(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2021 年 8 月 26 日 (26.08.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/164507 A1

(51) 国际专利分类号:
H04W 72/12 (2009.01)

(21) 国际申请号: PCT/CN2021/073764

(22) 国际申请日: 2021 年 1 月 26 日 (26.01.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010106750.2 2020年2月19日 (19.02.2020) CN

(71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 王坚 (WANG, Jian); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 徐晨 (XU, Chen); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 皇甫幼睿 (HUANGFU, Yourui); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 李榕 (LI, Rong); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 王俊 (WANG, Jun); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (CN)。

(54) **Title:** SCHEDULING METHOD, SCHEDULING ALGORITHM TRAINING METHOD AND RELATED SYSTEM, AND STORAGE MEDIUM

(54) **发明名称**: 调度方法、调度算法的训练方法及相关系统、存储介质

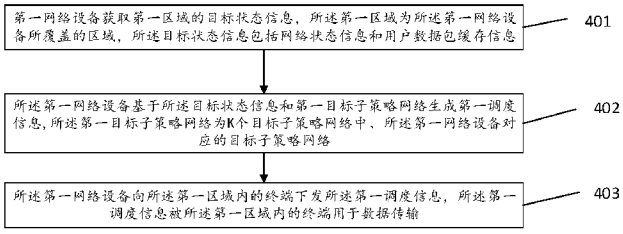

图 4

401 A first network device obtains target state information of a first area, the first area being an area covered by the first network device, and the target state information comprising network state information and user data packet cache information

402 The first network device generates first scheduling information on the basis of the target state information and a first target sub-policy network, the first target sub-policy network being a target sub-policy network corresponding to the first network device in K target sub-policy networks

403 The first network device issues the first scheduling information to a terminal in the first area, the first scheduling information being used by the terminal in the first area for data transmission

(57) **Abstract:** Embodiments of the present application provide a scheduling method, a scheduling algorithm training method and a related system, and a storage medium. The scheduling method is applied to a scheduling control system. The scheduling control system comprises K network devices, K being an integer greater than 1. The method comprises: a first network device obtains target state information of a first area, the target state information comprising network state information and user data packet cache information; the first network device generates first scheduling information on the basis of the target state information of the first area and a first target sub-policy network, the first target sub-policy network being a target sub-policy network corresponding to the first network device in K target sub-policy networks; and the first network device issues the first scheduling information to a terminal in the first area. The present solution improves the performance of the scheduling control system. Moreover, the feasibility of the scheduling control solution is improved by means of a distributed deployment policy network.

(57) **摘要**: 本申请实施例提供一种调度方法、调度算法的训练方法及相关系统、存储介质，所述调度方法应用于调度控制系统，所述调度控制系统包括K个网络设备，K为大于1的整数，所述方法包括：第一网络设备获取第一区域的目标状态信息，所述目标状态信息包括网络状态信息和用户数据包缓存信息；所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息，其中，所述第一目标子策略网络为K个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述第一网络设备向所述第一区域内的终端下发所述第一调度信息。本方案提升了调度控制系统的性能。且，通过分布式的部署策略网络，提高了调度控制方案的可行性。

WO 2021/164507 A1

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

调度方法、调度算法的训练方法及相关系统、存储介质

本申请要求于2020年2月19日提交中国专利局、申请号为202010106750.2、发明名称为“调度方法、调度算法的训练方法及相关系统、存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及通信技术领域，尤其涉及一种调度方法、调度算法的训练方法、调度控制系统、调度算法训练系统以及存储介质。

背景技术

在蜂窝网络中，媒体访问控制（Medium Access Control，MAC）层调度主要解决时频资源的分配、调制与编码策略（Modulation and Coding Scheme，MCS）选择、用户配对、预编码等问题。通过调度来实现系统吞吐和公平性的折中。

马尔可夫决策过程（MDP）是一种分析决策问题的数学模型。如图 1 所示，其假设环境具有马尔可夫性质（环境的未来状态的条件概率分布仅依赖于当前状态），决策者通过周期性地观察环境的状态，根据当前环境的状态做出决策，与环境交互后得到新的状态及奖励。

强化学习是机器学习中的一个领域，可以用于上述求解马尔科夫决策过程。如图 2 所示，强化学习强调智能体 Agent 通过和环境的交互过程，获得最大化的预期利益，学习得到最优的行为方式。智能体通过对环境的观察，得到当前状态 s，并按照一定的规则 π 决策一个动作 a 反馈给环境，环境将该动作实行后所得到的奖励 r 或惩罚反馈给智能体。通过多次的迭代训练，使智能体学会根据环境状态作出最优决策。

其中，将强化学习和深度学习相结合，就得到了深度强化学习（deep reinforcement learning，DRL），如图 3 所示。对比图 2 和图 3 可以发现，深度强化学习仍然符合强化学习中智能体和环境交互的框架。不同的是，智能体中使用深度神经网络进行决策。

为了实现在动态变化的无线传输环境中进行调度，现有技术采用深度强化学习 DRL 算法。该算法利用 DRL 中的智能体与无线传输环境的交互，不断更新其自身参数，以获得较优的决策策略。其中，智能体首先获取通信系统的当前状态，并根据此状态做出决策；执行决策后，通信系统进入下一状态，同时反馈收益。智能体根据收益情况对自身决策参数进行调整。智能体通过迭代式地与环境进行交互，不断调整自身参数以获得更大收益，最终收敛后即可得到较优的调度策略。由于现有技术采用一种中心式的调度方案，唯一的智能体负责全网所有小区/网络的决策。在多小区网络或多等级异构网络场景中，采用现有技术会导致动作空间过大，智能体所用神经网络的训练过程过慢，难以收敛。因此，在实际的系统中，部署这种中心式的调度方案可行性极低。

发明内容

本申请公开了一种调度方法、调度算法的训练方法及相关系统、存储介质，可以实现

基于多智能体的分布式的调度，提高了系统的性能。

第一方面，本申请实施例提供一种调度方法，所述方法应用于调度控制系统，所述调度控制系统包括K个网络设备，K为大于1的整数，所述方法包括：

第一网络设备获取第一区域的目标状态信息，其中，所述第一网络设备为所述K个网络设备中的任意一个，所述第一区域为所述第一网络设备所覆盖的区域，所述目标状态信息包括网络状态信息和用户数据包缓存信息；

所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息，其中，所述第一目标子策略网络为K个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述K个目标子策略网络与所述K个网络设备一一对应；

所述第一网络设备向所述第一区域内的终端下发所述第一调度信息，所述第一调度信息被所述第一区域内的终端用于数据传输。

本申请实施例基于K个网络设备中的第一网络设备通过获取第一区域的目标状态信息，然后基于目标状态信息和与该第一网络设备对应的第一目标子策略网络得到调度信息，进而向第一区域内的终端下发该调度信息，以便第一区域内的各终端根据该调度信息进行数据传输。采用该手段，其中，各个网络设备分别对应各自的策略网络进行调度控制，实现多智能体进行调度控制，提升了调度控制系统的性能。且，通过分布式的部署策略网络，提高了调度控制方案的可行性。

其中，所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述方法还包括：

所述第一网络设备对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络；

其中，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络，包括：

所述第一网络设备根据所述第一区域的目标状态信息S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，所述第一网络设备将第一子策略网络W_i确定为所述第一目标子策略网络，其中，i为正整数，所述第一区域的目标状态信息S_{i+1}为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为所述第一网络设备基于所述第一区域的目标状态信息S_i和所述第一子策略网络W_i生成的，所述目标状态信息S_i是第i次所述训练的目标状态信息；其中，当i=1时，所述第一子策略网络W_i为第一初始子策略网络。

该实施例以性能参数的值不低于预设值时停止训练。当然，本申请实施例并不限定上述条件。本申请实施例还可以以性能参数的值不高于预设值时停止训练。例如通过对上述预设值取反构成新的预设值等。

本申请还可以以迭代训练的次数达到预设次数时停止训练。或者，以更新参数的次数达到预设次数时停止训练等。

可替代的，本申请实施例还可以以策略网络对应的损失函数的值低于预设阈值时停止训练等。

其中，当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

作为另一种可选的实现方式，当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

作为又一种可选的实现方式，所述调度控制系统还包括集中式网元设备，所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述方法还包括：

所述第一网络设备接收所述集中式网元设备发送的第一目标子策略网络的参数，其中，所述 K 个目标子策略网络的参数均相同，其中，所述集中式网元设备为核心网设备或基站集中式单元 CU 设备。

第二方面，本申请实施例还提供一种调度算法的训练方法，所述方法应用于调度算法训练系统，所述调度算法训练系统包括 K 个网络设备，K 为大于 1 的整数；所述方法包括：

第一网络设备获取训练数据，其中，所述第一网络设备为所述 K 个网络设备中的任意一个；

所述第一网络设备根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络；其中，所述第一初始子策略网络为 K 个初始子策略网络中、所述第一网络设备对应的初始子策略网络；所述第一目标子策略网络为 K 个目标子策略网络中、

所述第一网络设备对应的目标子策略网络；所述 K 个初始子策略网络、所述 K 个目标子策略网络分别与所述 K 个网络设备一一对应。

其中，所述训练数据包括第一区域的目标状态信息 S_{i+1}，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述第一网络设备根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

所述第一网络设备根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，所述第一网络设备将第一子策略网络 W_i 确定为所述第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为所述第一网络设备基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次训练的价值网络得到的。

其中，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备分别对应的各子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

本申请实施例基于中心式的价值网络和分布式的策略网络构成的多智能体 MARL 架构进行训练，得到一个目标价值网络和多个分布式的目标策略网络。该分布式的目标策略网络可用于网络设备进行调度，避免了单智能体 DRL 完全中心式的调度，提高了方案可行性。

作为另一种可选的实现方式，当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对

应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

本申请实施例基于分布式的价值网络和分布式的策略网络构成的多智能体 MARL 架构进行训练，得到多个目标价值网络和多个分布式的目标策略网络。该分布式的目标策略网络可用于网络设备进行调度，避免了单智能体 DRL 完全中心式的调度，提高了方案可行性。

作为又一种可选的实现方式，所述方法还包括：

所述第一网络设备将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，当所述性能参数的值低于所述预设值时，所述第一网络设备将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；所述第一网络设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

作为再一种可选的实现方式，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述方法还包括：

所述集中式网元设备将价值网络 Q_i 确定为目标价值网络，其中，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

其中，当所述性能参数的值低于所述预设值时，所述集中式网元设备将所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理以得到 K 个评价价值，其中，所述 K 个评价价值与所述 K 个子策略网络一一对应；

所述集中式网元设备将所述 K 个评价价值分别发送至所述 K 个网络设备；

所述集中式网元设备调整所述价值网络 Q_i 中的参数，以得到用于下一次所述训练的价值网络。

作为又一种可选的实现方式，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述方法还包括：

所述集中式网元设备将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

当所述性能参数的值低于所述预设值时，所述集中式网元设备将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出

结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值；其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；

所述集中式网元设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

第三方面，本申请实施例还提供一种调度控制系统，所述调度控制系统包括 K 个网络设备，K 为大于 1 的整数，其中，第一网络设备为所述 K 个网络设备中的任意一个，所述第一网络设备用于：

获取第一区域的目标状态信息，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述目标状态信息包括网络状态信息和用户数据包缓存信息；

基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息，其中，所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述 K 个目标子策略网络与所述 K 个网络设备一一对应；

向所述第一区域内的终端下发所述第一调度信息，所述第一调度信息被所述第一区域内的终端用于数据传输。

其中，在基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述第一网络设备还用于：

对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络；

其中，对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络，具体包括：

根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，将第一子策略网络 W_i 确定为所述第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成的，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

其中，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据

传输后确定的。

作为另一种可选的实现方式，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

作为再一种可选的实现方式，所述调度控制系统还包括集中式网元设备，在基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述第一网络设备还用于：

接收所述集中式网元设备发送的第一目标子策略网络的参数，其中，所述 K 个目标子策略网络的参数均相同，其中，所述集中式网元设备为核心网设备或基站集中式单元 CU 设备。

第四方面，本申请实施例还提供一种调度算法训练系统，所述调度算法训练系统包括 K 个网络设备，K 为大于 1 的整数，第一网络设备为所述 K 个网络设备中的任意一个，所述第一网络设备用于：

获取训练数据；

根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络；其中，所述第一初始子策略网络为 K 个初始子策略网络中、所述第一网络设备对应的初始子策略网络；所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络；所述 K 个初始子策略网络、所述 K 个目标子策略网络分别与所述 K 个网络设备一一对应。

其中，所述训练数据包括第一区域的目标状态信息 S_{i+1}，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述第一网络设备具体用于：

根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，将第一子策略网络 W_i 确定为所述第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1

时，所述第一子策略网络 W_i 为第一初始子策略网络。

当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次训练的价值网络得到的。

其中，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备分别对应的各子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

作为另一种可选的实现方式，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

作为又一种可选的实现方式，所述第一网络设备还用于：

将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，当所述性能参数的值低于所述预设值时，所述第一网络设备还用于：

将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；所述第一网络设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

作为再一种可选的实现方式，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述集中式网元设备用于：

将价值网络 Q_i 确定为目标价值网络，其中，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

当所述性能参数的值低于所述预设值时，所述集中式网元设备用于：

将所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理以得到 K 个评价价值，其中，所述 K 个评价价值与所述 K 个子策略网络一一对应；

将所述 K 个评价价值分别发送至所述 K 个网络设备；

调整所述价值网络 Q_i 中的参数，以得到用于下一次所述训练的价值网络。

作为再一种可选的实现方式，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述集中式网元设备用于：

将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，当所述性能参数的值低于所述预设值时，所述集中式网元设备用于：

将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值；其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；

调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

第五方面，本申请提供了一种计算机存储介质，包括计算机指令，当所述计算机指令在电子设备上运行时，使得所述电子设备执行如第一方面任一种可能的实施方式和/或第二方面任一种可能的实施方式提供的方法。

第六方面，本申请实施例提供一种计算机程序产品，当计算机程序产品在计算机上运行时，使得计算机执行如第一方面任一种可能的实施方式和/或第二方面任一种可能的实施方式提供的方法。

可以理解地，上述提供的第三方面所述的装置、第四方面所述的装置、第五方面所述的计算机存储介质或者第六方面所述的计算机程序产品均用于执行第一方面中任一所提供的方法以及第二方面中任一所提供的方法。因此，其所能达到的有益效果可参考对应方法中的有益效果，此处不再赘述。

附图说明

下面对本申请实施例用到的附图进行介绍。

图 1 是现有技术中马尔可夫决策过程的示意图；

图 2 是现有技术中强化学习用于求解马尔科夫决策过程的示意图；

图 3 是现有技术中深度强化学习用于求解马尔科夫决策过程的示意图；

图 4 是本申请实施例提供的一种调度方法的流程示意图；

图 5 是本申请实施例提供的一种调度方法的应用场景示意图；

图 6 是本申请实施例提供的一种中心式价值网络+分布式策略网络的调度算法的训练方法的示意图；

图 7 是本申请实施例提供的一种中心式价值网络+分布式策略网络部署在多小区蜂窝网络中的场景示意图；

图 8 是本申请实施例提供的一种中心式价值网络+分布式策略网络部署在异构网络中的场景示意图；

图 9A 是本申请实施例提供的一种分布式价值网络+分布式策略网络的调度算法的训练方法的示意图；

图 9B 是本申请实施例提供的另一种分布式价值网络+分布式策略网络的调度算法的训练方法的示意图；

图 10 是本申请实施例提供的一种分布式价值网络+分布式策略网络部署在多小区蜂窝网络中的场景示意图；

图 11 是本申请实施例提供的一种分布式价值网络+分布式策略网络部署在异构网络中的场景示意图；

图 12 是本申请实施例提供的一种中心式价值网络+中心式策略网络的调度算法的训练方法的示意图；

图 13 是本申请实施例提供的一种中心式价值网络+中心式策略网络部署在多小区蜂窝网络中的场景示意图；

图 14 是本申请实施例提供的一种中心式价值网络+中心式策略网络部署在异构网络中的场景示意图。

具体实施方式

下面结合本申请实施例中的附图对本申请实施例进行描述。本申请实施例的实施方式部分使用的术语仅用于对本申请的具体实施例进行解释，而非旨在限定本申请。

参照图 4 所示，为本申请实施例提供的一种调度方法的流程示意图。其中，所述调度方法应用于调度控制系统，所述调度控制系统包括 K 个网络设备，K 为大于 1 的整数，如图 4 所示，其包括步骤 401-403，具体如下：

401、第一网络设备获取第一区域的目标状态信息，其中，所述第一网络设备为所述 K 个网络设备中的任意一个，所述第一区域为所述第一网络设备所覆盖的区域，所述目标

状态信息包括网络状态信息和用户数据包缓存信息；

其中，上述 K 个网络设备可以是 K 个基站。该基站可以是一种部署在无线接入网中为移动台（Mobile Station， MS）提供无线通信功能的装置。其中，上述基站可以为各种形式的宏基站、微基站（也称为小站）、中继站、接入点等。在采用不同的无线接入技术的系统中，具备基站功能的设备的名称可能会有所不同，例如，在 LTE 系统中，称为演进的节点 B（evolved NodeB，eNB 或者 eNodeB）；在第三代（3rd Generation，3G）系统中，称为节点 B（Node B）等。为方便描述，本申请所有实施例中，上述为 MS 提供无线通信功能的装置统称为基站。上述 MS 可以包括各种具有无线通信功能的手持设备、车载设备、可穿戴设备、计算设备或连接到无线调制解调器的其它处理设备。所述 MS 也可以称为终端（terminal）。还可以是用户单元（subscriber unit）、蜂窝电话（cellular phone）、智能手机（smart phone）、无线数据卡、个人数字助理（Personal Digital Assistant，PDA）电脑、平板型电脑、无线调制解调器（modem）、手持设备（handset）、膝上型电脑（laptop computer）、机器类型通信（Machine Type Communication, MTC）终端等。

第一网络设备可以是上述 K 个网络设备中的任意一个。如，该第一网络设备可以是基站 A，所述第一区域即为基站 A 所覆盖的区域。

上述目标状态信息可以是第一网络设备所覆盖小区内的各终端用户的状态信息。或者，对于某个基站所覆盖小区中存在一个宏站、多个微微站和家庭基站等时，所述目标状态信息还可以是上述宏站、微微站或者家庭基站中的任一个基站所覆盖范围内的注册用户状态信息。

其中，该目标状态信息包括网络状态信息和用户数据包缓存信息等。该网络状态信息包括信道状态信息、吞吐量信息和混合自动重传（Hybrid Automatic Repeat request，HARQ）信息等。上述用户数据包缓存信息包括缓存中数据包的数量、缓存中数据包的大小和缓存中数据包的时延等。

402、所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息，其中，所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述 K 个目标子策略网络与所述 K 个网络设备一一对应；

上述第一调度信息如可以是指示上述第一区域内的第一终端发送数据的方式的信息等。该发送数据的方式即发送数据时所使用的无线资源、调制编码策略、预编码策略等具体的配置。

其中，所述第一调度信息为第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成的。

可选的，上述第一网络设备可将上述第一区域的目标状态信息输入至第一目标子策略网络进行处理，并对所述第一目标子策略网络的输出结果进行处理以得到上述第一调度信息。

其中，Actor-Critic（演员-评判家）算法是一种常用的强化学习算法。如图 5 所示，采用 Actor-Critic 算法的强化学习架构中，智能体包括 Actor 和 Critic 两部分。其中，Actor

负责根据环境状态和 Critic 的输出做出决策，而 Critic 负责根据环境状态和收益来评估 Actor 做出的决策的好坏。在深度强化学习中，Actor 和 Critic 都可以采用深度神经网络来实现。此时，由于 Actor 神经网络负责做出决策，所以也叫策略网络。Critic 神经网络输出评价，也叫价值网络。

其中，上述调度控制系统包括 K 个网络设备。每个网络设备均对应一个目标子策略网络。第一网络设备对应上述第一目标子策略网络。如，每个网络设备上均部署有一个目标子策略网络。其中，第一网络设备上部署有上述第一目标子策略网络。

（1）在步骤 402 之前，作为第一种实现方式，所述方法还包括：

所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络。

其中，上述第一目标子策略网络可基于目标价值网络进行训练得到。

所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

所述第一网络设备根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，所述第一网络设备将第一子策略网络 W_i 确定为第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为所述第一网络设备基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成的；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

其中，上述以当前进行的训练为第 i 次训练进行说明。

具体地，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

S1、所述第一网络设备获取所述第一区域的目标状态信息 S_i，其中，i 为正整数；

S2、所述第一网络设备将所述目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述第一子策略网络 W_i 的输出结果进行处理以得到第二调度信息；

S3、所述第一网络设备向所述第一区域内的终端下发所述第二调度信息，所述第二调度信息被所述第一区域内的终端用于数据传输；

S4、所述第一网络设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

S5、所述第一网络设备根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值低于预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络 W_{i+1}；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次所述训练的价值网络 Q_{i-1} 得到的；令 i=i+1，并重复执行 S1-S5；当所述性能参数的值不低于所述预设值时，将所述第一子策略网络 W_i 确定为第一目标子策略网络；

其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

上述以当前进行的训练为第 i 次训练进行说明。其中，上述上一次所述训练即为第 i-1 次训练。上述下一次所述训练即为第 i+1 次训练。

其中，该实施例以性能参数的值不低于预设值时停止训练。当然，本申请实施例并不限定上述条件。本申请实施例还可以以性能参数的值不高于预设值时停止训练。例如通过对上述预设值取反构成新的预设值等。

作为第一种实现方式的第二种方案，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

所述第一网络设备对第一初始子策略网络进行 M 次迭代训练，以得到第一目标子策略网络，所述 M 为正整数；

其中，在进行第 i 次迭代训练时，所述第一网络设备获取所述第一区域的目标状态信息 S_i，i 为正整数；

所述第一网络设备将所述目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述子策略网络 W_i 的输出结果进行处理以得到第二调度信息；

所述第一网络设备向所述第一区域内的终端下发所述第二调度信息，所述第二调度信息被所述第一区域内的终端用于数据传输；

所述第一网络设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络 W_{i+1}；其中，所述第一子策略网络 W_i 的评价价值基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次所述训练的价值网络 Q_{i-1} 得到的；

其中，当 i=M 时，所述第一子策略网络 W_{i+1} 为第一目标子策略网络；当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

上述实施例以迭代训练的次数达到预设次数时停止训练。当然，本申请实施例并不限定上述条件。本申请实施例还可以以更新参数的次数达到预设次数时停止训练等。此处不作具体限定。

作为第一种实现方式的第三种方案，其中，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

A1、所述第一网络设备获取所述第一区域的目标状态信息 S_i，其中，i 为正整数；

A2、所述第一网络设备将所述目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述第一子策略网络 W_i 的输出结果进行处理以得到第二调度信息；

A3、所述第一网络设备向所述第一区域内的终端下发所述第二调度信息，所述第二调度信息被所述第一区域内的终端用于数据传输；

A4、所述第一网络设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

A5、所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结

果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络 W_{i+1}；其中，所述第一子策略网络 W_i 的评价价值基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次所述训练的价值网络 Q_{i-1} 得到的；

A6、所述第一网络设备获取所述第一子策略网络 W_{i+1} 对应的损失函数，当所述第一子策略网络 W_{i+1} 对应的损失函数的值不低于预设阈值时，令 i=i+1，并重复执行 A1-A6；当所述第一子策略网络 W_{i+1} 对应的损失函数的值低于所述预设阈值时，将所述第一子策略网络 W_{i+1} 确定为第一目标子策略网络；

其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

该实施例以策略网络对应的损失函数的值低于预设阈值时停止训练。当然，本申请实施例还可以以策略网络对应的损失函数的值高于预设阈值时停止训练等，此处不作具体限定。

其中，上述作为第一种实现方式的各个方案中，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息（包含目标状态信息 S_i、目标状态信息 S_{i+1}）、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

（2）在步骤 402 之前，作为第二种实现方式，所述方法还包括：

所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络。

其中，上述第一目标子策略网络可基于第一目标子价值网络进行训练得到。上述 K 个网络设备对应 K 个目标子价值网络。上述第一网络设备对应第一目标子价值网络。

所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

所述第一网络设备根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，所述第一网络设备将第一子策略网络 W_i 确定为第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为所述第一网络设备基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成的；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

其中，上述以当前进行的训练为第 i 次训练进行说明。

具体地，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

S6、所述第一网络设备获取所述第一区域的目标状态信息 S_i，其中，i 为正整数；

S7、所述第一网络设备将所述目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，

并对所述第一子策略网络 W_i 的输出结果进行处理以得到第二调度信息；

S8、所述第一网络设备向所述第一区域内的终端下发所述第二调度信息，所述第二调度信息被所述第一区域内的终端用于数据传输；

S9、所述第一网络设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

S10、所述第一网络设备根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值低于预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络 W_{i+1}；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应；令 i=i+1，并重复执行 S6-S10；当所述性能参数的值不低于所述预设值时，将所述第一子策略网络 W_i 确定为第一目标子策略网络；

其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

上述以当前进行的训练为第 i 次训练进行说明。其中，上述上一次所述训练即为第 i-1 次训练。上述下一次所述训练即为第 i+1 次训练。

其中，该实施例以性能参数的值不低于预设值时停止训练。当然，本申请实施例并不限定上述条件。本申请实施例还可以以性能参数的值不高于预设值时停止训练。例如通过对上述预设值取反构成新的预设值等。

作为第二种实现方式的第二种方案，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

所述第一网络设备对第一初始子策略网络进行 M 次迭代训练，以得到第一目标子策略网络，所述 M 为正整数；

其中，在进行第 i 次迭代训练时，所述第一网络设备获取所述第一区域的目标状态信息 S_i，i 为正整数；

所述第一网络设备将所述目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述子策略网络 W_i 的输出结果进行处理以得到第二调度信息；

所述第一网络设备向所述第一区域内的终端下发所述第二调度信息，所述第二调度信息被所述第一区域内的终端用于数据传输；

所述第一网络设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络 W_{i+1}；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值

网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应；

其中，当 i=M 时，所述第一子策略网络 W_{i+1} 为第一目标子策略网络；当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

上述实施例以迭代训练的次数达到预设次数时停止训练。当然，本申请实施例并不限定上述条件。本申请实施例还可以以更新参数的次数达到预设次数时停止训练等。此处不作具体限定。

作为第二种实现方式的第三种方案，其中，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

B1、所述第一网络设备获取所述第一区域的目标状态信息 S_i，其中，i 为正整数；

B2、所述第一网络设备将所述目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述第一子策略网络 W_i 的输出结果进行处理以得到第二调度信息；

B3、所述第一网络设备向所述第一区域内的终端下发所述第二调度信息，所述第二调度信息被所述第一区域内的终端用于数据传输；

B4、所述第一网络设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

B5、所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络 W_{i+1}；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应；

B6、所述第一网络设备获取所述第一子策略网络 W_{i+1} 对应的损失函数，当所述第一子策略网络 W_{i+1} 对应的损失函数的值不低于预设阈值时，令 i=i+1，并重复执行 B1-B6；当所述第一子策略网络 W_{i+1} 对应的损失函数的值低于所述预设阈值时，将所述第一子策略网络 W_{i+1} 确定为第一目标子策略网络；

其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

该实施例以策略网络对应的损失函数的值低于预设阈值时停止训练。当然，本申请实施例还可以以策略网络对应的损失函数的值高于预设阈值时停止训练等，此处不作具体限定。

其中，上述第二种实现方式中的各个方案中，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

（3）在步骤402之前，作为第三种实现方式，所述方法还包括：

所述第一网络设备接收集中式网元设备发送的第一目标子策略网络的参数，其中，所述K个目标子策略网络的参数均相同。

其中，上述调度控制系统还包括集中式网元设备。第一网络设备可接收集中式网元设备发送的第一目标子策略网络的参数，进而所述第一网络设备将所述第一区域的目标状态信息输入至第一目标子策略网络进行处理，并对所述第一目标子策略网络的输出结果进行处理以得到所述第一调度信息。

上述集中式网元设备为核心网设备或基站集中式单元CU设备。其中，核心网设备如可以是4G通信或5G通信中的核心网设备，也可以是未来通信网络中的核心网设备，本申请并不对实施该技术方案的核心网设备或者基站的通信技术代次或者应用领域进行限制。上述基站集中式单元CU设备如可以是5G通信中的基站集中式单元CU设备。

403、所述第一网络设备向所述第一区域内的终端下发所述第一调度信息，所述第一调度信息被所述第一区域内的终端用于数据传输。

上述调度方法可应用于如下场景。例如，蜂窝网络多小区调度问题，每个小区都需要针对本小区的用户进行调度决策。又如异构网络中，存在宏站Macrocell、微微站Picocell和家庭基站Femtocell等多个不同等级和覆盖范围的基站，这些基站需要针对其覆盖范围内注册在其名下的用户进行调度决策等。

如图5所示，基站4001可从所述基站4001所覆盖的区域中获取包含终端4002在内的各个终端的目标状态信息。其中，该目标状态信息包括网络状态信息和用户数据包缓存信息等。该各个终端可以是基站4001覆盖的小区内的各终端。或者，该各个终端也可以是某个基站覆盖的小区中的某个宏站、微微站或家庭基站所覆盖范围内的注册终端等。基站4001根据其所覆盖的区域中各个终端的目标状态信息得到调度信息，进而基站4001向包含终端4002在内的各个终端下发该调度信息，以便各终端根据该调度信息进行数据传输。

本申请实施例基于K个网络设备中的第一网络设备通过获取第一区域的目标状态信息，然后基于目标状态信息和与该第一网络设备对应的第一目标子策略网络得到调度信息，进而向第一区域内的终端下发该调度信息，以便第一区域内的各终端根据该调度信息进行数据传输。采用该手段，其中，各个网络设备分别对应各自的策略网络进行调度控制，实现多智能体进行调度控制，提升了调度控制系统的性能。且，通过分布式的部署策略网络，提高了调度控制方案的可行性。

下面具体介绍调度算法的训练方法。参照图6所示，为本申请实施例提供的一种调度算法的训练方法，该方法应用于调度算法训练系统，其中，该调度算法训练系统提供一种由中心式的价值网络（C网络）和分布式的策略网络（A网络）构成的多智能体强化学习MARL架构。通过该架构进行训练可得到一个目标价值网络和K个目标子策略网络。其中，该K个目标子策略网络与K个网络设备一一对应。上述K个网络设备可以基于所得

到的对应的目标子策略网络实现上述调度。

其中，上述中心式的价值网络可部署在集中式网元设备上。该集中式网元设备可以是核心网设备或基站的集中单元(Centralized Unit, CU)设备。上述分布式的子策略网络可部署在基站的分布单元(Distributed Unit，DU)设备上。

基于上述中心式的价值网络可部署在集中式网元设备上，本申请实施例提供一种调度算法的训练方法，包括步骤 601-602，具体如下：

601、集中式网元设备获取训练数据；

602、所述集中式网元设备根据所述训练数据对初始价值网络进行迭代训练，以得到目标价值网络。

其中，根据上述迭代训练的终止条件的不同，上述方法可包括至少三种实现方式。

作为第一种实现方式，所述训练数据包括 K 个网络设备所覆盖的 K 个区域中每个区域的目标状态信息 S_i，所述 K 个区域与所述 K 个网络设备一一对应，K 为大于 1 的整数，i 为正整数，所述集中式网元设备根据所述训练数据对初始价值网络进行迭代训练，以得到目标价值网络，包括：

S11、所述集中式网元设备获取所述 K 个网络设备所覆盖的 K 个区域中每个区域的目标状态信息 S_{i+1}，其中，所述 K 个区域中每个区域的目标状态信息 S_{i+1} 为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后得到的，所述 K 个区域分别对应的第二调度信息为将所述 K 个区域中每个区域的目标状态信息 S_i 分别输入至 K 个子策略网络 W_i 进行处理，并对所述 K 个子策略网络 W_i 的输出结果分别进行处理得到的，所述子策略网络 W_i 是基于上一次所述训练的子策略网络 W_{i-1} 得到的，所述 K 个子策略网络与所述 K 个网络设备一一对应；

S12、所述集中式网元设备根据所述 K 个区域的 K 个目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值低于预设值时，所述集中式网元设备将所述 K 个区域的 K 个目标状态信息 S_i、K 个子策略网络 W_i 的输出结果、K 个区域的 K 个状态信息 S_{i+1} 和所述 K 个区域对应的反馈收益均输入至价值网络 Q_i 进行处理，以得到 K 个子策略网络分别对应的评价价值；所述集中式网元设备调整所述价值网络 Q_i 中的参数，以得到用于下一次训练的价值网络 Q_{i+1}；令 i=i+1，并重复执行 S11-S12；当所述性能参数的值不低于所述预设值时，将所述价值网络 Q_i 确定为目标价值网络；

其中，当 i=1 时，所述价值网络 Q_i 为初始价值网络。

也就是说，本申请实施例中当系统性能参数达到设定的阈值时，则停止迭代训练，进而得到目标价值网络。上述系统性能参数可包括吞吐、公平性、丢包率、时延等。其中，可通过对目标状态信息进行处理，进而可得到性能参数。如基于目标状态信息中的网络状态信息和用户数据包缓存信息计算吞吐、公平性、丢包率、时延等系统性能参数。

作为第二种实现方式，当初始价值网络迭代训练的次数达到预设的 N 次时，则停止迭代训练，进而得到目标价值网络。

具体地，所述训练数据包括 K 个网络设备所覆盖的 K 个区域中每个区域的目标状态信息 S_i，所述 K 个区域与所述 K 个网络设备一一对应，i 为正整数，所述集中式网元设备

根据所述训练数据对初始价值网络进行迭代训练，以得到目标价值网络，包括：

所述集中式网元设备对初始价值网络进行 N 次迭代训练，以得到目标价值网络，所述 N 为正整数。

其中，在进行第 i 次迭代训练时，所述集中式网元设备获取所述 K 个网络设备所覆盖的 K 个区域中每个区域的状态信息 S_{i+1}，其中，所述 K 个区域的 K 个状态信息 S_{i+1} 为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后得到的，所述 K 个区域分别对应的第二调度信息为将所述 K 个区域的 K 个目标状态信息 S_i 分别输入至 K 个子策略网络 W_i 进行处理，并对所述 K 个子策略网络 W_i 的输出结果分别进行处理得到的，所述子策略网络 W_i 是基于子策略网络 W_{i-1} 得到的，所述 K 个子策略网络与所述 K 个网络设备一一对应；

所述集中式网元设备将所述 K 个网络设备所覆盖的 K 个区域中每个区域的状态信息 S_i、所述 K 个子策略网络 W_i 的输出结果、所述 K 个基站所覆盖的 K 个区域中每个区域的目标状态信息 S_{i+1} 和所述 K 个区域对应的反馈收益均输入至价值网络 Q_i 进行处理，以得到所述 K 个子策略网络 W_i 的评价价值；其中，所述 K 个区域对应的反馈收益为所述 K 个区域的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后得到的；

所述集中式网元设备调整所述价值网络 Q_i 中的参数，以得到价值网络 Q_{i+1}；

其中，当 i=N 时，所述价值网络 Q_{i+1} 为目标价值网络；当 i=1 时，所述价值网络 Q_i 为初始价值网络。

作为第三种实现方式，当得到的价值网络的损失函数低于预设阈值时，则停止迭代训练，进而得到目标价值网络。

具体地，所述训练数据包括 K 个网络设备所覆盖的 K 个区域中每个区域的目标状态信息 S_i，所述 K 个区域与所述 K 个网络设备一一对应，i 为正整数，所述集中式网元设备根据所述训练数据对初始价值网络进行迭代训练，以得到目标价值网络，包括：

C1、所述集中式网元设备获取所述 K 个网络设备所覆盖的 K 个区域中每个区域的目标状态信息 S_{i+1}，其中，所述 K 个区域的 K 个目标状态信息 S_{i+1} 为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后得到的，所述 K 个区域分别对应的第二调度信息为将所述 K 个区域的 K 个目标状态信息 S_i 分别输入至 K 个子策略网络 W_i 进行处理，并对所述 K 个子策略网络 W_i 的输出结果分别进行处理得到的，所述子策略网络 W_i 是基于子策略网络 W_{i-1} 得到的，所述 K 个子策略网络与所述 K 个网络设备一一对应；

C2、所述集中式网元设备将所述 K 个网络设备所覆盖的 K 个区域中每个区域的目标状态信息 S_i、所述 K 个子策略网络 W_i 的输出结果、所述 K 个基站所覆盖的 K 个区域中每个区域的目标状态信息 S_{i+1} 和所述 K 个区域对应的反馈收益均输入至价值网络 Q_i 进行处理，以得到所述 K 个子策略网络 W_i 的评价价值；其中，所述 K 个区域对应的反馈收益为所述 K 个区域的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后得到的；

C3、所述集中式网元设备调整所述价值网络 Q_i 中的参数，以得到价值网络 Q_{i+1}；

C4、所述集中式网元设备获取所述价值网络 Q_{i+1} 对应的损失函数，当所述价值网络 Q_{i+1} 对应的损失函数的值不低于预设阈值时，令 i=i+1，并重复执行 C1-C4；当所述价值网络 Q_{i+1} 对应的损失函数的值低于所述预设阈值时，将所述价值网络 Q_{i+1} 确定为目标价值网络；

其中，当 i=1 时，所述价值网络 Q_i 为初始价值网络。

上述各实施例以中心式的价值网络部署在集中式网元设备上为例进行调度算法的训练方法的说明。对于上述分布式的子策略网络部署在基站的分布单元(Distributed Unit，DU)设备上时，本申请实施例还提供一种调度算法的训练方法，该方法应用于调度算法训练系统，该调度算法训练系统包括 K 个基站，K 为大于 1 的整数，所述方法包括步骤 603-604，具体如下：

603、第一网络设备获取训练数据；其中，所述第一网络设备为所述 K 个网络设备中的任意一个；

604、所述第一网络设备根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，其中，所述第一初始子策略网络为 K 个初始子策略网络中、所述第一网络设备对应的初始子策略网络，所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述 K 个初始子策略网络、K 个目标子策略网络分别与所述 K 个网络设备一一对应。

相应地，根据上述迭代训练的终止条件的不同，上述方法可包括至少三种实现方式。

其中，针对第一网络设备对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络的具体实现方式，可参照在步骤 402 之前，作为第一种实现方式的三种方案的介绍，在此不再赘述。

基于上述各实施例得到的目标子策略网络，基站进而可实现调度。

本申请实施例基于中心式的价值网络和分布式的策略网络构成的多智能体 MARL 架构进行训练，得到一个目标价值网络和多个分布式的目标策略网络。该分布式的目标策略网络可用于网络设备进行调度，避免了单智能体 DRL 完全中心式的调度，提高了方案可行性。

如图 6 所示，其中，各 A 网络可从通信系统对应的环境中获取与该 A 网络对应的区域的目标状态信息 s。各 A 网络可基于上述目标状态信息 s 得到决策动作 a。环境执行完各决策动作 a 后反馈收益 r 给 C 网络。C 网络通过获取环境的总目标状态信息以及下一时刻全局状态，并基于上述反馈收益 r，各决策动作 a，确定出各 A 网络分别对应的评价价值 v。在上述架构中，如与基站 K 对应的策略网络 A_k 从环境获得基站 K 所覆盖的区域对应的目标状态信息 s_k，价值网络 C 从环境获得当前全局状态，即（s_1、s_2…s_k）。策略网络 A_k 作出决策 a_k。环境执行所有策略网络作出的决策后，反馈收益 r 给价值网络。价值网络根据当前全局状态 s，各策略网络的动作 a，反馈收益 r 以及下一时刻全局状态 s’，输出各策略网络的评价价值 v，并更新价值网络中的参数。上述下一时刻全局状态 s’为环境执行所有策略网络作出的决策后得到的全局状态信息。其中，策略网络 A_k 根据当前

状态 s_k，动作 a_k，下一时刻状态 s_k’和价值网络输出的评价价值 v_k，更新自身网络参数。

其中，策略网络和价值网络的参数更新可以同步，也可以是异步的。即可以同时更新，也可以某些调度周期内只更新价值网络或只更新策略网络等。此处不作具体限定。

示例性地，在如图 7 所示的多小区蜂窝网络场景中可部署上述 MARL 框架。以三个小区联合调度为例。参与联合调度的小区数可以根据小区间干扰情况进行设置，如将互相干扰较严重的多个小区放在一起联合调度。如图 7 所示，中心式的价值网络可部署在核心网设备或基站的集中单元(Centralized Unit, CU)上。分布式的策略网络部署在基站的分布单元(Distributed Unit，DU)上。各基站可以基于对应的目标策略网络实现调度控制。

如图 8 所示，上述 MARL 框架还可部署在多等级异构网络中。对于一个宏站覆盖范围内的小区，存在一个宏站、多个微微站和家庭基站等。此时，可以将价值网络部署在宏站上，策略网络部署在宏站、微微站和家庭基站上。

上述实施例以中心式的价值网络和分布式的策略网络组成的多智能体强化学习 MARL 架构进行说明。本申请实施例还提供一种分布式的价值网络和分布式的策略网络组成的多智能体强化学习 MARL 架构，如图 9A、图 9B 所示。其中，上述分布式的价值网络可部署在集中式网元设备上，该集中式网元设备可以是核心网设备或基站的集中单元(Centralized Unit, CU)设备上。上述分布式的策略网络可部署在基站的分布单元(Distributed Unit，DU)设备上。通过该架构进行训练可得到 K 个目标子价值网络和 K 个目标子策略网络。上述 K 个目标子价值网络、K 个目标子策略网络分别与 K 个网络设备一一对应。

上述 K 个网络设备可以基于所得到的对应的目标子策略网络实现上述调度。

基于上述分布式的价值网络可部署在集中式网元设备上，本申请实施例提供一种调度算法的训练方法，包括步骤 901-902，具体如下：

901、集中式网元设备获取训练数据；

902、所述集中式网元设备根据所述训练数据对第一初始子价值网络进行迭代训练，以得到第一目标子价值网络。

其中，所述第一初始子价值网络为 K 个初始子价值网络中、第一网络设备所对应的初始子价值网络，所述第一目标子价值网络为 K 个目标子价值网络中、所述第一网络设备对应的目标子价值网络，其中，所述第一网络设备为 K 个网络设备中的任意一个，所述 K 个初始子价值网络、所述 K 个目标子价值网络分别与所述 K 个网络设备一一对应。

进一步地，根据上述迭代训练的终止条件的不同，上述方法可包括至少三种实现方式。

作为第一种实现方式，所述训练数据包括所述第一网络设备所覆盖的第一区域的目标状态信息 S_i，i 为正整数，所述集中式网元设备根据所述训练数据对第一初始子价值网络进行迭代训练，以得到第一目标子价值网络，包括：

S13、所述集中式网元设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的用户根据所述第一区域对应的第二调度信息进行数据传输后得到的，所述第一区域对应的第二调度信息为将所述第一区域的目标状

态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述第一子策略网络 W_i 的输出结果进行处理得到的，所述第一子策略网络 W_i 是基于上一次所述训练的第一子策略网络 W_{i-1} 得到的；

S14、所述集中式网元设备根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值低于预设值时，所述集中式网元设备将所述第一区域的目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1}、所述第一区域对应的反馈收益以及除所述第一网络设备对应的第一子价值网络 q_i 之外的其他 K-1 个网络设备分别对应的子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的用户根据所述第二调度信息进行数据传输后得到的；所述集中式网元设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络 q_{i+1}；令 i=i+1，并重复执行 S13-S14；当所述性能参数的值不低于所述预设值时，将所述第一子价值网络 q_i 确定为第一目标子价值网络；

其中，当 i=1 时，所述第一子价值网络 q_i 为第一初始子价值网络。

也就是说，本申请实施例中当系统性能参数达到设定的阈值时，则停止迭代训练，进而得到目标子价值网络。

作为第二种实现方式，当第一初始子价值网络迭代训练的次数达到预设的 N 次时，则停止迭代训练，进而得到第一目标子价值网络。

具体地，所述训练数据包括所述第一网络设备所覆盖的第一区域的状态信息 S_i，i 为正整数，所述集中式网元设备根据所述训练数据对第一初始子价值网络进行迭代训练，以得到第一目标子价值网络，包括：

所述集中式网元设备对第一初始子价值网络进行 N 次迭代训练，以得到第一目标子价值网络，所述 N 为正整数。

其中，在进行第 i 次迭代训练时，所述集中式网元设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后得到的，所述第一区域对应的第二调度信息为将所述第一区域的目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述第一子策略网络 W_i 的输出结果进行处理得到的，所述第一子策略网络 W_i 是基于上一次所述训练的第一子策略网络 W_{i-1} 得到的；

所述集中式网元设备将所述第一网络设备所覆盖的第一区域的目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1}、所述第一区域对应的反馈收益以及除所述第一网络设备对应的第一子价值网络 q_i 之外的其他 K-1 个网络设备分别对应的子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

所述集中式网元设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络 q_{i+1}；

其中，当 i=N 时，所述第一子价值网络 q_{i+1} 为第一目标子价值网络；当 i=1 时，所述第一子价值网络 q_i 为第一初始子价值网络。

作为第三种可选的实现方式，当得到的第一子价值网络的损失函数低于预设阈值时，则停止迭代训练，进而得到第一目标子价值网络。通过对 K 个初始子价值网络分别进行迭代训练进而得到 K 个目标子价值网络。

具体地，所述训练数据包括所述第一网络设备 A 所覆盖的第一区域的目标状态信息 S_i，i 为正整数，所述集中式网元设备根据所述训练数据对初始子价值网络进行迭代训练，以得到目标子价值网络，包括：

E1、所述集中式网元设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后得到的，所述第一区域对应的第二调度信息为将所述第一区域的目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述第一子策略网络 W_i 的输出结果进行处理得到的，所述第一子策略网络 W_i 是基于上一次所述训练的第一子策略网络 W_{i-1} 得到的；

E2、所述集中式网元设备将所述第一区域的目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1}、所述第一区域对应的反馈收益以及除所述第一网络设备对应的第一子价值网络 q_i 之外的其他 K-1 个网络设备分别对应的子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

E3、所述集中式网元设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络 q_{i+1}；

E4、所述集中式网元设备获取所述第一子价值网络 q_{i+1} 对应的损失函数，当所述第一子价值网络 q_{i+1} 对应的损失函数的值不低于预设阈值时，令 i=i+1，并重复执行 E1-E4；当所述第一子价值网络 q_{i+1} 对应的损失函数的值低于所述预设阈值时，将所述第一子价值网络 q_{i+1} 确定为第一目标子价值网络；

其中，当 i=1 时，所述第一子价值网络 q_i 为第一初始子价值网络。

上述各实施例以分布式的价值网络部署在集中式网元设备上为例进行调度算法的训练方法的说明。对于上述分布式的子策略网络部署在基站的分布单元(Distributed Unit，DU)设备上时，本申请实施例还提供一种调度算法的训练方法，该方法应用于调度算法训练系统，该调度算法训练系统包括 K 个网络设备，K 为大于 1 的整数，所述方法包括步骤 903-904，具体如下：

903、第一网络设备获取训练数据；其中，所述基站 A 为所述 K 个基站中的任意一个；

904、所述第一网络设备根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，其中，所述第一初始子策略网络为 K 个初始子策略网络中、所述第一网络设备对应的初始子策略网络，所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述 K 个初始子策略网络、K 个目

标子策略网络分别与所述 K 个网络设备一一对应。

相应地，根据上述迭代训练的终止条件的不同，上述方法可包括至少三种实现方式。

其中，所述训练数据包括第一区域的目标状态信息 S_i，i 为正整数，其中，所述第一区域为所述第一网络设备所覆盖的区域。

所述第一网络设备根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络的具体实现方式，可参阅上述在步骤 402 之前，作为第二种实现方式的各个方案的介绍，在此不再赘述。

基于上述各实施例得到的目标子策略网络，基站进而可实现调度。

上述实施例以分布式的价值网络部署在集中式网元设备上，分布式的策略网络部署在基站的分布单元设备上为例进行说明。可替代的，上述分布式的价值网络和分布式的策略网络还可以均部署在基站的分布单元设备上。本申请实施例还提供一种调度算法的训练方法，所述方法应用于调度算法训练系统，所述调度算法训练系统包括 K 个网络设备，K 为大于 1 的整数，包括步骤 905-906，具体如下：

905、第一网络设备获取训练数据；其中，所述第一网络设备为所述 K 个网络设备中的任意一个；

906、所述第一网络设备根据所述训练数据对第一初始子价值网络、第一初始子策略网络分别进行迭代训练，以得到第一目标子价值网络、第一目标子策略网络，其中，所述第一初始子价值网络为 K 个初始子价值网络中与所述第一网络设备对应的初始子价值网络，所述第一目标子价值网络为 K 个目标子价值网络中与所述第一网络设备对应的目标子价值网络，所述第一初始子策略网络为 K 个初始子策略网络中与所述第一网络设备对应的初始子策略网络，所述第一目标子策略网络为 K 个目标子策略网络中与所述第一网络设备对应的目标子策略网络，所述 K 个初始子价值网络、K 个目标子价值网络、K 个初始子策略网络、K 个目标子策略网络分别与所述 K 个网络设备一一对应。

其中，所述训练数据包括第一区域的状态信息，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述第一网络设备根据所述训练数据对第一初始子价值网络、第一初始子策略网络分别进行迭代训练，以得到第一目标子价值网络、第一目标子策略网络，包括：

S26、所述第一网络设备将所述第一区域的目标状态信息 S_i 输入至第一子策略网络 W_i 进行处理，并对所述第一子策略网络 W_i 的输出结果进行处理以得到第二调度信息，其中，i 为正整数；

S27、所述第一网络设备获取所述第一区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；

S28、所述第一网络设备根据所述目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值低于预设值时，所述第一网络设备将所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和反馈收益均输入至第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值；所述第一网络设备调整所述第一子价值网络

q_i 中的参数，以得到用于下一次所述训练的第一子价值网络 q_{i+1}；其中，所述反馈收益为所述第一区域内的终端根据所述第二调度信息进行数据传输后得到的；所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络 W_{i+1}；令 i=i+1，并重复执行 S26-S28；当所述性能参数的值不低于所述预设值时，将所述第一子价值网络 q_i 确定为第一目标子价值网络，将所述第一子策略网络 W_i 确定为第一目标子策略网络；

其中，当 i=1 时，所述第一子价值网络 q_i 为第一初始子价值网络，所述第一子策略网络 W_i 为第一初始子策略网络。

上述仅以性能参数作为训练结束的判断条件。其中，以训练次数、网络的损失函数的值等作为训练结束的判断条件的具体处理手段可参阅上述介绍，在此不再赘述。

如图 9A 所示，该架构中有多个分布式的价值网络，每个价值网络单独为所对应的策略网络提供评价价值。多个价值网络之间通过一条通信总线连接，用于交互信息。在这个架构中，策略网络 A_k 从环境获得自身对应的状态 s_k，价值网络 C_k 从环境获得自身对应的状态 s_k。策略网络 A_k 作出决策 a_k，环境执行所有策略网络作出的决策后，反馈收益 r_k 给价值网络 C_k。价值网络根据当前状态 s_k、策略网络的动作 a_k、收益 r_k、下一时刻状态 s_k’，以及经过通信总线得到的其他价值网络的信息，输出策略网络 A_k 的评价价值 v_k，并更新自身网络参数。策略网络 A_k 根据当前状态 s_k、动作 a_k、下一时刻状态 s_k’和价值网络输出的评价价值 v_k，更新自身网络参数。其中，上述下一时刻状态 s_k’为相应环境执行策略网络作出的决策后得到的状态。上述多个价值网络经过通信总线交互的信息可以是各价值网络对应的状态 s_k，动作 a_k，收益 r_k，也可以是其他价值网络的输出结果、其他价值网络的参数或者是其他价值网络更新的梯度值等。同样，这个框架下，策略网络和价值网络参数的更新可以是同步的，也可以是异步的。

进一步地，如图 9B 所示，策略网络之间也可以互传信息。其中，经过通信总线可得到其他策略网络的信息。该其他策略网络的信息可包括其他策略网络的输出结果、其他策略网络的参数或者是其他策略网络更新的梯度值等。如策略网络 A_k 可根据当前状态 s_k、动作 a_k、下一时刻状态 s_k’、价值网络输出的评价价值 v_k 以及其他策略网络的信息进而来更新自身网络参数。

如图 10 所示，在多小区蜂窝网络场景中部署上述 MARL 框架。参与联合调度的小区数可以根据小区间干扰情况进行设置，如将互相干扰较严重的多个小区放在一起联合调度。以 3 小区联合调度为例。如图 10 所示，分布式的价值网络和分布式的策略网络均部署在基站的分布单元(Distributed Unit, DU)上。可替代的，分布式的价值网络也可以部署在核心网设备或基站的 CU 上。其中，分布式的价值网络部署在核心网设备或基站的 CU 上，有助于减少价值网络之间的通信开销。

上述布式价值网络和分布式策略网络对应的 MARL 框架也可以用于异构网络、认知无线网络等存在多等级网络的系统的调度。以异构网络为例，如图 11 所示，对于一个宏站覆盖范围内的小区，存在一个宏站、多个微微站和多个家庭基站。此时，可以将价值网

络和策略网络部署在宏站、微微站和家庭基站上。其中，也可以将多个分布式的价值网络部署在宏站上，以便减小价值网络间通信的开销。

本申请实施例基于分布式的价值网络和分布式的策略网络构成的多智能体 MARL 架构进行训练，得到多个目标价值网络和多个分布式的目标策略网络。该分布式的目标策略网络可用于网络设备进行调度，避免了单智能体 DRL 完全中心式的调度，提高了方案可行性。

本申请实施例还提供一种中心式的价值网络和中心式的策略网络组成的多智能体强化学习 MARL 架构。如图 12 所示，该架构包括中心式的价值网络和中心式的策略网络。通过该架构进行训练可得到目标价值网络和目标策略网络。训练结束后将目标策略网络下发给各个基站，可用于分布式地完成调度。

其中，中心式的价值网络和中心式的策略网络可以均部署在集中式网元设备上，如核心网设备或基站的集中单元(Centralized Unit, CU)上。为此，本申请实施例提供一种调度算法的训练方法，所述方法应用于调度算法训练系统，所述调度算法训练系统包括集中式网元设备，所述方法包括步骤 1201-1203，具体如下：

1201、所述集中式网元设备获取训练数据；

1202、所述集中式网元设备根据所述训练数据对初始价值网络、初始策略网络分别进行迭代训练，以得到目标价值网络、目标策略网络；

其中，所述训练数据包括 K 个网络设备所覆盖的 K 个区域的目标状态信息 S_i，其中，i 为正整数，所述集中式网元设备根据所述训练数据对初始价值网络、初始策略网络分别进行迭代训练，以得到目标价值网络、目标策略网络，包括：

S29、所述集中式网元设备将所述 K 个网络设备所覆盖的 K 个区域的目标状态信息 S_i 输入至策略网络 w_i 进行处理，并对所述策略网络 w_i 的输出结果进行处理以得到第二调度信息；

S30、所述集中式网元设备获取所述 K 个区域的目标状态信息 S_{i+1}，其中，所述目标状态信息 S_{i+1} 为所述 K 个区域内的终端根据所述第二调度信息进行数据传输后得到的状态信息；

S31、所述集中式网元设备根据所述 K 个区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值低于预设值时，所述集中式网元设备将所述 K 个区域的目标状态信息 S_i、所述策略网络 w_i 的输出结果、所述目标状态信息 S_{i+1} 和反馈收益均输入至价值网络 Q_i 进行处理，以得到所述策略网络 w_i 的评价价值；所述集中式网元设备调整所述价值网络 Q_i 中的参数，以得到价值网络 Q_{i+1}；其中，所述反馈收益为所述 K 个区域内的终端根据所述第二调度信息进行数据传输后得到的；所述集中式网元设备根据所述 K 个区域的目标状态信息 S_i、所述策略网络 w_i 的输出结果、所述目标状态信息 S_{i+1} 和所述评价价值，调整所述策略网络 w_i 中的参数，以得到策略网络 w_{i+1}；令 i=i+1，并重复执行 S29-S31；当所述性能参数的值不低于所述预设值时，将所述价值网络 Q_i 确定为目标价值网络，将所述策略网络 w_i 确定为目标策略网络；

其中，当 i=1 时，所述价值网络 Q_i 为初始价值网络，所述策略网络 w_i 为初始策略网络。

上述仅以性能参数作为训练结束的判断条件。其中，以训练次数、网络的损失函数的值等作为训练结束的判断条件的具体处理手段可参阅上述介绍，在此不再赘述。

1203、所述集中式网元设备将所述目标策略网络的参数发送至所述 K 个网络设备。

其中，集中式网元设备可将上述所得的目标策略网络下发给各个网络设备，进而可以实现分布式的调度，避免了单智能体 DRL 完全中心式的调度，提高了方案可行性。

如图 12 所示，中心式的 A 网络和 C 网络均获取全局的目标状态信息 s。然后，中心式的 A 网络为蜂窝网络中各小区或异构网络中各级基站做出决策动作 a_k。当动作 a_k 被执行后，系统状态更新，并反馈收益 r 给中心式的 C 网络。中心式的 C 网络根据收益 r、动作 a_k、目标状态信息 s、下一时刻全局状态 s’，给中心式的 A 网络计算评价价值 v，同时更新自身网络参数。中心式的 A 网络和 C 网络更新自身网络参数。若未到达训练终止条件则重复执行上述步骤。当到达训练终止条件，则将中心式的 A 网络下发至蜂窝网络中的各小区基站或异构网络中的各级基站。其中，上述终止条件包括当神经网络的更新次数达到设定的阈值，或系统性能（吞吐/公平性/丢包率/时延）达到设定的阈值，或神经网络的损失函数低于设定的阈值等。

如图 13 所示，中心式的价值网络和策略网络可以部署在多小区蜂窝网络的核心网设备或 CU 上。其中，经过上述训练后，可将中心式的策略网络复制下发给各小区基站，用于进行调度。中心式的价值网络和策略网络还可以部署在多等级异构网络的宏站上。如图 14 所示，其中，经过上述训练后，可将中心式的策略网络复制下发给各等级基站，用于进行调度。

进一步地，本申请实施例还提供一种调度控制系统，所述调度控制系统包括 K 个网络设备，K 为大于 1 的整数，其中，第一网络设备为所述 K 个网络设备中的任意一个，所述第一网络设备用于：

获取第一区域的目标状态信息，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述目标状态信息包括网络状态信息和用户数据包缓存信息；

基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息，其中，所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述 K 个目标子策略网络与所述 K 个网络设备一一对应；

向所述第一区域内的终端下发所述第一调度信息，所述第一调度信息被所述第一区域内的终端用于数据传输。

其中，在基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述第一网络设备还用于：

对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络；

其中，对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，具体包括：

根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于

预设值时，将第一子策略网络 W_i 确定为第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成的，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

进一步地，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

进一步地，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

作为另一种可选的实现方式，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

进一步地，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

作为另一种可选的实现方式，所述调度控制系统还包括集中式网元设备，在基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述第一网络设备还用于：

接收所述集中式网元设备发送的第一目标子策略网络的参数，其中，所述 K 个目标子策略网络的参数均相同，其中，所述集中式网元设备为核心网设备或基站集中式单元 CU 设备。

在一方面，本申请实施例还提供一种调度算法训练系统，所述调度算法训练系统包括

K 个网络设备，K 为大于 1 的整数，第一网络设备为所述 K 个网络设备中的任意一个，所述第一网络设备用于：

获取训练数据；

根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络；其中，所述第一初始子策略网络为 K 个初始子策略网络中、所述第一网络设备对应的初始子策略网络；所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络；所述 K 个初始子策略网络、所述 K 个目标子策略网络分别与所述 K 个网络设备一一对应。

其中，所述训练数据包括第一区域的目标状态信息 S_{i+1}，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述第一网络设备具体用于：

根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，将第一子策略网络 W_i 确定为所述第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

进一步地，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次训练的价值网络得到的。

进一步地，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备分别对应的各子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

作为另一种可选的实现方式，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子

价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

进一步地，所述第一网络设备还用于：

将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

进一步地，当所述性能参数的值低于所述预设值时，所述第一网络设备还用于：

将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；所述第一网络设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

作为再一种可选的实现方式，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述集中式网元设备用于：

将价值网络 Q_i 确定为目标价值网络，其中，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

其中，当所述性能参数的值低于所述预设值时，所述集中式网元设备用于：

将所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理以得到 K 个评价价值，其中，所述 K 个评价价值与所述 K 个子策略网络一一对应；

将所述 K 个评价价值分别发送至所述 K 个网络设备；

调整所述价值网络 Q_i 中的参数，以得到用于下一次所述训练的价值网络。

作为又一种可选的实现方式，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述集中式网元设备用于：

将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

其中，当所述性能参数的值低于所述预设值时，所述集中式网元设备用于：

将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值；其中，所述第一区域对应的反馈收益为所述第一区域

内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；

调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

本申请实施例基于 K 个网络设备中的第一网络设备通过获取第一区域的目标状态信息，然后基于目标状态信息和与该第一网络设备对应的第一目标子策略网络得到调度信息，进而向第一区域内的终端下发该调度信息，以便第一区域内的各终端根据该调度信息进行数据传输。采用该手段，其中，各个网络设备分别对应各自的策略网络进行调度控制，实现多智能体进行调度控制，提升了调度控制系统的性能。且，通过分布式的部署策略网络，提高了调度控制方案的可行性。

本申请实施例还提供了一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当其在计算机或处理器上运行时，使得计算机或处理器执行上述任一个方法中的一个或多个步骤。

本申请实施例还提供了一种包含指令的计算机程序产品。当该计算机程序产品在计算机或处理器上运行时，使得计算机或处理器执行上述任一个方法中的一个或多个步骤。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者通过所述计算机可读存储介质进行传输。所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线（例如同轴电缆、光纤、数字用户线）或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，（例如，软盘、硬盘、磁带）、光介质(例如，DVD)、或者半导体介质（例如，固态硬盘（solid state disk，SSD））等。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，该流程可以由计算机程序来指令相关的硬件完成，该程序可存储于计算机可读取存储介质中，该程序在执行时，可包括如上述各方法实施例的流程。而前述的存储介质包括：ROM 或随机存储记忆体 RAM、磁碟或者光盘等各种可存储程序代码的介质。

以上所述，仅为本申请实施例的具体实施方式，但本申请实施例的保护范围并不局限于此，任何在本申请实施例揭露的技术范围内的变化或替换，都应涵盖在本申请实施例的保护范围之内。因此，本申请实施例的保护范围应以所述权利要求的保护范围为准。

权 利 要 求

1、一种调度方法，其特征在于，所述方法应用于调度控制系统，所述调度控制系统包括K个网络设备，K为大于1的整数，所述方法包括：

第一网络设备获取第一区域的目标状态信息，其中，所述第一网络设备为所述K个网络设备中的任意一个，所述第一区域为所述第一网络设备所覆盖的区域，所述目标状态信息包括网络状态信息和用户数据包缓存信息；

所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息，其中，所述第一目标子策略网络为K个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述K个目标子策略网络与所述K个网络设备一一对应；

所述第一网络设备向所述第一区域内的终端下发所述第一调度信息，所述第一调度信息被所述第一区域内的终端用于数据传输。

2、根据权利要求1所述的方法，其特征在于，所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述方法还包括：

所述第一网络设备对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络；

其中，所述第一网络设备对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络，包括：

所述第一网络设备根据所述第一区域的目标状态信息S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，所述第一网络设备将第一子策略网络W_i确定为所述第一目标子策略网络，其中，i为正整数，所述第一区域的目标状态信息S_{i+1}为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为所述第一网络设备基于所述第一区域的目标状态信息S_i和所述第一子策略网络W_i生成的，所述目标状态信息S_i是第i次所述训练的目标状态信息；其中，当i=1时，所述第一子策略网络W_i为第一初始子策略网络。

3、根据权利要求2所述的方法，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息S_i、所述第一子策略网络W_i的输出结果、所述目标状态信息S_{i+1}和所述第一子策略网络W_i的评价价值，调整所述第一子策略网络W_i中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络W_i的评价价值是基于价值网络Q_i进行处理得到的，所述价值网络Q_i是基于上一次所述训练的价值网络得到的。

4、根据权利要求3所述的方法，其特征在于，所述第一子策略网络W_i的评价价值基于所述K个网络设备所覆盖的K个区域的各目标状态信息、所述K个网络设备对应的K个子策略网络的输出结果和所述K个区域对应的反馈收益均输入至所述价值网络Q_i进行处理得到，其中，所述K个区域对应的反馈收益为所述K个区域内的终端根据所述K

个区域分别对应的第二调度信息进行数据传输后确定的。

5、根据权利要求 2 所述的方法，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

6、根据权利要求 5 所述的方法，其特征在于，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

7、根据权利要求 1 所述的方法，其特征在于，所述调度控制系统还包括集中式网元设备，所述第一网络设备基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述方法还包括：

所述第一网络设备接收所述集中式网元设备发送的第一目标子策略网络的参数，其中，所述 K 个目标子策略网络的参数均相同，其中，所述集中式网元设备为核心网设备或基站集中式单元 CU 设备。

8、一种调度算法的训练方法，其特征在于，所述方法应用于调度算法训练系统，所述调度算法训练系统包括 K 个网络设备，K 为大于 1 的整数；所述方法包括：

第一网络设备获取训练数据，其中，所述第一网络设备为所述 K 个网络设备中的任意一个；

所述第一网络设备根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络；其中，所述第一初始子策略网络为 K 个初始子策略网络中、所述第一网络设备对应的初始子策略网络；所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络；所述 K 个初始子策略网络、所述 K 个目标子策略网络分别与所述 K 个网络设备一一对应。

9、根据权利要求 8 所述的方法，其特征在于，所述训练数据包括第一区域的目标状态信息 S_{i+1}，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述第一网络设备

根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络，包括：

所述第一网络设备根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，所述第一网络设备将第一子策略网络 W_i 确定为所述第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为所述第一网络设备基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

10、根据权利要求 9 所述的方法，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次训练的价值网络得到的。

11、根据权利要求 10 所述的方法，其特征在于，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备分别对应的各子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

12、根据权利要求 9 所述的方法，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

13、根据权利要求 12 所述的方法，其特征在于，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述

第一区域对应的第二调度信息进行数据传输后确定的。

14、根据权利要求 9 所述的方法，其特征在于，所述方法还包括：

所述第一网络设备将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

15、根据权利要求 14 所述的方法，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；所述第一网络设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

16、根据权利要求 9 所述的方法，其特征在于，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述方法还包括：

所述集中式网元设备将价值网络 Q_i 确定为目标价值网络，其中，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

17、根据权利要求 16 所述的方法，其特征在于，当所述性能参数的值低于所述预设值时，所述集中式网元设备将所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理以得到 K 个评价价值，其中，所述 K 个评价价值与所述 K 个子策略网络一一对应；

所述集中式网元设备将所述 K 个评价价值分别发送至所述 K 个网络设备；

所述集中式网元设备调整所述价值网络 Q_i 中的参数，以得到用于下一次所述训练的价值网络。

18、根据权利要求 9 所述的方法，其特征在于，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述方法还包括：

所述集中式网元设备将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

19、根据权利要求 18 所述的方法，其特征在于，当所述性能参数的值低于所述预设值时，所述集中式网元设备将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值；其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；

所述集中式网元设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

20、一种调度控制系统，其特征在于，所述调度控制系统包括 K 个网络设备，K 为大于 1 的整数，其中，第一网络设备为所述 K 个网络设备中的任意一个，所述第一网络设备用于：

获取第一区域的目标状态信息，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述目标状态信息包括网络状态信息和用户数据包缓存信息；

基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息，其中，所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络，所述 K 个目标子策略网络与所述 K 个网络设备一一对应；

向所述第一区域内的终端下发所述第一调度信息，所述第一调度信息被所述第一区域内的终端用于数据传输。

21、根据权利要求 20 所述的系统，其特征在于，在基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述第一网络设备还用于：

对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络；

其中，对第一初始子策略网络进行迭代训练，以得到所述第一目标子策略网络，具体包括：

根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，将第一子策略网络 W_i 确定为所述第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成的，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

22、根据权利要求 21 所述的系统，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息

S_{i+1}和所述第一子策略网络W_i的评价价值，调整所述第一子策略网络W_i中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络W_i的评价价值是基于价值网络Q_i进行处理得到的，所述价值网络Q_i是基于上一次所述训练的价值网络得到的。

23、根据权利要求 22 所述的系统，其特征在于，所述第一子策略网络W_i的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络Q_i进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

24、根据权利要求 21 所述的系统，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息S_i、所述第一子策略网络W_i的输出结果、所述目标状态信息S_{i+1}和所述第一子策略网络W_i的评价价值，调整所述第一子策略网络W_i中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络W_i的评价价值是基于第一子价值网络q_i进行处理得到的，所述第一子价值网络q_i是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络q_i为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

25、根据权利要求 24 所述的系统，其特征在于，所述第一子策略网络W_i的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息S_i以及目标状态信息S_{i+1}、所述第一网络设备对应的第一子策略网络W_i的输出结果、所述第一区域对应的反馈收益和除第一子价值网络q_i之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络q_i进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

26、根据权利要求 20 所述的系统，其特征在于，所述调度控制系统还包括集中式网元设备，在基于所述第一区域的目标状态信息和第一目标子策略网络生成第一调度信息之前，所述第一网络设备还用于：

接收所述集中式网元设备发送的第一目标子策略网络的参数，其中，所述 K 个目标子策略网络的参数均相同，其中，所述集中式网元设备为核心网设备或基站集中式单元 CU 设备。

27、一种调度算法训练系统，其特征在于，所述调度算法训练系统包括 K 个网络设备，K 为大于 1 的整数，第一网络设备为所述 K 个网络设备中的任意一个，所述第一网络设备用于：

获取训练数据；

根据所述训练数据对第一初始子策略网络进行迭代训练，以得到第一目标子策略网络；其中，所述第一初始子策略网络为 K 个初始子策略网络中、所述第一网络设备对应的初始子策略网络；所述第一目标子策略网络为 K 个目标子策略网络中、所述第一网络设备对应的目标子策略网络；所述 K 个初始子策略网络、所述 K 个目标子策略网络分别与所述 K 个网络设备一一对应。

28、根据权利要求 27 所述的系统，其特征在于，所述训练数据包括第一区域的目标状态信息 S_{i+1}，其中，所述第一区域为所述第一网络设备所覆盖的区域，所述第一网络设备具体用于：

根据所述第一区域的目标状态信息 S_{i+1}，得到性能参数，当所述性能参数的值不低于预设值时，将第一子策略网络 W_i 确定为所述第一目标子策略网络，其中，i 为正整数，所述第一区域的目标状态信息 S_{i+1} 为所述第一区域内的终端根据第二调度信息进行数据传输得到的；所述第二调度信息为基于所述第一区域的目标状态信息 S_i 和所述第一子策略网络 W_i 生成，所述目标状态信息 S_i 是第 i 次所述训练的目标状态信息；其中，当 i=1 时，所述第一子策略网络 W_i 为第一初始子策略网络。

29、根据权利要求 28 所述的系统，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于价值网络 Q_i 进行处理得到的，所述价值网络 Q_i 是基于上一次训练的价值网络得到的。

30、根据权利要求 29 所述的系统，其特征在于，所述第一子策略网络 W_i 的评价价值基于所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备分别对应的各子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理得到，其中，所述 K 个区域对应的反馈收益为所述 K 个区域内的终端根据所述 K 个区域分别对应的第二调度信息进行数据传输后确定的。

31、根据权利要求 28 所述的系统，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备用于：

根据所述目标状态信息 S_i、所述第一子策略网络 W_i 的输出结果、所述目标状态信息 S_{i+1} 和所述第一子策略网络 W_i 的评价价值，调整所述第一子策略网络 W_i 中的参数，以得到用于下一次所述训练的第一子策略网络；其中，所述第一子策略网络 W_i 的评价价值是基于第一子价值网络 q_i 进行处理得到的，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络

设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

32、根据权利要求 31 所述的系统，其特征在于，所述第一子策略网络 W_i 的评价价值基于所述第一网络设备所覆盖的第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理得到，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的。

33、根据权利要求 28 所述的系统，其特征在于，所述第一网络设备还用于：

将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

34、根据权利要求 33 所述的系统，其特征在于，当所述性能参数的值低于所述预设值时，所述第一网络设备还用于：

将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值，其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；所述第一网络设备调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

35、根据权利要求 28 所述的系统，其特征在于，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述集中式网元设备用于：

将价值网络 Q_i 确定为目标价值网络，其中，所述价值网络 Q_i 是基于上一次所述训练的价值网络得到的。

36、根据权利要求 35 所述的系统，其特征在于，当所述性能参数的值低于所述预设值时，所述集中式网元设备用于：

将所述 K 个网络设备所覆盖的 K 个区域的各目标状态信息、所述 K 个网络设备对应的 K 个子策略网络的输出结果和所述 K 个区域对应的反馈收益均输入至所述价值网络 Q_i 进行处理以得到 K 个评价价值，其中，所述 K 个评价价值与所述 K 个子策略网络一一对应；

将所述 K 个评价价值分别发送至所述 K 个网络设备；

调整所述价值网络 Q_i 中的参数，以得到用于下一次所述训练的价值网络。

37、根据权利要求 28 所述的系统，其特征在于，所述调度算法训练系统还包括集中式网元设备，当所述性能参数的值不低于所述预设值时，所述集中式网元设备用于：

将第一子价值网络 q_i 确定为第一目标子价值网络，其中，所述第一子价值网络 q_i 是基于上一次所述训练的第一子价值网络得到的，所述第一子价值网络 q_i 为 K 个子价值网络中、所述第一网络设备对应的子价值网络，所述 K 个子价值网络与所述 K 个网络设备一一对应。

38、根据权利要求 37 所述的系统，其特征在于，当所述性能参数的值低于所述预设值时，所述集中式网元设备用于：

将所述第一区域的目标状态信息 S_i 以及目标状态信息 S_{i+1}、所述第一网络设备对应的第一子策略网络 W_i 的输出结果、所述第一区域对应的反馈收益和除第一子价值网络 q_i 之外的其他 K-1 个子价值网络的信息均输入至所述第一子价值网络 q_i 进行处理，以得到所述第一子策略网络 W_i 的评价价值；其中，所述第一区域对应的反馈收益为所述第一区域内的终端根据所述第一区域对应的第二调度信息进行数据传输后确定的；

调整所述第一子价值网络 q_i 中的参数，以得到用于下一次所述训练的第一子价值网络。

39、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行以实现权利要求 1 至 7 任意一项所述的方法和/或 8 至 19 任意一项所述的方法。

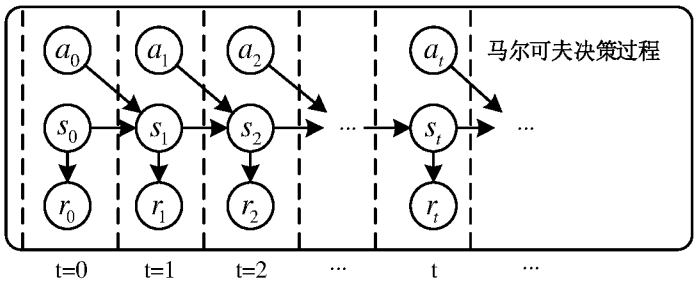
a_0
a_1
a_2
a_t
马尔可夫决策过程
s_0
s_1
s_2
s_t
r_0
r_1
r_2
r_t
t=0
t=1
t=2
t
...

图 1

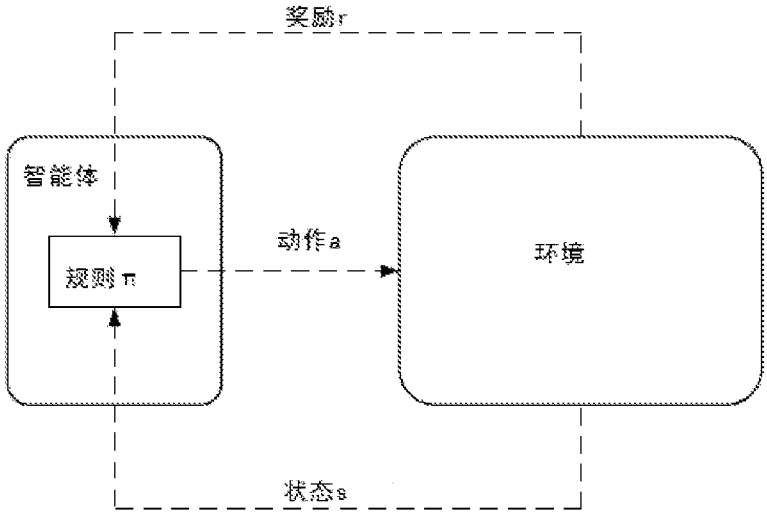
奖励r
智能体
规则π
动作a
环境
状态s

图 2

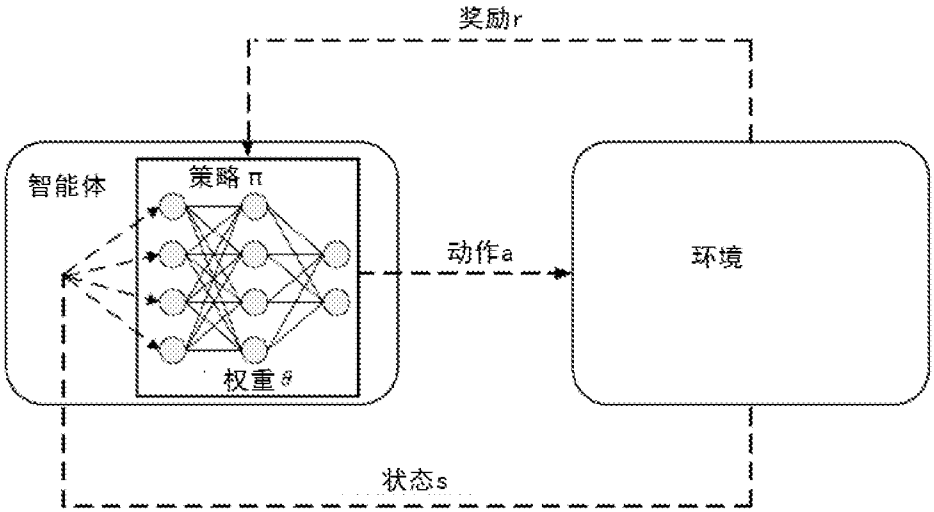
奖励r
智能体
策略π
权重θ
动作a
环境
状态s

图 3

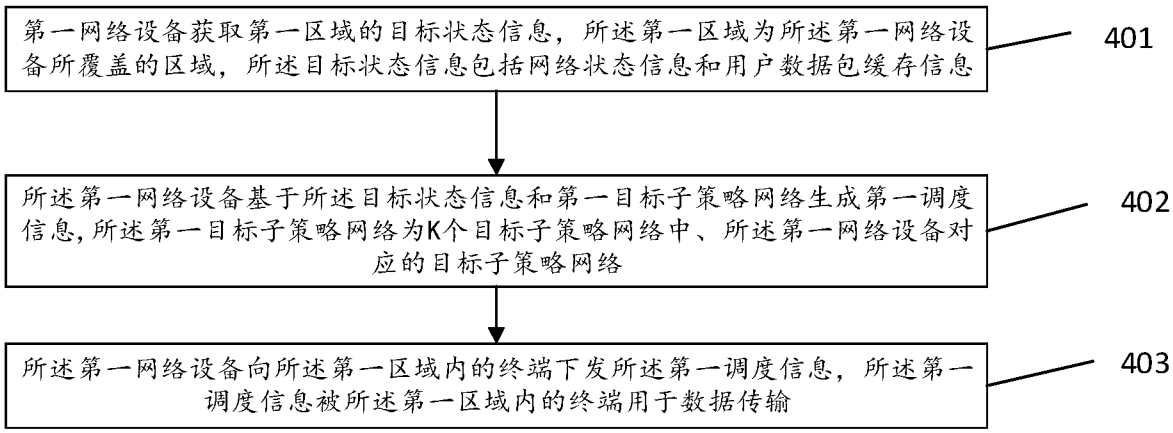
第一网络设备获取第一区域的目标状态信息，所述第一区域为所述第一网络设备所覆盖的区域，所述目标状态信息包括网络状态信息和用户数据包缓存信息
401
所述第一网络设备基于所述目标状态信息和第一目标子策略网络生成第一调度信息，所述第一目标子策略网络为K个目标子策略网络中、所述第一网络设备对应的目标子策略网络
402
所述第一网络设备向所述第一区域内的终端下发所述第一调度信息，所述第一调度信息被所述第一区域内的终端用于数据传输
403

图 4

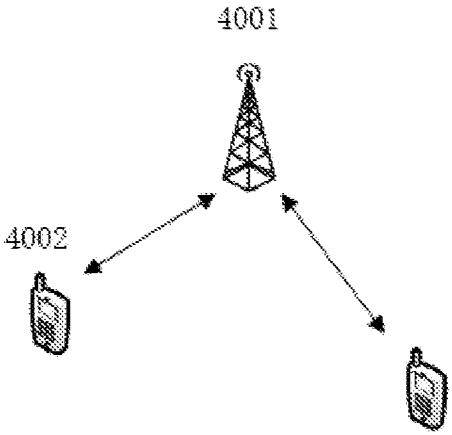
4001
4002

图 5

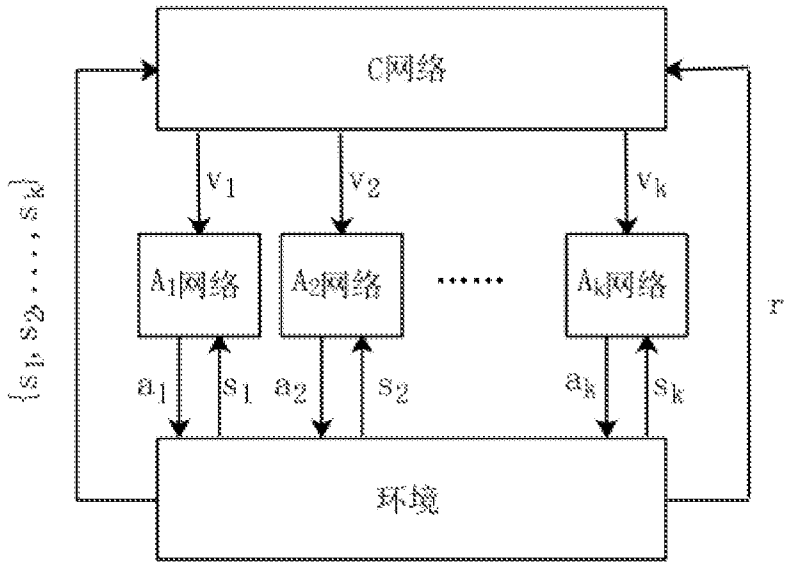
C网络
v_1
v_2
v_k
A_1网络
A_2网络
……
A_k网络
$\{s_1, s_2, \ldots, s_k\}$
r
a_1
s_1
a_2
s_2
a_k
s_k
环境

图 6

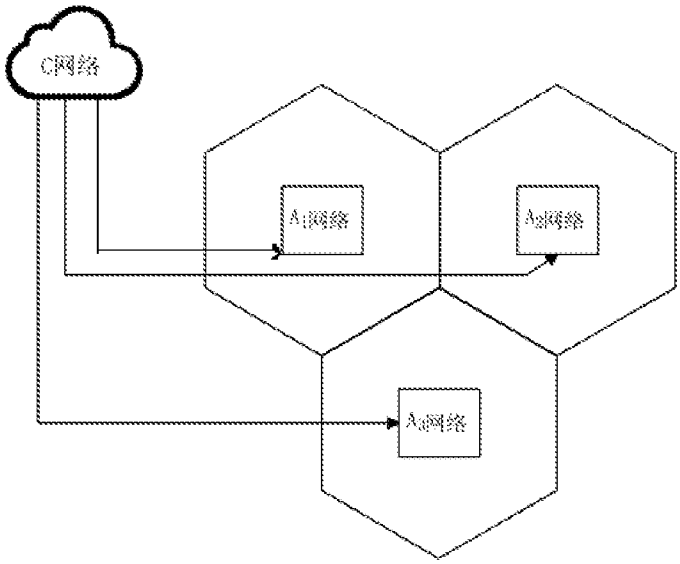

图 7

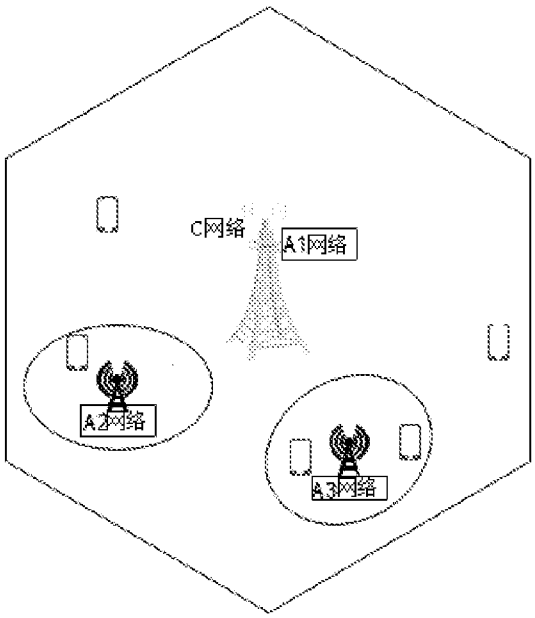

图 8

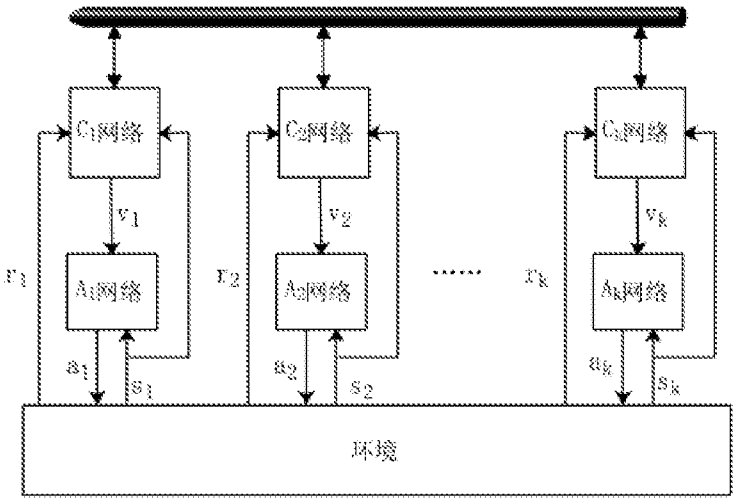
C1网络
C2网络
Ck网络
v1
v2
vk
r1
A1网络
r2
A2网络
.....
rk
Ak网络
a1
s1
a2
s2
ak
sk
环境

图 9A

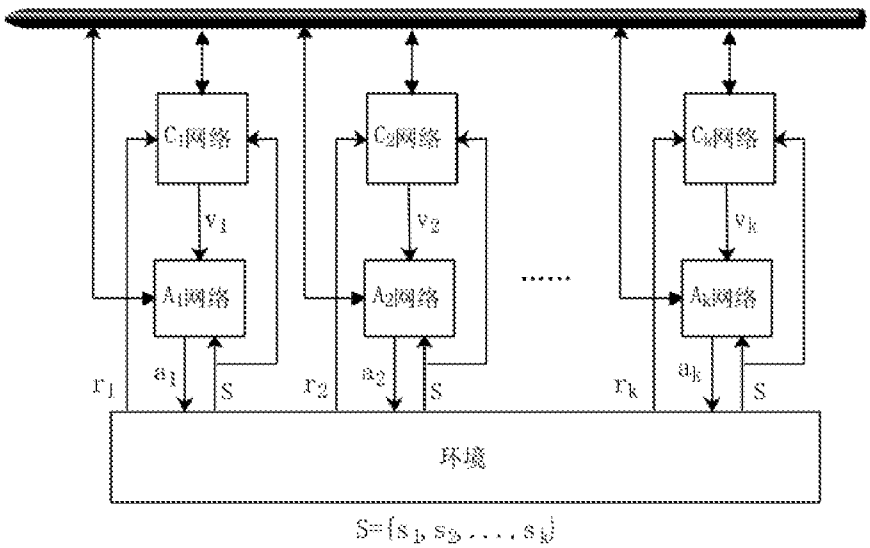
C1网络
C2网络
Ck网络
v1
v2
vk
A1网络
A2网络
.....
Ak网络
r1
a1
S
r2
a2
S
rk
ak
S
环境
S=(s1, s2, . . . , sk)

图 9B

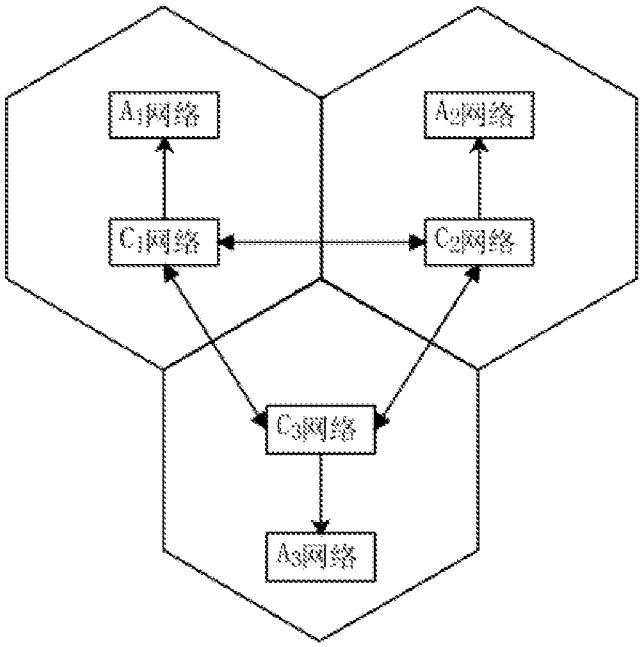

图 10

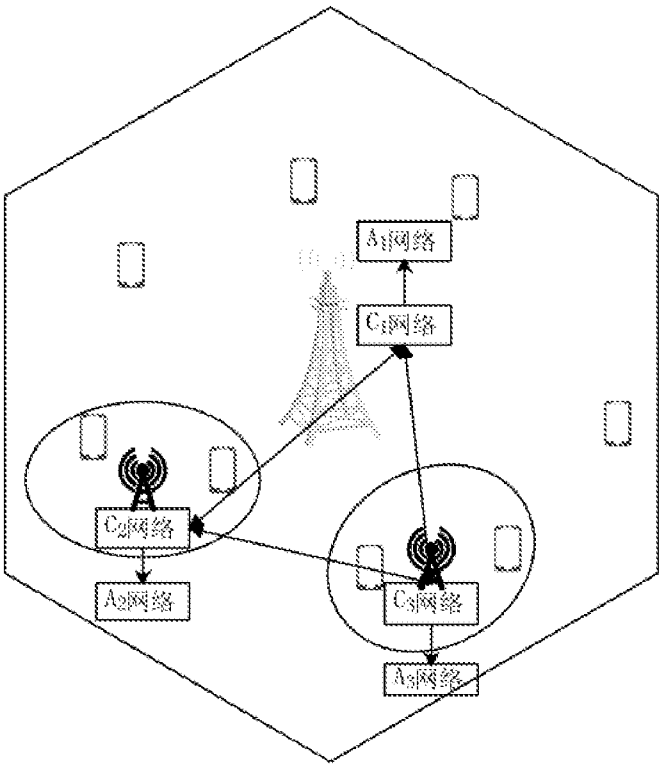

图 11

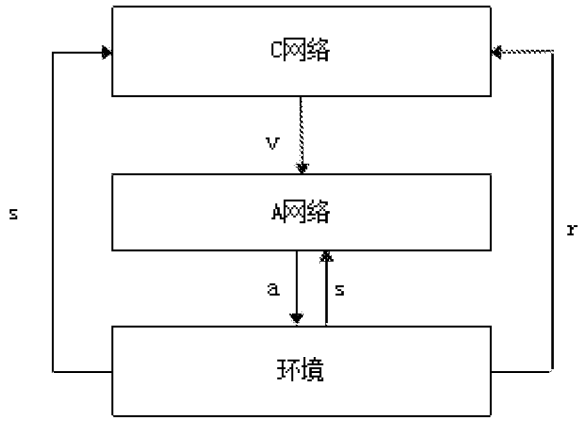

图 12

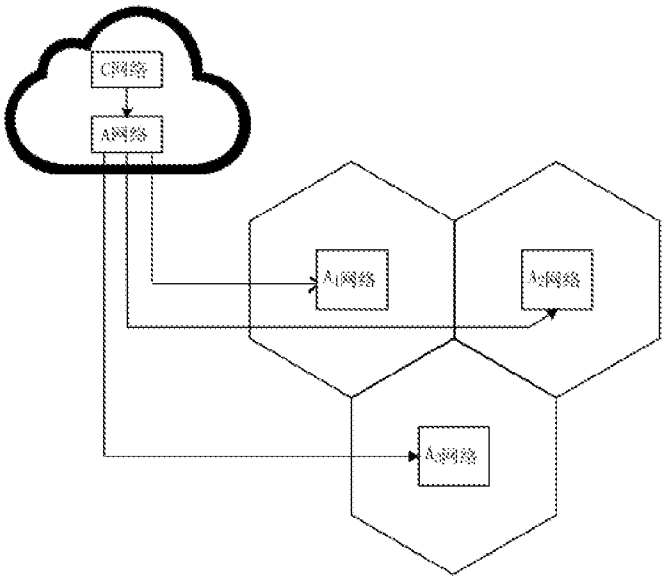

图 13

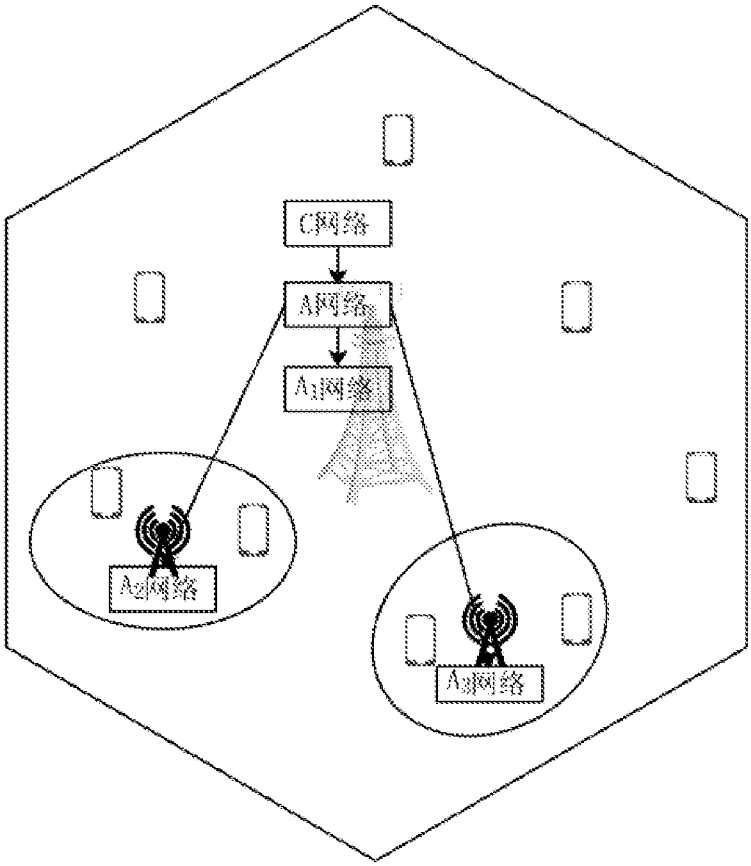

图 14

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/073764

A. CLASSIFICATION OF SUBJECT MATTER

H04W 72/12(2009.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04W; H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; VEN; EPTXT; USTXT; WOTXT; 3GPP; IEEE: 策略, 决策, A网络, 价值网络, C网络, 分布式, 调度, 训练, 迭代, 叠代, 深度强化学习, DRL, 智能体, 基站, 区域, 覆盖, 信道, 状态, 吞吐量, 缓存, strategy, policy, actor, network, critic, distribut+, schedul+, learn+, iterative, deep reinforcement learning, intelligent agent+, base station, area, cover, channel, state, throughout, cach+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110012547 A (UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA) 12 July 2019 (2019-07-12) see entire document	1-39
A	CN 110662238 A (NANJING UNIVERSITY) 07 January 2020 (2020-01-07) see entire document	1-39
A	CN 110708259 A (JIANGSU FUTURE NETWORKS INNOVATION INSTITUTE et al.) 17 January 2020 (2020-01-17) see entire document	1-39
A	CN 108966352 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 07 December 2018 (2018-12-07) see entire document	1-39
A	CN 110278149 A (NANJING UNIVERSITY) 24 September 2019 (2019-09-24) see entire document	1-39

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 April 2021	**30 April 2021**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/073764

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 110012547 A	12 July 2019	None	
CN 110662238 A	07 January 2020	CN 110662238 B	25 August 2020
CN 110708259 A	17 January 2020	None	
CN 108966352 A	07 December 2018	CN 108966352 B	27 September 2019
CN 110278149 A	24 September 2019	None	

国际检索报告

国际申请号
PCT/CN2021/073764

A. 主题的分类

H04W 72/12(2009.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04W; H04L

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS;CNTXT;CNKI;VEN;EPTXT;USTXT;WOTXT;3GPP;IEEE:策略，决策，A网络，价值网络，C网络，分布式，调度，训练，迭代，叠代，深度强化学习，DRL，智能体，基站，区域，覆盖，信道，状态，吞吐量，缓存，strategy, policy, actor, network, critic, distribut+, schedul+, learn+, iterative, deep reinforcement learning, intelligent agent+, base station, area, cover, channel, state, throughout, cach+

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 110012547 A (电子科技大学) 2019年 7月 12日 (2019-07-12) 参见全文	1-39
A	CN 110662238 A (南京大学) 2020年 1月 7日 (2020-01-07) 参见全文	1-39
A	CN 110708259 A (江苏省未来网络创新研究院 等) 2020年 1月 17日 (2020-01-17) 参见全文	1-39
A	CN 108966352 A (北京邮电大学) 2018年 12月 7日 (2018-12-07) 参见全文	1-39
A	CN 110278149 A (南京大学) 2019年 9月 24日 (2019-09-24) 参见全文	1-39

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2021年 4月 11日	2021年 4月 30日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 吴旭 电话号码 86-010-62089859

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/073764

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	110012547	A	2019年 7月 12日		无		
CN	110662238	A	2020年 1月 7日	CN	110662238	B	2020年 8月 25日
CN	110708259	A	2020年 1月 17日		无		
CN	108966352	A	2018年 12月 7日	CN	108966352	B	2019年 9月 27日
CN	110278149	A	2019年 9月 24日		无		