PR #40669 完整报告

vllm-project/vllm

[Build] Bump CUDA to 13.0.2 to match PyTorch 2.11.0

合并时间：2026-04-24 18:27

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40669

执行摘要

该 PR 将 vllm 构建基础设施中的 CUDA 版本统一升级至 13.0.2，以对齐 PyTorch 2.11.0 的构建环境。变更涉及 Dockerfile、BuildKite 发布流水线、夜间构建脚本和文档，消除了之前多个位置版本漂移的问题。review 中讨论了架构列表的一致性，已解决。本次变更风险低，建议合并。

功能与动机

PyTorch 2.11.0（在 requirements/cuda.txt 中锁定）是基于 CUDA 13.0.2 构建的，但 vllm 的 Dockerfile 默认使用 13.0.0、BuildKite 发布流水线覆盖为 13.0.1、GB300 文档也使用 13.0.1，存在版本不一致。本 PR 将所有 CUDA 13 相关引脚统一为 13.0.2，确保构建工具链与 PyTorch 完全一致，避免潜在的兼容性问题。

实现拆解

1. 更新 Dockerfile 默认版本 docker/Dockerfile 中 ARG CUDA_VERSION=13.0.0 → 13.0.2，这是所有 Docker 构建的基础。
2. 同步元数据文件 docker/versions.json 通过自动生成工具同步更新，CUDA_VERSION、BUILD_BASE_IMAGE、FINAL_BASE_IMAGE 的默认值随之改变。
3. 更新 BuildKite 发布流水线 .buildkite/release-pipeline.yaml 中，aarch64 wheel、x86_64 wheel 和 x86_64 release image 三个 CUDA 13.0 构建步骤的 CUDA_VERSION 和 BUILD_BASE_IMAGE 从 13.0.1 更新为 13.0.2。CUDA 12.9 步骤保持不变。
4. 更新夜间构建脚本 .buildkite/image_build/image_build_torch_nightly.sh 中的 NIGHTLY_CUDA_VERSION 从 13.0.0 改为 13.0.2。
5. 更新文档与依赖图 - docs/getting_started/installation/gpu.cuda.inc.md 中 GB300 示例的版本号与基础镜像同步更新。 - docs/assets/contributing/dockerfile-stages-dependency.png 自动重新生成（仅标签变化，拓扑不变）。

评论区精华

- gemini-code-assist[bot] 指出aarch64 CUDA 13.0 wheel 构建的 torch_cuda_arch_list 缺少架构 12.1，与 release image 构建不一致。作者回复“fixed.”并实际修复，将 12.1 加入 aarch64 arch list。
- gemini-code-assist[bot] 同时建议x86_64 release image 应显式指定 arch list 以包含 12.1。作者解释 12.1 仅用于 GH10 芯片（无 x86_64 变体），无需修改。该解释被接受。

- Harry-Chen 提醒注意 #39878 已涵盖架构部分，建议 rebase；作者随后将 PR 范围缩小为仅版本升级。

风险与影响

- 回归风险低：CUDA 13.0.2 是 PyTorch 2.11.0 的官方构建版本，已充分测试。Dockerfile 仅更改默认 ARG，不影响显式传参的构建。
- 构建一致性提升：消除版本漂移后，所有 CUDA 13 构建产物均基于同一基础镜像，降低因工具链差异导致的潜在问题。
- 文档准确性：用户参考 GB300 示例时不会因版本号错误而构建失败。
  - 无代码逻辑变更，不影响运行时行为。

关联脉络

本 PR 与 #39878（切换默认 CUDA 为 13.0 并更新架构列表）紧密相关。#39878 完成了主要迁移，本 PR 将版本修正确保完全对齐 PyTorch 2.11.0。两者共同推进 vllm 的 CUDA 13 构建链到稳定状态。