PR #44036 完整报告

vllm-project/vllm

[CI/Build] Bump flashinfer to v0.6.12

合并时间：2026-06-03 06:19

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44036

执行摘要

本次 PR 将 flashinfer 依赖从 v0.6.11.post2 升级至 v0.6.12，核心变更涉及 Docker 构建配置和 Python 依赖声明。升级后 CI 测试无新增失败，与主分支表现一致，属于常规依赖更新。

功能与动机

flashinfer 是 vLLM 中用于高性能 Attention 和 MoE 推理的 CUDA 算子库。本次升级旨在跟进上游最新版本，获取 bug 修复和性能优化。

实现拆解

1. 更新 Docker 版本配置：在 docker/versions.json 中将 FLASHINFER_VERSION 默认值改为 0.6.12。
2. 更新 Dockerfile：在 docker/Dockerfile 和 docker/Dockerfile.nightly_torch 中同步更新版本号，确保构建时使用正确版本。
3. 更新 Python 包要求：在 requirements/cuda.txt 中更新 flashinfer-python 和 flashinfer-cubin 版本，使 pip 安装时使用新版本。

关键源码片段

docker/Dockerfile中的版本参数更新：

```
ARG FLASHINFER_VERSION=0.6.12 # 从 0.6.11.post2 升级
```

docker/versions.json中的配置项： `{ "FLASHINFER_VERSION":{ "default":"0.6.12"// 版本升级 } }`

评论区精华

无 review 讨论。作者在 PR 评论中说明“三个失败作业在主分支上也同样失败”，确认升级未引入新问题。

风险与影响

- 风险：依赖版本升级可能导致不兼容的 API 变更，但本 PR 仅改版本号，且 CI 通过，风险低。
- 影响：所有使用 flashinfer 的部署和运行环境将使用新版本，可能获得性能提升或 bug 修复。

关联脉络

与历史 PR #43669 均涉及 flashinfer，该 PR 为 NVFP4 不支持架构添加快速失败。本次升级可能包含相关修复。