PR #39233 完整报告

vllm-project/vllm

[NVIDIA] Add sm_110 (Jetson Thor) to CUDA 13.0 build targets

合并时间：2026-04-24 03:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39233

执行摘要

- 一句话：新增 sm_110 (Jetson Thor) 到 CUDA 13.0 构建目标
- 推荐动作：此 PR 变更简单明确，解决了 Jetson Thor 用户无法使用 vLLM 官方镜像的关键问题。建议快速合并，并注意保持配置文件间的架构列表一致性。如果未来有其他架构加入，应同步更新此处。

功能与动机

在 PR body 中明确指出：官方 vllm/vllm-openai:v0.19.0-cu130 镜像未包含 sm_110 内核，导致 Jetson Thor（GB10）上 Marlin（GPTQ-INT4）、CUTLASS（cutlass_scaled_mm、FP8）和 flash-attn（BF16）均报 CUDA error: no kernel image is available for execution on the device。因此需要将 11.0 加入 CUDA 13.0+ 的构建架构列表，以使镜像能在 Jetson Thor 上直接运行。

实现拆解

变更涉及 4 个配置文件，分为三步：

1. Dockerfile（docker/Dockerfile）：在构建基础镜像和 KV connectors 两个构建阶段中，将 torch_cuda_arch_list 默认值由 '7.5 8.0 8.6 8.9 9.0 10.0 12.0+PTX' 修改为 '7.5 8.0 8.6 8.9 9.0 10.0 11.0 12.0+PTX'，新增 11.0。
2. Docker Bake 和版本配置：
   - docker/docker-bake.hcl：将 TORCH_CUDA_ARCH_LIST 默认值从 "8.0 8.9 9.0 10.0" 更新为 "8.0 8.9 9.0 10.0 11.0 12.0"，同时补充了此前缺失的 12.0（Blackwell）。
   - docker/versions.json：同步更新默认架构列表，添加 11.0。
3. FlashInfer 构建脚本（tools/flashinfer-build.sh）：在 CUDA 13.0+ 分支的 FI_TORCH_CUDA_ARCH_LIST 中添加 11.0，并将 12.0 修正为 12.0f（包含 PTX），确保 FlashInfer AOT 编译为 Jetson Thor 生成内核。

所有变更均仅针对 CUDA ≥13.0 的构建环境；对于 CUDA 12.x 环境，CMake 中的 cuda_archs_loose_intersection() 会自动过滤掉 11.0，因此对旧版本构建无影响。

关键文件：

- docker/Dockerfile（模块 Docker 构建；类别 infra；类型 infrastructure）：在构建基础镜像和 KV connectors 两个阶段中，向 torch_cuda_arch_list 添加 11.0，是确保 Jetson

Thor 内核编译的核心变更。

- docker/docker-bake.hcl（模块 Docker 构建；类别 infra；类型 configuration）：修改 TORCH_CUDA_ARCH_LIST 默认值以包含 11.0 和 12.0，与 Dockerfile 保持一致，避免构建选项遗漏关键架构。
- docker/versions.json（模块 Docker 构建；类别 infra；类型 configuration）：存储构建配置的默认值，与 Dockerfile 同步，确保一致性。
- tools/flashinfer-build.sh（模块 构建工具；类别 other；类型 configuration）：在 FlashInfer AOT 编译的 arch list 中增加 11.0，确保 Jetson Thor 拥有对应的 FlashInfer 预编译内核。

关键符号：未识别

关键源码片段

docker/Dockerfile

在构建基础镜像和 KV connectors 两个阶段中，向 torch_cuda_arch_list 添加 11.0，是确保 Jetson Thor 内核编译的核心变更。

```
# 构建基础镜像时指定 CUDA 架构列表，添加 sm_110 (11.0) 以支持 Jetson Thor
ARG torch_cuda_arch_list='7.5 8.0 8.6 8.9 9.0 10.0 11.0 12.0+PTX'
ENV TORCH_CUDA_ARCH_LIST=${torch_cuda_arch_list}

# 安装 KV connectors 的构建阶段同样添加 sm_110
ARG torch_cuda_arch_list='7.5 8.0 8.6 8.9 9.0 10.0 11.0 12.0+PTX'
ENV TORCH_CUDA_ARCH_LIST=${torch_cuda_arch_list}
```

评论区精华

Review 中主要讨论了以下几点：

- 关于移除 FlashInfer TRTLLM BMM headers 预下载块：gemini-code-assist[bot] 指出该移除可能导致 air-gapped 环境回归，因为之前该块用于在构建时缓存 headers，避免运行时从 edge.urm.nvidia.com 下载。但最终提交未包含该移除（相关 commits 已被 revert），离线部署不受影响。
- 关于 docker-bake.hcl 缺少 sm_120 (12.0)：gemini-code-assist[bot] 指出 Docker Bake 的默认架构列表缺少 12.0，与 Dockerfile 不一致。最终提交已添加 12.0，问题已解决。
- 关于错误添加 sm_70 (7.0) 和移除 sm_86 (8.6)：mgoin 评论“We don't support 7.0”和“Why did you add 7.0 and remove 8.6?”，指出不应添加 7.0 且不应移除 8.6。最终提交已纠正，仅添加 11.0，保留 7.5、8.0、8.6 等。
  - 移除 FlashInfer TRTLLM BMM headers 预下载可能导致离线部署失败 (question)：最终提交已 revert 了该移除（相关 commit 撤销），预下载块得以保留，离线部署不受影响。
  - docker-bake.hcl 缺少 sm_120 (12.0) 架构 (correctness)：最终提交已将默认值更新为 8.0 8.9 9.0 10.0 11.0 12.0，补充了 12.0，问题已解决。
  - 错误添加 sm_70 (7.0) 和移除 sm_86 (8.6) (correctness)：最终提交修正了这两处错误，仅添加 11.0，保留原有架构列表（7.5、8.0、8.6、8.9、9.0、10.0、12.0+PTX），

问题已解决。

风险与影响

- 风险：主要风险包括：
  - 构建时间增加：新增 sm_110 架构会导致 CUDA 内核编译时间延长，但只影响 CUDA 13.0+ 的 Docker 构建，且增量有限。
  - 离线部署兼容性：虽然 review 中提及的 FlashInfer headers 预下载移除已 revert，但若后续合并其他变更导致回归，可能影响离线部署。目前无此风险。
  - 旧版本不影响：由于 CMake 的 cuda_archs_loose_intersection 会在 CUDA <13.0 时自动过滤 11.0，因此现有 CUDA 12.x 构建完全不受影响，兼容性风险极低。
  - 影响：
  - 用户影响：Jetson Thor 用户现在可以直接使用官方 CUDA 13.0 Docker 镜像，无需自行编译内核，开箱即用。对普通用户透明，无破坏性变更。
  - 系统影响：Docker 镜像体积略增（因包含更多架构内核），构建时间稍长。
  - 团队维护：维护成本低，后续只需在新增架构时同步更新这些配置文件的列表即可。
  - 风险标记：构建时间增加，兼容性风险低，离线部署潜在风险（已解决）

关联脉络

- PR #40669 [Build] Bump CUDA to 13.0.2 to match PyTorch 2.11.0: 此 PR 将 CUDA 版本提升至 13.0.2，并为构建基础设施（如 Dockerfile、release-pipeline）奠定了基础，与本 PR 同属构建系统改进，且本 PR 的变更依赖 CUDA 13.0+ 构建环境。