

PR #1950 完整报告

THUDM/slime

fix: drop incorrect critic GPU add to rollout_num_gpus in colocate mode

合并时间: 2026-05-30 11:59

原文链接: <http://prhub.com.cn/THUDM/slime/pull/1950>

执行摘要

- 一句话: 修复 colocate 模式下 critic GPU 被错误计入 rollout 资源
- 推荐动作: 该 PR 是推荐精读的 bugfix 样例: 仅 2 行代码的删除, 但背后有完整的根因分析、历史追溯和测试验证。值得学习的是如何将运行时错误追溯到资源计算逻辑的不一致性, 并给出根治方案而非继续加防御性检查。建议合并。

功能与动机

修复 Issue #1896 中报告的 IndexError: 在 colocate 模式下使用 critic (如 ppo 优势估计器) 时, rollout_num_gpus 被错误地累加了 critic 的 GPU 数量, 导致 ServerGroup.start_engines 中索引 reordered_gpu_ids 越界。PR #1934 仅增加了边界检查将静默错误转为显式错误, 但未解决根本的资源计算问题。本 PR 直接从根源修正资源数学。

实现拆解

1. 定位错误代码: 在 slime/utils/arguments.py 的 slime_validate_args 函数中, colocate 分支 (约第 1827-1828 行) 存在 if args.use_critic: args.rollout_num_gpus += args.critic_num_gpus_per_node * args.critic_num_nodes 两行代码。
2. 删除错误累加: 直接移除这两行代码。因为 colocate 模式下, critic 复用 actor 的 placement group (见 slime/ray/placement_group.py: create_placement_group 的 colocate 分支), 并不贡献额外的 GPU 插槽, 所以 rollout_num_gpus 应该只等于 actor 的 GPU 数量。
3. 测试验证: 使用 tests/test_qwen2.5_0.5B_ppo_critic_only_short.py (同时设置 --colocate 和 --use_critic) 在 4-GPU H200 节点上验证: 修改前重现 IndexError, 修改后完整通过 2 步 RL 循环。同时验证非 critic 的 colocate 测试 tests/test_qwen2.5_0.5B_short.py 无回归。

关键文件:

- slime/utils/arguments.py (模块 参数校验; 类别 source; 类型 core-logic): 核心修复文件: 删除了 colocate 分支下错误地将 critic GPU 数量累加到 rollout_num_gpus 的两行代码, 从根本上修复了资源计算不一致的 bug。

关键符号: 未识别

关键源码片段

slime/utils/arguments.py

核心修复文件：删除了 colocate 分支下错误地将 critic GPU 数量累加到 rollout_num_gpus 的两行代码，从根本上修复了资源计算不一致的 bug。

```
# slime/utils/arguments.py
# colocate 模式下, critic 复用 actor 的 placement group, 不贡献额外 GPU 插槽
# 因此 rollout_num_gpus 应等于 actor 的 GPU 总数, 不应累加 critic 的 GPU
if args.colocate:
    if args.offload_train is None:
        args.offload_train = True
    if args.offload_rollout is None:
        args.offload_rollout = True
    if args.rollout_num_gpus != args.actor_num_gpus_per_node * args.actor_num_nodes:
        logger.info(
            f"rollout_num_gpus {args.rollout_num_gpus} != actor_num_gpus_per_node {args.actor_num_gpus_per_node} "
            f"* actor_num_nodes {args.actor_num_nodes}, overriding rollout_num_gpus to match."
        )
        args.rollout_num_gpus = args.actor_num_gpus_per_node * args.actor_num_nodes
    # 以下两行已被删除 (PR #1950) :
    # if args.use_critic:
    #     args.rollout_num_gpus += args.critic_num_gpus_per_node * args.critic_num_nodes
```

评论区精华

该 PR 没有 review 评论。唯一的审核来自 zhuzilin，状态为 APPROVED，但没有留下文字评论。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 回归风险：修改仅删除两行代码，影响范围限于 colocate + use_critic 组合下的 rollout_num_gpus 计算。作者已通过两个测试用例验证了 critic 和非 critic 路径，降低了回归风险。
 2. 未完成的 CI 矩阵：PR body 中声明未运行完整的 PPO/colocate CI 矩阵，可能存在未覆盖的边缘配置。
- 影响：
 1. 用户影响：修复了 colocate 模式下使用 critic（如 ppo）时报 IndexError 的问题，用户现在可以正常使用该配置进行训练。
 2. 系统影响：仅修改了参数校验逻辑，不涉及训练流程或推理引擎改动，影响面小。
 3. 团队影响：修正了 PR #394 引入的资源计算不一致问题，clear up 了技术债。 - 风险标记：核心路径变更，未完成全量 CI 覆盖

关联脉络

- PR #394 [Fix] ppo rollout engines for distribute: 引入 bug: 在 arguments.py 中添加了错误累加 critic GPU 的逻辑, 但未同步修改 placement_group.py 为 critic 分配独立 PG 资源。
- PR #1896 [Bug] test_qwen2.5_0.5B_ppo_critic_only_short.py fails with IndexError at start_engines after #1866 (multi-role megatron config + stale sglang defaults): 本 PR 修复的 bug 的具体 Issue 报告。
- PR #1934 Add GPU placement validation before starting rollout engines: 之前的 workaround PR, 通过增加边界检查将静默 IndexError 转为显式 ValueError, 但未解决根本的资源计算问题。