PR #39878 完整报告

vllm-project/vllm

[Build] Switch default CUDA to 13.0, update CUDA architecture lists, clean up stale build-args

合并时间：2026-04-23 15:51

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39878

执行摘要

- 一句话：默认 CUDA 版本 12.9 → 13.0，重构架构列表
- 推荐动作：建议仔细阅读架构列表调整部分，特别是关于 SM86 和 SM89 纳入的决策理由。同时关注 Volta 用户迁移路径的文档说明是否充分。

功能与动机

PyTorch 2.11 发布后默认 CUDA 版本升级至 13.0，vLLM 需要对齐以保持默认构建一致性。同时需更新 CUDA 架构列表以反映 PyTorch 上游变化，并清理遗留的构建参数引用。

实现拆解

1. 切换默认 CUDA 版本：在 vllm/envs.py 中将 VLLM_MAIN_CUDA_VERSION 默认值从 12.9 改为 13.0，运行时环境变量默认值同步变更。
2. 更新 CUDA 架构列表：遵循 PyTorch RELEASE.md 重新定义架构列表。默认（CUDA 13.0）x86_64 为 7.5 8.0 8.6 8.9 9.0 10.0 12.0+PTX，aarch64 新增 11.0（Thor）并移除 12.1。CUDA 12.9 变体使用无 +PTX 的对应列表。列表在 release-pipeline.yaml 中定义为顶层环境变量 CUDA_ARCH_X86 等，所有构建步骤通过 ${...} 引用，消除硬编码重复。
3. 清理 FLASHINFER_AOT_COMPILE：在 docker/docker-bake.hcl 和构建命令中移除该残留参数，该参数早在 PR #32627 中已从 Dockerfile 删除。
4. 调整发布脚本：更新 annotate-release.sh 中的默认 wheel 名、镜像标签和下载提示，使 CUDA 13.0 成为无后缀默认版本。同时为 aarch64 CUDA 12.9 构建显式添加 manylinux_2_31 参数保证一致。
5. 同步 CMakeLists.txt 架构检查：更新 CUDA 编译器版本 >= 13.0 时支持的架构列表，移除 Volta（7.0/7.5 中的 7.0），新增 10.0 和 12.0。同时更新版本声明以匹配实际使用的 CUDA 12.9+。

关键文件：

- .buildkite/release-pipeline.yaml（模块 CI 流水线；类别 config；类型 configuration）：核心 CI 配置文件，提取架构列表为环境变量，消除重复并作为所有构建步骤的单一来源。
- vllm/envs.py（模块 环境配置；类别 source；类型 core-logic）：定义运行时 CUDA 版本默认值，直接影响 pip install vllm 的变体选择。

- docker/Dockerfile（模块 Docker 构建；类别 infra；类型 infrastructure）：Torch CUDA 架构列表默认值更新，影响所有 Docker 构建。

关键符号：未识别

关键源码片段

vllm/envs.py

定义运行时 CUDA 版本默认值，直接影响 pip install vllm 的变体选择。

```
# vllm/envs.py

class EnvironmentVariables:
    # ...
    # 默认 CUDA 版本，用于运行时 PCI ID 匹配和变体选择
    VLLM_MAIN_CUDA_VERSION: str = "13.0" # 原来为 "12.9"
    # ...

# 运行时环境变量定义部分
environment_variables: dict[str, Callable[[], Any]] = {
    # ...
    "VLLM_MAIN_CUDA_VERSION": lambda: (
        os.getenv("VLLM_MAIN_CUDA_VERSION", "").lower() or "13.0"
        # 原来默认值为 "12.9"
    ),
    # ...
}
```

docker/Dockerfile

Torch CUDA 架构列表默认值更新，影响所有 Docker 构建。

```
# docker/Dockerfile

# 构建阶段：编译 torch 扩展所需的架构列表
ARG torch_cuda_arch_list='7.5 8.0 8.6 8.9 9.0 10.0 12.0+PTX'
# 原来为 '7.0 7.5 8.0 8.9 9.0 10.0 12.0'
# 变更：移除 7.0（Volta），新增 8.6（Ampere GA106/107），添加 +PTX 前向兼容
ENV TORCH_CUDA_ARCH_LIST=${torch_cuda_arch_list}
```

评论区精华

- Volta支持移除争议：gemini-code-assistbot指出移除SM70（V100）可能导致用户回归。Harry-Chen 回应 PyTorch 已从 12.8 起放弃 Volta 支持，vLLM 跟随上游决定。
- aarch64 架构完整性：dmitry-tokarev-nv 建议为 aarch64 添加 SM86（8.6）和 SM103（10.3）以覆盖更多 GPU。Harry-Chen 解释 aarch64 此前未包含 8.6，且 10.3 可通过 CUDA 13 的 family specifier 10.0f 被覆盖，无需显式添加。
- CMakeLists.txt 版本条件：dmitry-tokarev-nv 指出 CUDA 12.8 不支持 11.0 和 12.1 架构，Harry-Chen 确认实际仅使用 CUDA 12.9，风险可控。

- Volta (SM70) 支持移除 (design): 接受移除，因为上游 PyTorch 已放弃，且 Volta 用户可通过 CUDA 12.9 变体获得支持。
- aarch64 架构列表完整性 (correctness): 未添加 8.6 和 10.3，但确认了架构兼容性理由。
- CMakeLists.txt 版本条件准确性 (correctness): 理解存在偏差，但实际风险低，因为只使用 CUDA 12.9+。
- aarch64 CUDA 12.9 缺少 manylinux 参数 (correctness): 已添加 manylinux_2_31 参数。

风险与影响

- 风险:
  - Volta 用户兼容性：移除 SM70 意味着 Tesla V100 等 GPU 无法在默认 CUDA 13.0 构建下运行。用户必须切换到 CUDA 12.9 变体，可能造成部署中断。
  - 架构列表偏离上游：部分架构（如 SM86、SM89）超出 PyTorch 默认列表，虽便于覆盖更多 GPU，但可能引入未充分测试的代码路径。
  - CMakeLists.txt 条件偏移：CUDA 13.0 分支的架构列表包含 11.0（Thor），但 CUDA 12.9 版本分支未包含 11.0，可能导致用户在使用 CUDA 12.9 构建时缺少 Thor 支持。
- 影响:
  - 用户：从 pip install vllm 默认获得 CUDA 13.0 构建；Volta 用户需改用 vllm+cu129 变体。
  - 构建团队：发布流程需同步调整，确保两套变体正确生成。
  - 系统兼容性：需要 NVIDIA 驱动程序支持 CUDA 13.0（R570+）。
  - 风险标记：默认 CUDA 版本升级，Volta 支持移除，架构列表偏离上游

关联脉络

- PR #32627 [Build] Remove FLASHINFER_AOT_COMPILE from Dockerfile: 本 PR 清理了该 PR 遗留的 FLASHINFER_AOT_COMPILE 构建参数引用。