

PR #48041 完整报告

vllm-project/vllm

[Build/CI] Build arm64 PR and postmerge image builds for Blackwell SM10x and SM110

合并时间: 2026-07-10 22:12

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48041>

执行摘要

- 一句话: arm64 CI 镜像新增 Blackwell GPU 架构支持
- 推荐动作: 建议合并, 改动明确且经过审核。关注后续 CI 构建是否成功, 以及是否需要在文档中更新支持的硬件列表。

功能与动机

当前无法在 PR 的 CI 中构建可直接部署到 GB200 或 GB300 设备的 arm64 镜像, 该 PR 旨在解决此问题。PR body 明确说明: "The main benefit here is to deploy an image built in CI for a pull request directly onto a GB200 or GB300, which is not possible today."

实现拆解

1. 修改构建脚本参数: 在 `.buildkite/image_build/image_build_arm64.sh` 中, 将 `torch_cuda_arch_list` 从 `"9.0 12.0"` 扩展为 `"9.0 10.0 11.0 12.0"`, 新增对 CUDA 架构 10.0 (SM100/SM103) 和 11.0 (SM110) 的支持。
2. 更新注释: 同步更新脚本注释, 明确列出新增的 Blackwell/Thor 架构。
3. 无其他文件或逻辑变更: 该 PR 仅此一处修改, 无需测试、配置或部署配套调整。

关键文件:

- `.buildkite/image_build/image_build_arm64.sh` (模块 CI 构建脚本; 类别 `infra`; 类型 `configuration`): 唯一变更文件, 修改 `torch_cuda_arch_list` 构建参数以支持 Blackwell SM10x 和 SM110 GPU 架构。

关键符号: 未识别

关键源码片段

`.buildkite/image_build/image_build_arm64.sh`

唯一变更文件, 修改 `torch_cuda_arch_list` 构建参数以支持 Blackwell SM10x 和 SM110 GPU 架构。

```
# 文件: .buildkite/image_build/image_build_arm64.sh
# 该脚本负责构建 arm64 架构的 Docker 镜像
# 修改点: 扩展 torch_cuda_arch_list 以支持 Blackwell GPU

if docker manifest inspect "$IMAGE" >/dev/null 2>&1; then
```

```
    echo "Image found"
else
    echo "Image not found, proceeding with build..."
    # build for arm64 GPU targets: Grace/GH200 (sm_90),
    # Blackwell/Thor (sm_100/sm_103/sm_110), and DGX Spark/GB10
    # (sm_121, family-covered by 12.0 under CUDA 13)
    docker build --file docker/Dockerfile \
        --platform linux/arm64 \
        --build-arg max_jobs=16 \
        --build-arg nvcc_threads=4 \
        --build-arg torch_cuda_arch_list="9.0 10.0 11.0 12.0" \
        --build-arg USE_SCCACHE=1 \
        --build-arg buildkite_commit="$BUILDKITE_COMMIT" \
        --tag "$IMAGE" \
    .
fi
```

评论区精华

无实质性技术讨论。LucasWilkinson 直接批准，claude[bot] 自动评论因来自 fork 而跳过审查。Mergify 机器人提示存在合并冲突，但最终通过 rebase 解决。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅扩展构建参数，不会影响现有 arm64 镜像构建流程（GH200 和 DGX Spark/GB10 仍受支持）。新增架构可能导致镜像构建时间略微增加，但影响可忽略。
- 影响：影响范围小：仅影响 arm64 镜像构建流程，用户无感知。团队可在 CI 中直接获取支持 Blackwell GPU 的 arm64 镜像，简化部署流程。
- 风险标记：仅 CI 镜像构建变更，无功能影响

关联脉络

- 暂无明显关联 PR