

PR #7293 完整报告

verl-project/verl

[ci] chore: Update ci image

合并时间: 2026-08-13 19:03

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7293>

执行摘要

- 一句话: 更新 Ascend CI 镜像标签, 调整 nightly 调度与依赖
- 推荐动作: 该 PR 属于纯 CI 治理, 逻辑机械, 不建议逐文件精读; 但如果你负责 CI 或 NPU 硬件支持, 值得关注两点: 一是镜像标签从长版本号收敛为「后端 + 硬件」短标签的治理思路, 让版本升级从改 20 个文件变成换一个 latest 标签; 二是 crane CLI 从 tag delete 迁移到 delete 的踩坑点。可与 #7345、#7338、#7311 结合起来看 Ascend CI 镜像体系的完整演进。

功能与动机

PR body 未提供实质背景 (仅模板内容), 动机需从提交与文件变更推断: 作者需要将 Ascend CI 的镜像引用整体切换到新构建的基础镜像。旧标签直接编码了 CANN、torch_npu、Python 版本等细节 (例如 latest-cann9.0.0-torch_npu2.9.0post2-910b-ubuntu22.04-py3.11-vllm), 每次底层组件升级都要改动所有 workflow; 新标签按 vllm/sglang + a3/910b 硬件类别收敛, 把版本细节下沉到镜像发布侧。配套的 crane delete 修正 (crane tag delete 已废弃)、PYTHONPATH 补充和 triton-ascend 依赖, 都是为了配合新镜像在 CI 中正确、可维护地运行。

实现拆解

变更入口是 20 个文件的批量镜像标签替换, 核心动作可按 5 步拆解:

1. 镜像标签统一替换: 在 model_ascend.yml、e2e_ascend.yml、reward_model_sglang_ascend.yml、reward_model_vllm_ascend.yml、e2e_ppo_trainer_megatron_vllm_2_ascend.yml、e2e_v1_separate_async_ascend.yml、nightly_ascend_multinode.yml、npu_unit_tests.yml、sgl_ascend.yml、vllm_ascend.yml 等全部 workflow 中, container.image 从完整版本号标签改为短标签。vllm 系列统一换成 latest-vllm-910b-ubuntu, sglang 系列换成 latest-sglang-a3-ubuntu 或 latest-sglang-910b-ubuntu。新标签只保留推理后端与硬件维度, 版本信息移到镜像仓库侧。
2. Nightly 任务调度调整: nightly_ascend.yml 中 4 个 job 的 cron 从 0 17 * * * 改为 0 18 * * *; quick_start_qwen3_0_6b_megatron_sglang_ascend 与 quick_start_qwen3_0_6b_fsdp2_sglang_ascend 的 runner 从 linux-aarch64-a3-8 换为 linux-aarch64-a2b3-8, 镜像同步换为 latest-sglang-910b-ubuntu, 并删除手动 pip install --no-deps TransferQueue==0.1.8 步骤, 新镜像已内置该包。

3. crane 清理命令修正: docker-build-ascend-a2.yml / docker-build-ascend-a3.yml 中删除临时标签的命令从 crane tag delete 改为 crane delete, 适配新版 crane CLI (tag delete 子命令已移除), 并保留 ll true 兜底。
4. 运行环境与依赖补充: e2e_ppo_trainer_megatron_sglang_ascend.yml 在克隆 /Megatron-LM、/MindSpeed 后新增 export PYTHONPATH=\$PYTHONPATH:/Megatron-LM:/MindSpeed, 保证后续步骤能直接 import 源码; requirements-npu.txt 增加 triton-ascend, 为 NPU 提供 Triton 后端支持。
5. 文档同步: docs/ascend_tutorial/get_start/install_guidance.rst 与 quick_start.rst 随镜像与依赖变化做了相应更新, 确保用户安装步骤与新版 CI 镜像保持一致。提交记录还显示清理了部分废弃 workflow, 属于本次镜像体系的收尾工作。

测试与配置配套: 本 PR 无新增单元测试, CI 行为由后续 Ascend 工作流实跑验证; 镜像构建工作流 docker-build-ascend-a2 / a3 与各消费方工作流同步更新, 避免新标签尚未发布时出现拉取失败。

关键文件:

- .github/workflows/nightly_ascend.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : 集中承载 nightly 调度时间、runner、镜像标签与依赖安装的调整, 是本 PR 中改动最复杂的工作流
- .github/workflows/e2e_ppo_trainer_megatron_sglang_ascend.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : 除了镜像标签外, 还补上了 Megatron-LM/MindSpeed 的 PYTHONPATH, 是运行环境修正的关键文件
- .github/workflows/model_ascend.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : 模型测试工作流, 镜像标签从固定 CANN 版本改为 latest-vllm-910b-ubuntu 短标签
- .github/workflows/e2e_ascend.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : e2e 主流程工作流, 三类镜像 (vllm-a3、sglang-a3、vllm-910b) 全部切换到短标签, 覆盖最广
- .github/workflows/docker-build-ascend-a2.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : 镜像构建侧同步修正 crane 命令从 tag delete 到 delete, 保证临时 tag 清理在新 CLI 下可用
- .github/workflows/docker-build-ascend-a3.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : 与 a2 对应的 a3 平台构建工作流, 同样修正 crane 命令
- requirements-npu.txt (模块 依赖清单; 类别 config; 类型 dependencies) : 新增 triton-ascend 依赖, 配合新版镜像为 NPU 提供 Triton 后端支持
- docs/ascend_tutorial/get_start/install_guidance.rst (模块 文档; 类别 docs; 类型 documentation) : 随镜像与依赖变化同步更新 NPU 安装指导, 保证用户安装步骤与新版 CI 镜像一致

关键符号: 未识别

关键源码片段

[.github/workflows/nightly_ascend.yml](https://github.com/Ascend-Workflows/nightly_ascend.yml)

集中承载 nightly 调度时间、runner、镜像标签与依赖安装的调整，是本 PR 中改动最复杂的工作流

```
# nightly_ascend.yml —— Ascend 夜间回归任务运行环境调整示意
# 其一：计划执行时间从 UTC 17:00 调整为 UTC 18:00，错开白天 CI 高峰
# 其二：sglang 相关 job 的 runner 由 linux-aarch64-a3-8 换成 linux-aarch64-a2b3-8
# 其三：镜像统一使用新的短标签 latest-sglang-910b-ubuntu，便于镜像仓库侧滚动更新
quick_start_qwen3_0_6b_megatron_sglang_ascend:
  if: ${{ github.repository_owner == 'verl-project' && github.event.schedule == '0 18 * * *' }}
  runs-on: linux-aarch64-a2b3-8
  timeout-minutes: 180
  container:
    image: swr.ap-southeast-1.myhuaweicloud.com/base_image/ascend-ci/verl/verl:latest-sglang-910b-ubuntu
    options: >-
      --shm-size 16g
  steps:
    # 新镜像已内置 TransferQueue 等依赖，不再需要手动安装
    - name: Check initial pip list from image
      run: pip list
```

[.github/workflows/e2e_ppo_trainer_megatron_sglang_ascend.yml](#)

除了镜像标签外，还补上了 Megatron-LM/MindSpeed 的 PYTHONPATH，是运行环境修正的关键文件

```
# e2e_ppo_trainer_megatron_sglang_ascend.yml —— Megatron 训练环境准备步骤
# 克隆 Megatron-LM 与 MindSpeed 后手动导出 PYTHONPATH,
# 保证后续步骤的 Python 进程能直接 import 这两个源码仓库的模块
- name: Prepare source code
  run: |
    git clone --depth 1 --branch core_r0.18.0 https://github.com/NVIDIA/Megatron-LM.git /
    Megatron-LM
    rm -rf /MindSpeed
    git clone --depth 1 --branch core_r0.18.0 https://gitcode.com/ascend/MindSpeed.git /
    MindSpeed
    export PYTHONPATH=$PYTHONPATH:/Megatron-LM:/MindSpeed
```

评论区精华

该 PR 没有任何 review 评论 (review_comments_count=0)。审核记录方面：ji-huazhong 直接 APPROVED；wucong25 先 DISMISSED、后在最终版本上 APPROVED，推测因提交历史中的多次 merge 与删除废弃 workflow 等中间改动触发了重新审核。最终审核通过并由 wucong25 合并，没有遗留未解决的讨论。

- 审核流程：DISMISSED 后再次 APPROVED (other)：最终审核全部通过，由 wucong25 合并，无未解决疑虑。

风险与影响

- 风险：具体风险点如下：
- 镜像 latest 标签漂移：从固定版本标签改为 latest 系列短标签后，CI 拉取到的镜像内容会随发布侧更新而变动。若 docker-build-ascend-a2 / a3 上游构建失败或标签未及时推送，所有 Ascend workflow 可能拉取到旧镜像，造成 CI 结果漂移。
- crane 命令语义变化：crane tag delete 删除某个 tag，而 crane delete 删除整个 manifest（含所有 tag）。本 PR 仅用于清理 tmp-a2 / tmp-a3 临时标签，且带 ll true 兜底，风险可控，但误用在正式标签上会影响镜像可用性。
- nightly 调度时间变更：cron 从 UTC 17:00 改为 18:00，夜间回归产物生成时间顺延（北京时间次日凌晨 2 点），关注 nightly 结果的用户需要调整预期。
- 硬编码 PYTHONPATH：export PYTHONPATH=\$PYTHONPATH:/Megatron-LM:/MindSpeed 依赖工作流中 git clone 目录固定为 /Megatron-LM 与 /MindSpeed，若未来目录变化需同步修改。
- 新镜像内置依赖假设：移除手动安装 TransferQueue 的步骤，依赖新镜像已预置该包；若镜像构建与 workflow 更新节奏不一致，相关 job 会直接失败。
- triton-ascend 依赖：requirements-npu.txt 新增 triton-ascend，与既有 torch_npu / triton 版本可能存在约束冲突，需实装验证。
- 影响：对用户：无训练或推理功能影响，仅影响 Ascend 安装文档读者，安装步骤随 requirements-npu.txt 变化略有调整。对系统：全部 Ascend CI 任务切换至新镜像标签，CI 环境更统一；nightly 任务时间从 UTC 17:00 后移至 18:00。对团队：镜像维护者只需更新镜像仓库的最新标签即可完成底层 CANN / torch_npu 升级，不再需要改动 20 个 workflow，但对最新标签的发布纪律要求更高。整体影响范围限于 CI 基础设施，影响程度低。
- 风险标记：CI 镜像标签漂移风险，crane 命令语义变化，nightly 调度时间变更，硬编码 PYTHONPATH，依赖新镜像内置包

关联脉络

- PR #7338 [ci] chore: Update npu docker image cann version: 同为 Ascend CI 镜像体系更新，将 CANN 升级到 9.1.0，与本 PR 的镜像标签切换同属镜像治理
- PR #7345 [ci] chore: Fix npu nightly ci: 直接修改 nightly_ascend.yml，与本 PR 中的 nightly 调度调整相关
- PR #7311 [ci] chore: Fix docker image upload: 修复 docker-build-ascend-a2/a3 的镜像上传，与本 PR 的 crane 命令修正同文件
- PR #7305 [ci] chore: Update npu docker build file: 调整同一批 docker-build-ascend 工作流的构建细节
- PR #7282 [ci] chore: Update npu ci docker image: 更新 Ascend CI 镜像安装方式（--no-deps），与本 PR 的镜像与依赖调整同脉络