# PR #22602 完整报告

sgl-project/sglang

ci: use local NVIDIA wheels to avoid re-downloading ~2GB every CI run

合并时间：2026-04-12 12:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22602

## 执行摘要

该 PR 通过修改 CI 脚本 cache_nvidia_wheels.sh，引入本地 NVIDIA wheel 缓存机制，避免每次 CI 运行重复下载约 2GB 依赖，显著提升 5090 等网络较慢 runner 的效率，节省约 7 分钟安装时间。变更风险低，但 review 中提到的 uv 包管理器支持问题未解决。

## 功能与动机

动机：pypi.nvidia.com 的 Cache-Control: no-store 策略导致每次 CI 运行都重新下载所有 NVIDIA torch 依赖（如 cublas、cufft 等，总计约 2GB），这不仅浪费带宽，还使网络较慢的 5090 runner 面临超时风险（下载占用约 7 分钟，总安装超时 20 分钟）。PR body 明确指出目标是“避免重复下载”，优化 CI 性能。

## 实现拆解

仅修改一个文件 scripts/ci/cuda/cache_nvidia_wheels.sh：

- 注释更新：从仅缓存 cudnn 和 nvshmem 扩展为缓存所有 NVIDIA 依赖，并说明预缓存目录位置（/opt/ci-cache/nvidia-pip-wheels/）。
- 核心逻辑： bash NVIDIA_PIP_WHEELS="/root/.cache/nvidia-pip-wheels" if [ -d "$NVIDIA_PIP_WHEELS" ] && ls "$NVIDIA_PIP_WHEELS"/*.whl &>/dev/null; then export PIP_FIND_LINKS="${PIP_FIND_LINKS:+$PIP_FIND_LINKS }$NVIDIA_PIP_WHEELS" fi 条件判断确保目录存在且包含 wheel 文件时才设置 PIP_FIND_LINKS，使 pip 优先从本地安装，不影响无缓存目录的 runner。
- 原有功能保留：继续缓存和安装 cudnn、nvshmem wheel。

## 评论区精华

review 中仅有一条来自 gemini-code-assist[bot] 的评论，指出潜在问题：

> “PIP_FIND_LINKS covers runners using standard pip... many other CI runners use uv... uv does not respect PIP_FIND_LINKS; it uses UV_FIND_LINKS instead.”

建议同时导出两个环境变量以确保所有 runner 受益。此评论未得到回复，PR 已合并，可能被视为优化补充项而非阻塞问题。

## 风险与影响

风险：

- 本地缓存 wheel 版本可能与 pip 要求不匹配，但脚本无版本检查，依赖预缓存机制的正确性。
- 未支持 UV_FIND_LINKS，使用 uv 的 runner 无法享受优化，但不会导致功能错误。

影响：

- 正面：减少 CI 下载时间和网络负载，提升 5090 runner 稳定性。
- 范围：仅影响 CI 基础设施，对用户功能和系统逻辑无影响。

## 关联脉络

从近期历史 PR 看，该 PR 属于一系列 CI 优化的一部分：

- PR 22609 调整 B200 测试时间防止超时。
- PR 22608 重命名 GB200 工作流文件。
- PR 22228 修复 AMD CI 超时配置。

这些 PR 共同反映了团队对 CI 效率和稳定性的持续改进，特别是在多硬件平台（如 Blackwell、AMD）和网络环境下的优化趋势。本 PR 通过本地缓存解决依赖下载瓶颈，是基础设施优化的重要一环。