

PR #36929 完整报告

sgl-project/sglang

Update CUDA 13.4 image to flashinfer 0.6.18rc10, cutedsl 4.8. Fix sgl- wheel unpinning

合并时间: 2026-08-29 08:52

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36929>

执行摘要

- 一句话: 更新 cu134 镜像依赖, 修复 sgl- wheel 固定问题
- 推荐动作: 该 PR 是基础设施调整, 值得关注其解决 wheel 版本冲突的思路, 以及如何通过环境变量和 sed 逻辑管理依赖版本。

功能与动机

PR 动机明确: 一是修复本地构建的 sgl-* wheel 被 pip 安装版本替换的问题; 二是升级依赖以支持新硬件和 JIT 缓存。具体地, FlashInfer 0.6.18 为 CUDA 13.4 提供了 JIT 缓存, cutedsl 4.8 支持 sm_107 架构, 因此需要更新镜像中的依赖版本。

实现拆解

本 PR 仅修改 `docker/Dockerfile.cu134` 文件, 具体变更如下:

1. 移除对 cutedsl 4.6.2 的架构环境变量设置: 删除了 CUTE_DSL_ARCH 和 CUTE_EXPERIMENTAL_DSL_ARCH 的环境变量设置, 因为旧版本 cutedsl 没有 sm_107 架构, 需要设置这些变量才能运行, 而新版本 4.8 已经原生支持 sm_107, 不再需要。
2. 更新依赖版本参数: 将 FLASHINFER_VERSION 从 0.6.17 更新为 0.6.18rc10, 新增 CUTLASS_DSL_VERSION=4.8.0.dev0 参数, 并默认启用 INSTALL_FLASHINFER_JIT_CACHE=1。
3. 修改 pyproject.toml 的 sed 逻辑: 在 CUDA 版本为 13.4.0 时, 现在会将 flashinfer_python[cu13]==* 替换为指定的 FLASHINFER_VERSION, 而不是直接解除版本固定。同时删除了注释中关于 sgl-* wheel 的说明, 因为现在不再解除版本固定, 本地构建的 wheel 会保持其版本。
4. 新增对 CUTLASS_DSL_VERSION 的处理: 在安装 cutedsl 时, 使用指定的版本号, 而不是通过 pip install 自动解析。

关键文件:

- `docker/Dockerfile.cu134` (模块 Dockerfile; 类别 infra; 类型 infrastructure): 唯一变更文件, 包含依赖版本升级和 wheel 固定逻辑调整。

关键符号: 未识别

评论区精华

无 review 评论，但 Fridge003 已批准该 PR。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于 Dockerfile 的 sed 命令是否准确匹配了 pyproject.toml 中的版本模式，如果版本号模式不匹配，可能导致安装失败。此外，升级 cutedsl 到 dev 版本（4.8.0.dev0）可能存在稳定性风险，但鉴于这是为支持 sm_107 架构的必要升级，风险可控。
- 影响：影响范围仅限于使用 CUDA 13.4 镜像的部署环境，主要是依赖版本升级和 wheel 固定策略的调整，不会影响其他平台。对于依赖该镜像的用户，升级后 FlashInfer 和 cutedsl 的版本会发生变化，可能影响相关功能的稳定性，但整体可控。
- 风险标记：缺失测试覆盖

关联脉络

- PR #36912 Use kernel build node for cu134 image: 同样涉及 CUDA 13.4 镜像的构建基础设施。