

PR #6043 完整报告

verl-project/verl

[ci] chore: change some npu ci test yml machine

合并时间: 2026-04-21 17:21

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6043>

执行摘要

- 一句话: 迁移 NPU CI 测试至 A3 机型并增强设备兼容
- 推荐动作: 值得 PR 审核者重点审阅 device.py 的容错设计, 其余为常规基础设施搬迁移。建议 merge 后观察 CI 运行状态。

功能与动机

PR body 指出 'change npu ci machine', 实际驱动力是昇腾 NPU CI 集群从 A2B3 升级至 A3, 要求同步更新 CI 工作流的运行环境和镜像。同时, K8s 环境下的设备编号不一致和 A3 的 one-card-two-die 架构导致 npu-smi 查询失败, 需要增强容错。

实现拆解

步骤

1. CI 工作流机型切换: 修改 .github/workflows/ 下 5 个 Ascend 相关 YAML 文件的 runs-on 字段, 将 linux-aarch64-a2b3-8 替换为 linux-aarch64-a3-8 或 a3-16, 并将容器镜像标签从 verl-8.5.0-910b 更新为 verl-8.5.0-a3, 确保 CI 运行在新集群上。
2. NPU 版本查询容错: 在 verl/utils/device.py 的 get_npu_versions() 函数中, 将原来直接调用 npu-smi info -t board -i 1 的代码改为多层 try-except: 先尝试默认 ID 1, 失败后读取 ASCEND_VISIBLE_DEVICES 获取首个设备 ID, 再失败则推测物理卡号 (npu_id // 2) 以适配 one-card-two-die 架构。
3. 移除过时调用: 在 verl/experimental/one_step_off_policy/ray_trainer.py 的 fit_step 方法中删除 self._fit_torch_memory(), 该函数在新机器上可能不必要或不存在。
4. 配套改动: 以上变更均无新增测试文件, 依赖于已有的 CI 流水线验证。

关键文件:

- verl/utils/device.py (模块 设备工具; 类别 source; 类型 core-logic; 符号 get_npu_versions): 核心设备工具模块, 增强 get_npu_versions 对 K8s 和 one-card-two-die 的容错, 是 NPU 兼容性的关键改进。
- .github/workflows/e2e_ascend.yml (模块 CI 配置; 类别 infra; 类型 infrastructure): 主 Ascend RL 训练 CI 工作流, 变更机器和镜像, 影响最大。
- .github/workflows/e2e_fully_async_policy_ascend.yml (模块 CI 配置; 类别 infra; 类型 infrastructure): fully_async_policy 的 NPU CI 工作流, 同步迁移机器和镜像。

- `.github/workflows/e2e_one_step_off_policy_ascend.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : `one_step_off_policy` 的 NPU CI 工作流, 同步迁移机器和镜像。
- `.github/workflows/reward_model_vllm_ascend.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : `reward model` 的 NPU CI 工作流, 更新镜像标签 (无机器变更)。
- `.github/workflows/vllm_ascend.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : `vllm` 的 NPU CI 工作流, 更新镜像标签。
- `verl/experimental/one_step_off_policy/ray_trainer.py` (模块 离策略训练器; 类别 source; 类型 core-logic; 符号 `fit_step`) : 移除过时的 `torch_memory` 调用, 避免在新机器上出错。

关键符号: `get_npu_versions`, `fit_step`

关键源码片段

`verl/utils/device.py`

核心设备工具模块, 增强 `get_npu_versions` 对 K8s 和 `one-card-two-die` 的容错, 是 NPU 兼容性的关键改进。

```
def get_npu_versions() -> tuple[str, str]:
    """Get the software version and CANN toolkit version for NPU devices.
    Returns: (software_version, cann_version)
    Raises: RuntimeError if version cannot be determined.
    """
    # 尝试查询第一个 NPU 设备 (ID=1), 常见于非容器环境
    try:
        result = subprocess.run(
            ["npu-smi", "info", "-t", "board", "-i", "1"],
            capture_output=True, text=True, check=True
        )
    except subprocess.CalledProcessError:
        # 设备 ID 1 不存在, 常见于 K8s 容器中设备编号不连续
        visible_devices = os.environ.get("ASCEND_VISIBLE_DEVICES")
        if not visible_devices:
            raise # 无环境变量则重新抛出原始错误
        try:
            npu_id = int(visible_devices.split(",")[0])
        except (ValueError, IndexError):
            raise # 环境变量格式无效则重新抛出
        # 使用第一个可见设备 ID 重试
        try:
            result = subprocess.run(
                ["npu-smi", "info", "-t", "board", "-i", str(npu_id)],
                capture_output=True, text=True, check=True
            )
        except subprocess.CalledProcessError:
            # A3 机器上一卡双芯, 设备 ID 是 die 索引, 物理卡号应为 npu_id // 2
            physical_card_id = npu_id // 2
            result = subprocess.run(
```

```

        ["npu-smi", "info", "-t", "board", "-i", str(physical_card_id)],
        capture_output=True, text=True, check=True,
    )
# 从 npu-smi 输出中解析 Software Version
software_version = None
for line in result.stdout.split("\n"):
    if "Software Version" in line:
        parts = line.split(":")
        if len(parts) > 1:
            software_version = parts[1].strip().lower()
            break
if not software_version:
    raise RuntimeError("Could not find Software Version in npu-smi output")
# 后续解析 CANN 版本（与原始逻辑相同，此处省略）
...

```

评论区精华

无，PR 被 wucong25 直接批准，未产生 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：

1. CI 中断风险：新 A3 机器和镜像可能与环境不完全兼容，若镜像配置有误，所有 Ascend CI 会失败，影响团队合并流程。
2. 设备兼容逻辑缺陷：device.py 中 npu_id // 2 的退避假设可能只对特定 A3 配置有效，若后续机器拓扑变化，可能误判物理卡号。
3. 回归风险低：ray_trainer.py 移除一行不影响核心训练流程，风险极小。- 影响：用户：无直接影响，该 PR 仅修改 CI 和内部工具。系统：所有 5 条 Ascend CI 流水线将运行于 A3 机器，性能可能略有变化。团队：需确认新 CI 能通过，否则需回滚。设备兼容逻辑提升在 K8s 和 A3 上的鲁棒性。

- 风险标记：CI 环境变更，K8s 设备编号兼容，新硬件适配

关联脉络

- PR #6020 [ci] fix: engine_mindspeed_llm_rl_job switch to A3.: 同为 Ascend CI 机器迁移至 A3 的变更，与本 PR 构成同一基础设施演进。