

PR #1996 完整报告

radixark/miles

Drop external router and fixed router ip mode

合并时间: 2026-09-04 13:44

原文链接: <http://prhub.com.cn/radixark/miles/pull/1996>

执行摘要

- 一句话: 移除路由器外部模式和固定 IP 模式, 简化路由器启动逻辑并强制 Miles 自行启动路由器。
- 推荐动作: 本 PR 是路由器管理重构的关键一步, 建议精读。重点关注 `start_router` 接口简化和强制内部路由器的断言。注意 reviewer 提出的端口检查问题, 以及后续 PR 如何解决。

功能与动机

本 PR 是 Issue #1837 中跟踪的重构和增强的一部分, 旨在简化路由器管理逻辑, 移除遗留的外部路由器模式, 为后续更规范的路由器架构 (如 k8s-native 模式) 铺路。

实现拆解

1. 入口逻辑简化 (`miles/ray/rollout/router_manager.py`): `start_router` 函数移除了 `force_new` 参数和相关的条件分支 (当 `args.sglang_router_ip` 已设置时直接返回)。函数现在总是获取主机 IP 并分配一个可用端口, 然后启动路由器。移除了端口占用检查和相应的 `RuntimeError`。
2. 强制内部路由器 (`miles/ray/rollout/rollout_server.py`): 在 `start_rollout_servers` 函数开头添加了断言, 确保 `args.sglang_router_ip` 为 `None`, 否则抛出错误。这阻止了外部路由器模式的使用。同时移除了 `start_router` 调用中的 `force_new` 参数。
3. 测试覆盖调整:
 - 新增 `tests/fast/examples/test_example_launcher_flags.py`, 包含两个测试:
`test_example_launchers_only_emit_registered_miles_flags` (确保示例脚本中的标志已注册) 和 `test_example_launchers_do_not_request_an_external_router` (确保示例脚本不传递 `--sglang-router-ip` 或 `--sglang-router-port`)。
 - 更新 `tests/fast/ray/rollout/test_router_manager.py`: 移除了测试外部路由器返回现有值的测试 (`test_returns_existing_when_already_configured`) 和端口冲突测试 (`test_port_conflict_raises_runtime_error`), 并调整了 `patch_low_level` fixture 以返回固定的占位地址而非使用参数中的值。
 - 更新 `tests/fast/ray/rollout/real_ray/test_inference_controller.py`: 调整 `patch_low_level` fixture 中 `start_router` 的模拟, 直接返回固定占位地址 ("127.0.0.1", 30000), 并移除了测试参数中设置 `sglang_router_ip` 和 `sglang_router_port` 的代码。

4. 文档和示例更新：更新 docs/user-guide/cli-reference.md 和 docs/user-guide/training-backend.md，说明外部路由器模式已移除，并更新了 --sglang-router-port 的描述。同时，从 examples/swe-agent-harbor-docker/run.py 和 run-glm47-flash-agentic-async.py 中删除了传递 --sglang-router-port 参数的行。

关键文件：

- miles/ray/rollout/router_manager.py（模块 路由器管理；类别 source；类型 entrypoint；符号 start_router）：核心路由器启动逻辑简化，移除外部路由器模式
- miles/ray/rollout/rollout_server.py（模块 部署入口；类别 source；类型 core-logic）：添加断言强制内部路由器模式，是变更的入口点
- tests/fast/examples/test_example_launcher_flags.py（模块 示例测试；类别 test；类型 test-coverage；符号 test_example_launchers_only_emit_registered_miles_flags, test_example_launchers_do_not_request_an_external_router）：新增测试确保示例脚本不使用已弃用的路由器标志

关键符号：start_router

关键源码片段

miles/ray/rollout/rollout_server.py

添加断言强制内部路由器模式，是变更的入口点

```
async def start_rollout_servers(args, pg) -> dict[str, "RolloutServer"]:  
    """Start rollout servers: one per model, each with its own router.  
  
    Returns a dict mapping model name -> ``RolloutServer``.  
    """  
  
    # 强制内部路由器模式，外部路由器模式已移除  
    assert args.sglang_router_ip is None, (  
        "external router mode was removed: miles always starts its own routers "  
        "(expected to return with the k8s-native mode)"  
    )  
  
    config = resolve_sglang_config(args)  
    servers: dict[str, RolloutServer] = {}  
    engine_offset = 0  
    port_allocator = PortAllocator()  
  
    for model_idx, model_cfg in enumerate(config.models):  
        has_pd = model_cfg.has_pd_disaggregation  
        # 不再传递 force_new 参数，每个模型独立启动路由器  
        router_ip, router_port = start_router(args, has_pd_disaggregation=has_pd)  
        if model_idx == 0:  
            args.sglang_router_ip = router_ip  
            args.sglang_router_port = router_port  
        # ... 后续逻辑不变
```

评论区精华

在 `miles/ray/rollout/router_manager.py` 的评论中，reviewer guapisolo 指出了一个潜在问题：移除端口检查后，静态路由器端口（通过 `--sglang-router-port` 固定）可能绕过节点本地可用性检查，导致陈旧监听器干扰启动。评论建议在启动前检查端口占用（或保留并验证所有权），并更新文档以更精确地描述哪些标志被移除。此问题被标记为未解决，并被后续 PR 部分修复（如 #2051、#2053、#2096）。

- 端口检查缺失与文档准确性 (correctness): 此问题未在本 PR 完全解决，后续 PR（#2051、#2053、#2096）部分修复。

风险与影响

- 风险：
 1. 回归风险：移除外部路由器模式可能导致依赖此模式的用户无法直接升级。但该模式被视为遗留功能，文档已明确说明。
 2. 安全风险：移除端口占用检查可能增加端口冲突或陈旧进程干扰的风险。后续 PR 部分缓解了此问题，但当前 PR 未完全解决。
 3. 兼容性风险：更新了 CLI 参考文档，明确 `--sglang-router-ip` 必须保持未设置，`--sglang-router-port` 现在仅用于固定路由器启动端口。这可能会混淆依赖旧行为的用户。
 - 影响：用户影响：移除了外部路由器模式，用户不再能指定外部路由器 IP 和端口。这简化了配置，但可能影响现有部署。路由器端口固定行为变化（现在用于内部路由器）。系统影响：简化了路由器启动逻辑，减少了条件分支。强化了 Miles 自行管理路由器的模式。团队影响：为后续路由器重构（如 k8s-native 模式）清理了代码路径。
- 风险标记：端口检查缺失可能引发陈旧进程干扰，移除外部路由器模式可能影响现有部署

关联脉络

- PR #1997 Add worker specs for the sglang and miles routers: 为路由器引入 Worker Spec，是路由器重构的下一步
- PR #1999 Implement a worker spec for the session server: 为会话服务器实现工作者规范，统一启动逻辑
- PR #2050 Stop checking actor liveness while waiting for a tcp port: 移除 `wait_tcp_ready` 的 actor 存活检查参数，简化端口就绪探测