Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Intelligent radio access networks: architectures, key techniques, and experimental platforms*

Zeyu WANG, Yaohua SUN†‡, Shuo YUAN

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China

†E-mail: sunyaohua@bupt.edu.cn

Received June 29, 2021; Revision accepted Oct. 19, 2021; Crosschecked Nov. 8, 2021

Abstract: Intelligent radio access networks (RANs) have been seen as a promising paradigm aiming to better satisfy diverse application demands and support various service scenarios. In this paper, a comprehensive survey of recent advances in intelligent RANs is conducted. First, the efforts made by standard organizations and vendors are summarized, and several intelligent RAN architectures proposed by the academic community are presented, such as intent-driven RAN and network with enhanced data analytic. Then, several enabling techniques are introduced which include AI-driven network slicing, intent perception, intelligent operation and maintenance, AI-based cloud-edge collaborative networking, and intelligent multi-dimensional resource allocation. Furthermore, the recent progress achieved in developing experimental platforms is described. Finally, given the extensiveness of the research area, several promising future directions are outlined, in terms of standard open data sets, enabling AI with a computing power network, realization of edge intelligence, and software-defined intelligent satellite-terrestrial integrated network.

Key words: Intelligent network architecture; Artificial intelligence; Experimental platforms
https://doi.org/10.1631/FITEE.2100305 **CLC number:** TN929.5

1 Introduction

At present, with the development of communication networks, application types tend to be diversified and service scenarios become more and more complex. Moreover, with the involvement of new technologies, such as edge computing (Liu YQ et al., 2020), intelligent reflecting surface (Ding and Poor, 2020), and network slicing (Zhang HJ et al., 2017), it is challenging to effectively realize network management and optimization. To better deal with the above issues, the industry and academia have recognized artificial intelligence (AI) as one of the potential key techniques in the sixth generation (6G) era.

To promote deep integration of AI and radio access networks (RANs), researchers have proposed various AI-enabled architectures. In the space-air-ground-aqua integrated network (SAGAIN), the problems of heterogeneous network convergence, unbalanced load, and large latency should be addressed. In Liu J et al. (2020), a task-oriented intelligent network architecture for the SAGAIN was proposed to provide personalized services that can meet user needs. Through edge-cloud computing and network domain division, intelligent networking was achieved, along with reduced response latency. Specifically, a task-oriented intelligent networking requirement extraction method was designed according to different types of tasks and network conditions, thereby providing personalized networking solutions.

‡ Corresponding author
* Project supported by the Beijing Natural Science Foundation, China (No. JQ18016), the National Natural Science Foundation of China (No. 62001053), and the Fundamental Research Funds for the Central Universities, China (No. 24820202020RC11)
ORCID: Zeyu WANG, https://orcid.org/0000-0003-2372-7600; Yaohua SUN, https://orcid.org/0000-0002-8200-5010

In terms of management and control, an intelligent network management and control architecture was proposed in Yu P et al. (2020), in which the intelligent management and control unit completed the functions of perception, analysis, decision-making, execution, and network evaluation through closed-loop control. To better support the deployment of federated learning in network, Lu et al. (2020) paid more attention to data security and user privacy, and a data-sharing structure based on blockchain and federated learning was designed which incorporates local differential privacy into a gradient descent training scheme.

In contrast, some researchers have conducted extensive studies on specific AI-enabled RAN technologies. Focusing on the physical layer, He HT et al. (2019) discussed the applications of model-driven deep learning (DL) in the transmission scheme, receiver design, and channel state information recovery. In the medium access control (MAC) layer, to lower the latency in scenarios involving Internet of Vehicles, a distributed resource allocation scheme based on deep reinforcement learning (DRL) methods was proposed which can be used for both unicast and broadcast communications (Ye et al., 2019). Significant progress has also been made toward AI-enabled mobility management. By applying an unsupervised learning method, the handover rate and system throughput were well balanced (Wang et al., 2018). In addition, user positions can be predicted using supervised learning based on the historical trajectory information, which helps realize seamless handover (Yu C et al., 2017). Intelligent network slicing is a promising approach that can be used to satisfy customized service demands in a cost-efficient way. Bega et al. (2020) discussed AI applications in network slice management, which includes the scheduling of slice traffic, network resource orchestration, and admission control of slice requests. As another key technique that lowers network operating expense, AI-enabled self-optimization and self-healing of RANs is attractive. In Asghar et al. (2018), a self-healing framework for cellular networks was introduced, outlining the solutions of anomaly detection, fault diagnosis, and performance compensation.

Recent progress in intelligent RANs has encouraged researchers to survey related research and provide useful opinions. Integrating AI and network technologies, the "intellicise" wireless network operation paradigm proposed by Zhang P et al. (2022) emphasizes endogenous intelligence using semantic information and a primitive-concise paradigm. Mao et al. (2018) and Sun et al. (2019c) focused mainly on the applications of machine learning (ML) and DL in wireless networks respectively, and Zhao et al. (2020) summarized federated-learning-enabled intelligent fog radio access networks (F-RANs), including the theory, techniques, and future trends. Combining AI and RANs, Xia et al. (2020) paid attention to the interplay between AI and F-RANs, including how AI makes F-RANs smarter and how F-RANs enable AI deployment. From the data-driven perspective, an overview of AI in wireless networks, including sensing, network device, access, user device, and data provenance, was presented (Nguyen et al., 2021).

Unlike most previous reviews, here we consider the combination of AI and RANs in many aspects, and are not limited by key techniques and specific AI tools. First, the progress from industry on intelligent networks is outlined, and the network architectures proposed by the academic community are presented. Then the key techniques related to intelligent RANs are summarized. Furthermore, two experimental platforms are presented that facilitate the implementation of intelligent networks, on which network functions are demonstrated and performance advantages are evaluated. Finally, we discuss the future challenges related to the intelligent RANs.

2 Industrial progress

In this section, we present state-of-the-art advances in intelligent networks deployed in the industry. The standardization organizations including the 3rd Generation Partnership Project (3GPP), International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), European Telecommunications Standards Institute (ETSI), China Communications Standards Association (CCSA), International Mobile Telecommunication 2020 (IMT-2020), and China Institute of Communications (CIC) have actively promoted the development of AI in communication networks. Major organizations, equipment vendors, and operators also take the initiative to conduct research on intelligent RAN architectures.

2.1 3GPP

In 2017, the network data analysis function (NWDAF) was introduced into the core network (CN). Later on, in 2020, it was further enhanced to collect data from other fifth generation (5G) functions, and the results were fed back to the network functions requesting data analysis service for network management and optimization (3GPP, 2019a). In the RAN domain, 3GPP RAN3 launched the RAN-centric research work for data collection and application, which is oriented toward the automation and intelligence of RANs. It provides data support for various AI applications, such as AI-enabled mobility management. In June 2020, 3GPP RAN3 approved the research project of "Study on Further Enhancement for Data Collection" to further extend AI functions to RANs, aiming at improving network energy saving, load balancing, mobility management, and coverage optimization (3GPP, 2020). As for SA5, it began to study AI in 2018 and defined a new management function: management data analytics function (MDAF) (3GPP, 2019b). In 2020, research on the enhancement of management data analytics service was completed, rendering that it is feasible to apply AI technology to data analysis.

2.2 ETSI

In 2017, ETSI established an Experiential Networked Intelligence (ENI) group, which is committed to providing intelligent services for network operation and maintenance, network security, and equipment management (ETSI, 2017). To achieve the joint orchestration of network resources and network services based on service level agreement, the logical and physical connections between AI and network functions should be clarified. To this end, ETSI ENI provided a reference architecture in 2019, in which the functions of the environment awareness component, data processing component, knowledge management component, and policy management component were described, along with the related application programming interface. Recently, ENI has defined more advanced applications, such as energy saving based on the intent network and data processing mechanisms. Based on the concept of intelligence-defined network, mobile intelligent network decision-making entities were introduced for the RAN domain, each of which is responsible for data collection, data analysis, data modeling, decision making, and verification. Afterwards, ETSI proposed the zero-touch network and service management architecture, in which the AI-driven domain intelligence was introduced to realize intelligent closed-loop automation. Moreover, focusing on AI security, ETSI established the Industry Specification Group on securing AI, which includes three aspects: using AI to enhance security, mitigating against attacks that use AI, and protecting AI itself from attacks.

2.3 ITU-T

In 2017, ITU-T established the Focus Group on Machine Learning for Future Networks including 5G (FG-ML5G). In June 2019, FG-ML5G described 5G intelligent application scenarios and requirements. In terms of the AI functional framework and process, the framework standards of ML and ML-enabled data processing were released in 2019 and 2020, respectively. In February 2020, FG-ML5G also released the framework for evaluating the intelligence level of future networks (ITU-T, 2020). In addition, owing to the influence of ITU-T, some technical specifications have been translated into standards, including the end-to-end network slicing management framework based on ML in the multi-domain environment and the design of the ML function orchestrator.

2.4 CCSA, IMT-2020, and CIC

Since 2019, CCSA has carried out the research on AI, including mainly the intelligent capability classification of mobile communication networks, applications of AI and big data in wireless communication, and 5G intelligent CN slicing technology. In November 2020, CCSA TC5 held a Network Intelligence Seminar to promote the integration of 5G and AI. In 2020, TC3WG1 launched the project of Intelligent Communication Networks Based on Software Defined Networks (SDN)/Network Function Virtualization (NFV) and Intent-Based Networking Architectures. In July 2019, IMT-2020 (5G) released the white paper "Intelligent Slicing Management and Collaboration Based on AI," which outlined the typical applications and requirements of intelligent slicing, together with an intelligent network slicing architecture. In 2020, CIC organized the "From Cloud

AI to Network AI: Building 6G Network Architecture" Seminar and established the 6G Alliance of Network AI (6GANA), focusing on 6G networks and AI technologies.

2.5 Huawei

In 2020, Huawei introduced the solution to the autonomous driving network (ADN) (Huawei, 2020). Fig. 1 shows the architecture which involves real-time perception and AI inference capabilities. Through a local knowledge library and an AI inference framework, the network management and control unit automatically converts upper-layer services and application intentions into network operations, thereby achieving single-domain autonomy and closed-loop management. In addition, the data interaction between the network management and control unit and the cloud can continuously enrich the local AI model and knowledge base, which can optimize and enhance the local intelligent perception and decision-making capabilities. In the cloud, the AI network unit is responsible for the continuous training of the AI models, along with the extraction of the network data that is aggregated at the cloud. Through unified management, the complete sharing and reuse of the AI models and the knowledge base can be realized, hence reducing the need for repetitive training.

2.6 O-RAN

Under the trends of endogenous intelligence and network openness, the O-RAN Alliance, in which many operators participate, has proposed an open intelligent wireless network reference architecture (RAN Alliance, 2018). In this architecture, an AI-enabled software-defined RAN intelligent controller (RIC) is designed to realize embedded intelligence. RIC includes the non-real-time part and the near-real-time part. The main goal of the non-real-time RIC is to support non-real-time intelligent radio resource management, higher-layer process optimization, and strategy optimization of RAN. Near-real-time RIC is responsible for load balancing, radio resource block management, interference management, etc. At the same time, it provides new functions that use embedded intelligence, such as quality of service (QoS) management and seamless handover control. As shown in Fig. 2, through the A1 interface, the non-real-time RIC collects data from the central unit (CU) and distributed unit (DU), and distributes trained AI models to the near-real-time RIC. In addition, through the open E2 interface (between the near-real-time RIC, multi-RAT CU protocol stack, and RAN DU), the near-real-time RIC can not only obtain the near-real-time network conditions but also issue configuration commands.

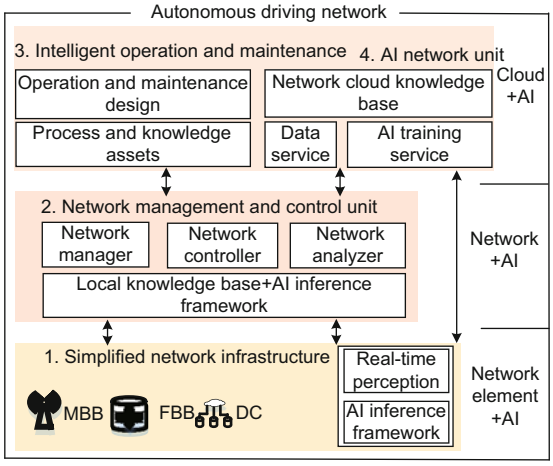

Fig. 1 Solution to the autonomous driving network (Huawei, 2020)

3 Academic progress

3.1 Intelligent and concise RAN

To adaptively meet the differentiated requirements of various communication application scenarios, an AI-enabled 6G intelligent and concise radio

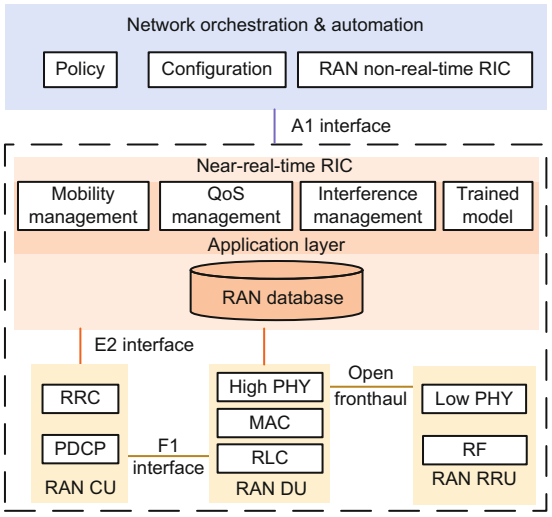

Fig. 2 O-RAN architecture (RAN Alliance, 2018)

RAN: radio access network; RIC: RAN intelligent controller; QoS: quality of service

access network architecture was proposed (Peng et al., 2020), which features the air-space-ground-underground integrated networking and the collaboration of communication, computing, caching, and control resource (Fig. 3). Moreover, it achieves deep integration of AI and a flexible network reconfiguration, along with the decoupling of network functions from dedicated hardware, with the aid of network slicing, SDN, and NFV technologies.

To provide wide area seamless communication in remote areas, in the space and terrestrial ground communication layer, real-time wireless signal processing and resource management can be implemented on satellites and unmanned aerial vehicle (UAV) platforms. As communication nodes, they can share resources and cooperate with each other through wireless links. In addition, shipborne base stations (BSs) support the maritime communication service. In the terrestrial cellular mobile communication system, to suppress serious interference in the high-capacity hotspots, the intelligent and concise network adaptively chooses the cloud RAN mode, which can achieve high capacity by massive remote radio units and centralized baseband processing. For the intelligent manufacturing scenario, the distributed computing power at edge can be exploited through cloud-edge collaboration ability, and the collaborative computing of multiple fog access points (F-APs) can help execute computation-intensive and latency-sensitive industrial applications.

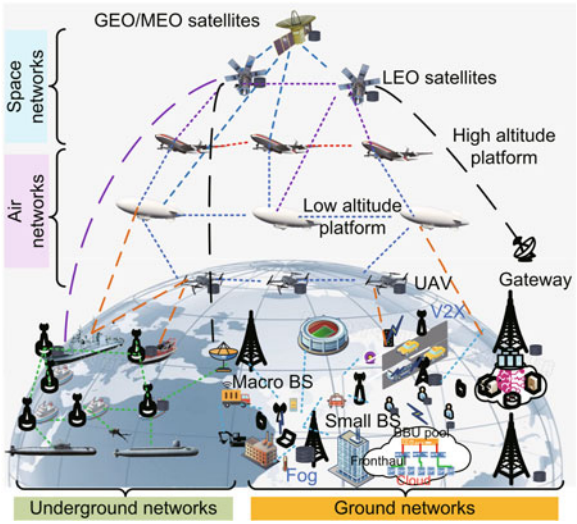

Fig. 3 Intelligent and concise RAN architecture (Peng et al., 2020)

BS: base station; RAN: radio access network; UAV: unmanned aerial vehicle

3.2 Intent-driven RAN

In Zhou et al. (2020), an intent-driven radio access network (ID-RAN) was proposed. As shown in Fig. 4, the intent-driven radio network controller (ID-RNC) is deployed at the centralized cloud, the BS controller, and the macro BS, and it is responsible for capturing network operation and maintenance data, wireless transmission data, and terminal data, and for issuing the networking configuration and optimization instructions to network entities. Each intent received by ID-RAN will sequentially go through five modules in its life cycle, namely, intent translation, conflict resolution, network orchestration, configuration activation, and strategy optimization.

1. Intent translation

The intent generated from the operator or operation and maintenance management office, including networking intent, performance intent, and business intent, can be identified and extracted through natural language processing (NLP). Then, based on networking experience, the intent will be transformed into corresponding network configuration statements. Specifically, networking intent can be expressed in the form of control commands, while business intent and performance intent can be transformed into optimization problems with additional constraints through mathematical modeling.

2. Conflict resolution

Before network configuration requirements enter the orchestration module, the conflict resolution module needs to determine the entry sequence according to the type of intent corresponding to the network configuration requirements. Intents are characterized by varying priorities. In general, networking intent has the highest priority, and performance intent and business intent have lower priority. For network configuration requirements transformed from the same type of intent, multi-objective joint optimization can be designed, in which network resources can be rationally allocated by AI-based algorithms, so as to meet as many intent requests as possible.

3. Network orchestration

The structured network configuration requirements outputted from the intent translation module can be used as reference objectives for network optimization, and based on the multi-dimensional network environment data, networking strategies of

RAN that match the intent needs can be derived with the aid of DRL.

4. Configuration activation

Through SDN, NFV, and other technologies, networking strategies can be transformed into control commands, which can then be used for the programmatic configuration on the function and parameters of the RAN infrastructure.

5. Strategy optimization

After the implementation of intent networking, network performance is monitored in real time for index evaluation and failure prediction. If the difference between the expected performance and the actual feedback performance is too large, the current network configuration scheme should be further optimized through DRL.

3.3 Intelligent endogenous trusted network

In AI-enabled RANs, data collected by the massive network entities is facing the security problem of data leakage. As a distributed ledger, blockchain can help solve hidden security problems. Blockchain is responsible for tracking and recording information and data sharing among network entities, which can then maintain a safe and credible network ecology. In El Azzaoui et al. (2020), blockchain and AI technology were combined, and an intelligent and secure architecture "Block5GIntell" for data analysis was proposed. As shown in Fig. 5, the architecture is divided mainly into four layers, namely, device layer, access layer, fog layer, and cloud layer.

In the device layer, blockchain is responsible for collecting a large amount of private personal information and storing it anonymously in the real-time sharing ledger to ensure its security. In the access layer, the data such as network status information will also be stored in the blockchain, which is generated by macro BSs, small BSs, and BS controllers.

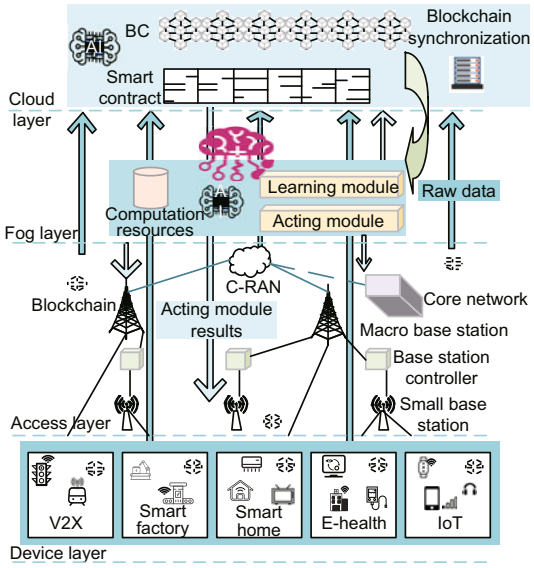

Fig. 5 Intelligent endogenous trusted network architecture (El Azzaoui et al., 2020)

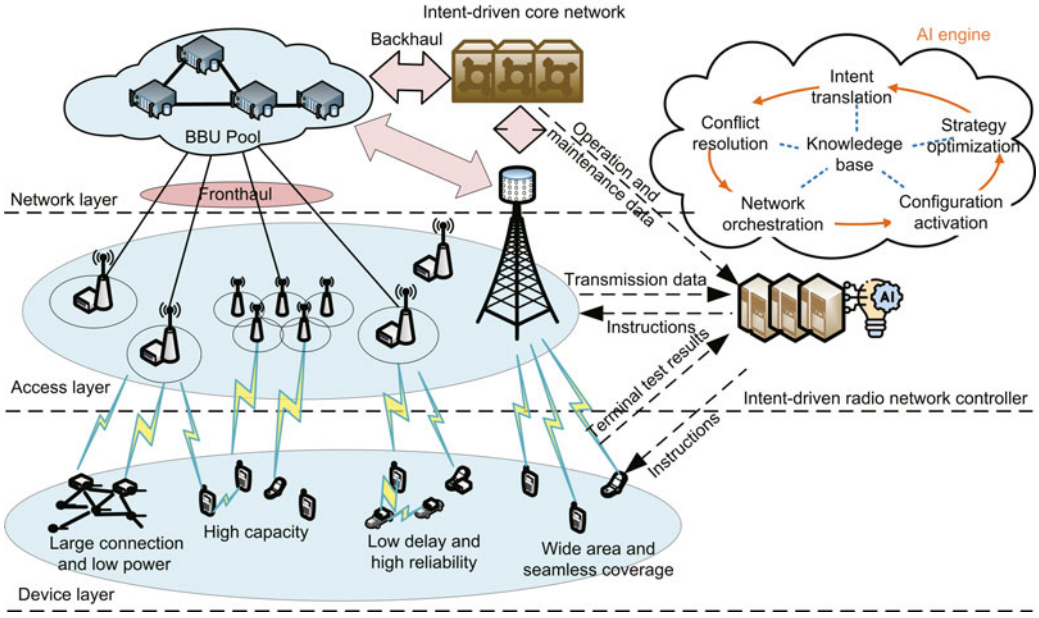

Fig. 4 Intent-driven radio access network architecture (Zhou et al., 2020)

In the fog layer, the blockchain contains computing resources and AI-driven fog nodes with learning and execution capabilities. Through centralized learning and feature analysis of the data stored in the cloud layer, network organization and planning is strengthened, and the generated strategies are fed back to the access layer for execution.

In the cloud layer, the blockchain is responsible for centralized storage of all data forwarded from other layers. Meanwhile, a blockchain synchronization module is added to ensure the real-time transmitted data updating. AI takes charge of organizing and clustering the scattered data in the blockchain to facilitate efficient use in the fog layer. The strategies that have been successfully outputted from the fog layer will also be transmitted back to the cloud layer and stored in the form of smart contracts, which can be automatically executed in the network without passing through the fog layer.

3.4 Enhanced data analytic network

The service-based architecture (SBA) for 5G has been specified in the 3GPP, and service-based interface has been introduced to expose the service of control plane function, including network data analytics service (NWDAS) and management data analysis service (MDAS). On this basis, an enhanced integrated data analysis framework was proposed (Pateromichelakis et al., 2019) to perform multi-level data analysis in different domains and configure or parameterize end-to-end analysis function in a slice-customized manner. As shown in Fig. 6, to make data analysis for real-time operation and allow customers to closely manage some operations vertically, the framework integrates new data analytics functionality (DAF) blocks in RAN, data network (DN), and application function (AF) domains, corresponding to RAN-DAF, DN-DAF, and AF-DAF, respectively.

To achieve rapid decision-making in RAN, real-time analysis function is required to be directly performed locally. Within RAN, RAN-DAF can fulfill the role of control and management, which can be achieved through the inter domain message bus interface. The operation, administration, maintenance (OAM) system can provide an RAN configuration for RAN, in which MDAS uses network management data to make corresponding analysis to improve the deployment and optimization of the network slice. In addition, performance management (PM) and fault management (FM) are responsible for performance improvement and failure prevention, respectively.

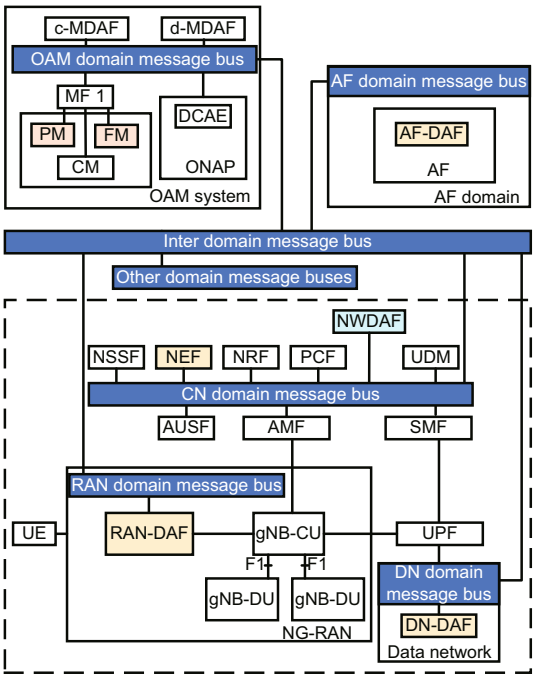

Fig. 6 Enhanced integrated data analysis framework (Pateromichelakis et al., 2019)

CN: core network; FM: fault management; NWDAF: network data analysis function; PM: performance management

3.5 AI agent based software-defined network

Large-scale access by mobile devices imposes a huge communication burden upon the network, and with merely the traditional service resource scheduling and allocation mechanism, it is difficult to meet the growing demands of users. To address the above problems, Cao et al. (2020) defined AI agents in SDN. Through the deployment of AI agents in the device layer, BS layer, and SDN controller layer, network service prediction, resource scheduling, and other functions are easy to implement.

As shown in Fig. 7, the main task of the AI agent in the user device layer is the perception and integration of the bottom history environmental information, such as the application resource consumption, terminal consumption, used application types, and user locations. Then, by applying AI models (such as the neural network, long short-term memory network, and support vector machine), the resource type and quantity requested by users in a future period can be predicted. In the BS layer, the

AI agent can effectively schedule communication resources according to the user resource requests and the available resources of resource providers. On one hand, communication resources are used mainly for edge servers to provide services for users. On the other hand, resource providers sink resources from the remote server to the edge server. For the decision-making of communication resource scheduling, the learning process based on trial and error can be adopted, and at the same time, the decision information in each environment can be stored in the knowledge base, which conveniently allows ready improvement of the learning process performance.

The AI agent in the edge server layer is responsible mainly for the resource deployment on the edge server. Specifically, by comprehensively considering the concrete location of users, resource requests, usage and quantity of application, and computing resources in the edge server, the AI agent can make decisions related to mobile service caching and offloading, thus reducing the latency of data service acquisition.

4 Key techniques

4.1 AI-driven network slicing

To adapt to diversified business scenarios, with the help of SDN and NFV, network slicing technology can divide a substrate physical network into multiple independent logical networks (Xiang et al., 2017). Considering the complexity of slicing network resources, AI-driven network slicing technology is attractive, and can handle two key issues, namely, the allocation of communication and computing resources and the scheduling and deployment of network functions, with the advantageous results that the capability of responding to the dynamic network environment on demand is gained and that there is an improvement in the utilization efficiency of network resources. Furthermore, reasonable resource allocation strategies can be generated by the slice manager to realize the differentiated customization of network functions and ensure the service capabilities of slices. As for resource allocation, based on the RAN slicing framework, the problem of content caching and mode selection optimization was formulated in Xiang et al. (2020), which considers the differences of user demands and resource constraints. The real-time content caching and user mode selection strategy can be obtained by DRL in the centralized cloud, which uses historical data to learn content popularity. Compared to other slicing solutions, the DRL method can help improve the cache hit rate and maximize system performance.

4.2 Intent perception and translation

As the core procedure in the intent-driven RAN, the intent translation process transforms wireless

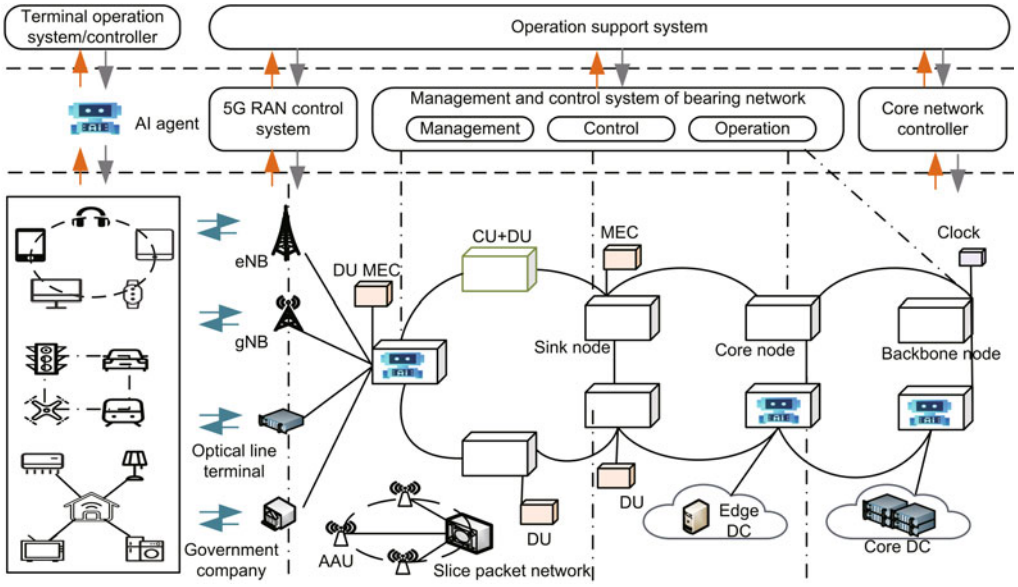

Fig. 7 System architecture of the artificial intelligence agent based software-defined network (Cao et al., 2020)

network intent into corresponding network configuration strategies according to current network conditions, realizing the accurate identification and transformation of the dynamic intent: (1) By the aid of "entity recognition" belonging to the lexical analysis method in NLP, along with segment, part-of-speech tagging, and dictionary query processes, the received wireless network intent can be used to identify the keywords, which would be the same as, or related to, intent words compared with the network knowledge base. (2) The extracted keywords are classified according to the tag categories of the key elements in the intent translation language model. As shown in Fig. 8, the intent language model is an expression of intent, which is presented in the format of objects-operations-results. Specifically, for the current infrastructure and network resources (objects), corresponding networking strategies (operations) are formulated to achieve the users' expected business demands (results). (3) According to the expressions of key elements in the network knowledge base and the mapping relationship between the elements, the sequence-to-sequence model in NLP can be applied. With the aid of the recurrent neural network (RNN) coding and decoding framework, the intent keywords are converted into structured configuration statements. The statements can be described in the form of network optimization problems, including three elements: optimization objectives (performance indicators), optimization objects (network resources), and constraints (network resource and space-time constraints). (4) These problems can be solved quickly based on DRL models, which are trained from the historical configuration experience stored in the network knowledge base, and then the corresponding networking strategies can be generated.

4.3 Intelligent operation and maintenance

To reduce the cost of network operation and maintenance, AI-enabled fault diagnosis and recovery techniques need to be introduced to solve anomaly more quickly and efficiently, such as weak coverage, signal interruption, and strong interference. In contrast with the active detection method of sending signaling to the network to monitor network faults, a passive fault identification and location method was proposed by Srinivasan et al. (2019) to directly capture the packet loss rate, round trip time, and other traffic characteristics from the network, thus avoiding additional communication delay and overhead. The traffic behavior captured was trained through ML algorithms of support vector machine, multi-layer perceptron, and random forest. In addition, the performance advantages of this method were verified through comparative experiments; it can improve the fault detection accuracy to up to 97% and greatly reduce the time for fault location.

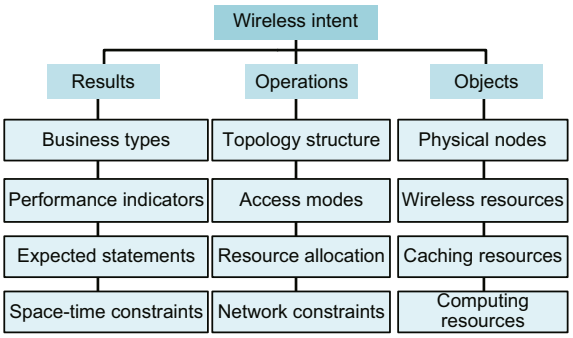

Fig. 8 Wireless intent language model

Aiming at the problem of data scarcity and difficulty in obtaining effective labels, Wu et al. (2020) proposed a fault diagnosis method based on unsupervised deep transfer learning. Through the convolutional neural network, data features can be extracted from the source and target domains and classified, and these features are then sent to the domain adversarial neural network to learn domain-invariant features and distribution, so as to solve the problem of unlabeled data classification. Compared with the method of directly using labeled data from the source domain for supervised learning, this method can effectively improve the accuracy of fault detection.

4.4 AI-based cloud-edge collaborative networking

As a promising network architecture, F-RAN can make full use of local caching and resource management capabilities at edge devices, which significantly reduces end-to-end latency and relieves the fronthaul load (Peng et al., 2016). To improve the spectrum efficiency of F-RANs, Sun et al. (2019b) proposed a method of joint cache and radio resource management based on game theory and reinforcement learning (RL) to raise the opportunity of joint transmission among F-APs. However, since there is no explicit objective expression for long-term cache resource optimization, a model-free multi-agent RL

based caching scheme has been designed. With collected historical channel information and user service requests, the cloud uses the caching scheme to approach the optimal caching strategy, based on which contents are pushed to each F-AP at a large time scale. On a small time scale, F-APs are self-organized into cooperative clusters and the interference among F-APs is significantly reduced by time division based on non-coherent joint transmission.

Considering that F-RAN users can select different communication modes (i.e., each user can acquire services by accessing multiple remote radio units in C-RAN mode or operate in device-to-device mode to directly enjoy local services), Sun et al. (2019a) proposed a joint mode selection and resource management method based on DRL to help users realize adaptive mode selection. With DRL, the network controller in the cloud took current user communication modes, on-off states of baseband units, and cache state at each user device as the input of the deep Q-network. Based on the output Q-values, intelligent decisions were made on user communication mode selection and the on-off control of baseband units. Simulation results showed that the proposal can reduce the system power consumption and has significant advantages compared with other methods, which may help alleviate the problem of tremendous power consumptions faced in AI-enabled F-RAN.

4.5 Intelligent multi-dimensional resource allocation

To better meet the needs of applications such as ultra-low latency and ultra-dense connections with limited resource, more intelligent optimization methods for multi-dimensional resource allocation are desirable. There is a need to consider the time scale characteristics of resources in different dimensions and the differences in resource granularity. In Chen et al. (2019), under the dynamic changes in channel quality and computing resources, the problem of random offloading in sliced RANs was modeled as a Markov decision process. Double deep Q-network algorithms were designed to obtain the optimal computation offloading strategy and allocate the integrated computation and communication resources, so as to maximize the cost of energy consumption and latency. In addition, they designed an online deep state-action-reward-state-action based RL algorithm. Experimental results showed that the proposed algorithms achieve the best offloading performance.

To overcome the problems of high energy consumption in industrial Internet of Things (IIoT) and high complexity of traditional computing offloading methods, an F-AP selection method based on multi-agent DRL was proposed in Ren et al. (2021), which aims at minimizing the system energy consumption. With the trained DRL model for each IIoT device, each F-AP can identify the appropriate F-AP for its currently associated device by taking dynamic computation task requests and channel states into account. After that, a low-complexity greedy algorithm was performed at each F-AP to decide the offloading requests that need to be further offloaded to the cloud. Simulation results showed that the proposed method can achieve the lowest system energy consumption.

5 Experimental platforms

5.1 FlexRAN

Issa et al. (2019) reported an experiment of network slicing management by running Mosaic 5G FlexRAN software on an OpenAirInterface (OAI) platform. OAI is an RAN technology simulation platform that is able to implement 4G and 5G RAN protocol stacks together with 4G evolved packet core as well as 5G core functions. Being a flexible and programmable platform for software-defined RANs, FlexRAN decouples the user plane and control plane, and the control plane is further consolidated into a centralized controller, which is named the real-time controller (RTC). The RTC can perform coordinated intelligent control of multiple RANs and support real-time RAN control applications. Those applications can be developed using RTC SDK, allowing RAN infrastructure monitoring and coordinated control. As a local agent of RTC, RAN runtime is responsible for virtualizing resources in the underlying RAN and providing SDK that supports distributed control applications. The FlexRAN protocol is used to realize the interactions between RTC and RAN runtime.

To test the RAN slicing function, as shown in Fig. 9, two CNs and an IP multimedia subsystem (IMS) server are virtually deployed in a personal

computer (PC). Two user terminals are connected to the OAI BS, and the BS is connected to OAI CNs and the ClearWater IP IMS server through an Ethernet cable (S1-flex). To realize intelligent resource management and network slicing, the FlexRAN RTC is deployed on the PC running the OAI BS. The WiFi-AP is physically connected to the CN, connecting to the ClearWater IMS server to access the Internet. To transmit audio and video streams on the IP network, the Session Initiation Protocol (SIP) is used to connect the ClearWater IMS server.

To test the IMS server, the SIP client application is installed on the UE side, and the Zoiper software IP phone is used. Through RFBENCHMARK, which is an application detecting mobile networks, the existence of two CNs and their performance can be detected to verify the good deployment of the network. In addition, the Iperf testing tool is used to generate Transmission Control Protocol (TCP) streams to measure the network throughput, and User Datagram Protocol (UDP) streams are used to measure the delay jitter.

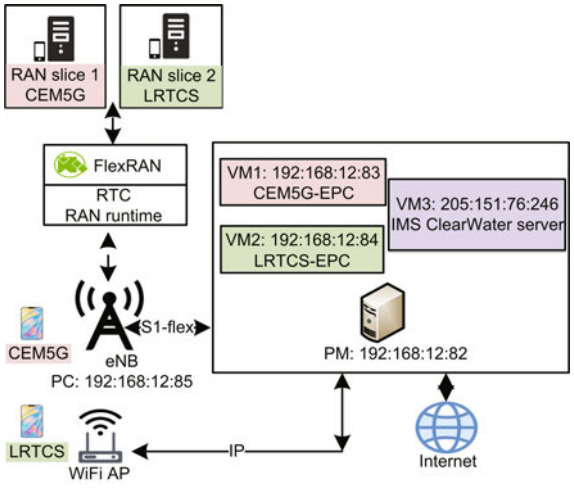

Fig. 9 System structure of the experimental scenario (Issa et al., 2019)

RAN: radio access network; RTC: realtime controller

5.2 O-RAN testbed

In Bonati et al. (2020), based on the architecture proposed by O-RAN, data-driven closed-loop control was integrated and demonstrated. Through the O-RAN open interface, the data at the edge of the network was collected, and the xApp deployed at the RAN near-real-time RIC can realize the optimization of scheduling strategies for network slicing. Specifically, experiments were conducted on Colosseum, which is the world's largest closed-loop (including hardware) wireless network simulation platform, including 128 computing nodes, and is equipped with USRP X310 SDRs that can run on a general protocol stack. In addition, Colosseum contains a data center with the storage capacity of 900 TB and the data processing capacity of 52 TB/s, which is used for large-scale data processing and ML algorithm testing in heterogeneous networks.

In Fig. 10, a 5G cellular network in a dense city scenario is simulated on the Colosseum, where there are four BSs. Furthermore, each BS is divided into three slices, which are responsible for three kinds of services of eMBB/URLLC/MTC generated by the Colosseum traffic generator. These three

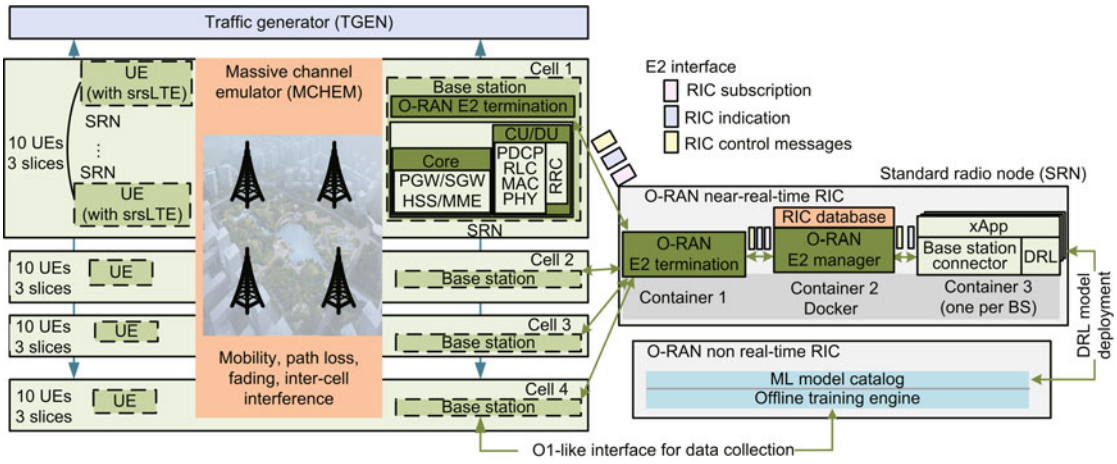

Fig. 10 O-RAN integration in Colosseum (Bonati et al., 2020)

DRL: deep reinforcement learning; ML: machine learning; RAN: radio access network; RIC: RAN intelligent controller

service slices serve 3/3/4 users respectively within each cell. After 63 h of continuous experiments on the Colosseum platform, 6-GB training data is generated. Through the O-RAN E2 interface, the DRL agents can obtain the real-time performance testing results of the slice, including RIC subscription, indication, and control messages. Then data dimensionality is further reduced by an encoder, and the agent can recognize the system state based on the output. Based on a fully connected neural network with five layers and 30 neurons in each layer, the DRL agent can select its optimal scheduling strategy for each RAN slice. The experimental results of the platform show that, in comparison with the dedicated scheduling strategies of round robin, proportionally fair, and waterfilling, DRL-based data-driven optimization can achieve higher eMBB spectrum efficiency, and that the gain is as high as 20%. As for the URLLC slice, DRL can achieve the best performance and reduce the buffer occupancy rate by 20% in the downlink of the BS.

6 Future directions

6.1 Standard open data set

To facilitate the research community to study and fairly evaluate the performance of key techniques for RAN intelligence, a standard and open data set is essential. Intuitively, it is more convenient for mobile operators and vendors to collect real network data from their operational support systems and equipment databases. Although massive data sets contribute to higher model training performance, storage of huge volumes of data and their pre-processing are time-consuming. In addition, data sets should be continuously updated to cope with network dynamics and new scenarios. If not, the inconsistent data distribution between the training samples and the latest network situations can degrade AI model performance significantly. In the case where real network data cannot be easily obtained or the amount of the data obtained is insufficient, the schemes of generating high-quality training samples need to be investigated, including the way of establishing a realistic simulation environment based on digital twin (Sun et al., 2021).

6.2 Enabling AI with a computing power network

To better satisfy the huge demand for computing power to better support AI model training within RANs, one can use the core idea of computing power network, which is gathering idle computing power through the network and implementing the global management and scheduling of it (He T et al., 2020). At present, research on the computing power network is popular in the industry and many related standards have been set up. In November 2019, China Unicom issued a white paper on the computing power network, which expounds the view of integrating computing and networks (China Unicom, 2019). In June 2021, at the 5$^{\text{th}}$ Future Network Development Conference, China Unicom pointed out four developing directions of the computing power network, namely, satisfying high requirements of high performance metrics, perception and intelligent scheduling of computing power, flexible deployment of network functions, and simplification of the network architecture.

6.3 Realization of edge intelligence

Edge intelligence moves AI services from the remote data center to the network edge, which raises the intelligence level of network edge and reduces AI service latency. However, since it is difficult for edge devices to gather enough data, few-shot learning with quick learning ability is needed. In addition, federated learning is an effective paradigm to mix local AI models trained at edge devices. In this way, a model with a better global performance can be obtained, and meanwhile, the data privacy of each edge device is preserved. Currently, there have been many works discussing the performance improvement of federated learning in RANs, in terms of model compression, device scheduling, and training strategies. Finally, to achieve efficient deployment and migration of edge AI services, further study needs to be done on the integration of virtualization technology and AI frameworks, such as TensorFlow, Torch, and Caffe.

6.4 Software-defined intelligent satellite-terrestrial integrated network

With software-defined networking and AI, software-defined intelligent satellite-terrestrial

networks can provide much flexibility in networking and resource management. However, for a large-scale satellite-terrestrial network, relying on a single centralized SDN controller for network management and control results in the need to satisfactorily tackle the issues of single point fault and large control plane latency (Yuan et al., 2021). Therefore, it is necessary to design a distributed multi-layer management architecture for future satellite-terrestrial networks, to realize collaborative management and control with a coarse-grained manner for the global network and a fine-grained manner for each regional network. In addition, AI can help make intelligent decisions on virtual network function deployment among satellites and terrestrial nodes, as well as routing selection by taking the dynamics of network topology, resource limitation at satellites, and differentiated service QoS into account.

7 Conclusions

To efficiently support differentiated business scenarios and give full play to the performance advantages of RANs, RAN intelligence has attracted more and more attention. In line with this trend, we make a comprehensive introduction of the research progress from the view of industry and academia, which includes standardization progress, architectures, enabling techniques, and experimental platforms. Finally, we put forward deep thoughts on the development in this direction.

Contributors

Yaohua SUN outlined the paper. Zeyu WANG and Shuo YUAN collected the materials. Zeyu WANG drafted the paper. Zeyu WANG, Yaohua SUN, and Shuo YUAN revised and finalized the paper.

Acknowledgements

The authors would like to give special thanks to their colleague Xiqing LIU for his valuable suggestions on paper organization.

Compliance with ethics guidelines

Zeyu WANG, Yaohua SUN, and Shuo YUAN declare that they have no conflict of interest.

References

3GPP, 2019a. TR23.791 V16.2.0: Study of Enablers for Network Automation for 5G.

3GPP 2019b. TR28.805 V1.1.0: Study on Management Aspects of Communication Services.

3GPP, 2020. TR28.809 V0.3.0: Study on Enhancement of Management Data Analytics (MDA).

Asghar A, Farooq H, Imran A, 2018. Self-healing in emerging cellular networks: review, challenges, and research directions. *IEEE Commun Surv Tutor*, 20(3):1682-1709. https://doi.org/10.1109/COMST.2018.2825786

Bega D, Gramaglia M, Garcia-Saavedra A, et al., 2020. Network slicing meets artificial intelligence: an AI-based framework for slice management. *IEEE Commun Mag*, 58(6):32-38. https://doi.org/10.1109/MCOM.001.1900653

Bonati L, D'Oro S, Polese M, et al., 2020. Intelligence and learning in O-RAN for data-driven nextG cellular networks. https://arxiv.org/abs/2012.01263.

Cao Y, Wang R, Chen M, et al., 2020. AI agent in software-defined network: agent-based network service prediction and wireless resource scheduling optimization. *IEEE Int Things J*, 7(7):5816-5826. https://doi.org/10.1109/JIOT.2019.2950730

Chen XF, Zhang HG, Wu C, et al., 2019. Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning. *IEEE Int Things J*, 6(3):4005-4018. https://doi.org/10.1109/JIOT.2018.2876279

China Unicom, 2019. Computing Power Network. White Paper.

Ding ZG, Poor HV, 2020. A simple design of IRS-NOMA transmission. *IEEE Commun Lett*, 24(5):1119-1123. https://doi.org/10.1109/LCOMM.2020.2974196

El Azzaoui A, Singh SK, Pan Y, et al., 2020. Block5GIntell: blockchain for AI-enabled 5G networks. *IEEE Access*, 8:145918-145935. https://doi.org/10.1109/ACCESS.2020.3014356

ETSI, 2017. Improved Operator Experience Through Experiential Networked Intelligence (ENI).

He HT, Jin S, Wen CK, et al., 2019. Model-driven deep learning for physical layer communications. *IEEE Wirel Commun*, 26(5):77-83. https://doi.org/10.1109/MWC.2019.1800447

He T, Cao C, Tang XY, et al., 2020. Research on computing power network technology for 6G requirements. *Mob Commun*, 44(6):131-135 (in Chinese). https://doi.org/10.3969/j.issn.1006-1010.2020.06.020

Huawei, 2020. Autonomous Driving Network (ADN) Solution. White Paper.

Issa A, Hakem N, Kandil N, 2019. Wireless SDN architecture testbed to support IP multimedia subsystem. 4th Int Conf on Advances in Computational Tools for Engineering Applications, p.1-6. https://doi.org/10.1109/ACTEA.2019.8851107

ITU-T, 2020. Framework for Evaluating Intelligence Levels of Future Networks Including IMT.

Liu J, Du XQ, Cui JH, et al., 2020. Task-oriented intelligent networking architecture for the space-air-ground-aqua integrated network. *IEEE Int Things J*, 7(6):5345-5358. https://doi.org/10.1109/JIOT.2020.2977402

Liu YQ, Peng MG, Shou GC, et al., 2020. Toward edge intelligence: multiaccess edge computing for 5G and Internet of Things. *IEEE Int Things J*, 7(8):6722-6747. https://doi.org/10.1109/JIOT.2020.3004500

Lu YL, Huang XH, Dai YY, et al., 2020. Differentially private asynchronous federated learning for mobile edge computing in urban informatics. *IEEE Trans Ind Inform*, 16(3):2134-2143. https://doi.org/10.1109/TII.2019.2942179

Mao Q, Hu F, Hao Q, 2018. Deep learning for intelligent wireless networks: a comprehensive survey. *IEEE Commun Surv Tutor*, 20(4):2595-2621. https://doi.org/10.1109/COMST.2018.2846401

Nguyen DC, Cheng P, Ding M, et al., 2021. Enabling AI in future wireless networks: a data life cycle perspective. *IEEE Commun Surv Tutor*, 23(1):553-595. https://doi.org/10.1109/COMST.2020.3024783

Pateromichelakis E, Moggio F, Mannweiler C, et al., 2019. End-to-end data analytics framework for 5G architecture. *IEEE Access*, 7:40295-40312. https://doi.org/10.1109/ACCESS.2019.2902984

Peng MG, Yan S, Zhang KC, et al., 2016. Fog-computing-based radio access networks: issues and challenges. *IEEE Netw*, 30(4):46-53. https://doi.org/10.1109/MNET.2016.7513863

Peng MG, Sun YH, Wang WB, 2020. Intelligent-concise radio access networks in 6G: architecture, techniques and insight. *J Beijing Univ Posts Telecommun*, 43(3):1-10 (in Chinese). https://doi.org/10.13190/j.jbupt.2020-079

RAN Alliance, 2018. O-RAN: Towards an Open and Smart RAN. White Paper. https://www.coursehero.com/file/93485199/O-RANWPFInal181017pdf/

Ren YJ, Sun YH, Peng MG, 2021. Deep reinforcement learning based computation offloading in fog enabled industrial Internet of Things. *IEEE Trans Ind Inform*, 17(7):4978-4987. https://doi.org/10.1109/TII.2020.3021024

Srinivasan SM, Truong-Huu T, Gurusamy M, 2019. Machine learning-based link fault identification and localization in complex networks. *IEEE Int Things J*, 6(4):6556-6566. https://doi.org/10.1109/JIOT.2019.2908019

Sun YH, Peng MG, Zhou YC, et al., 2019a. Application of machine learning in wireless networks: key techniques and open issues. *IEEE Commun Surv Tutor*, 21(4):3072-3108. https://doi.org/10.1109/COMST.2019.2924243

Sun YH, Peng MG, Mao SW, 2019b. Deep reinforcement learning-based mode selection and resource management for green fog radio access networks. *IEEE Int Things J*, 6(2):1960-1971. https://doi.org/10.1109/JIOT.2018.2871020

Sun YH, Peng MG, Mao SW, 2019c. A game-theoretic approach to cache and radio resource management in fog radio access networks. *IEEE Trans Veh Technol*, 68(10):10145-10159. https://doi.org/10.1109/TVT.2019.2935098

Sun YH, Wang ZY, Yuan S, et al., 2021. The sixth-generation mobile communication network with endogenous intelligence: architectures, use cases and challenges. *Appl Electron Tech*, 47(3):8-13, 17 (in Chinese). https://doi.org/10.16157/j.issn.0258-7998.211392

Wang Z, Li LH, Xu Y, et al., 2018. Handover control in wireless systems via asynchronous multiuser deep reinforcement learning. *IEEE Int Things J*, 5(6):4296-4307. https://doi.org/10.1109/JIOT.2018.2848295

Wu WB, Peng MG, Chen WY, et al., 2020. Unsupervised deep transfer learning for fault diagnosis in fog radio access networks. *IEEE Int Things J*, 7(9):8956-8966. https://doi.org/10.1109/JIOT.2020.2997187

Xia WC, Zhang XR, Zheng G, et al., 2020. The interplay between artificial intelligence and fog radio access networks. *China Commun*, 17(8):1-13. https://doi.org/10.23919/JCC.2020.08.001

Xiang HY, Xiao YW, Zhang X, et al., 2017. Edge computing and network slicing technology in 5G. *Telecommun Sci*, 33(6):54-63 (in Chinese). https://doi.org/10.11959/j.issn.1000-0801.2017200

Xiang HY, Yan S, Peng MG, 2020. A realization of fog-RAN slicing via deep reinforcement learning. *IEEE Trans Wirel Commun*, 19(4):2515-2527. https://doi.org/10.1109/TWC.2020.2965927

Ye H, Li GY, Juang BF, 2019. Deep reinforcement learning based resource allocation for V2V communications. *IEEE Trans Veh Technol*, 68(4):3163-3173. https://doi.org/10.1109/TVT.2019.2897134

Yu C, Liu Y, Yao DZ, et al., 2017. Modeling user activity patterns for next-place prediction. *IEEE Syst J*, 11(2):1060-1071. https://doi.org/10.1109/JSYST.2015.2445919

Yu P, Li WJ, Feng L, et al., 2020. Intelligent network management and control architecture and key technologies for future 6G networks. *Front Data Comput*, 2(3):32-44 (in Chinese). https://doi.org/10.11871/jfdc.issn.2096-742X.2020.03.003

Yuan S, Ren YJ, Wang ZY, et al., 2021. Software defined intelligent satellite-terrestrial integrated wireless network. *Telecommun Sci*, 37(6):66-77 (in Chinese). https://doi.org/10.11959/j.issn.1000-0801.2021123

Zhang HJ, Liu N, Chu XL, et al., 2017. Network slicing based 5G and future mobile networks: mobility, resource management, and challenges. *IEEE Commun Mag*, 55(8):138-145. https://doi.org/10.1109/MCOM.2017.1600940

Zhang P, Peng MG, Cui SG, et al., 2022. Theory and techniques for "intellicise" wireless networks. *Front Inform Technol Electron Eng*, 23(1):1-4. https://doi.org/10.1631/FITEE.2210000

Zhao ZY, Feng CY, Yang HH, et al., 2020. Federated-learning-enabled intelligent fog radio access networks: fundamental theory, key techniques, and future trends. *IEEE Wirel Commun*, 27(2):22-28. https://doi.org/10.1109/MWC.001.1900370

Zhou YC, Yan S, Peng MG, 2020. Intent-driven 6G radio access network. *Chin J Int Things*, 4(1):72-79 (in Chinese). https://doi.org/10.11959/j.issn.2096-3750.2020.00146