# PR #21778 完整报告

sgl-project/sglang

Cache nvidia wheels locally to skip repeated 830 MB downloads in CI

合并时间：2026-04-01 07:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21778

## 执行摘要

本 PR 通过新增本地缓存脚本，解决了 CI 中因 pypi.nvidia.com 的 no-store 缓存策略导致的重复下载问题，将 Nvidia wheels 下载时间从每次 830 MB 减少到零，显著提升了 CI 运行效率，属于基础设施优化。

## 功能与动机

由于 pypi.nvidia.com 返回 Cache-Control: no-store，pip 在每次 CI 运行中都会重新下载约 830 MB 的 Nvidia wheels（包括 nvidia-cudnn-cu12 和 nvidia-nvshmem-cu12），造成时间和带宽浪费。PR body 明确指出："pypi.nvidia.com returns Cache-Control: no-store, causing pip to re-download cudnn (~707 MB) and nvshmem (~125 MB) on every CI run." 目标是通过本地缓存避免重复下载，缩短 CI 安装时间。

## 实现拆解

实现主要包括两个文件变更：

1. 新增脚本 scripts/ci/cuda/cache_nvidia_wheels.sh：
   - 使用 curl 下载 wheels 到持久化目录 /root/.cache/nvidia-wheels/。
   - 通过 unzip -t 验证文件完整性，避免使用损坏的缓存。
   - 预安装 wheels 使后续 pip install 看到 "Requirement already satisfied"。
   - 代码示例：
2. 修改脚本 scripts/ci/cuda/ci_install_dependency.sh：
   - 提取版本变量 NVIDIA_CUDNN_VERSION 和 NVIDIA_NVSHMEM_VERSION，避免硬编码。
   - 在安装主包前调用缓存脚本：source "$(dirname "$0")/cache_nvidia_wheels.sh"。
   - 更新依赖检查逻辑，使用变量版本号进行比较。

## 评论区精华

review 评论为空，但 commit 历史揭示了设计迭代过程：

> 作者尝试了多种方案，从初始的 pip 约束到本地缓存，最终采用 curl 下载和完整性检查。例如，commit 消息显示："Use curl with file-exists check instead of pip download pip download also respects no-store and re-downloads every time." 这表明了绕过

no-store 限制的实用技巧。

## 风险与影响

风险：

- 依赖版本硬编码：脚本中 wheels 的 URL 包含固定版本号，若 Nvidia 更新版本，需手动更新脚本，否则可能导致兼容性问题。
- 缓存完整性风险：尽管使用 unzip -t 验证，但极端情况下缓存文件可能损坏，脚本会重新下载，增加失败风险。
- 环境依赖性：缓存目录 /root/.cache/nvidia-wheels/ 假设特定权限和路径，在不同 CI 环境中可能不适用。

影响：

- 正面影响：CI 安装时间大幅减少，从约 4.5 分钟缩短到 1.5 分钟，提升开发效率，减少带宽消耗。
- 范围：影响所有使用该 CI 脚本的运行器，但对最终用户透明，无需额外配置。

## 关联脉络

与近期历史 PR 关联：

- PR #21751和 PR #21753：同属 CI 优化类别，分别修复测试超时和测试套件检测问题，与本 PR 共同构成 CI 基础设施的持续改进脉络。
- 整体趋势：团队正专注于提升 CI 效率和可靠性，减少等待时间和资源浪费，本 PR 是这一方向的具体实践。