PR #6222 完整报告

verl-project/verl

[docker] feat: bump aarch64 vllm 0.17->0.18

合并时间：2026-05-08 11:42

原文链接：http://prhub.com.cn/verl-project/verl/pull/6222

执行摘要

- 一句话：升级 aarch64 vLLM 至 0.18.0，简化 Dockerfile。
- 推荐动作：该 PR 为常规基础设施升级和测试修复，影响范围有限。值得关注的是测试中 Ray 初始化模式的改进，可作为团队测试编写的参考。建议尽快合并以解除 aarch64 的 vLLM 版本锁定。

功能与动机

aarch64 vLLM 0.17.0 因性能问题被锁定，现该问题已解决，可升级至 0.18.0。同时简化 Dockerfile 中的条件逻辑，确保构建一致。修复测试中的 Ray 初始化问题，避免跨测试污染和在单 GPU 机器上挂起。

实现拆解

1. Dockerfile 升级与简化：在 docker/Dockerfile.stable.vllm 中将 aarch64 的 vLLM 从 0.17.0 升级到 0.18.0，移除之前的 TODO 注释。同时取消 transformers 安装的架构判断，统一安装 5.3.0 版本。
2. 测试中 Ray 初始化修复：在多个测试文件（如 test_colocated_workers.py, test_nested_worker.py, test_worker_group_basics.py, test_special_server_adapter.py）中添加 ray.shutdown() 调用，确保每个测试开始时关闭之前的 Ray 实例，避免资源冲突。对于需要 GPU 的测试，显式传递 num_gpus 参数以正确注册 GPU 资源。
3. 测试 GPU 资源适配：在 test_special_server_adapter.py 中，将硬编码的 n_gpus_per_node=4 改为动态计算 max(2, torch.cuda.device_count() - 2)，以适应不同 GPU 数量的环境。
4. 回滚与审校：后续提交 (f06028f) 回滚了部分测试变更，但保留了 Dockerfile 的升级和 test_special_server_adapter.py 中的 address="auto" 修复。

关键文件：

- docker/Dockerfile.stable.vllm（模块 部署脚本；类别 infra；类型 infrastructure）：核心变更文件，升级 aarch64 vLLM 至 0.18.0 并简化 transformers 安装条件。
- tests/checkpoint_engine/test_special_server_adapter.py（模块 测试；类别 test；类型 test-coverage；符号 init_config, test_server_adapter）：修复了 Ray 初始化问题和 GPU 资源动态计算，确保测试兼容不同 GPU 数量的环境。

- tests/single_controller/test_colocated_workers.py（模块 测试；类别 test；类型 test-coverage；符号 test_colocated_workers）：演示了测试隔离的修复模式：添加 ray.shutdown() 避免跨测试污染。

关键符号：未识别

关键源码片段

docker/Dockerfile.stable.vllm

核心变更文件，升级 aarch64 vLLM 至 0.18.0 并简化 transformers 安装条件。

```
# vllm: x86_64=0.18.0, aarch64=0.18.0
FROM nvidia/cuda:12.9.1-devel-ubuntu24.04

# ... 前置构建步骤 ...

# 移除之前的 TODO 注释，统一使用 vLLM 0.18.0
RUN if [ "$(uname -m)" = "aarch64" ]; then apt-get remove -y python3-jwt; fi && \
    pip install vllm==0.18.0

# 取消 transformers 的架构判断，统一版本 5.3.0
RUN pip install transformers==5.3.0
```

tests/single_controller/test_colocated_workers.py

演示了测试隔离的修复模式：添加 ray.shutdown() 避免跨测试污染。

```
# 初始修复（最终被回滚）
ray.shutdown()
ray.init(num_gpus=torch.cuda.device_count())

# 最终版本（仅添加 shutdown）
ray.shutdown()
ray.init()
```

评论区精华

1. GPU 资源分配争议：gemini-code-assist[bot] 指出 max(2, torch.cuda.device_count() - 2) 在 1-GPU 机器上会请求 2 块 GPU 导致挂起，建议用 min 限制。但 wuxibin89 认为 CI 机器是 8-GPU，应改为 //2。最终未采纳 bot 建议，保留了原始逻辑。
2. Ray 初始化争议：gemini-code-assist[bot] 建议在 test_special_server_adapter.py 的 ray.init(address='auto') 前添加 ray.shutdown()，但实际提交中未添加，而是通过 ray.shutdown() + ray.init(num_gpus=...) 方式处理。wuxibin89 指出某些测试也被 NPU 使用，应让 Ray 自动检测设备，作者 kaixih 同意并移除了手动设置。
3. 测试隔离方案：作者通过添加 ray.shutdown() 来解决跨测试污染问题，避免了在 ray.init() 中手动指定 GPU 资源的需要。

- GPU 资源计算 1-GPU 挂起风险 (correctness): 保留了原始逻辑，未修改。风险在 CI 环境下较低。

- Ray 初始化测试隔离 (testing): 作者通过添加 ray.shutdown() 解决了 hang 问题，但最终未在 test_special_server_adapter.py 中添加 shutdown，而是添加了 address='auto'。
- 手动指定 GPU 资源 vs Ray 自动检测 (design): 移除手动 num_gpus 设置，仅保留 ray.shutdown()。

风险与影响

- 风险：
  1. Dockerfile 风险低：仅涉及版本升级和条件分支简化，已被验证通过多级测试。
  2. 测试风险中等：部分测试 GPU 资源计算可能不适用于单 GPU 环境（如 max(2, ...) 逻辑），但 CI 环境为多 GPU，影响有限。此外，ray.shutdown() 的添加解决了潜在 hang 问题但可能影响测试顺序。
  3. 无安全风险。
- 影响：
  1. 对用户：构建镜像时 aarch64 架构将使用 vLLM 0.18.0，享受新特性与性能改进。
  2. 对系统：Dockerfile 更简洁，维护成本降低。
  3. 对团队：测试更加健壮，减少因 Ray 初始化问题导致的随机失败，提升 CI 稳定性。 - 风险标记：测试 GPU 资源计算可能不兼容单 GPU 环境，测试变更涉及多处 Ray 初始化调整

关联脉络

- PR #6262 [ci] chore: bump trtllm to 1.3.0rc14 and pin mbridge: 类似的 Dockerfile 版本升级 PR，反映了团队在持续更新基础设施依赖。
- PR #5631 [rollout] feat: enable Async RL for trtllm rollout: vLLM 是 rollout 引擎的关键依赖，版本升级可能影响 rollout 功能。