译, 未定义 CUDACC_VER_MAJOR) 也能正确找到 cccl 路径。
3. 配合 pin 升级, 已将 flashmla_ops.so 中新增了 HEAD_DIM_K=512 的显式实例化。

关键文件:

- sgl-kernel/cmake/flashmla.cmake (模块 构建配置; 类别 infra; 类型 core-logic): 唯一变更文件, 包含 FlashMLA pin 升级和 CUDA 13 cccl 修复, 是 sgl-kernel 构建系统核心配置。

关键符号: 未识别

关键源码片段

[sgl-kernel/cmake/flashmla.cmake](#)

唯一变更文件, 包含 FlashMLA pin 升级和 CUDA 13 cccl 修复, 是 sgl-kernel 构建系统核心配置。

```

# flash_mla
# 更新 FlashMLA pin: 支持 sm90 dense decode 的 HEAD_DIM_K=512 (来自 sgl-project/FlashMLA#9) 。
# 新 pin 位于 rebase 分支, 合并了相关变更。
FetchContent_Declare(
    repo-flashmla
    URL https://${GITHUB_ARTIFACTORY}/sgl-project/FlashMLA/archive/05e26647fe840b8baedae486c2d86d5ce4efeb7c.tar.gz
    URL_HASH SHA256=ce369489bbfc42cdfbba9aa949de0270e64469d530748dea9f4f60b3c69dea9b
)
FetchContent_Populate(repo-flashmla)

# ... 中间省略其他配置 ...

# CUDA 13 将 cuda/std/* 移到了 cccl/cuda/std/* 下。
# 项目使用的 cutlass 在 __CUDACC_VER_MAJOR__ >= 13 时会将 <cuda/std/...>
# 重定向到 <cccl/cuda/std/...>。但 host 端 C++ 文件由 g++ 编译时该宏未定义,
# 导致找不到头文件。因此需要显式添加 cccl 的包含路径。
if(CMAKE_CUDA_COMPILER_VERSION VERSION_GREATER_EQUAL "13.0")
    find_path(FLASHMLA_CCCL_INCLUDE NAMES cuda/std/utility
        HINTS ${CMAKE_CUDA_TOOLKIT_INCLUDE_DIRECTORIES}
            ${CMAKE_CUDA_TOOLKIT_INCLUDE_DIRECTORIES}/cccl)
endif()

target_include_directories(flashmla_ops PRIVATE
    ${repo-flashmla_SOURCE_DIR}/csrc
    ${repo-flashmla_SOURCE_DIR}/csrc/kerutils/include
    ${repo-flashmla_SOURCE_DIR}/csrc/sm90
    ${repo-flashmla_SOURCE_DIR}/csrc/extension/sm90/dense_fp8/
    ${repo-flashmla_SOURCE_DIR}/csrc/cutlass/include
    ${repo-flashmla_SOURCE_DIR}/csrc/cutlass/tools/util/include
    ${FLASHMLA_CCCL_INCLUDE} # 新增: 将 cccl 路径加入搜索列表, 仅在 CUDA 13 时有值
)

```

评论区精华

本 PR 的 review 过程中没有产生实质性的技术讨论。唯一的审批来自 Fridge003 (仓库维护者), 直接 Approved, 无评论。PR 作者在 body 中详细说明了变更动机、验证过程和测试结果, 包括 bit-exact 数值验证和三个 e2e 测试全部通过。

- 暂无高价值评论线程

风险与影响

- 风险:

1. 回归风险: 低。pin 升级引入的是线性代数内核参数化变更和新增实例化, 不会影响现有 576 路径。作者已验证 bit-exact (HEAD_DIM_K=512 输出与 576 零填充 RoPE tail

版本完全一致)，且三个 e2e 测试全部通过。

2. 兼容性风险：低。仅修改 CMake 构建逻辑，不影响运行时行为和 API。CUDA 13 修复是向后兼容的，对小于 13.0 的 CUDA 版本无影响。

3. 性能风险：无。新内核与原有性能水平一致。

• 影响：

1. 用户：无直接用户可见影响。此 PR 为基础设施变更，后续可让使用 HEAD_DIM_K=512 的模型（如某些 DeepSeek 变体）受益于 sm90 加速。

2. 系统：不影响生产环境运行时行为，仅影响 sgl-kernel 的构建。

3. 团队：降低了 CUDA 13 环境下开发的门槛，避免编译失败。 - 风险标记：依赖升级

关联脉络

- PR #29356 fix(bench): pass DCP_RANK/DCP_WORLD_SIZE to set_mla_kv_buffer_kernel: 同为 MLA 相关变更，且涉及 flashmla 路径。
- PR #14194 [feature] implement dcp for deepseek_v2: DeepSeek MLA 相关，可能受益于 flashmla 内核改进。