PR #25576 完整报告

sgl-project/sglang

[Deps] Use cu13 extra for nvidia cutlass dsl

合并时间：2026-05-21 10:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25576

执行摘要

- 一句话：升级 cutlass-dsl 至 4.5.1 并添加 CUDA 13 额外依赖标记
- 推荐动作：该 PR 是一个常规依赖升级，架构影响极小，但关联了 B300 硬件兼容性问题。建议结合 PR #25564 和后续的 pyproject.toml 重构一同追踪。对于主要使用 CUDA 12 的用户，可留意后续是否引入条件化依赖机制。

功能与动机

关联 Issue #25564 报告 NVIDIA B300 (sm_103) 上 flash-attn-4 的 cute 内核架构检查断言失败，导致 Qwen-3.5-VL 模型首次预热前向传播崩溃。上游修复已在 Dao-AILab/flash-attention#2572 中完成，sglang 需要升级 flash-attn-4 依赖并同步更新 nvidia-cutlass-dsl 以获取兼容的 cutlass 内核。本 PR 升级 nvidia-cutlass-dsl 到 4.5.1 版本并添加 [cu13] extra，以匹配 CUDA 13 的默认构建环境。

实现拆解

1. 修改依赖声明：在 python/pyproject.toml 第 40 行，将 nvidia-cutlass-dsl==4.5.0 改为 nvidia-cutlass-dsl[cu13]==4.5.1。
2. 添加 CUDA 13 extra 标记：通过 [cu13] 显式标记该依赖在 CUDA 13 环境下需要安装配套的 cu13 变体，确保与 flashinfer_python[cu13] 等其他 CUDA 13 依赖一致。
3. 版本升级：从 4.5.0 升级到 4.5.1，以包含上游修复和 B300 兼容性改进。
4. 测试与验证：通过 CI 运行确认变更通过所有基础测试，PR 状态显示最新 PR Test 通过。

关键文件：

- python/pyproject.toml（模块 项目配置；类别 config；类型 configuration）：唯一变更文件，升级 nvidia-cutlass-dsl 到 4.5.1 并添加 [cu13] extra 标记，直接影响 CUDA 13 环境下的 cutlass 依赖解析。

关键符号：未识别

关键源码片段

python/pyproject.toml

唯一变更文件，升级 nvidia-cutlass-dsl 到 4.5.1 并添加 [cu13] extra 标记，直接影响 CUDA 13 环境下的 cutlass 依赖解析。

```
# 在 dependencies 列表中，将旧版本 4.5.0 升级为 4.5.1，
# 并添加 [cu13] extra 标记以匹配 CUDA 13 默认构建环境。
# 这样 nvidia-cutlass-dsl 在 CUDA 13 下会安装 cu13 变体，
# 提供 B300 (sm_103) 所需的兼容性修复。
dependencies = [
    ...
    "nvidia-cutlass-dsl[cu13]==4.5.1", # 升级并添加 extra
    ...
]
```

评论区精华

Review 中审核者 trevor-m 提出关键问题：是否可以在仅针对 CUDA 13 构建时应用此依赖？并询问 我们是否仍为 CUDA 12 构建，还是已完全切换到 CUDA 13？。作者 mmangkad 回应：对于 Docker 构建可以处理，但对于发布的 wheel 包，目前没有干净的方式使依赖条件化于 CUDA 版本。最终 Kangyan-Zhou 决定：先合并此变更用于后续发布，稍后重构 toml 文件使其感知 CUDA 版本。 该讨论已达成决策并被标记为已解决。

- 条件化依赖 CUDA 版本 (design): 决定先合并此变更用于后续发布，稍后重构 toml 文件使其感知 CUDA 版本。

风险与影响

- 风险：
  1. CUDA 12 兼容性风险：[cu13] *extra* 标记会导致在 CUDA 12 环境下 pip install sglang 时尝试安装不兼容的 nvidia-cutlass-dsl[cu13] 版本，可能引发安装失败或运行时错误。但评论中已确定当前项目已默认使用 CUDA 13，且 Docker 构建可以通过剥离 [cu13] 来兼容 CUDA 12。
  2. 依赖锁定风险：版本从 ==4.5.0 升级到 ==4.5.1，若新版本存在未发现的回归，可能影响依赖于 cutlass 内核的模块（如 flash-attn-4 和其他 GPU 内核）。
- 影响：
  1. 对用户：修复了 NVIDIA B300 (sm_103) 用户启动 flash-attn-4 相关模型时的崩溃问题。现有 CUDA 12 用户如果通过非 Docker 方式安装，可能面临 pip install 失败的风险，但官方 Docker 镜像不受影响。
  2. 对系统：仅修改 Python 包依赖声明，无运行时逻辑变更。
  3. 对团队：需在后续版本中考虑将 pyproject.toml 重构为支持 CUDA 版本条件化依赖，以避免类似问题。 - 风险标记：条件化依赖缺失，兼容性风险

关联脉络

- PR #25564 [Bug] Qwen-3.5 on B300 *crashes in flash-attn-4 cute kernel*: 直接关联的 Issue，本 PR 修复了其描述的 cutlass-dsl 依赖版本问题。