PR #26854 完整报告

sgl-project/sglang

[Deps] Bump FI to 0.6.12 and cutedsl to 4.5.2

合并时间：2026-06-04 03:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26854

执行摘要

- 一句话：升级 FlashInfer 到 0.6.12，CUTLASS DSL 到 4.5.2
- 推荐动作：建议合入，但需确认 CI 中无关失败不会影响后续主线。

功能与动机

根据 PR 标题和描述，此变更为常规依赖升级，以跟进上游更新，获取可能的 bug 修复和性能提升。

实现拆解

1. 更新 pyproject.toml 中的依赖版本约束：将 flashinfer_python 和 flashinfer_cubin 从 0.6.11.post1 改为 0.6.12，nvidia-cutlass-dsl 从 4.5.1 改为 4.5.2。
2. 更新 Dockerfile 中的构建参数 FLASHINFER_VERSION 从 0.6.11.post1 改为 0.6.12。
3. 在 engine.py 中将运行时版本断言的最低版本从 0.6.11.post1 改为 0.6.12。
4. 在 common.py 中更新 check_pkg_version_at_least 函数文档字符串的示例版本号。

关键文件：

- python/pyproject.toml（模块 项目配置；类别 config；类型 configuration）：核心依赖版本约束更新，定义项目安装时所需的最低版本。
- python/sglang/srt/entrypoints/engine.py（模块 引擎入口；类别 source；类型 core-logic）：运行时版本断言更新，确保启动时加载的 FlashInfer 版本 >= 0.6.12。
- python/sglang/srt/utils/common.py（模块 工具库；类别 source；类型 documentation）：check_pkg_version_at_least 的文档字符串示例版本更新，与主代码保持一致。
- docker/Dockerfile（模块 容器构建；类别 infra；类型 infrastructure）：构建参数 FLASHINFER_VERSION 更新，确保 Docker 镜像使用正确版本。

关键符号：未识别

关键源码片段

python/pyproject.toml

核心依赖版本约束更新，定义项目安装时所需的最低版本。

```
[project]
```

```
dependencies = [
    # ...
    "flashinfer_python[cu13]==0.6.12", # 从 0.6.11.post1 升级
    "flashinfer_cubin==0.6.12", # 同步升级
    # ...
    "nvidia-cutlass-dsl[cu13]==4.5.2", # 从 4.5.1 升级
    # ...
]
```

python/sglang/srt/entrypoints/engine.py

运行时版本断言更新，确保启动时加载的 FlashInfer 版本 >= 0.6.12。

```
# Check flashinfer version
if not get_bool_env_var("SGLANG_SKIP_SGL_KERNEL_VERSION_CHECK"):
    if server_args.attention_backend == "flashinfer":
        assert_pkg_version(
            "flashinfer_python",
            "0.6.12", # 从 0.6.11.post1 升级到 0.6.12
            "Please uninstall the old version and "
            "reinstall the latest version by following the instructions "
            "at https://docs.flashinfer.ai/installation.html.",
        )
    if _is_cuda:
        assert_pkg_version(
            "sglang-kernel",
            "0.4.3", # 保持不变
            "Please reinstall the latest version with `pip install sglang-kernel --force-reinstall`",
        )
```

评论区精华

作者 mmangkad 在评论中指出唯一失败的 CI 测试（Gemma 4 26B MTP GSM8K）已在主分支上禁用，与本 PR 无关。

- CI 失败是否与 PR 相关 (other): 确认 CI 失败不是由本 PR 引入，可以忽略。

风险与影响

- 风险：属于小版本依赖升级，兼容性风险低。但需注意新版本可能引入行为变化，CI 已覆盖大部分场景。
- 影响：影响所有使用 FlashInfer attention backend 和 CUTLASS DSL 的用户，需要重新安装或构建镜像。不涉及 API 变更。
- 风险标记：低风险版本升级，CI 非关键失败

关联脉络

- 暂无明显关联 PR