

PR #2774 完整报告

radixark/miles

feat(session): configure session server workers explicitly

合并时间: 2026-08-28 06:09

原文链接: <http://prhub.com.cn/radixark/miles/pull/2774>

执行摘要

- 一句话: 端口与并发数解耦, 新增 workers 参数
- 推荐动作: 值得精读。这是一个有清晰设计意图的破坏性 CLI 重构: 默认 32 workers 的取舍、拒绝旧语法而非兼容、TODO(#1837) 标记临时端口分配, 都体现了对“配置属主唯一”和“迁移可见性”的坚持。建议重点关注 `_resolve_session_server_ports` 的实现与 review 中未解决的连续端口占用问题, 后续很可能被 #1837 的统一端口分配重构取代。

功能与动机

PR body 明确说明: 把实例数量编码为 `--session-server-port START END` 的终点会让一个 flag 重载两个独立设置。分离起始端口与 worker 数输入, 让操作员改并发度时无需重新计算端口区间端点; 同时刻意拒绝旧的双值形式, 让起始端口和实例数量各有一个唯一属主。

实现拆解

1. CLI 参数定义变更 (`miles/utils/arguments.py`): `--session-server-port` 从 `nargs="+"` 改为单值 `type=int`、`default=None`, 语义从“端口或区间”收敛为“起始端口, 不传则自动分配”; 新增 `--session-server-workers`, 默认 32。这是默认行为变化的源头, 也是 review 争议所在。
2. 端口解析函数重写 (`miles/ray/rollout/router_manager.py`):
`_resolve_session_server_ports(start, workers)` 保留“None 时自动分配”路径, 删除 `start >= end` 空区间校验和“超过两个值报错”分支, 新增 `workers < 1` 校验并在 spawn 任何进程之前失败; 调用点 `start_session_server` 改为传入 `args.session_server_port` 与 `args.session_server_workers`。该改动直接决定并发规模, 且移除了旧端口的兼容路径。
3. 测试配套: `tests/fast/ray/rollout/test_router_manager.py` 用 `test_one_worker_uses_the_starting_port`、`test_workers_expand_from_the_starting_port`、`test_non_positive_workers_raise` 替换旧的区间用例;
`tests/fast/utils/test_arguments.py` 新增 `TestSessionServerScalingArguments`, 锁定默认 32、显式覆盖, 以及旧 `end-port` 形式被 `argparse` 拒绝。
4. 示例与文档联动: 所有仓库内 `session-server` 启动脚本显式追加 `--session-server-workers 32` (`nemo-gym`、`openenv`、`terminus-compaction`、`swe-agent-harbor-docker` 等), 并同步 README 网络图到端口 30000-30031 + router 31000, 最后通过 `sync-example-docs` 重新生成文档页, 保证 CI 漂移检查一致。

关键文件：

- miles/ray/rollout/router_manager.py（模块 会话服务；类别 source；类型 core-logic；符号 `_resolve_session_server_ports`, `start_session_server`）：核心逻辑变更：
`_resolve_session_server_ports` 从 `[start, end)` 区间语义改为 `start + workers` 展开，破坏性移除旧双值语法；是后续会话服务器并发调控的入口。
- miles/utils/arguments.py（模块 参数配置；类别 source；类型 core-logic；符号 `add_session_arguments`）：CLI 参数定义变更：`--session-server-port` 从 `nargs='+'` 改为单值，新增 `--session-server-workers` 默认 32；是默认行为变化的源头，直接命中 review 中的静默迁移争议。
- tests/fast/ray/rollout/test_router_manager.py（模块 路由管理；类别 test；类型 test-coverage；符号 `test_none_auto_allocates_one_port`, `test_one_worker_uses_the_starting_port`, `test_workers_expand_from_the_starting_port`, `test_non_positive_workers_raise`）：单元测试同步新签名，覆盖单 worker、多 worker、非正 workers 报错等边界，验证解析函数行为。
- tests/fast/utils/test_arguments.py（模块 参数配置；类别 test；类型 test-coverage；符号 `TestSessionServerScalingArguments`, `test_defaults_to_32_workers_and_an_auto_port`, `test_parses_starting_port_and_worker_count`, `test_rejects_the_removed_end_port_form`）：新增 `TestSessionServerScalingArguments` 验证默认 32 workers、显式覆盖，以及旧 end-port 形式被 `argparse` 拒绝。
- examples/experimental/nemo-gym/run.py（模块 示例脚本；类别 source；类型 example；符号 `execute`）：示例启动器显式声明 `--session-server-workers 32`，避免依赖新默认值造成静默行为变化，代表六处示例迁移的典型模式。
- examples/swe-agent-harbor-docker/run.py（模块 示例脚本；类别 infra；类型 infrastructure）：SWE-Agent 部署脚本显式声明 32 workers；连同其 README 一起验证端口区间 30000-30031 与启动器配置同步。

关键符号：`_resolve_session_server_ports`, `start_session_server`, `add_session_arguments`

关键源码片段

miles/ray/rollout/router_manager.py

核心逻辑变更：`_resolve_session_server_ports` 从 `[start, end)` 区间语义改为 `start + workers` 展开，破坏性移除旧双值语法；是后续会话服务器并发调控的入口。

```
# miles/ray/rollout/router_manager.py
def _resolve_session_server_ports(start: int | None, workers: int) -> list[int]:
    """Return the requested number of consecutive ports from the configured or auto-selected
    start."""
    if workers < 1:
        # 在生成任何进程之前失败，避免出现 0 个或“负数量”端口列表。
        raise ValueError("--session-server-workers must be at least 1.")
    # TODO(#1837): Refactor IP/port allocation; keep this naive for now.
    if start is None:
```

```

# 仅探测单个可用端口，后续连续端口不逐一检查占用（review 已指出）。
start = find_available_port(random.randint(5000, 6000))
return list(range(start, start + workers))

# miles/ray/rollout/router_manager.py —— start_session_server 中的调用点
ip = args.session_server_ip
ports = _resolve_session_server_ports(args.session_server_port, args.session_server_workers)
for port in ports:
    if not is_port_available(port):
        raise RuntimeError(
            f"Port {port} is already in use — a stale session server may still be running. "
            f"Run 'pskill -q python' to kill it, then retry."
        )
# The canonical driver-side value; rollout code picks from this list.
args.session_server_ports = ports

```

miles/utils/arguments.py

CLI 参数定义变更：--session-server-port 从 nargs='+' 改为单值，新增 --session-server-workers 默认 32；是默认行为变化的源头，直接命中 review 中的静默迁移争议。

```

# miles/utils/arguments.py —— add_session_arguments 中的参数定义
parser.add_argument(
    "--session-server-port",
    type=int,
    default=None,
    help="Starting port for standalone session servers. Auto-allocated if not set.",
)
parser.add_argument(
    "--session-server-workers",
    type=int,
    default=32, # 旧语义下单端口 = 1 个服务器，新默认 32 是刻意的并发选择。
    help="Number of standalone session servers to launch on consecutive ports.",
)

```

评论区精华

围绕默认值发生了本轮最有价值的交锋：claude[bot] 指出默认 32 会静默改变所有只传 --session-server-port 的启动脚本（旧语义下单端口 = 单实例）；guapisolo 回应 32 是刻意的，单个 session-server 进程固定开销太大，默认 1 会让操作员被迫手动选择高效配置，且所有仓库内 launcher 已显式声明 32，迁移可见。另一条未解决的是 auto-port 路径只探测 1 个端口再派生 31 个连续端口，没有校验占用，claude[bot] 建议复用 get_free_port(start_port, consecutive)，作者仅加 TODO(#1837)。Shi-Dong 的 README 同步请求已在 0cdda4e87 完成。

- 默认 32 workers 的静默行为变更 (design): 默认保持 32；仓库内启动器全部显式声明以避免静默继承，仓库外调用者视为有意的默认采纳。

- 自动端口分配未探测连续端口块 (correctness): 未解决; 当前保留朴素实现, 连续端口冲突风险由 `is_port_available` 逐端口前置检查兜底, 但仍可能全部不可用导致启动失败。
- 示例 README 与启动器端口同步 (documentation): 已同步并通过 `sync-example-docs` 校验。

风险与影响

- 风险:
 - CLI 破坏性变更: 移除 `--session-server-port A B` 旧形式, 旧脚本会 `argparse` 报错退出; 仓库内已全部迁移, 外部脚本需同步。
 - 默认并发从 1 升到 32: 省略 `--session-server-workers` 的仓库外调用方会静默拉起 32 个进程, 端口与内存资源消耗陡增; `start_session_server` 虽有逐端口 `is_port_available` 前置检查做快速失败, 但首个端口可用时仍可能因后续端口冲突在中途抛错。
 - 自动端口分配只探测单端口 (`router_manager.py` 的 `find_available_port` 分支): 派生连续端口不检查占用, 与代码库已有的 `get_free_port(start_port, consecutive)` 不一致; 该问题在 review 中提出但未闭合。
 - 32 个 spawn 子进程各自承担约 10s 的 `transformers` import, 启动墙钟时间和内存占用成倍增长。
 - 影响: 对操作员是直接的 CLI 语义变更: 需要理解 `--session-server-port` 现在是“起始端口”、并发数由 `--session-server-workers` 控制, 仓库外脚本必须迁移。对系统资源, 默认 32 个 `session-server` 进程意味着更高的常驻内存与端口占用, 小规模实验需显式调小。对团队, 所有示例与生成文档已同步, CI 有测试锁定默认值和破坏性行为, 回归风险主要在外部调用方。
 - 风险标记: CLI 破坏性变更, 默认并发 1 -> 32, `auto-port` 未校验连续端口, 仓库外脚本需迁移

关联脉络

- PR #2604 fix(rollout): raise server readiness timeout to 120s: 同为 `miles/ray/rollout/router_manager.py` 的改动, 围绕 `session-server` / `router` 启动可靠性的演进主线。
- PR #2748 fix(swe-agent): flush all session-server instances on abort: 同一批 `examples/swe-agent-harbor-docker` 会话服务器部署脚本; 本 PR 在其中显式声明 `worker` 数, 两者构成 `session-server` 生命周期与并发规模的可运维化演进。