

PR #7238 完整报告

verl-project/verl

[doc] feat: Update vllm/vllm-ascend and torch version

合并时间: 2026-08-04 15:46

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7238>

执行摘要

- 一句话: vLLM/vLLM-Ascend 升到 0.23.0, 同步 NPU 文档、镜像与 CI 配置
- 推荐动作: 值得快速扫读而非精读: 没有算法或核心逻辑变化, 核心价值在于版本配套表和 CI tag 对应关系。建议维护 NPU 环境的同学核对: install_guidance.rst 与 install_guidance_A5.rst 版本表一致性、夜间 CI 镜像 tag 是否已推送、是否需要在后续 PR 统一 NPU 侧与 GPU 侧的 vLLM 大版本。

功能与动机

PR body 仅说明 'update vllm/vllm-ascend torch and transformer version in npu install docs and dockerfile'。其动机是跟随 vLLM-Ascend 上游 0.23.0 版本, 使 NPU 安装文档、镜像构建和 CI 环境保持同一版本基线; A5 文档还将 vLLM-Ascend 从固定 commit 改为 release tag, 降低手工 cherry-pick 的维护成本。

实现拆解

1. 升级安装脚本 scripts/install_vllm_mcore_npu.sh: 将 vLLM 克隆分支从 v0.18.0 改为 v0.23.0, vLLM-Ascend 从 releases/v0.18.0 改为 releases/v0.23.0, 同时将 torchvision 从 0.24.0 升到 0.25.0、torchaudio 从 2.9.0 升到 2.10.0、transformers 从 5.5.3 升到 5.10.4 (脚本内两处同步)。这是新版本组合可复现安装的依据。
2. 更新两个 Ascend 基础镜像 Dockerfile: docker/ascend/Dockerfile.ascend_9.0.0_a2 与 _a3 在克隆依赖时把 vLLM / vLLM-Ascend 分别指向 v0.23.0 与 releases/v0.23.0, 并在 x86_64 分支下将 torch 2.9.0+cpu / torchvision 0.24.0+cpu 升为 2.10.0+cpu / 0.25.0+cpu; Megatron-LM 与 MindSpeed 仍锁定 core_r0.16.0, 推理引擎与训练框架版本互相独立。
3. 同步两份安装文档: docs/ascend_tutorial/get_start/install_guidance.rst 追加 2026/08/03 关键更新日志并刷新完整依赖表 (torch 2.10.0 / torch_npu 2.10.0.post2 / vLLM 0.23.0 等); install_guidance_A5.rst 将 vLLM-Ascend 安装从固定 commit + cherry-pick 简化为 git checkout releases/v0.23.0, vLLM 也同步升到 0.23.0。
4. 切换夜间 CI 镜像引用: .github/workflows/nightly_ascend.yml 中约 8 处 job 的 container.image 从带 CANN/torch_npu 版本后缀的旧 tag 统一改为简化 tag (如 latest-vllm-910b-ubuntu、latest-sglang-a3-ubuntu), 与 PR #7201 的镜像重命名保持一致, 避免 CI 继续拉取旧镜像。

5. 测试与提交演进：本 PR 未新增任何单元或端到端测试，验证完全依赖 Ascend 夜间 CI；6 个 commit 中曾有一次 ruff format 误伤后 revert，最终以 format rst file 收尾，说明格式化工具与文档文件的边界问题值得后续关注。

关键文件：

- .github/workflows/nightly_ascend.yml（模块 CI 流水线；类别 infra；类型 infrastructure）：Ascend 夜间 CI 全部 job 的容器镜像引用切换到简化 tag，是本次版本升级后 CI 能否跑通的关键配套，直接与 PR #7201 的镜像重命名联动。
- scripts/install_vllm_mcore_npu.sh（模块 安装脚本；类别 infra；类型 deps-upgrade）：NPU 手动安装脚本，集中体现 vLLM/vLLM-Ascend 0.23.0 与 torch 2.10.0、torchvision 0.25.0、torchaudio 2.10.0、transformers 5.10.4 的版本配套关系，是文档宣称版本可复现安装的依据。
- docker/ascend/Dockerfile.ascend_9.0.0_a2（模块 镜像构建；类别 infra；类型 infrastructure）：Ascend A2 基础镜像构建文件，升级 vLLM/vLLM-Ascend 源码分支到 0.23.0，并把 x86_64 下的 torch/torchvision 升级到 2.10.0/0.25.0；新镜像将作为 CI 新 tag 的来源。
- docker/ascend/Dockerfile.ascend_9.0.0_a3（模块 镜像构建；类别 infra；类型 infrastructure）：与 a2 Dockerfile 对应，面向 A3 机型镜像做相同的依赖版本升级，保证不同硬件型号的镜像版本一致。
- docs/ascend_tutorial/get_start/install_guidance.rst（模块 安装文档；类别 docs；类型 documentation）：Ascend 主安装文档，以依赖版本表的形式给用户最直接的版本指引，并将「关键更新」日志推进到 2026/08/03。
- docs/ascend_tutorial/get_start/install_guidance_A5.rst（模块 安装文档；类别 docs；类型 documentation）：A5 机型安装指南，将 vLLM-Ascend 从固定 commit + cherry-pick 改为 release 分支 releases/v0.23.0，显著简化安装流程。

关键符号：未识别

关键源码片段

scripts/install_vllm_mcore_npu.sh

NPU 手动安装脚本，集中体现 vLLM/vLLM-Ascend 0.23.0 与 torch 2.10.0、torchvision 0.25.0、torchaudio 2.10.0、transformers 5.10.4 的版本配套关系，是文档宣称版本可复现安装的依据。

```
# NPU 环境安装脚本关键片段（vLLM-Ascend 0.23.0 配套）
# 版本配套关系：vLLM-Ascend 0.23.0 需要 torch 2.10.0 / torchvision 0.25.0 / torchaudio 2.10.0
echo "1. install basic packages"
pip uninstall -y triton triton-ascend
# 基础包版本与 vLLM-Ascend 0.23.0 对齐
pip install torchvision==0.25.0
pip install torchaudio==2.10.0
pip install triton-ascend==3.2.1 --extra-index-url https://triton-ascend.osinfra.cn/pypi/simple/ --
trusted-host triton-ascend.osinfra.cn
pip install "transformers==5.10.4"
```

```
pip install setuptools-scm
```

```
echo "2. install vllm & vllm-ascend"
```

```
# vllm 使用 v0.23.0 分支, vllm-ascend 使用 releases/v0.23.0 分支, 确保版本严格对应
```

```
git clone --depth 1 --branch v0.23.0 https://github.com/vllm-project/vllm.git
```

```
cd vllm
```

```
VLLM_TARGET_DEVICE=empty pip install -v -e .
```

```
cd ..
```

```
git clone -b releases/v0.23.0 https://github.com/vllm-project/vllm-ascend.git
```

```
cd vllm-ascend
```

```
git submodule update --init --recursive
```

```
pip install -v -e . --no-build-isolation --extra-index-url https://triton-ascend.osinfra.cn/pypi/simple/  
--trusted-host triton-ascend.osinfra.cn
```

```
echo "5. May need to check other necessary packages"
```

```
# transformers 与 xgrammar 需要额外确认版本, transformers 与上文保持一致
```

```
pip install transformers==5.10.4 xgrammar==0.1.33
```

docker/ascend/Dockerfile.ascend_9.0.0_a2

Ascend A2 基础镜像构建文件, 升级 vLLM/vLLM-Ascend 源码分支到 0.23.0, 并把 x86_64 下的 torch/torchvision 升级到 2.10.0/0.25.0; 新镜像将作为 CI 新 tag 的来源。

```
# Ascend A2 镜像关键版本固定逻辑: 推理引擎锁 vLLM / vLLM-Ascend 0.23.0
```

```
# 训练框架仍锁定 Megatron-LM / MindSpeed core_r0.16.0, 与推理引擎版本相互独立
```

```
RUN git clone --depth 1 --branch v0.23.0 https://github.com/vllm-project/vllm.git && \
```

```
git clone -b releases/v0.23.0 https://github.com/vllm-project/vllm-ascend.git && \
```

```
git clone https://gitcode.com/Ascend/MindSpeed.git && \
```

```
cd MindSpeed && git checkout core_r0.16.0 && cd .. && \
```

```
git clone --depth 1 --branch core_r0.16.0 https://github.com/NVIDIA/Megatron-LM.git && \
```

```
cd Megatron-LM && git checkout core_r0.16.0 && cd ..
```

```
# x86_64 架构下安装 CPU 版 torch 2.10.0 与 torchvision 0.25.0, 与 NPU 侧版本对齐
```

```
RUN export PIP_EXTRA_INDEX_URL=https://triton-ascend.osinfra.cn/pypi/simple/ && \
```

```
cd vllm && pip install -r requirements/build.txt && \
```

```
VLLM_TARGET_DEVICE=empty pip install -v -e. && cd .. && \
```

```
if [ "$ARCH" = "x86_64" ]; then \
```

```
pip install torch==2.10.0+cpu --index-url https://download.pytorch.org/whl/cpu/; \
```

```
pip install torchvision==0.25.0+cpu --index-url https://download.pytorch.org/whl/cpu/; \
```

```
fi && \
```

```
cd vllm-ascend && pip install -r requirements.txt
```

评论区精华

本 PR 没有任何 review 评论与讨论线程, wucong25 直接批准 (APPROVED), 说明变更被视为低风险的常规版本维护。提交历史中出现过一次 'style: apply ruff format fixes' 后紧跟 'revert format', 说明格式化工具曾误伤文档 / 脚本文件, 最终由 'format rst file' 提交统一收尾。

- 暂无高价值评论线程

风险与影响

- 风险:

1. 依赖兼容风险: vLLM 0.23.0 相对 0.18.0 属于大版本跳跃, verl 与 vLLM-Ascend 的交互路径 (如 vllm_rollout、delta 权重同步) 可能存在未覆盖的兼容问题; 本 PR 无新增测试, 风险敞口完全落在 .github/workflows/nightly_ascend.yml 对应的 e2e job 上。
2. CI 镜像强耦合: nightly_ascend.yml 引用的简化 tag (如 latest-vllm-910b-ubuntu) 依赖新镜像已成功构建并推送至 SWR 仓库; 若 Dockerfile 构建产物命名与 tag 不一致或推送延迟, CI 会立即失败。
3. 版本矩阵分叉: GPU 侧已随 PR #7101 升到 vLLM 0.24.0 / Megatron core_v0.18.0, NPU 侧停在 vLLM 0.23.0 / Megatron core_r0.16.0, 后续跨端复用代码时需留意依赖差异。
4. 文档内部不一致: install_guidance_A5.rst 的 transformers 为 4.57.6, 而主文档为 5.10.4, 两套版本表并存可能对用户产生误导。
5. 无自动化测试覆盖本次版本组合, 安装脚本与 Dockerfile 的可复现性依赖人工验证。 - 影响: 对用户: Ascend/NPU 用户按新文档安装将获得 vLLM-Ascend 0.23.0 + torch 2.10.0 组合, 已有环境不受自动影响。对系统: Ascend 夜间 CI 将切换到新镜像 tag, CI 结果与镜像发布流程强绑定。对团队: 文档维护者需关注两份文档版本表同步; 后续 NPU 相关 PR 需以 0.23.0 为基线复现。影响程度中低, 不涉及 verl 运行时源码, 不影响 GPU 用户。 - 风险标记: 依赖大版本升级, CI 镜像 tag 强耦合, 无新增自动化测试, NPU/GPU 版本矩阵分叉

关联脉络

- PR #7201 [doc] refactor: rename latest ascend docker name: 本 PR 的 nightly_ascend.yml 正是切换到其定义的简化镜像 tag, 两者构成镜像重命名到引用切换的完整链路。
- PR #7176 [ci] fix: update Ascend nightly CI & docker: 同为 Ascend CI 与镜像维护线, 本 PR 是其在 vLLM 0.23.0 时代的延续。
- PR #7101 [docker] feat: upgrade vllm and megatron version, add packages to support DeepSeek-V4: GPU 侧 vLLM 0.24.0 升级, 与本 PR 的 NPU 侧 0.23.0 升级互为对照, 构成跨硬件后端的版本演进矩阵。
- PR #7190 [vllm] refactor: drop support for vLLM older than 0.18.0: 仓库将最低 vLLM 支持版本抬到 0.18.0, 本 PR 进一步把 NPU 侧推到 0.23.0, 符合持续收紧 vLLM 版本区间的策略。