JYX

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Chang, Zheng; Chen, Tao

Title: Virtual Resource Allocation for Wireless Virtualized Heterogeneous Network with Hybrid Energy Supply

Year: 2022

Version: Accepted version (Final draft)

Copyright: © 2022, IEEE

Rights: In Copyright

Rights url: http://rightsstatements.org/page/InC/1.0/?language=en

Please cite the original version:

Chang, Z., & Chen, T. (2022). Virtual Resource Allocation for Wireless Virtualized Heterogeneous Network with Hybrid Energy Supply. IEEE Transactions on Wireless Communications, 21(3), 1886-1896. https://doi.org/10.1109/twc.2021.3107867

Virtual Resource Allocation for Wireless Virtualized Heterogeneous Network with Hybrid Energy Supply

Zheng Chang, *Senior Member, IEEE,* Tao Chen, *Senior Member, IEEE*

***Abstract*—In this work, two novel virtual user association and resource allocation algorithms are introduced for a wireless virtualized heterogeneous network with hybrid energy supply. In the considered system, macro base stations (MBSs) are supplied by the grid power and small base stations (SBSs) have the energy harvesting capability in addition to the grid power supplement. Multiple infrastructure providers (InPs) own the physical resources, i.e., BSs and radio resources. The Mobile Virtual Network Operators (MVNOs) are able to recent these resources from the InPs and operate the virtualized resources for providing services to different users. In particular, aiming to maximize the overall utility for the MVNOs, a joint resource (spectrum and power) allocation and user association problem is presented. First, we present an alternating direction method of multipliers (ADMM)-based algorithm solution to find the near-optimal solution in a static manner. Moreover, we also utilize deep reinforcement learning to design the optimal policy without knowing a priori knowledge of the dynamic nature of networks. We have conducted extensive simulation and the performance evaluation demonstrate the advantages and effectiveness of the proposed schemes.**

***Index Terms*—energy harvesting, ADMM, reinforcement learning, deep learning, wireless network virtualization, resource allocation.**

I. Introduction

A. Background and Motivation

The future mobile communication system is expected to provide the ubiquitous connectivity and unprecedented services over billions of devices. However, the increased network density and demand for diverse service and applications also introduces significant challenges for higher capacity, reduced energy consumption and low latency and for trillions of devices. To provide the ubiquitous and unlimited Internet and data access, some recent developed platforms, such as Software Defined Network (SDN) and Network Function Virtualization (NFV), have attracted significant research interests and the research outcome also sheds the light on revisiting the current cellular networks [1]. Incorporating with the wireless network, the advanced SDN/NFV architectures are applied to Radio Access Networks (RANs), which creates the Wireless Virtualized Networks (WVNs) framework. In the WVN, the RAN functions can be executed on commoditized platforms owned by multiple Infrastructure Providers (InPs), instead of dedicated telecommunications hardware owned by the specific companies.

In particular, the concept of virtualization allows the customized WVNs for particular applications on top of a physical network. The WVN makes physical infrastructure and radio resources being abstracted, sliced and shared, which makes them well suited to address the diverse requirements of future wireless network and ease the network management. After virtualization, the virtual slices containing radio resources are offered to the service provider based on their demands. The mobile operator and service provider can rent the virtual resources, instead of owning them, to provide services to the users. Consequently, the overall expenses of network deployment and operation can be significantly decreased [2].

It can be found that WVN has many advantages in effectively utilizing the radio resources in the future wireless network. However, successfully merging the NFV with recent advances of RAN to reduce the operation cost requires dedicated efforts [3]. In the meantime, one promising solution for ubiquitous connectivity is to deploy small based stations (SBSs) to complement the traditional macro BS (MBS)-based cellular architecture. Accordingly, the future wireless network emerges with dense and heterogeneous features, and it is expected that the heterogeneous wireless networks are able to meet the stringent requirements of end-user's Quality of Experience (QoE) and Quality of Service (QoS).

Nevertheless, the wide and dense installation of the SBSs faces many different challenges in reality as it will increase the complexity of network planning and optimization, and bring additional deployment expenditure. In particular, providing a large number of SBSs with stable grid power in a cost-efficient way is one of the major concerns. Comparing with the MBSs, the high density and irregular location of SBSs make them difficult to access the power grid. Therefore, applying the energy harvesting (EH) technology to the development of wireless network receives many interests. As the EH technology is able to utilize the renewable energy (e.g., solar and wind), the EH-powered SBSs are able to accelerate the deployment of dense small cells [4]. In addition, considering the random variations of the wireless channel and the energy arrival in the time domain, dynamic schemes, instead of static ones, are preferred for network optimization [38]. In this context, Reinforcement Learning (RL) has been utilized recently as an promising solution to address the problems caused by network dynamics [6]. The RL agent is able to adopt its policy to the system dynamic to obtain the best long-term utility by receiving feedback (reward) from the environment, which is desirable for the time-variant systems [7].

Z. Chang is with School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, and also with Faculty of Information Technology, University of Jyväskylä, P.O. Box 35, FIN-40014 Jyväskylä, Finland. T. Chen is with VTT Technical Research Centre of Finland, Espoo, Finland. Part of this work has been presented in WCNC 2018 [5]. This work is partly supported by NSFC (No. 62071105).

To this end, there are no doubts that in order to establish a heterogeneous network (Hetnets) architecture, both EH and virtualization technologies should be carefully explored to be integrated to the SBS. WVN may offer us a novel view in the small cell development from the network operation perspective. As for the future networks, a wireless virtualized Hetnet with hybrid power supply is full of possibility. Then, how to efficiently operate it is also of profoundly importance. Therefore, this work aims at presenting user association, power, and spectrum allocation algorithms for such a complex system to provide efficient solutions and useful insights. To solve the formulated problem, we first propose a static and distributed alternating direction method of multipliers (ADMM)-based optimization solution. In addition, we also dedicate to investigate a machine learning-based scheme for addressing the formulated problem in a dynamic manner. In particular, we apply Q-learning in the algorithm design to learn the policy and decide the action. Moreover, as the amount of actions may dramatically increase with the growing number of users and BSs in Q-learning, we utilize deep learning and form a Deep Reinforcement Learning (DRL)-based framework to estimate the Q-value to obtain efficient solution of the formulated resource allocation and user association problem.

B. Related works

EH is considered as a promising technology for prolonging the life of the battery-limited device. Recently, applying EH technology for providing BS with renewable energy has received great research interests. In particular, the investigation on utilizing EH for small cell networks (SCNs) emerges as one of the key research area since the SBSs may not be easy to access the electric grid power. So far, there are some works dedicated on investigating the resource allocation and user association problems in the EH-based Hetnets [4], [9]-[11]. In [8], the authors propose time, power and rate allocations in a multi-user EH network. The main objective of this work is to maximize system throughput with the consideration of different channel status among multiple users and ensuring the user fairness. Considering a energy constrained Hetnet, The authors of [9] investigate the backhaul-aware resource allocation problem. An optimization problem is then presented to optimize the defined system utility with proportional fairness consideration. The authors of [10] explore the energy provisioning problem aiming at minimizing the cost of EH system deployment in the cellular networks. A solution consisting of load balancing scheme and system sizing is proposed. In [11], a traffic load balancing scheme is presented, where the relations between EH utilization and latency are studied.

Meanwhile, the development of WVN has also received many research interests [2]. Successfully integrating the virtualization into wireless network faces several main challenges, including virtual resource allocation, abstraction, isolation and signaling overhead issues, etc [3]. In [12], the authors propose a wireless resource slicing scheme to flexibly divide spectrum into different slices. Accordingly, complete resource abstraction is obtained in dense SCNs. A performance comparison of different network sharing schemes is provided in [14]. In [15], the authors present a virtual resource allocation scheme to optimize the network utility of a virtualized information centric network. The authors of [16] propose a resource provisioning algorithm for a WVN with massive antenna BS. A joint optimization problem of antenna, spectrum and transmit power allocations is introduced to optimize the defined utility while maintaining the user fairness. In our previous works [5], we also consider the hybrid power empowered WVN and present resource allocation solution. Moreover, there are some works considering to utilize the features of virtualization to efficiently manage the SCNs. In [17], the authors propose an user association algorithm for saving energy consumption and limiting interference in a Cloud-RAN-based SCN. The authors of [18] utilize the resource allocation scheme and analyze the downlink of virtualized cellular networks with large scale multiple antenna. With the consideration of full-duplex self-backhaul, the virtual resource allocation problem has been studied for the SCN in [19].

Recently, utilizing machine learning (ML) framework for resource management in wireless networks receives increasing research interests. ML-based algorithms have great success in supporting big data analytics, efficient parameter estimation and interactive decision making in many application areas and have shown its great potential in advancing the wireless networks [20]. Specifically, applying the RL-based scheme for optimization in heterogeneous wireless networks is able to reach the Pareto-optimal solution which achieves the trade-off among different objectives [21]. In [22], the authors present DRL-based scheme for mode selection and resource allocation to optimize the energy usage of fog radio access networks. In [23], the authors investigate the joint design of beamforming matrix at the BS and analog beamforming matrices at the intelligent surfaces, by leveraging the DRL-based algorithm to combat the propagation loss. In [24], how to utilize the RL-based scheme for D2D communication is investigated. The authors advocate the RL for investigating the D2D coalition formation game, where the coalition are formed to maximize long-term rewards of the D2D users. The authors of [25] apply DRL framework to address resource allocation and user association problem in heterogeneous cellular networks with the objective to maximize network utility and satisfy the QoS requirements of users over a long-term. In [26], the authors propose to allow a central unit in wireless virtualization to learn to configure radio resources autonomously with the goal of minimizing a network cost function. DRL-based algorithm is presented to solve the formulated problem. The authors of [27] decompose the complex network virtulization function into function components to make more effective decisions for a virtual and heterogeneous IoT network. The authors propose a DRL-based scheme with experience replay and target network as a solution that can efficiently handle complex and dynamic service function chain in IoT. In [28], the authors investigate the network slicing realization problem in a virtualized fog-RAN environment. The framework for network slicing is formulated as an joint optimization problem of content caching and mode selection, and addressed by DRL-based scheme.

As can be observed, the problem of effectively utilizing the virtual resources of wireless virtualized Hetnets is still under-investigation. Moreover, there are spare works utilizing machine learning-based schemes for addressing the associated resource allocation and user association problems to effectively and efficiently operate WVN, which, however, is significant for the development of future wireless communication system.

C. *Contribution*

Our primary target in this work is to propose different optimization algorithms to find effective user association and resource allocation solutions for hybrid power supplied wireless virtualized Hetnets, in both static and dynamic manners. The InPs in the WVN own the physical infrastructure, such as MBS, SBSs and spectrum resources. The MBS are empowered by the grid power, and SBSs are supplied by EH in addition to the grid power. The Mobile Virtual Network Operators (MVNOs) need to rent radio resources, virtualize and then operate them to provide services to a number of users. Comparing with the existed works, we can briefly summarize the main contributions as follows.

- We introduce a virtualized and hybrid power suppled Hetnets architecture with multiple InPs and MVNOs. In this system, the InPs own MBS, EH-SBS, and various type of radio resources. The physical network owned by the InPs can be virtualized and flexibly shared for different MVNOs to purchase.
- Based on the presented system model, an optimization problem related to user association and resource allocation is formulated. The aim is to maximize the utility of all the MVNOs, which concerns both the revenue earned from the subscriber/user and the cost paid to the InPs. In order to address the formulated problem, power and spectrum allocation, and user association over different time slots should be jointly optimized with explicit consideration of EH limitations.
- Directly addressing the formulated non-convex mixed integer programming problem induce a very high computational complexity. In this context, we aim at addressing the formulated problem in a static manner, i.e., we turn to find the solution in a certain time slot. We can then divide the original problem into two subproblems and proposed an ADMM-based scheme to find the solution with a fast convergence rate.
- We also utilize the machine learning and propose a DRL-based resource allocation and user association algorithm to investigate the optimal policy. Particularly, we adopt DQN to address the formulated problem by learning an optimal policy without a priori information. According to the utility function, the action space, state space and reward functions in the DRL are carefully defined.
- Performance evaluations are presented to examine the proposed schemes. It is found that both of the proposed schemes have fast convergence performance. In addition, by utilizing the proposed schemes, the overall utility of MVNOs can be maximized. The proposed schemes can obtain superior performance comparing with other schemes as well.

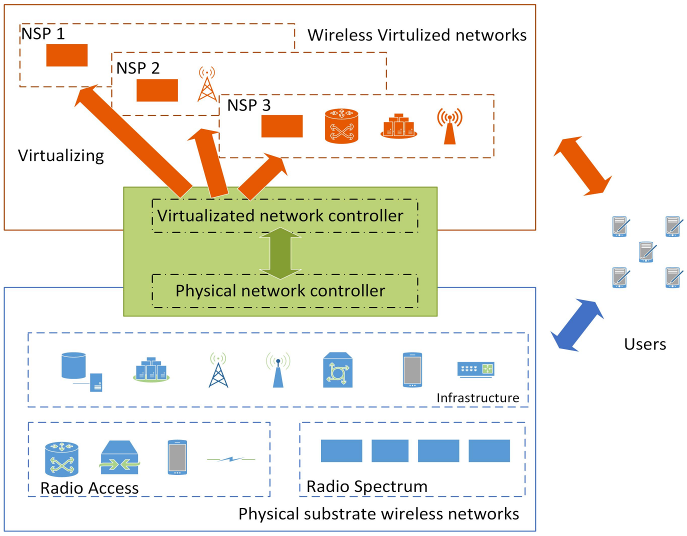

NSP 1
Wireless Virtulized networks
NSP 2
NSP 3
Virtualizing
Virtualizated network controller
Physical network controller
Users
Infrastructure
Radio Access
Radio Spectrum
Physical substrate wireless networks

Fig. 1. Wireless Network Virtualization

The reminder of this paper is organized as follows. The system model is depicted in Section II. The problem formulation is given in Section III. Section IV present the proposed ADMM-based. In Section V, a DRL-based algorithm is introduced. The simulation study is conducted in VI and Section VII concludes the work.

II. Wireless Virtualized Networks with Hybrid Energy Supply

A. *Wireless Network Virtualization*

Virtualization has mainly been studied for computing server to deal with the computing resources. While talking about the virtualization in the wireless networks, radio resources and physical infrastructure of the wireless network are abstracted and allocated into virtual slices with certain functionalities. The virtual slices are then ready to be shared by different parties after resource isolation [2]. Therefore, after the virtualization process, the resources will be provided to different Network Service Providers (NSPs). Then, the virtual resources are used for service provisioning.

In Fig. 1, we have illustrated the concept of Wireless Network Virtualization (WNV). When the NSPs receive the demand of service from the users, they can request the radio resources from InPs. Then, the radio resources and physical infrastructures own by different InPs will be processed by the network controller. After being isolation, abstract and virtualization, the virtual slices containing radio resources are offered to the NSPs based on their demands. The MVNO is able to virtualize the physical resources according to the demand of NSPs. The users is able to logically connect to the virtual network through service subscription and communicates with the cellular network physically.

B. *System Model*

In Fig. 2, we present an example of wireless virtualized SCN with hybrid energy supply. The SBS is empowered by both renewable energy via EH and power grid. The harvested energy

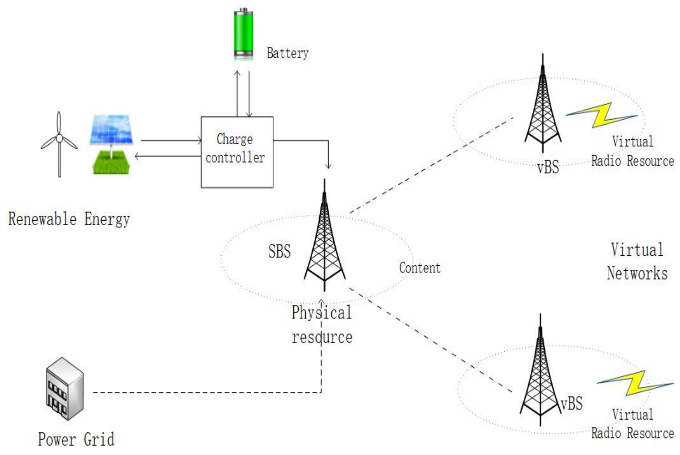


Fig. 2. Wireless virtualized networks with hybrid energy supply

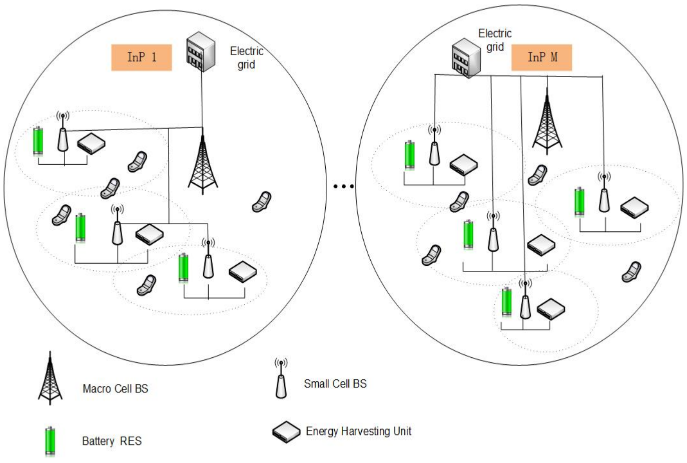


Fig. 3. System Model

can be stored in the battery for further usage. The physical infrastructure and radio resource are virtualized based on its application and after virtualization, they can be dynamically used according to the demand.

In the considered system, there are M InPs and each InP owns a certain area containing J SBSs and one MBS, as shown in Fig. 3. The set of InP is denoted as $\mathcal{M}$. We also assume there are I MVNOs and the set of MVNOs is denoted as $\mathcal{I}$. The MBSs and SBSs are connected to the power grid for constant power supply. In addition, the SBSs are also with EH capabilities and can be powered by renewable energy supply (RES). We use S_m^0 to denote the MBS of InP m, $S_m^j, j \neq 0$ to denote SBS j of InP m, and $\mathcal{S}_m$ to represent the set of BSs of InP m. The set of users who is served by BS S_m^j is denoted as $\mathcal{U}_m^j$, the set of users who connects with the InP m is denoted as $\mathcal{U}_m$, and the set of users who subscribes to the MVNO i is $\mathcal{U}_i$. It is assumed that one user only connects with one BS in a certain time slot. At time slot l, the user association indicator $\beta_u^{m,j}$ for user-BS connection is defined as,

$$\beta_u^{m,j}(l) = \begin{cases} 1, & \text{if user } u \text{ associates with BS } S_m^j, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

C. Transmission Model

In this work, OFDM-based network is assumed and the whole spectrum usage of one InP is separated by the MBS and SBSs to avoid the interference. We consider the spectrum used by MBS is $f_m = \alpha_m W_m$ and the one used by the SBS is denoted as $f_m^s = (1 - \alpha_m) W_m$, where W_m is the spectrum bandwidth of InP m. $R_u^{m,j}, j \in \{0, ..., J\}$ is denoted as the throughput that user u can obtain when communicating with the BSs. On the DL of MBS m, the throughput of user u at time slot l is expressed as

$$R_u^{m,0}(l) = f_m \log_2 \left(1 + \frac{p^{m,0}(l) h_u^{m,0}(l)}{\sigma^2} \right). \quad (2)$$

where $p^{m,0}(l)$ is the transmit power of MBS of InP m, σ^2 is the noise variance and $h_u^{m,0}(l)$ is the channel gain from the MBS to user u. Without loss of generality, the noise is with zero-mean and unit-variance. When connecting with SBS, the throughput $R_u^{m,j}(l), j \in \{1, ..., J\}$ is

$$R_u^{m,j}(l) = f_m^s L \log_2 \left(1 + \frac{p^{m,j}(l) h_u^{m,j}(l)}{\sum_{k=1, k \neq j}^{J} p^{m,k}(l) h_u^{m,k}(l) + \sigma^2} \right), \quad (3)$$

where $p^{m,j}(l)$ is the transmit power of SBS S_m^j. As the SBS can be supplied by the harvested energy, we consider $p^{m,j}(l) = p_e^{m,j}(l) + p_g^{m,j}(l)$, where $p_e^{m,j}(l)$ is the power consumption from harvested energy stored in the battery and $p_g^{m,j}(l)$ is the one from grid power. $h_u^{m,j}(l)$ is the channel gain from the SBS S_m^j to user u. We also denote $\delta_u^{m,0}(l)$ as the resource ratio that MBS is used for transmitting to user u and $\delta_u^{m,j}(l), j \in \{1, ..., J\}$ as the resource ratio that SBS S_m^j is used for transmitting to user u. Therefore, for user u, the total achievable throughput is

$$\begin{aligned} C_u^m(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{p}) &= \beta_u^{m,0}(l) \delta_u^{m,0}(l) R_u^{m,0}(l) \\ &\quad + \sum_{j=1}^{J} \beta_u^{m,j}(l) \delta_u^{m,j}(l) R_u^{m,j}(l) \\ &= \sum_{j \in \mathcal{S}_m} \beta_u^{m,j}(l) \delta_u^{m,j} R_u^{m,j}(l), \end{aligned} \quad (4)$$

where $\boldsymbol{\beta} = \{\beta_u^{m,j}(l)\}$ is the set of user association indicators. $\boldsymbol{\alpha} = \{\alpha_m(l)\}$ and $\boldsymbol{\delta} = \{\delta_u^{m,j}(l)\}$ is the resource allocation policy for SBS and user, respectively. $\boldsymbol{p} = \{p^{m,j}(l)\}$ is the power allocation policy.

D. Energy Harvesting Model

We use $B_m^j(l)$ to denote the battery letter at the beginning of time slot l. The energy packets in EH is assumed to arrive at the beginning of time slot and it is denoted as $E_m^j(l)$. The EH is modeled as a proper random process which is related to its type of energy source. Such an assumption on the renewable energy availability is reasonable as the resource allocation and user association are decided based on the available information at the beginning of time slot [9]. We also assume that the battery capacity is infinite at the SBS to facilitate the study of the algorithm design.

Then, the battery level at time slot $l+1$, is expressed as

$$B_m^j(l+1) = \chi(B_m^j(l), E_m^j(l), \zeta_{m,e}^j(l)), \tag{5}$$

where $\zeta_{m,e}^j(l)$ is the energy consumption from the SBS's battery. $\chi(.)$ is the relation function shows the usage of the battery and the harvested energy, and it depends on the type of battery, such as storage efficiency and memory effects. The consumed energy has the following constraint:

$$\zeta_{m,e}^j(l) \le \psi_m^j(l) B_m^j(l), \tag{6}$$

where ψ_m^j is the ratio of the battery that can be used for serving the users. For SBS S_m^j, $p_e^{m,j}$ is the consumed power from battery, and thus, $\zeta_{m,e}^j = \sum_{u\in\mathcal{U}_m^j} \beta_u^{m,j} p_e^{m,j} L$. For simplicity, we assume length of a time slot $L=1$.

III. Problem Formulation

A. Utility Function

Given the system model and assumptions, the utility function of a MVNO is defined in (7). The first term of the right side of (7) is the benefit of the MVNO, which can be expressed as follows,

$$\begin{aligned} U_u^m(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}) &= \varsigma_u U\left(C_u^m\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}\right)\right) \\ &= \sum_{j\in\mathcal{S}_m} \beta_u^{m,j}(l) \log\left(\delta_u^{m,j}(l) R_u^{m,j}(l)\right). \end{aligned} \tag{8}$$

where ς_u is the profit per user per rate unit. Υ_i^m is defined as cost for using spectrum and power resource of a MVNO, which is

$$\begin{aligned} \Upsilon_i^m(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}) = \psi_{m,1} &\sum_{u\in\mathcal{U}_i} \beta_u^{m,0}(l) f_m p^{m,0}(l) \\ + \psi_{m,2} &\sum_{u\in\mathcal{U}_i}\sum_{j\in\mathcal{S}_m} \beta_u^{m,j}(l) f_m^s(l) p_g^{m,j}(l) \\ + \psi_{m,3} &\sum_{u\in\mathcal{U}_i}\sum_{j\in\mathcal{S}_m} \beta_u^{m,j}(l) f_m^s(l) p_e^{m,j}(l). \end{aligned} \tag{9}$$

We use a bandwidth-power product to quantify the resource consumption of the BS. The coefficients $\psi_{m,1}$ $\psi_{m,2}$ and $\psi_{m,3}$ specify the cost unit of MBS, the grid power and RES of SBS, respectively. In order to prompt the usage of RES, we can assume $\psi_{m,1} > \psi_{m,2} > \psi_{m,3} > 0$ so that more users would prefer to choose the RES and SBS. The expense of the MVNOs for the usage of wireless backhaul is also considered in (7). Such a expense depends on the amount of data transmission on backhaul, which is given as

$$Q_i^m(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}) = \omega_m \sum_{u\in\mathcal{U}_i} C_u^m(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}), \tag{10}$$

where ω_m is the unit charge (per bit) for backhaul transmission, which is related to the type of backhaul.

B. Problem formation

In this work, transmit power allocation $\boldsymbol{p}$, spectrum resources $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, and user associations $\boldsymbol{\beta}$, are jointly optimized. Correspondingly, $\mathbf{P1}$ can be formulated as follows,

$$\mathbf{P1}: \quad \max_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}} \quad \Sigma_{mvno}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}), \tag{11}$$

$$\begin{aligned} s.t. \quad &\mathbf{C1}: && \beta_u^{m,j}(l) \in \{0,1\}, \sum_{m\in\mathcal{S}_m}\sum_{j\in S_m} \beta_u^{m,j}(l) \le 1, \\ &\mathbf{C2}: && \beta_u^{m,j}(l) \in \{0,1\}, \sum_{m\in\mathcal{S}_m}\sum_{j\in S_m} \beta_u^{m,j}(l) \le 1, \\ &\mathbf{C3}: && 0 \le p^{m,j}(l) \le p_{max}^j, j\in\{1,...,J\}, \\ &\mathbf{C4}: && \sum_{u\in\mathcal{U}_m^j} \zeta_{m,e}^j(l) \le \psi_m^j(l) B_m^j(l), j\in\{1,...,J\}, \\ &\mathbf{C5}: && 0 \le \delta_u^{m,j}(l) \le 1, \\ &\mathbf{C6}: && \beta_u^{m,j}(l)\delta_u^{m,j}(l) R_u^{m,j}(l) \ge r_u, \\ &\mathbf{C7}: && 0 \le \alpha_m(l) \le 1. \end{aligned} \tag{12}$$

In (12), $\hat{p}_u^{m,j} = \hat{p}_{u,e}^{m,j} + \hat{p}_{u,g}^{m,j}$. The constraints in $\mathbf{C1}$ and $\mathbf{C2}$ are to ensure each user can only be served by one BS at a time. $\mathbf{C3}$ ensures the transmit power can not exceed the maximum power and $\mathbf{C4}$ makes sure that the energy consumed up to any time cannot exceed the accumulatively harvested energy before this moment. $\mathbf{C5}$-$\mathbf{C7}$ are to ensure that resource allocation strategy remains at an acceptable level and the minimum data requirement. Moreover, in this work, we consider the transmit power of MBS is constant which is reasonable as the MBS is supplied by grid power. Thus, the focus on power allocation, is in turn to optimize the usage of transmit power of SBS. In the following, we first present an ADMM-based algorithm to address the formulated problem in a certain time slot. Then, a DRL-based scheme is explored in order to solve the problem in a dynamic manner.

IV. Proposed ADMM-based Solution

In the following, we aim to present distributed resource allocation and user association scheme in each time slot. Thus, in this section, time representative l is omitted to ease the presentation unless being specified. In general, it can be found that $\mathbf{P1}$ is a non-convex problem with a combinatorial integer programming structure. To obtain an optimal solution for such a NP hard problem, exhaustive search approaches can be used. However, it requires high computational cost, which is infeasible for a large scale system. In addition, insightful discussions cannot be made accordingly for system design.

In order to reduce the complexity of finding the solution, in this section, $\mathbf{P1}$ is divided into two subproblems. First, assuming power allocation $\boldsymbol{p}$ is fixed, the user association and spectrum allocation problems are addressed. Then, by variable transformation and relaxation, $\mathbf{P1}$ is converted into a convex problem to find $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$. After addressing the user association and spectrum allocation, optimal $\boldsymbol{p}$ can be addressed accordingly. Through iterative scheme, the final solution of $\boldsymbol{p}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\delta}$ will converge.

$$\Sigma_{mvno}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p})=\sum_{i\in\mathcal{I}}\sum_{m\in\mathcal{M}}\sum_{u\in\mathcal{U}_i}U_u^m(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p})-\sum_{i\in\mathcal{I}}\sum_{m\in\mathcal{M}}\Upsilon_i^m(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p})-\sum_{i\in\mathcal{I}}\sum_{m\in\mathcal{M}}Q_i^m(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{p}). \tag{7}$$

A. Proposed Solution for Solving $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$

1) Problem transformation: First, we will present solution for solving $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$. The original problem then becomes:

$$\mathbf{P2}:\max_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}\quad\Sigma_{mvno}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}\right), \tag{13}$$

$$s.t.\quad \mathbf{C1},\mathbf{C3}-\mathbf{C7}. \tag{14}$$

Nevertheless, due to the binary variable $\beta_u^{m,j}$ and its objective function, $\mathbf{P2}$ still has a non-convex structure after transformation. We can first relax $\beta_u^{m,j},\forall j\in\{1,...,J\}$ in the formulated problems so that $0\leq\beta_u^{m,j}\leq 1$ [32]. Accordingly, it can be interpreted as the ratio of time that user u can connect with BS S_m^j. However, because of its objective function, $\mathbf{P2}$ remains non-convex after variable relaxation. To address this problem, an auxiliary continuous variable is introduced, i.e. $\widetilde{\alpha_u^{m,j}}=\beta_u^{m,j}\alpha_{u,n}^{m,j}$. If $\beta_u^{m,j}=0$, $\alpha_u^{m,j}=0$ certainly holds which is due to the fact that if user does not associate with certain BS, it will not receive any resource from it. Correspondingly, $\mathbf{P2}$ can be transformed into $\mathbf{P3}$ with the objective in (15) and the constraints in (16).

$$\mathbf{P3}:\max_{\boldsymbol{\beta},\widetilde{\boldsymbol{\alpha}},\boldsymbol{\delta}}\quad\Sigma_{mvno}\left(\boldsymbol{\beta},\widetilde{\boldsymbol{\alpha}},\boldsymbol{\delta}\right), \tag{15}$$

$$\begin{aligned} s.t.\qquad &\mathbf{C1},\mathbf{C3},\\ &\widetilde{\mathbf{C2}}:\quad 0\leq\widetilde{\alpha_u^{m,j}}\leq 1,\\ &\widetilde{\mathbf{C4}}:\quad 0\leq\widetilde{\delta_u^{m,j}}\leq 1,\\ &\widetilde{\mathbf{C6}}:\quad \sum_{i\in\mathcal{I},u\in\mathcal{U}_i}z_u^m\widetilde{\gamma_u^{m,j}}\leq Z_m^j. \end{aligned} \tag{16}$$

We can see that $\mathbf{P3}$ is now a convex problem and we present ADMM-based scheme to solve $\mathbf{P3}$ in a distributed manner in the following.

2) ADMM-based solution algorithm: To utilize ADMM to solve the formulated problem, we introduce a set of new variables representing local copies of the global optimal solutions [29] [31]. Firstly, local copy $\boldsymbol{\beta}_m^{\diamond}$ of the related global user association factor $\boldsymbol{\beta}$ is defined for the InP m and $\boldsymbol{\beta}_m^{\diamond}$ can be roughly interpreted as the opinion of InP m about the global variable. Similarly, $(\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond})$ are the local variables of resource allocation strategies $(\boldsymbol{\alpha},\boldsymbol{\delta})$ of InP m. Correspondingly, the feasible local variable set of InP m is denoted as $\pi_m=(\boldsymbol{\beta}_m^{\diamond},\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond})$. Accordingly, we are able to express the local utility function which is shown in (17). Then, $\mathbf{P3}$ is reformed as

$$\mathbf{P4}:\min_{\boldsymbol{\beta}_m^{\diamond},\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond}}\sum_{m}^{M}\varpi_m(\boldsymbol{\beta}_m^{\diamond},\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond}) \tag{18}$$

$$s.t.\quad \boldsymbol{\beta}_m^{\diamond}-\boldsymbol{\beta}=0. \tag{19}$$

One can see that $\mathbf{P4}$ is a global consensus problem, as discussed in [31]. Utilizing ADMM to address a global consensus problem needs to establish an augmented Lagrangian with corresponding global consensus constraint. Accordingly, λ_m is defined as the Lagrange multiplier associate with the corresponding consensus constraint in $\mathbf{P4}$. Then, the augmented Lagrangian can be expressed as

$$\begin{aligned} &\mathcal{L}_{\varrho}\left(\{\boldsymbol{\beta}_m^{\diamond},\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond}\},\boldsymbol{\beta},\boldsymbol{\lambda}\right)=\sum_{m=1}^{M}\varpi_m(\boldsymbol{\beta}_m^{\diamond},\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond})\\ &+\sum_{m=1}^{M}\lambda_m(\boldsymbol{\beta}_m^{\diamond}-\boldsymbol{\beta})+\frac{\varrho}{2}\sum_{m}^{M}\|\boldsymbol{\beta}_m^{\diamond}-\boldsymbol{\beta}\|_2^2, \end{aligned} \tag{20}$$

where $\boldsymbol{\lambda}=\{\lambda_m\}$ and $\rho\in R_{++}$ is a positive constant parameter which is used to adjust the convergence rate [31]. Similar to other applied ADMM-based scheme [31], our proposed ADMM method for addressing $\mathbf{P4}$ with consensus constraints has the following iterative optimization steps.

1) Updating $\{\boldsymbol{\beta}_m^{\diamond},\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond}\}$: As we can see, the local decisions for resource allocation and user association are decoupled across differents BSs. Therefore, Finding $\{\boldsymbol{\beta}_m^{\diamond},\boldsymbol{\alpha}_m^{\diamond},\boldsymbol{\delta}_m^{\diamond}\}$ is able to be decomposed into M subproblems, each of which is addressed locally at BS level. At each iteration, the optimization problem in (21) is solved by InP m.
2) Updating $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$: Generally, updating $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is an unconstrained quadratic optimization problem, and the specific updating process can be found in (22) and (23).
3) Convergence performance: Due to the fact that our objective function is closed, proper and convex, and the Lagrangian $\mathcal{L}_{\varrho}$ has saddle point, as proved in [31], the proposed ADMM-based scheme satisfies residual convergence, objective convergence and dual variable convergence as iteration $t\rightarrow\infty$, .

As the objective function of $\mathbf{P3}$ is convex, we are able to reach optimal solution. However, as $\boldsymbol{\beta}$ is defined as a set of binary variables, the relaxed value should be recovered to boolean. Similar to one widely applied approach, we can first compute the marginal benefits for each $\beta_u^{m,j}$, i.e., $H_u^{m,j}=\partial\mathcal{L}_{\varrho}/\partial\beta_u^{m,j}$ [33]. Then, user association decision is able to be obtained as follows,

$$\beta_u^{m,j}=\begin{cases}1, & \text{if } H_u^{m,j}=\max_j H_u^{m,j},\ \text{and } H_u^{m,j}>0;\\ 0, & \text{Otherwise;}\end{cases} \tag{23}$$

After obtaining $\beta_u^{m,j*}$, we can obtain the optimal solution of $\widetilde{\theta_u^{m,j*}}$ accordingly and then $\omega_u^{m,j*}$.

B. Proposed Solution for $\boldsymbol{p}$

After obtaining $\boldsymbol{\beta},\boldsymbol{\delta}$, and $\boldsymbol{\alpha}$, and their values are fixed, we then turn to achieve the optimal solution of $\boldsymbol{p}$. Then, $\mathbf{P1}$

$$\varpi_m(\boldsymbol{\beta}_m^\diamond, \boldsymbol{\theta}_m^\diamond, \boldsymbol{\delta}_m^\diamond) = \begin{cases} -\sum_{i\in\mathcal{I}}\sum_{u\in\mathcal{U}_i} \varsigma_u U_u^m(\boldsymbol{\beta}_m^\diamond, \boldsymbol{\theta}_m^\diamond) + \sum_{i\in\mathcal{I}} \Upsilon^m(\boldsymbol{\beta}_m^\diamond, \boldsymbol{\theta}_m^\diamond) + & \\ \sum_{i\in\mathcal{I}} Q_i^m(\boldsymbol{\beta}_m^\diamond, \boldsymbol{\theta}_m^\diamond), & \boldsymbol{\beta}_m^\diamond, \boldsymbol{\alpha}_m^\diamond, \boldsymbol{\delta}_m^\diamond \in \pi_m \\ +\infty, & \text{else;} \end{cases} \tag{17}$$

$$\{\boldsymbol{\beta}_m^\diamond, \boldsymbol{\alpha}_m^\diamond, \boldsymbol{\delta}_m^\diamond\}^{[t+1]} : = \arg\min \left\{ \varpi_m(\boldsymbol{\beta}_m^\diamond, \boldsymbol{\alpha}_m^\diamond \boldsymbol{\delta}_m^\diamond) + \lambda_m \left(\boldsymbol{\beta}_m^\diamond - \boldsymbol{\beta}^{[t]}\right) + \frac{\rho}{2}\|\boldsymbol{\beta}_m^\diamond - \boldsymbol{\beta}^{[t]}\|_2^2 \right\} \tag{21}$$

$$\boldsymbol{\beta}^{[t+1]} \ : = \arg\min \left\{ \sum_{m=1}^{M} \lambda_m^{[t]} \left(\boldsymbol{\beta}_m^{\diamond[t+1]} - \boldsymbol{\beta}\right) + \frac{\rho}{2} \sum_{m=1}^{M} \|\boldsymbol{\beta}_m^{\diamond[t+1]} - \boldsymbol{\beta}\|_2^2 \right\} \tag{22}$$

$$\boldsymbol{\lambda}^{[t+1]} \ : = \boldsymbol{\lambda}^{[t]} + \rho \left(\boldsymbol{\beta}_m^{\diamond[t+1]} - \boldsymbol{\beta}^{[t+1]}\right) \tag{23}$$

becomes

$$\begin{aligned} \mathbf{P5} : \max_{\boldsymbol{p}} \quad & \Sigma_{mvno}(\boldsymbol{p}) \\ s.t. \quad & \mathbf{C2}, \mathbf{C3} \end{aligned} \tag{24}$$

We can see that the objective function of **P5** is convex with respect to $\boldsymbol{p}$. It can be observed **P5** is then a convex optimization problem with unique global optimal solution. There are some classical approaches that can address such a problem. We advocate the steepest descent method for fast convergence to solve it in this work.

V. Proposed DRL-based Solution

In this section, we will adopt the DRL framework into the development of virtual resource management in WVN and present a DRL-based scheme to address the formulated problem. As we can see, over a certain amount of time slots, the formulated problem should have the complete knowledge about the future time slots to reach the optimal solution of the next time slot. Therefore, absence of prior information about channel state and energy arrival may lead to a degraded system performance. Correspondingly, we will use a RL framework to address such a problem without prior knowledge. In the following, the basics of DRL are first presented, including the defined state, action and reward strategies. To avoid high dimensionality problem, DNN is applied and a DQN scheme is used to perform Q-learning action-value function estimation.

A. RL Framework Formulation

In the RL framework, the agent can chose actions to interact with the environment. In general, there are 3 basic elements: state, action and reward. In the presented WVN, the network controller is the agent and all the other entities can be considered as the environment. Within the action space, the network controller chooses an action in each time slot from the action space, which decides the user association and resource allocation, and then emerges to next state. After action execution, a reward or punishment can be obtained from the environment. Such a framework aims to maximize the cumulative received rewards of the system during the interactions with the environment.

B. State, Action and Reward

As presented, it is important to properly define the state space, action space and reward for applying the DRL to solve the formulated problem. In the following, the specific definitions are given.

1) State: In the WVN, the network controller is able to have the necessary information of BSs, e.g., battery level, maximum transmit power, EH capability, and channel information. Thus, we are able to define the state at time slot l as follows,

$$s_l = [E_l^1, B_l^1, C_l^1, D_l^1, ..., E_l^N, B_l^N, C_l^N, D_l^N], \tag{25}$$

which indicates the controller will know the battery levels, harvested energy levels, throughput and transmission delays of the BSs. N is the total amount of BSs, i.e., $N = M(J+1)$.

2) Action: In the presented WVN, the action space contains different strategies, i.e., the resource allocation factors $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, power allocation $\mathbf{p}$ and user association $\boldsymbol{\beta}$. Then, the action space $\mathcal{A}$ comprises of all the possible strategies.

3) Reward: The network controller can obtain a reward after executing action. The definition of reward is crucial as it can enforce the network controller to take proper action. As shown in **P1**, the main target of the formulated problem is to optimize the utility of all MVNOs while satisfying each user's QoS. In order to link the reward to the objective function, following points are explicitly considered.

- Since the goal of RL framework is maximizing the reward, there should be a positive relations between the objective function and the defined reward;
- In order to meet the the users' QoS requirements,the reward will be decreased if there is any loss of the QoS.

Therefore, we can define the immediate reward as follows,

$$r(s_l, a_l) = \varphi_a \Sigma_{mvno} + \varphi_b (R_u^{m,j} - r_u), \tag{26}$$

where φ_a and φ_b are the weights of objective function and QoS loss, respectively.

Algorithm 1 The proposed Q-learning scheme

```
Initialize $Q(\mathbf{s}, \mathbf{a})$
for each episode do
    Initialize s of each BS randomly.
    for each time do
        Select $a_l$ from all actions of state $s_l$;
        Execute selected $a_l$, observe reward and next state $s_{l+1}$;
        $Q(s_l, a_l) \leftarrow \kappa r(s_l, a_l) + \kappa\xi \max_{a_{l+1}} Q(s_{l+1}, a_{l+1}) + (1-\kappa)Q(s_l, a_l)$;
        Let $s_l \leftarrow s_{l+1}$.
    end for
end for
```

C. Q-Learning Method

At time slot l, the network controller first watches the state $s_l \in \mathcal{S}$ of all the BSs, and then chooses an action $a_l \in \mathcal{A}$ according to a stochastic policy π. After selecting an action, the network controller can transmit the action information to the BSs via control signaling and a reward $r(s_l, a_l)$ can be achieved. Then the network will take a transition to s_{l+1}.

Therefore, each pair of state-action has a value $Q(s_l, a_l)$ for time slot l. $Q(s_l, a_l)$ is the expected cumulative future discounted reward at state s_l and action a_l, which can be expressed as

$$Q(s_l, a_l) = \mathbb{E}\left[\hat{r}_l | s_l, a_l\right], \tag{27}$$

where $\hat{r}_l = \sum_{t=l}^{T} \xi r(s_t, a_t)$ and $\xi, 0 \leq \xi \leq 1$ is a discount parameter. If $\xi \rightarrow 1$, the future is the main focus and if $\xi \rightarrow 0$, the immediate reward will be mainly considered. The network controller computes $Q(s_l, a_l)$, the value of which is stored in a Q-table for each time slot.

By considered a learning rate κ, the presented Q-learning scheme is shown in Alg. 1. In this algorithm, the value of $Q(s_l, a_l)$ is iterated in each step. When the optimal policy $\pi(s_l) = \max_{a_l} Q(s_l, a_l)$ is satisfied, the optimal function $Q^*(s_l, a_l)$ for action a_l is obtained and it should follow the Bellman optimality equation:

$$Q^*(s_l, a_l) = r(s_l, a_l) + \xi \max_{a_{l+1}} Q^*((s_{l+1}, a_{l+1}) | s_l, a_l), \tag{28}$$

D. Proposed DRL-based Solution

In the Q-learning, there should be a Q-table which consists of all possible states as its rows and actions as its columns for each BS. This Q-table will be the reference for the network controller to select the proper action according to the Q-value. Although using such a table relieves the dependence on full network statistics information, but Q-learning still needs to confront the problem of a huge state space.

In the considered WVN, as many different entities are involved, the possibility that a very large amount of states and actions co-exist will be very high. Then, the dimension of Q-table will be very high if all the state and actions are stored. Consequently, the algorithm may not be working properly as it is difficult to get enough samples to traverse each state. Therefore, we can utilize the Q-learning with neural network (NN) to estimate $Q(\mathbf{s}, \mathbf{a})$ instead of calculating each pair's Q-value, which leads to the concept of Deep Q-Network (DQN).

Denoting θ as the weight, we can use a NN $Q(\mathbf{s}, \mathbf{a}; \theta)$ to represent Q-function, which results in the Q-network. We can train the Q-network to approximate the real Q-values by updating the value of θ at each iteration. When incorporating with the NN, the performance of the Q-learning on flexibility is able to be improved [7]. When it comes to the DQN, DNN is used in stead of NN in Q-network and it has been proved as a robust learning approach with better performance[36]. Comparing with the Q-network, there are three major improvements in the DQN [36][37].

The first one is that DNN can replace the ordinary NN with a multiple layer structure. In the DNN, the multiple layers of convolution filters are used to explore the local spatial correlations. Therefore, DQN is able to extract the high-level features of input raw data. In addition, the experience replay in DQN is able to save the experience tuple $e(l) = (s_l, a_l, r_l, s_{l+1})$ into a replay memory $\mathcal{O}$. Then from the memory, a randomly sample batch $\hat{\mathcal{O}}$ can be used to train the DNN. In this way, DQN can learn from past experience instead of only from the current one. Moreover, a second network is adopted and it can compute target Q-values which can be used to compare with the estimated Q-values to obtain the loss of each action. Using one network for the target Q-values and estimated ones can fall into feedback loops between the target and estimated values.

As explained, in each iteration, a DNN is used to represent $Q(s_l, a_l)$ in the DQN. Ba sed on the sample batch $\hat{\mathcal{O}}$ taken from experience memory, policy π and θ are updated to train the DQN in a online manner. DQN can be optimized by minimizing $\mathcal{L}(\theta)$. Denoting Ω_l as the target Q-value,$\mathcal{L}(\theta)$ is expressed as

$$\mathcal{L}(\theta) = \mathbb{E}[\Omega_l - Q(s_l, a_l; \theta)^2], \tag{29}$$

where

$$\Omega_l = r(s_l, a_l) + \max_{a_{l+1}} Q^*(s_{l+1}, a_{l+1}.\theta^-). \tag{30}$$

In (30), when the online network $-Q(\mathbf{s}, \mathbf{a}; \theta)$ is updated by gradient descent, target network parameter θ^- is frozen for some iterations. Specially, according to (28), the network controller selects action a_l, gets reward r_l at time slot l and then transitions to s_{l+1}. A experience replay memory $\mathcal{O}$ is used by network controller to save (s_l, a_l, r_l, s_{l+1}). In order to achieve the balance between exploration and exploitation, the ϵ-greedy policy can be applied. That is, we aim at balancing the reward maximization according to the known information with selecting new actions to obtain unknown information. We present the proposed DRL-based scheme in Alg. 2 and its flow is illustrated in Fig. 4. As the network controller in the WVN can manage the radio resources, and collect the corresponding information from the environment via e.g., physical networks, the proposed DRL solution can be executed in the network controller in the centralized way.

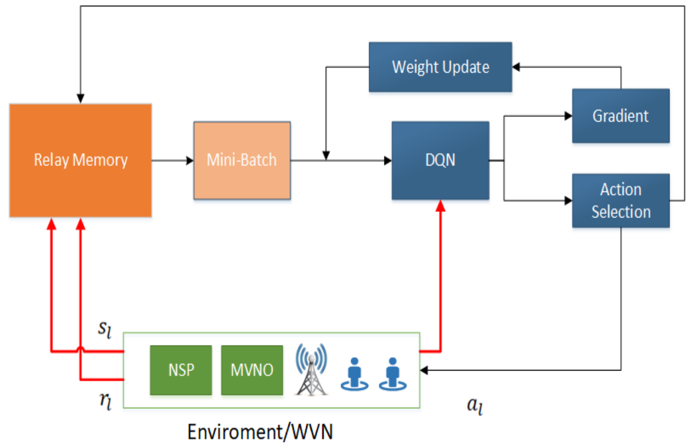


Fig. 4. Flow of the proposed scheme

Algorithm 2 DRL-based user association and resource allocation method

```
Initialize experience replay memory $\mathcal{O}$ and parameter of the DNN $\theta$ with random weights
for each episode do
    Initialize the parameters of WVN scenario
    Obtain observations on the state $s_1$.
    for each time slot $l$ do
        Choose $a_l$ randomly with probability $\epsilon$, otherwise, select $a_l = \arg\max_a Q(x, a, \theta)$;
        Execute selected $a_l$, observe reward and $s_{l+1}$;
        Save $(s_l, a_l; r_l, s_{l+1})$ in replay memory $\mathcal{O}$;
        Sample a random batch of $Y$ vectors $(s_i, a_i; r_i, s_{i+1})$ from $\mathcal{O}$;
        Calculate the target Q-value $\Omega_i$ from the target DQN, as follows,
            $\Omega_i = r_i + \xi \max_{a_{l+1}} Q(s_{i+1}, \arg\max_{a'} Q(s_{i+1}, a', \theta), \theta^-)$   (31)
        Update the main DQN by minimizing $\mathcal{L}(\theta_i)$,
            $\mathcal{L}(\theta) = \frac{1}{Z} \sum_i (\Omega_i - Q(s_i, a_i, \theta))^2).$   (32)
        Execute a gradient descent step on $\mathcal{L}(\theta)$ with respect to $\theta$.
    end for
end for
Output: the optimal user association strategy $\boldsymbol{\beta}$, resource allocation factors $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, and power allocation $\mathbf{p}$
```

The information exchange and updates (e.g. reward and action) in the DRL can be realized by the signalling exchange in the virtualization process.

VI. Performance Evaluations and Discussions

In this section, we will present the simulation results to illustrate the proposed schemes. Here, we consider there are 2 InPs, 4 SBSs, and 15 users unless specified. The profit of the user is set to 5 unit/b/s. The price of backhaul of InP 1 is 1 unit/b/s, while the price of backhaul of InP 2 is 1.2 units/b/s. The prices for requesting virtual resources from InP 1 are 20 units/w/Hz, 18 units/w/Hz, and 15 units/w/Hz, and from InP 2 are 18 units/w/Hz, 16 units/w/Hz, and 13 units/w/Hz, respectively. The transmit power of 49 dBm is assumed for MBS and the maximum transmit power of SBS is 20 dBm. The channel bandwidth is 20 MHz. The parameter setting of price is basically according to [15]. To implement the DRL, TensorFlow is used. First, we examine the impact of various factors on the system utility to see their impact. Then we evaluate the proposed schemes and show their performance.

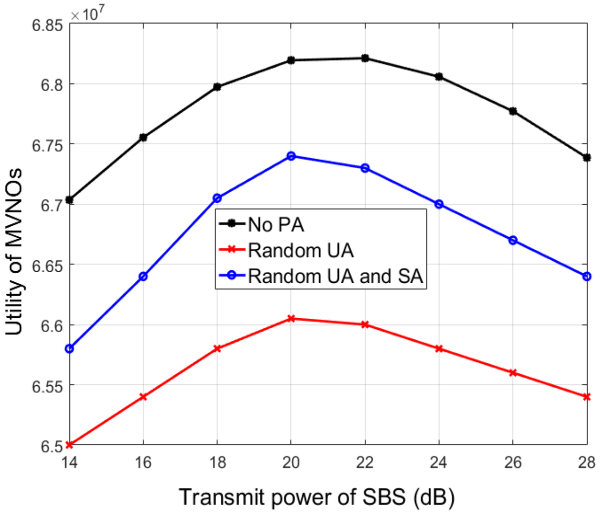


Fig. 5. Impact of transmit power of BS on the utility performance.

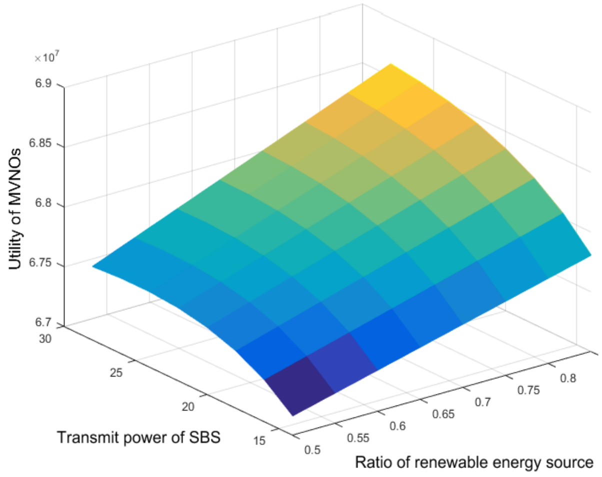


Fig. 6. Impact of transmit power and ratio of RES of BS on the utility performance.

First, to examine the impact of transmit power of SBS, we vary the transmit power and plot utility of the MVNOs in Fig. 5 using the proposed ADMM-based scheme without power allocation ('No PA'). In addition, we also evaluate the

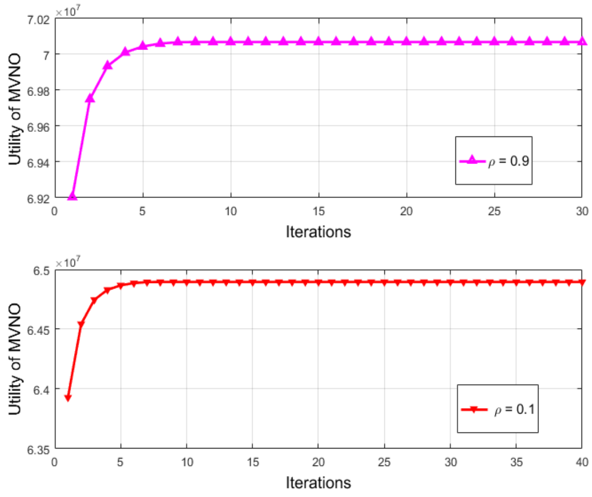


Fig. 7. Convergence performance of proposed ADMM-based scheme.

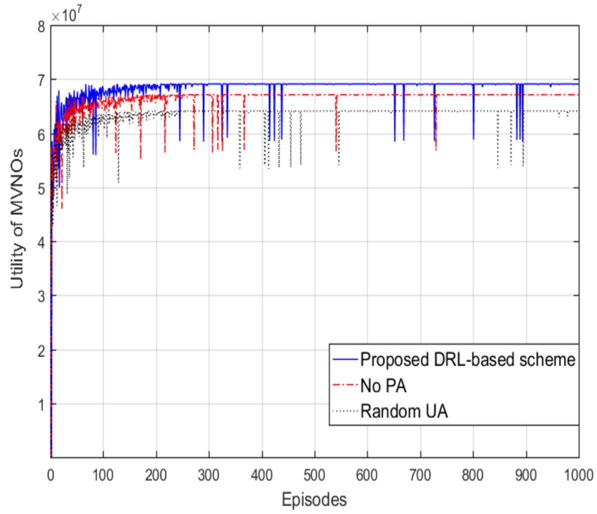


Fig. 8. Convergence performance of different DRL-based schemes.

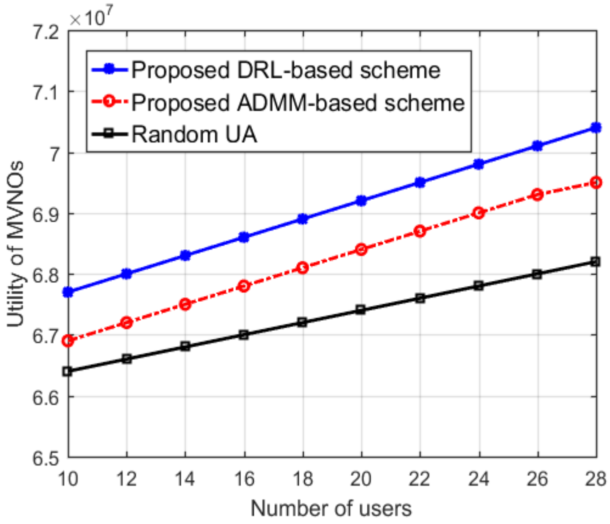


Fig. 9. Utility of MVNO vs. number of users.

proposed ADMM scheme without power allocation and with random user association ('Random UA') and the proposed ADMM scheme without power allocation and with random user association and spectrum allocation ('Random UA and SA'). From this figure, we can observe that the system utility first becomes larger as the transmit power increases, then reaches its maximum and decreases. Such an observation evidences the necessity of investigating the power allocation scheme for the considered system. Moreover, we can also find the proposed user allocation and spectrum can effectively improve the utility of MVNOs.

In Fig. 6, we jointly change the transmit power and the usage of RES and plot the utility performance of MVNOs in three dimension figure. It can be observed that with the variation of the transmit power, the utility of MVNOs has an optimal value. Such a observation also evidences the findings of Fig. 5. We can also find that incremental of ratio of RES usage results in the increase of system utility. Such a phenomenon indicates RES usage should be encouraged, which conforms to the pricing strategies in (9). In Fig. 7, the convergence performance of the proposed ADMM-based algorithm is examined, where there are 30 users. It can be found by properly choosing the parameters, the proposed ADMM-based algorithm can converge with a fast rate.

We present the convergence performance of DRL-based scheme in Fig. 8. In this figure, we also show the effectiveness of power allocation and user association, by comparing the proposed DRL-based scheme with the one without power allocation ('No PA') and the one with random user association (random UA). One can observe that the proposed DRL-based scheme has a good convergence rate and outperform the others. From this figure, it can also be found that utility of MVNOs of all cases are low at the beginning. As the time goes on, the utility all three cases become larger until reaching a relatively stable value. In addition, the proposed DRL-based scheme outperform the other two, which shows the necessity of investigating the power allocation and user association.

To evaluate the effects of the number of users, Fig. 9 and Fig. 10 vary the number of users and plots the utility performance and system throughput, respectively. Moreover, we compare the proposed DRL-based scheme and the proposed ADMM-based scheme. In addition, we also plot the proposed ADMM-based scheme with random user association (Random UA). We can see our proposed DRL scheme outperform the ADMM-based scheme in both cases. This is mainly because the proposed ADMM-based scheme may fail to perform accurate resource allocation in a dynamic scenario and then results in performance loss.

VII. Conclusion and Future

We have investigated resource allocation and user association schemes for a wireless virtualized heterogeneous network with hybrid energy supply in this work. In particular, in order to maximize the utility for all the MVNOs, a joint spectrum and power allocation, and user association problem is introduced. We first present a ADMM-based optimization

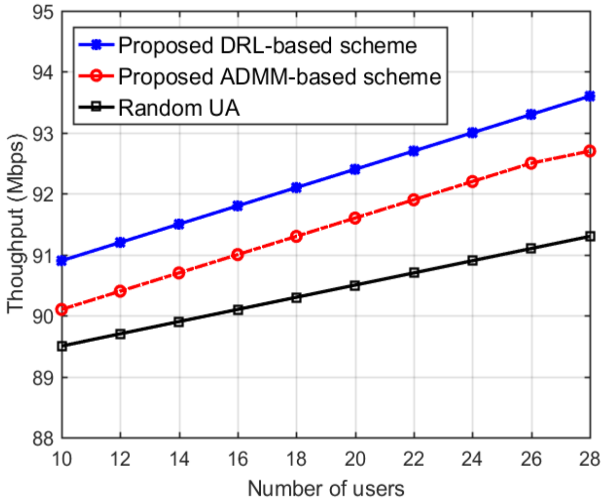


Fig. 10. Throughput vs. number of users

scheme to find the solution in a static manner. Moreover, based on deep reinforcement learning, we also propose to learn the optimal strategy without having the priori information of network dynamics. Extensive simulation studies have been conducted and the performance evaluation demonstrates the advantages of our proposed schemes. In the future, we will take multi-antenna effects into consideration when designing the virtual resource allocation scheme. When considering multi-antenna effects, the overhead of estimating all involved channels should be carefully considered and the problems related to CSI uncertainty should be addressed [39] [40]. We will focus on developing learning-based scheme to address the induced problems.

References

[1] Y. L. Lee, D. Qin, L. -C. Wang and G. H. Sim, "6G Massive Radio Access Networks: Key Applications, Requirements and Challenges," *IEEE Open Journal of Vehicular Technology*, vol. 2, pp. 54-66, 2021.

[2] C. Liang and F. R. Yu, "Wireless network virtualization for next generation mobile cellular networks," *IEEE Comm. Mag.*, vol. 22, no. 1, pp. 61-69, March 2015.

[3] M. Kamel, L. B. Le, and A. Girard, "LTE multi-cell dynamic resource allocation for wireless network virtualization," *Proc. IEEE WCNC'2015*, New Orleans, LA, March 2015.

[4] G. Piro, M. Miozzo, G. Forte, N. Baldo, L. A. Grieco, G. Boggia, and P. Dini, "Hetnets powered by renewable energy sources: sustainable next-generation cellular networks," *IEEE Internet Comp.*, vol. 17, no. 1, pp. 32-39, 2013.

[5] Z. Chang, C. Jing, X. Guo, Z. Han, and T. Ristaniemi, "Resource allocation for wireless virtualized hetnet with caching and hybrid energy supply", *Proc. IEEE WCNC'2018*, Barcelona, Spain, April 2018.

[6] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan and D. Matolak, "A machine learning approach for power allocation in hetnets considering qos," *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, 2018.

[7] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.

[8] N. Tekbiyik, T. Girici, E. Uysal-Biyikoglu, and K. Leblebicioglu, "Proportional fair resource allocation on an energy harvesting downlink," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1699-1711, Apr. 2013.

[9] Q. Han, B. Yang, G. Miao, C. Chen, X. Wang and X. Guan, "Backhaul-aware user association and resource allocation for energy-constrained Hetnets," *IEEE Trans. Veh. Tech.*, vol. 66, no. 1, pp. 580-593, Jan. 2017.

[10] T. Han and N. Ansari, "Provisioning green energy for base stations in heterogeneous networks," *IEEE Trans. Veh. Tech.*, vol. 65, no. 7, pp. 5439-5448, Jul. 2016.

[11] T. Han and N. Ansari, "A traffic load balancing framework for software-defined radio access networks powered by hybrid energy sources," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 1038-1051, Apr. 2016.

[12] S. Hong, J. Mehlman, and S. Katti, "Picasso: flexible RF and spectrum slicing," *Proc. ACM SIGCOMM'12*, Helsinki, Finland, Aug. 2012.

[13] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "Nvs: a substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333-1346, Oct. 2012.

[14] J. Panchal, R. Yates, and M. Buddhikot, "Mobile network resource sharing options: performance comparisons," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4470-4482, Sep. 2013.

[15] C. Liang, F. R. Yu, H. Yao, and Z. Han, "Virtual resource allocation in information-centric wireless networks with virtualization," *IEEE Tran. on Veh. Tech.*, vol. 65, no. 12, pp. 9902-9914, Dec. 2016.

[16] Y. Yu, X. Bu, K. Yang, H. K. Nguyen and Z. Han, "Network function virtualization resource allocation based on joint benders decomposition and admm," *IEEE Tran. on Veh. Tech.*, vol. 69, no. 2, pp. 1706-1718, Feb. 2020.

[17] H. Zhang, W. Wang, X. Li and H. Ji, "User association scheme in Cloud-RAN based small cell network with wireless virtualization," *Proc. of INFOCOM'2015*, Hong Kong, China, April 2015.

[18] Z. Chang, Z. Han and T. Ristaniemi, "Energy efficient optimization for wireless virtualized small cell networks with large-scale multiple antenna," *IEEE Trans. Commun.*, vol. 65, no. 4, April 2017.

[19] L. Chen, F. R. Yu, H. Ji, G. Liu, and V.C.M. Leung, "Distributed virtual resource allocation in small cell networks with full duplex self-backhauls and virtualization," *IEEE Trans. Veh. Tech.*, vol. 65, no. 7, pp. 5410-5423, July 2016.

[20] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. -C. Chen and L. Hanzo, "Thirty Years of Machine Learning: The Road to Pareto-Optimal Wireless Networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1472-1514, thirdquarter 2020.

[21] J. Wang, C. Jiang, K. Zhang, X. Hou, Y. Ren and 0Y. Qian, "Distributed Q-Learning Aided Heterogeneous Network Association for Energy-Efficient IIoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2756-2764, April 2020.

[22] Y. Sun, M. Peng and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1960-1971, April 2019.

[23] C. Huang, Z. Yang, G. C. Alexandropoulos, Kai Xiong, Li Wei, Chau Yuen, Zhaoyang Zhang, "Hybrid Beamforming for RIS-Empowered Multi-hop Terahertz Communications: A DRL-based Method," arXiv preprint, arXiv:2009.09380, https://arxiv.org/abs/2009.09380.

[24] A. Asheralieva, "Bayesian reinforcement learning-based coalition formation for distributed resource sharing by device-to-device users in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5016-5032, Aug. 2017.

[25] N. Zhao, Y. Liang, D. Niyato, Y. Pei, M. Wu and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141-5152, Nov. 2019.

[26] J. S. Pujol Roig, D. M. Gutierrez-Estevez and D. Gündüz, "Management and Orchestration of Virtual Network Functions via Deep Reinforcement Learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 2, pp. 304-317, Feb. 2020.

[27] X. Fu, F. R. Yu, J. Wang, Q. Qi and J. Liao, "Dynamic Service Function Chain Embedding for NFV-Enabled IoT: A Deep Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 507-519, Jan. 2020.

[28] H. Xiang, S. Yan and M. Peng, "A Realization of Fog-RAN Slicing via Deep Reinforcement Learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2515-2527, April 2020.

[29] H. K. Nguyen, Y. Zhang, Z. Chang, and Z. Han, "Parallel and distributed resource allocation with minimum traffic disruption for network virtualization," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1162-1175, Mar. 2017.

[30] S. Boyd and L. Vandenberghe, "Convex optimization," *Cambridge university press*, 2009.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, Jan. 2011.

[32] W. Yu, and R. Lui, "Dual methods for nonconvex spectrum optimization in multicarrier system," *IEEE Trans. on Commun.*, vol. 54, no. 7, pp. 1310-1322, Jul. 2006.

[33] D. W. K. Ng, E. S. Lo and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292-3304, Sep. 2012.

[34] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," *in Proc. AAAI*, Phoenix, AZ, Feb. 2016.

[35] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvari, "Convergence results for single-step on-policy reinforcement- learning algorithms," *Mach. Learn.*, vol. 38, no. 3, pp. 287-308, Mar. 2000.

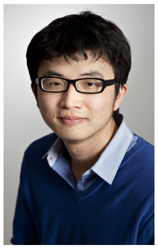
[36] V. Minh et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-533, 2015.

[37] H. Y. Ong, K. Chavez, and A. Hong, "Distributed deep Q-learning," arXiv:1508.04186, 2015.

[38] E. Vlachos, G. C. Alexandropoulos and J. Thompson, "Wideband MIMO Channel Estimation for Hybrid Beamforming Millimeter Wave Systems via Random Spatial Sampling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 1136-1150, Sept. 2019.

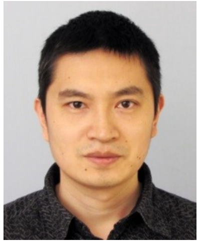
[39] G. C. Alexandropoulos, P. Ferrand, J. Gorce and C. B. Papadias, "Advanced coordinated beamforming for the downlink of future LTE cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 54-60, July 2016, doi: 10.1109/MCOM.2016.7509379.

[40] G. C. Alexandropoulos, P. Ferrand and C. B. Papadias, "On the Robustness of Coordinated Beamforming to Uncoordinated Interference and CSI Uncertainty," *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1-6, doi: 10.1109/WCNC.2017.7925853.



Zheng Chang (S'10-M'13-SM'17) received the B.Eng. degree from Jilin University, Changchun, China in 2007, M.Sc. (Tech.) degree from Helsinki University of Technology (Now Aalto University), Espoo, Finland in 2009 and Ph.D degree from the University of Jyväskylä, Jyväskylä, Finland in 2013. Since 2008, he has held various research positions at Helsinki University of Technology, University of Jyväskylä and Magister Solutions Ltd in Finland. He was a visiting researcher at Tsinghua University, China, from June to August in 2013, and at University of Houston, TX, from April to May in 2015. He has been awarded by the Ulla Tuominen Foundation, the Nokia Foundation and the Riitta and Jorma J. Takanen Foundation for his research excellence. He has been awarded as 2018 IEEE Communications Society best young researcher for Europe, Middle East and Africa Region.

He has published over 100 papers in Journals and Conferences, and received received best paper awards from IEEE TCGCC and APCC in 2017. He serves as an editor of IEEE Wireless Communications Letters, Springer Wireless Networks and International Journal of Distributed Sensor Networks, and a guest editor for IEEE Networks, IEEE Wireless Communications, IEEE Communications Magazine, IEEE Internet of Things Journal, IEEE Transactions on Industrial Informatics, Physical Communications, EURASIP Journal on Wireless Communications and Networking, and Wireless Communications and Mobile Computing. He was selected as an exemplary reviewer of IEEE Wireless Communication Letters in 2018. He has participated in organizing workshop and special session in Globecom' 19, WCNC'18-22, SPAWC'19 and ISWCS'18. He also serves as TPC member for many IEEE major conferences, such as INFOCOM, ICC, and Globecom. His research interests include IoT, cloud/edge computing, security and privacy, vehicular networks, and green communications.



Tao Chen (M'10-SM'13) received his Ph.D. degree in telecommunications engineering from the University of Trento, Italy in 2007. Since then he joined VTT Technical Research Centre of Finland and currently he is a senior researcher in the Connectivity research area. He is the docent (adjunct professor) in University of Jyväskylä, the project coordinator of the EU H2020 COHERENT project, and the board member of EU 5G PPP Steering Board. His current research interests include software defined networking for 5G mobile networks, massive IoT in 5G, dynamic spectrum access, energy efficiency and resource management in heterogeneous wireless networks, and social-aware mobile networks.