

PR #7609 完整报告

verl-project/verl

[ci] fix: pin ROCm e2e CI containers to a single NUMA node

合并时间: 2026-08-31 10:01

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7609>

执行摘要

- 一句话: ROCm CI 容器固定 NUMA 节点, 解决 e2e 超时
- 推荐动作: 值得快速阅读, 但不需要精读。它展示了一个很有价值的运维经验: 在虚拟化环境 (QEMU VM) 下, 容器自由调度导致的跨 NUMA 启动开销可能被误判为计算性能问题; `--cpuset-cpus/--cpuset-mems` 组合是恢复 NUMA 局部性的标准手段。对于负责 CI 基础设施的工程师, 建议关注: 1) NUMA 拓扑感知的容器调度实践; 2) YAML folded scalar 中反斜杠陷阱 (作者在 PR body 中特别提醒); 3) 这类硬编码硬件假设应记录到 workflow 注释或文档, 以便环境迁移时更新。

功能与动机

ROCm MI300 e2e CI 任务持续超时。PR body 指出: ROCm runners 位于 rocm7.14 QEMU VM 中, 暴露 160 vCPU 分布在两个 NUMA 节点 (`node0 = CPUs 0-79`, `node1 = CPUs 80-159`)。容器可在全部 CPU 上自由调度时, 主机侧 launch 路径严重退化: 进程和线程引导在节点间扇出, 跨节点流量主导任务 wall clock, 任务大部分时间消耗在启动而非工作负载上, 最终命中 60 分钟超时。PR 明确这是主机侧 / 虚拟化问题而非 verl 问题, 训练脚本本身不受影响。

实现拆解

本 PR 是纯 CI 配置修复, 变更集中在单个 workflow 文件, 实现步骤如下:

1. 修改目标文件: `.github/workflows/e2e_ppo_trainer_megatron_vllm_rocm.yml`, 涉及两个任务 `e2e_ppo_trainer_fsdp_vllm_rocm` 和 `e2e_ppo_trainer_megatron-moe-expert-parallel_rocm`。
2. 添加容器 NUMA 固定选项: 在两个任务的 `container.options` 中追加 `--cpuset-cpus 0-31` (限制容器使用 NUMA node 0 的 32 个 CPU) 和 `--cpuset-mems 0` (将内存分配固定到同一节点)。PR body 特别说明 options 是 YAML folded scalar (`>-`), 因此每个选项单独一行、不加尾部反斜杠——若加 `\` 会在折叠后以字面参数传给 `docker create`。
3. 保留其他配置不变: GPU 数量、并行度配置、超时均未修改, 仅动容器调度约束。
4. 验证方式: 无新增测试 (CI 配置本身就是测试)。作者在应用前已在同一 rocm7.14 MI300 VM 上验证 CPU 固定方案有效。

关键文件:

- `.github/workflows/e2e_ppo_trainer_megatron_vllm_rocm.yml` (模块 CI 工作流; 类别 infra; 类型 infrastructure): 唯一变更文件, 为 ROCm e2e 两个任务容器添加 NUMA 固定选项, 解决 CI 超时回归。

关键符号: 未识别

评论区精华

该 PR 无 review 评论和讨论线程 (`review_comments_count = 0`, `comments_count = 0`), 唯一审核由 wuxibin89 给出 APPROVED 且无文字说明。因此没有可提炼的争议或权衡交锋; PR body 中作者自己强调了 YAML folded scalar 的换行细节 (不能写尾部反斜杠), 这是唯一值得注意的实现注意事项。

- 暂无高价值评论线程

风险与影响

- 风险: 技术风险如下:
 - CI 环境依赖风险: 修复依赖特定硬件拓扑假设 (160 vCPU 两 NUMA 节点的 QEMU VM)。若未来 ROCm runner 迁移到不同拓扑 (如单 NUMA 或不同 CPU 分布), `--cpuset-cpus 0-31` 的硬编码可能不适用甚至无效, 需要随环境变更同步更新。
 - 性能容量降低: 容器可调度 CPU 从 160 降至 32, 若未来训练脚本或 vLLM 工作负载需要更多 CPU 资源 (如更大并行度或更多数据加载线程), 可能成为瓶颈。
 - 回归风险极低: 仅修改容器调度参数, 不触碰 `verl` 代码路径、GPU 数量或超时配置, 且改动经作者在真实 VM 上验证。
 - 无测试覆盖: PR 未添加自动化验证, NUMA 固定效果依赖人工确认; 这是 CI-only 变更的固有局限。
- 影响:
 - 对 CI 稳定性: 直接影响 ROCm e2e 两个任务 (FSDP vLLM 与 Megatron MoE 并行) 的超时问题, 预期从频繁超时恢复为稳定通过, 属于正面影响。
 - 对用户 / 系统: 无生产代码影响, 不涉及 `verl` 训练路径、API 或配置。
 - 对团队: 减少了 ROCm 相关 CI 噪音, 降低维护者排查超时的时间成本。
 - 影响范围: 单一 workflow 文件、两个 job 容器, 影响面小且明确。
 - 风险标记: 硬编码硬件拓扑假设, 缺少自动化测试覆盖, 可用 CPU 容量降低

关联脉络

- PR #7629 [ci] chore: fix ci failure: 同为 CI 稳定性修复, 处理 E2E 超时与不稳定测试, 反映近期 CI 维护主线。
- PR #7585 [ci] chore: Remove Ascend CI: 同类 CI 基础设施调整 (临时停用不稳定 job), 展示跨硬件 CI 维护策略。
- PR #7584 [ci] fix: drop stale `enable_chunked_prefill=False` from Ascend NPU scripts: 同为硬件相关 CI 配置清理, 体现 CI 配置与硬件环境匹配的持续维护。