IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 28, 2019, accepted August 8, 2019, date of publication September 11, 2019, date of current version September 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940435

ARMA-Prediction-Based Online Adaptive Dynamic Resource Allocation in Wireless Virtualized Network

LUN TANG, XIAOYU HE, XIXI YANG, YANNAN WEI, XIAO WANG, AND QIANBIN CHEN, (Senior Member, IEEE)
School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
Key Laboratory of Mobile Communication, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Qianbin Chen (cqb@cqupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571073, and in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJZD-M201800601.

ABSTRACT Wireless network virtualization (WNV) provides a novel paradigm shift in the fifth-generation (5G) system, which enables to utilize network resources more efficiently. In this paper, by jointly considering cache space and time-frequency resource allocation in wireless virtualized networks, we first formulate an optimization programming to investigate the minimization problem of network overheads while satisfying the quality of service (QoS) requirements of each virtual network on overflow probability. Then, with diverse demands of virtual networks for different kinds of resources taken into consideration, an online adaptive virtual resource allocation algorithm with multiple time-scales based on auto regressive moving average (ARMA) prediction method is proposed to solve the formulation, which could eliminate the irrationalities existed in traditional approaches caused by the uncertainty of traffic and information feedback delay. More specifically, in the proposed resource scheduling mechanism with multiple time-scales, on the one hand, a reservation strategy of cache space is developed according to the ARMA prediction information under long time-scales. On the other hand, virtual networks are sorted by the overflow probabilities derived by the large-deviation principle and dynamic time-frequency resource scheduling under short time-scales. Simulation results reveal that our proposal can provide tangible gains in reducing the bit loss rate and improving the utilization of physical resources.

INDEX TERMS Wireless virtualized networks, resource allocation, multiple time-scales, ARMA, large-deviation principle.

I. INTRODUCTION

With the rapid development of intelligent terminals, the flourish of diversified applications which have different demands for delay, reliability and throughput, has brought great challenges to the existing network [1]. For 5G system, the wireless network virtualization technology has emerged as a key concept to provide an effective solution to network sharing [2]–[5]. By leveraging spectrum sharing, infrastructure virtualization, and hollow virtualization technologies and so forth, wireless networks can realize the unified management of resources, and improve the flexibility of network deployment and reduce the capital expenditures (CAPEX) and operating expenditures (OPEX), meanwhile meeting the QoS of different application scenarios [6], [7]. The essence of wireless network virtualization is to virtualize and reconstitute into multiple kinds of virtual network resources, such as virtual spectrum resources, cache resources and so forth. After virtualization, traditional operators are decoupled into two separate roles, i.e. infrastructure provider (InP) and service provider (SP). The InP is responsible for abstracting and slicing the physical resources, while the SP provides end-to-end services to users by leasing virtual resources from the InP [8]. Appropriate virtual resources should be provided for SPs based on their respective demands, so as to form multiple virtual networks which coexist in the same physical network but are logically independent of each other. This brings huge benefits of improving resource utilization and

The associate editor coordinating the review of this manuscript and approving it for publication was Noor Zaman.

reducing the CAPEX and OPEX [9], [10]. However, considering the limitation of physical resources and diversity of virtual network requirements, it is critical to design an effective virtual network resource allocation mechanism to improve network virtualization performance [11].

A lot of efforts have been dedicated to this issue. Jiang *et al.* [12] designed priorities for different virtual networks and service users respectively, and proposed a heuristic control mechanism to solve access control and dynamic resource scheduling problem of users in virtual networks. Later, to achieve efficient allocation of network resources, in [13], a game mechanism was presented to investigate buffer-space and wireless bandwidth scheduling problem in wireless virtualized network. A joint resource allocation and content caching problem was studied in [14], which aimed to efficiently utilize the radio and content storage resources in the highly congested backhaul scenario. Sciancalepore *et al.* [15] designed a virtual resource allocation algorithm based on service prediction which can satisfy the service level agreement (SLA) of different network slices and improve the resource utilization. Moreover, in [16], L.Tang et al. proposed an integrated virtualization framework with the frequency division duplexing (FDD) self-backhaul mechanism, and formulated a stochastic optimization model to investigate the average total utility maximization problem in the wireless virtualized networks. In [17], a resource management scheme was proposed by introducing two types of slices, namely the rate-based slices and the resource-based slices, which require the minimum data rate and the minimum network resources, respectively, and the result of [17] had been extended to multi-cell scenario in [18]. Lu *et al.* [19] proposed a multi-step dynamic optimization to achieve efficient resource utilization in the case of limited transmission power.

The concept of auction game was also applied for the interactions among slices, network operators and users in wireless virtualized networks. For example, in [20]–[22], the network operators managed the spectrum resources and each slice was responsible for its own users with different QoS requirements. Furthermore, there were some other works focusing on the combination of wireless network virtualization and other key technologies to provide better services. For instance, a resource sharing scheme in C-RAN was proposed in [23]. Later, Ahmadi *et al.* [24] proposed a virtualization solution in cloud radio access network (C-RAN) scenarios, and studied the complementarity and partial substitutability between spectrum resources and antenna resources based on massive multiple input and multiple output (MIMO) technology. Zhou *et al.* [25] applied the massive MIMO to extend the feasibility condition of wireless virtualized networks. Moreover, a virtual network isolation scheme in single cellular scenario was studied in [26], where the system allocates orthogonal spectral bands for virtual networks to avoid mutual interferences, and dynamically adjusted the transmission power on each spectrum block to prevent the data packet accumulation on the backhaul links with limited capacity. The authors in [27] proposed an optimal virtual resource allocation strategy in the information-centric heterogeneous virtualized networks, where the gains of not only virtualization but also caching and computing are taken into consideration.

Although some excellent works have been done on virtual resource allocation algorithm in wireless virtualization scenarios, most of them focused on solving the resource scheduling problem within a single time interval, while neglecting the network dynamics in time domain. Consider that the stochasticity of network states and delay caused by information feedback may lead to unreasonable virtual resource allocation in the networks, in this paper, we investigate the joint cache space and time-frequency resource blocks (RBs) allocation problem in dynamic wireless virtualized networks with large-deviation principle and auto regressive moving average (ARMA) prediction, then propose an adaptive virtual resource allocation algorithm to solve the formulation. The main contributions of this paper are summarized as follows:

- By jointly considering cache space and time-frequency RBs in the wireless virtualized network, we formulate an optimization problem to minimize the network overheads while satisfying the QoS requirement of each virtual network on the overflow probability.
- Leveraging the large-deviation principle and the ARMA prediction method, we propose an online adaptive virtual resource allocation algorithm with multiple time-scales to solve the formulation while taking diverse demands of virtual networks for different types of resources into consideration, which could eliminate the unreasonableness existed in traditional approaches.
- In the proposed resource scheduling mechanism with multiple time-scales, a reservation strategy of caching space is designed according to the ARMA prediction information under long time-scales, and virtual networks are sorted by the overflow probabilities derived by the large-deviation principle and dynamic time-frequency RBs scheduling under short time-scales.
- Both theoretical analyses and simulation results are given to validate the effectiveness of our proposed algorithm and show that our proposal can provide significant gains in reducing the bit loss rate and improving the utilization of physical resources.

The rest of this paper is organized as follows. In section II, we introduce the system model and problem formulation. The proposed dynamic virtual resource allocation with multiple time-scales is elaborated in Section III. In Section IV, simulation results are given to demonstrate the effectiveness of our proposed algorithm. Finally the conclusion and future works are drawn in Section V.

II. SYSTEM MODEL

A. SYSTEM FRAMEWORK

Fig. 1 shows the architecture of dynamic resource configuration based on virtualization technology, which consisting of wireless virtual network users, virtual network

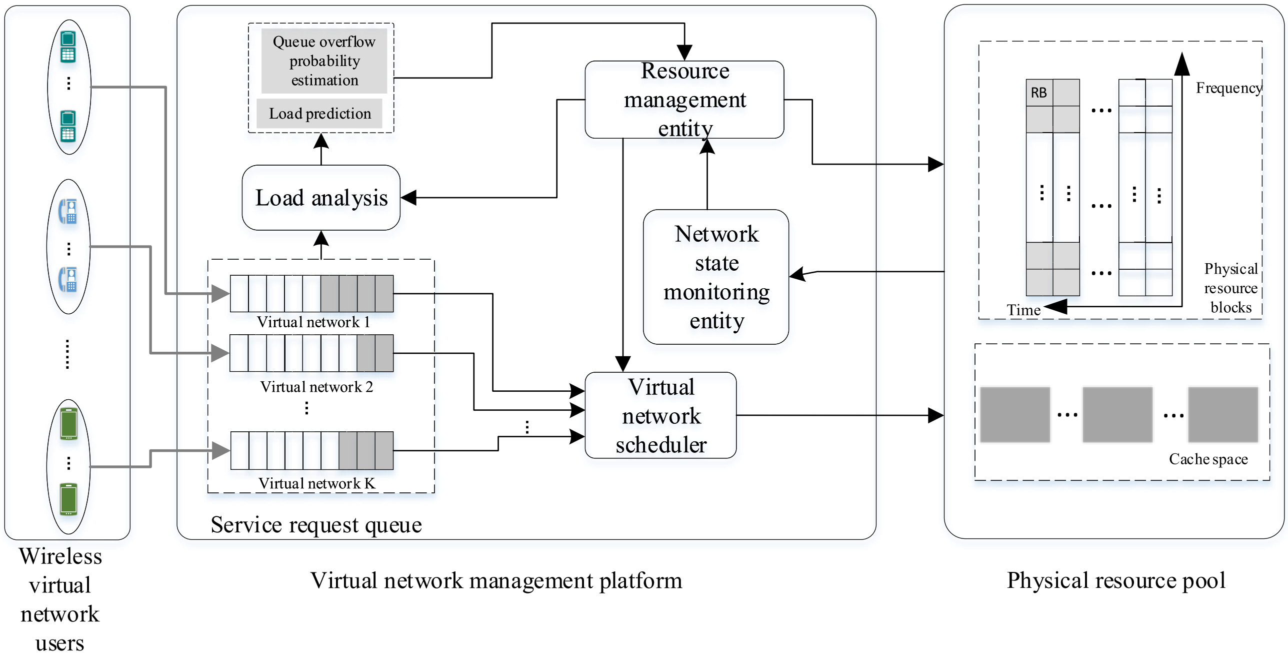

Queue overflow probability estimation
Load prediction
Resource management entity
Load analysis
Network state monitoring entity
Virtual network 1
Virtual network 2
Virtual network K
Virtual network scheduler
Service request queue
RB
Frequency
Time
Physical resource blocks
Cache space
Wireless virtual network users
Virtual network management platform
Physical resource pool

FIGURE 1. System framework.

management platform, and physical resource pool. In this system, the physical resource pool provides multiple kinds of physical resources, including computing resources, cache, wireless bandwidth resources and so forth. Virtual network management platform allocates adequate physical resources for each virtual network, based on service states, instantaneous channel conditions, QoS requirements, etc, of each virtual network user. In order to allocate physical resources more effectively, and achieve an efficient utilization of physical resources, in this section, we design a virtual network management platform which is comprised of service request units, load analysis module, resource management entity, network state monitoring entity, and virtual network scheduler. The service request units are used for caching newly arrived and failed processed services of virtual network users. The load analysis module is utilized to analyze the load characteristics of each virtual network, and predict the load state in next period and estimate the queue overflow probability. The resource management entity determines the optimal physical resources for each virtual network on the basis of the evaluation of the load analysis module, thus ensuring the QoS requirements of each virtual network. The network state monitoring entity is to observe the real-time state of each kind of physical resources. Moreover the basic function of the network scheduler is to dispatch the available virtual resources for each virtual network user.

In this paper, K represents a set of virtual networks in this system. The main work is to design a dynamic allocation strategy of cache space and time-frequency RBs for the resource management entity in the virtual network management platform, with the purpose of decreasing the physical resource leasing cost and simultaneously guaranteeing the QoS requirements of each virtual network.

B. PROBLEM STATEMENT

As mentioned above, each virtual network rents a certain amount of space to cache data of its service users, and at the same time leases RBs to provide users with data transmission services. As for virtual network $k, k \in K$, $A_k(t) \in K \triangleq \{0, \ldots, A_{\max\_k}\}$ indicates the number of data packets of service reached within the scheduling period t where $A_{\max\_k}$ denotes the maximum amount of packets reached in a single period. Due to randomness of data generated by non-periodic applications of virtual network users, we assume that the arrival process of data packet $A_k(t)$ is random independent identical distributed. $D_k(t) \in D \triangleq \{0, 1, \ldots, D_{\max\_k}\}$ indicates the number of packets of virtual network k left during scheduling period t, where $D_{\max\_k}$ denotes the maximum amount of packets left in a single period. In addition, $Q_k(t)$ is defined as the queue length of virtual network k at the beginning of scheduling period t. Therefore, the queue dynamics of virtual network k can be expressed as:

$$Q_k(t+1) = \max\{Q_k(t) - D_k(t) + A_k(t), 0\}. \tag{1}$$

Through Little theorem, the relationship between the average queue length of virtual network and the average queuing delay of service can be described as:

$$\overline{Q}_k = \lambda_k \overline{L}_k, \tag{2}$$

where $\overline{Q}_k$ is the average queue length of virtual network k, $\overline{L}_k$ is the average queuing delay, λ_k is the arrival rate of service. In equation (2), if the arrival rate λ_k is given, it can be concluded that the longer the average queuing length of virtual network in system, the greater the waiting time of cached data. To this end, the delay performance can be directly affected by controlling the caching queue of virtual network. In this work, the objective is to select an appropriate mechanism of physical resource allocation for each virtual network to control the growth rate of its caching queue effectively, so as to ensure the delay requirements of each virtual network. In order to describe the matching degree between the service request and the allocated cache space, the queue overflow probability of each virtual network can be defined as:

$$P_{overflow}^{k} = P(Q_k(t) > B_k), \quad \forall k, \tag{3}$$

where B_k represents the size of cache space that virtual network k rents during current scheduling period. Since the service arrival process in this paper is bit-sized, the size of cache space is also described as bit-sized storage capability. It is worth mentioning that equation (3) can also on behalf of the bit loss of each virtual network. When the transmission rate or caching capacity of virtual network is insufficient, queue overflow means that data will be lost. For that reason, the QoS requirements of each virtual network can be described as the following queue overflow probability constrained problem:

$$P(Q_k(t) > B_k) < \varepsilon_k, \quad \forall k, \tag{4}$$

If the virtual network management platform allocates enough time-frequency RBs for virtual network k to obtain sufficient service rate, or the virtual network k rents plenty of cache, thus the platform have prominent service caching capacity and its overflow probability will be lower than the threshold ε_k, therefore the QoS requirements of each virtual network will be satisfied. Nevertheless, due to scarcity of physical resources, physical network will not be capable to provide infinite resources for each virtual network. In addition, leasing appropriate rather than excessive physical resources will provide a certain service assurance for its users, and bring better economic benefits to the virtual network. For the sake of guaranteeing the QoS requirements optimally and reducing the total cost of virtual network services, the dynamic resource allocation problem in this paper can be established as the following mathematical model:

$$\begin{aligned} &\min_{B,x} \sum_{k=1}^{K} \left[\rho_k B_k + \alpha_k \sum_{n=1}^{N} x_{n,k} \right] \\ &C1: \sum_{k=1}^{K} x_{n,k} = 1, \quad \forall n \\ &C2: P\left(Q_k\left(t\right) > B_k\right) < \varepsilon_k, \quad \forall k \\ &C3: \sum_{k=1}^{K} B_k \leq B_{tot} \end{aligned} \tag{5}$$

where the overhead of each virtual network leasing resources consists of two portions, and the first represents the cost of cache space, while the second represents the cost of employing time-frequency RBs. In equation (5), ρ_k denotes the unit price of virtual network k for leasing cache space, and α_k indicates the unit price for RBs. K is the number of virtual networks and N is the total quantity of RBs. In C1, $x_{n,k}$ is a binary assignment indication of RBs, which $x_{n,k} = 1$ denotes that the RB n is configured for virtual network k, and it is assumed that a RB can only be configured for one virtual network. C3 is the constraint for the physical cache space limit where B_{tot} is the upper limit of cache space.

III. DYNAMIC SCHEDULING MECHANISM WITH MULTIPLE TIME-SCALES

On account of the existence of constraint C2 in equation (5), a strong dependence has been formed between two physical resources. Moreover, in realistic scenario, the service request of each virtual network arrives randomly, in addition, the configuration of virtual cache space has a certain delay [29]. Thus, it will be unrealistic and inaccurate to determine the scheduling strategy of cache space and time-frequency RBs in the same period. As a consequence, we design a resource allocation mechanism with multiple time-scales, which allocates cache space for each virtual network in long period and time-frequency RBs in short period respectively. The specific process is shown in Fig 2. T_s and T_F denote the running time of long period and short period severally, and it contains $M(T_s/T_F)$ short periods in one long period. For purpose of simplicity, it is assumed that M is an integer.

Since network states vary dynamically, we allocate cache space for each virtual network based on load forecasting during long period, so as to improve the cache utilization and reduce the bit loss rate. To be specific, on the one hand, in long period, the load analysis module predicts possible variations in next long period according to the load variation characteristics of each virtual network, and configure appropriate cache space for each virtual network based on the predicted results. On the other hand, in short period, each virtual network achieves user's caching function of service in accordance with the determined size of cache space during long period, moreover, in order to guarantee the service rate of each virtual network, the resource management entity schedule time-frequency RBs for each virtual network on the strength of queue overflow probability evaluation.

A. ARMA PREDICTION OF CACHE SPACE RESERVATION STRATEGY

The variation characteristics of each network load directly affect the efficiency of physical resource scheduling, thus, based on the multiple time-scales resource configuration framework proposed in this paper, it is critical to design an effective prediction mechanism to realize the accurate forecasting of each virtual network load, and assist allocation of cache space in advance. There are some common algorithms, such as time series prediction, neural network, Markov model, gray prediction model, etc., 28]. The neural network prediction algorithm needs to provide training

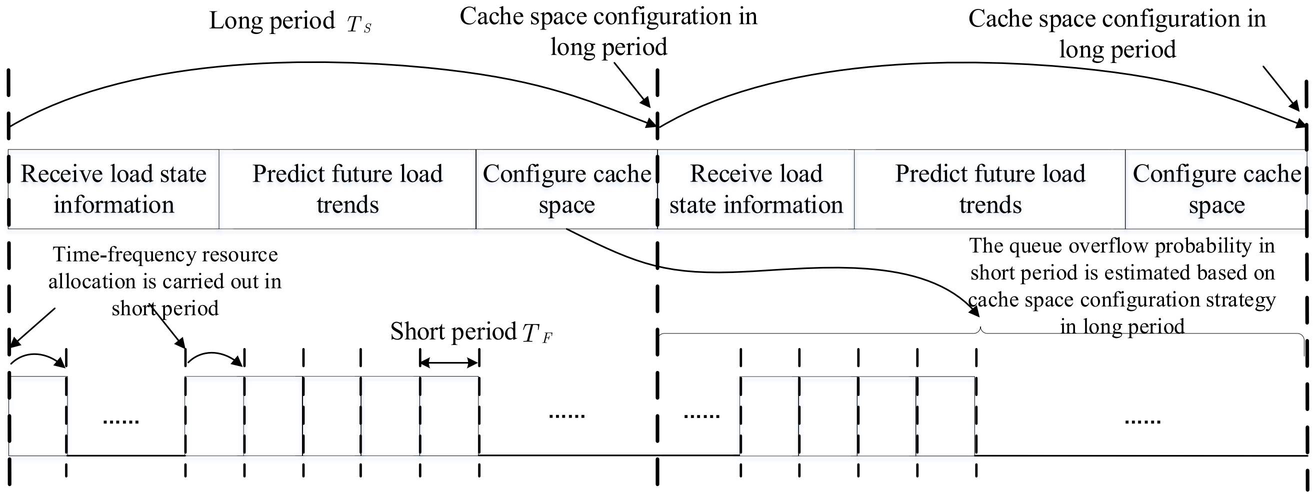

Long period T_S
Cache space configuration in long period
Cache space configuration in long period
Receive load state information
Predict future load trends
Configure cache space
Receive load state information
Predict future load trends
Configure cache space
Time-frequency resource allocation is carried out in short period
The queue overflow probability in short period is estimated based on cache space configuration strategy in long period
Short period T_F
......
......
......
......

FIGURE 2. **Multiple time-scale resource allocation.**

samples, which exists problems such as slow convergence and high complexity. Markov prediction model has low precision, big error and narrow scope of applications. The prediction accuracy of grey prediction model is related to the grading regularity of predicted object and the smoothness of data sequence. Only when the predicted sequence has the characteristic of exponential growth can it be highlighted. Considering that ARMA integrates the functions of regression analysis and time series analysis, the prediction error variance is smaller [29]. In this paper, the future load state of each virtual network is predicted by constructing the ARMA model.

The virtual network management platform adjusts the cache space configuration in advance by predicting the load fluctuation of each virtual network. In terms of each virtual network load and actual needs, the load state is divided into L levels, which can be described as a finite state space $B_k(t) \in \{B^1, B^2, \ldots, B^{L-1}, B^L\}$. In this scenario, we predict the average load in long period, and the future average load state of each virtual network based on ARMA model can be represented as a linear combination of the last p long periods of historical average load and q long periods of white noise, which can be expressed as:

$$\begin{aligned} y_k(t) = \varphi_1 y_k(t-1) + \ldots + \varphi_p y_k(t-p) + \xi(t) \\ - \theta_1 \xi(t-1) - \ldots - \theta_q \xi(t-q), \end{aligned} \tag{6}$$

where $y_k(t-i)|i = 1, \ldots, p\}$ denotes the average load state of virtual network k in past p long periods. $\xi(t-i)|i = 1, \ldots, q\}$ denotes gaussian white noise with mean zero and variance σ^2, moreover, $\xi(t)$ is unrelated to the historical observation sequence. $\{\varphi_i|i = 1, \ldots, p\}$ and $\{\theta_i|i = 1, \ldots, q\}$ are model parameters to be estimated.

In the process of establishing the predicted ARMA model, it is necessary to start from a stationary data sequence with mean zero. Therefore, in this scheme, historical data are first logarithmically processed and mean-subtracted, and then predicted depending on equation (6). Considering that each virtual network may have different load characteristics, an independent prediction model will be constructed according to the historical data of each virtual network. The specific process is detailed as follows:

Step 1: Estimate $\{\varphi_i|i = 1, \ldots, p\}$ and $\{\theta_i|i = 1, \ldots, q\}$ of each virtual network prediction model. The self-covariance of each virtual network can be obtained by using the logarithmically processed and mean-subtracted observation sequence $y_k{}'$:

$$\gamma_k(i) = \frac{1}{I} \sum_{j=1}^{I-i} y_k{}'(i) y_k{}'(i+j), \tag{7}$$

where I denotes the size of observation sequence. We use the preprocessed observation sequence $y_k{}'(t)$ to replace $y_k(t)$ in equation (6), i.e., $y_k{}'(t) = \varphi_1 y_k{}'(t-1) + \ldots + \varphi_p y_k{}'(t-p)$ $+\xi(t) - \theta_1 \xi(t-1) - \ldots - \theta_q \xi(t-q)$. Next, through multiplying both sides of it by $y_k{}'(t-i)$, then taking mean value of it, we can obtain the relational expression of auto-covariance:

$$\begin{aligned} \gamma_k(i) &= E\{y_k{}'(t) y_k{}'(t-i)\} \\ &= \varphi_1 E\{y_k{}'(t-1) y_k{}'(t-i)\} + \ldots \\ &\quad + \varphi_p E\{y_k{}'(t-p) y_k{}'(t-i)\} \\ &\quad + E\{\xi(t) y_k{}'(t-i)\} - \theta_1 E\{\xi(t-1) y_k{}'(t-i)\} \\ &\quad - \ldots - \theta_q E\{\xi(t-q) y_k{}'(t-i)\}, \end{aligned} \tag{8}$$

where $E\,[\cdot]$ represents the expected factor.

As previously mentioned, $\xi(t)$ is unrelated to the historical observation sequences, i.e., $E\{y_k{}'(s)\xi(t)\} = 0, s > t$. Therefore, when $i > q$, $E\{\xi(t) y_k{}'(t-i)\} - \theta_1 E\{\xi(t-1)$ $y_k{}'(t-i)\}$ $- \ldots - \theta_q E\{\xi(t-q) y_k{}'(t-i)\} = 0$. In addition, since $y_k{}'(t)$ is preprocessed, i.e., logarithmetics and mean-subtraction, hence it is a stationary data sequence with mean zero. According to the properties of auto-convariance of stationary sequence, it is easily obtained that

$E\{{y_k}'(t-1){y_k}'(t-i)\} = \gamma_k[t-1-(t-i)] = \gamma_k(i-1)$. As a consequence, equation (8) can be rewritten as:

$$\gamma_k(i) = \varphi_1\gamma_k(i-1) + \varphi_2\gamma_k(i-2) + \ldots + \varphi_p\gamma_k(i-p). \quad (9)$$

Based on equation (9), for $i = q+1, q+2, \ldots, q+p$, the following equations can be obtained:

$$\begin{cases} \gamma_k(q+1)=\varphi_1\gamma_k(q) + \varphi_2\gamma_k(q-1)\ldots + \varphi_p\gamma_k(q+1-p) \\ \gamma_k(q+2)=\varphi_1\gamma_k(q+1) + \varphi_2\gamma_k(q)\ldots + \varphi_p\gamma_k(q+2-p) \\ \vdots \\ \gamma_k(q+p)=\varphi_1\gamma_k(q+p-1)+\varphi_2\gamma_k(q+p-2)\ldots+\varphi_p\gamma_k(q). \end{cases} \quad (10)$$

Therefore, by leveraging the extended Yule-Walker equation, the estimation of $\varphi_i|i = 1, \ldots, p\}$ can be determined by:

$$\begin{bmatrix} \gamma_{q+1} \\ \gamma_{q+2} \\ \vdots \\ \gamma_{q+p} \end{bmatrix} = \Upsilon \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{bmatrix}, \quad (11)$$

where

$$\Upsilon = \begin{bmatrix} \gamma_k(q) & \gamma_k(q-1) & \ldots & \gamma_k(q+1-p) \\ \gamma_k(q+1) & \gamma_k(q) & \ldots & \gamma_k(q+2-p) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_k(q+p-1) & \gamma_k(q+p-2) & \ldots & \gamma_k(q) \end{bmatrix}.$$

The next step is to estimate $\theta_i|i = 1, \ldots, q\}$ and noise variance based on the $\varphi_i|i = 1, \ldots, p\}$ solved by above equations. Because $y_k^*(t) \triangleq y_k'(t) - \sum_{i=1}^{p} \varphi_i y_k'(t-i)$ satisfies moving average (MA) model, $y_k^*(t)$ is approximatively regarded as the observations of MA(q), which is specifically expressed as:

$$y_k^*(t) = y_k'(t) - [\varphi_1 y_k'(t-1) + \ldots + \varphi_p y_k'(t-p)], \\ t = p+1, \ldots I. \quad (12)$$

Similarly, in accordance with the approximate observations, the corresponding auto-convariance function can be determined:

$$\begin{aligned} y_k^*(i) &= E\{y_k^*(t)y_k^*(t-i)\} \\ &= E\left\{\left(-\sum_{j=0}^{p}\varphi_j y_k'(t-j)\right)\left(-\sum_{l=0}^{p}\varphi_l y_k'(t-i-l)\right)\right\} \\ &= \sum_{j,l=0}^{p}\varphi_j\varphi_l\gamma_k(i+l-j). \end{aligned} \quad (13)$$

By means of the caculation of equation (13) and the inverse correlation function of MA model, the estimations of $\theta_i|i = 1, \ldots, q\}$ and the variance $\hat{\sigma}_\xi^2$ of white noise $\xi(t)$ can be acquired.

Step 2: Determine the order of prediction model for each virtual network. The forecasting performance of the model which is combined with different p, q has a certain difference. In order to realize the future load forecasting optimally, Akaike information criterion (AIC) is adopted to set the order for each virtual network prediction model. AIC can be defined as:

$$AIC\ (s) = \ln \hat{\sigma}_\xi^2 + \frac{2s}{I}, \quad (14)$$

where s is the amount of parameters of the prediction model, which is the sum of the variance of $\xi(t)$, p and q, that is $s = p+q+1$. The values of AIC(s) with different order combinations are calculated within a certain range, and the corresponding combination is the model order when the minimum value is reached.

After establishing the prediction model for each virtual network, the cache space reservation policy will be implemented according to the prediction results. To avoid a large number of data loss at the beginning of a new long period, in this scheme, we adopt a cache space reservation mechanism with combination of static and dynamic. By comparing the average load state $y_k(t)$ in next long period t predicted by ARMA model with the load state interval, the reserved static portion B^{st} of virtual network k can be confirmed, i.e. $B^{st} = B^l$ where B^l satisfies $B^{l-1} < y_k(t) \le B^l$. If at the end of current long period, the instantaneous queue length $Q_k(t)$ of virtual network k is much greater than B^l, the dynamic portion B^{dy} should be reserved. Let $B^{dy} = B^{l'} - B^l$, where $B^{l'}$ follows $B^{l'-1} < Q_k(t) \le B^{l'}$, therefore, the cache space reserved by virtual network k is $B_k(t) = B^{st} + B^{dy}$. At the same time, considering detecting the utilization of dynamic cache space in each long period, 60 short periods are taken as a detection period. If the utilization of dynamic cache space is less than 50% in the detection period, half of the dynamic portion will be released. If the bit loss rate during the detection period is more than 15%, we consider adding $B^* = B^{l'-1}$ to dynamic portion, where $B^{l'-1} \le \overline{Q}_{lost} < B^{l'}$, and $\overline{Q}_{lost}$ denotes the average lost data during the detection period. The specific process is given in Fig. 3.

B. TIME-FREQUENCY RESOURCE SCHEDULING STRATEGY WITH QoS CONSTRAINTS

In each short period, we evaluate the queue overflow probability of each virtual network in terms of the size of cache space reserved during long period, so as to obtain the scheduling strategy of time-frequency RBs. Since the distribution of $Q_k(t)$ is unknown, the closed-form expression of equation (3) can not be acquired directly. Therefore, for the sake of performing the proactive time-frequency RBs scheduling, we leverage the large-deviation principle to estimate the queue overflow probability in $t+T$ period based on the historical data which is as of short period t, where T is the forecasting period.

1) QUEUE OVERFLOW PROBABILITY ESTIMATION MODEL

$\Delta_k\ (t) = A_k\ (t) - D_k\ (t)$ is defined as queue increment in single period, where the range of $\Delta_k(t)$ is $\Delta_k(t) \in \{-D_k, \cdots,$

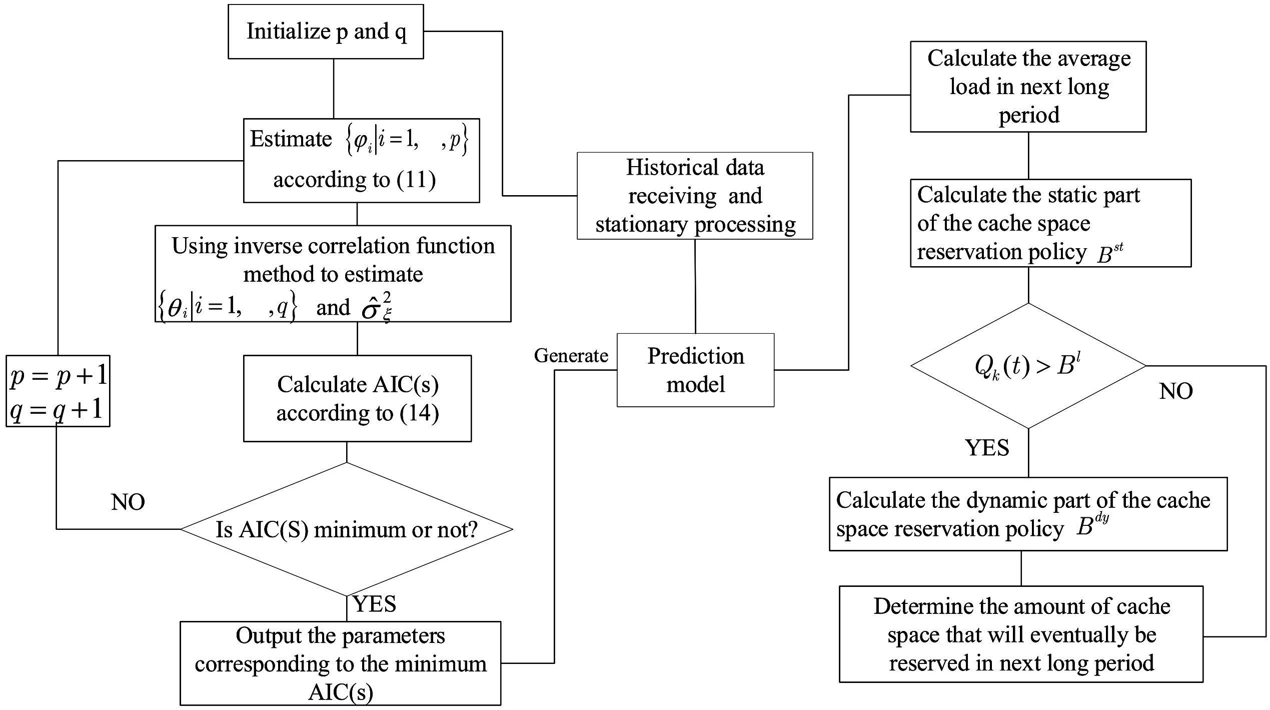

Initialize p and q
Estimate $\{\varphi_i | i=1, \quad ,p\}$ according to (11)
Using inverse correlation function method to estimate $\{\theta_i | i=1, \quad ,q\}$ and $\hat{\sigma}_\xi^2$
$p=p+1$
$q=q+1$
Calculate AIC(s) according to (14)
NO
Is AIC(S) minimum or not?
YES
Output the parameters corresponding to the minimum AIC(s)
Historical data receiving and stationary processing
Generate
Prediction model
Calculate the average load in next long period
Calculate the static part of the cache space reservation policy B^{st}
$Q_k(t) > B^l$
NO
YES
Calculate the dynamic part of the cache space reservation policy B^{dy}
Determine the amount of cache space that will eventually be reserved in next long period

FIGURE 3. The cache space reservation policy based on ARMA prediction in long period.

$0, 1, \cdots, A_k\}$. $\pi_k^d = P(\Delta_k(t) = d)$ is the probability distribution of queue variations of virtual network k. Due to $A_k(t)$ is determined by bit arrival rate and $D_k(t)$ is the amount of packets transmitted successfully, the difference $\Delta_k(t)$ indicates the matching degree between service rate and arrived data packets of virtual network k. $\Delta_k(t) < 0$ indicates that the service rate of current virtual network k is relatively high, while $\Delta_k(t) > 0$ indicates that the service rate of virtual network k can not meet the requests in current short period.

In summary, the queue increment of virtual network k from short period t to $t+T$ is given by:

$$\Delta_k(t+T) = \sum_{i=1}^{T} \Delta_k(t+i). \tag{15}$$

Therefore, in short period $t+T$, the instantaneous queue length of virtual network k is as follows:

$$Q_k(t+T) = Q_k(t) + \Delta_k(t+T). \tag{16}$$

In accordance with equation (15) (16), the queue overflow probability of virtual network k in short period t can be derived as:

$$\begin{aligned} P_{overflow}^k&(t+T) \\ &= P\left(Q_k(t+T) > B_k\right) \\ &= P\left(Q_k(t) + \sum_{i=1}^{T} \Delta_k(t+i) > B_k\right) \\ &= P\left(\frac{\sum_{i=1}^{T} \Delta_k(t+i)}{T} > \frac{B_k - Q_k(t)}{T}\right) \\ &= P\left(\frac{\sum_{i=1}^{T} \Delta_k(t+i)}{T} > a_k\right), \end{aligned} \tag{17}$$

where $a_k = (B_k - Q_k(t))\big/T$ represents the acceptable average queue growth rate of virtual network k in next T short periods, and $m_k = E\left[\sum_{i=1}^{T} \Delta_k(t+i)\Big/T\right]$ indicates the expected average queue growth rate of virtual network k in next T short periods, and $E[\cdot]$ is the expected factor. Since $\sum_{i=1}^{T} \Delta_k(t+i)\Big/T$ in equation (17) depends on the resource configuration and bit arrival process, and a_k is determined by the instantaneous queue length of virtual network k in current short period, equation (17) can be regarded as the service capability of the resource configuration mode in current scheduling period for future service requests. Specifically, the larger the value of equation (17) is, the queue overflow will be more likely to occur, meanwhile a higher priority the corresponding virtual network will have to participate in the configuration of time-frequency RBs to meet its QoS requirements.

$A_k(t)$ is assumed as an independent identically distributed process, hence $\Delta_k(t)$ is also an independent identically distributed random variable, and follows a finite instantaneous moment generating function $G(\omega) = E\left[e^{\omega\,\Delta_k(t)}\right]$. If $E\left[\Delta_k(t)\right] < a_k$, the sequence $\Delta_k(t)$ conforms to the large-deviation principle in terms of literature [30], [32]. As a result, it can be given by using Cramer's theorem [32] when $a_k > m_k$:

$$\lim_{T\to\infty}\frac{1}{T}\log P\left(\frac{\sum_{i=1}^{T}\Delta_k(t+i)}{T} > a_k\right) = -f(a_k), \tag{18}$$

where $f(a_k)$ is the rate function, and is specifically expressed as:

$$f(a_k) = \sup_{\omega>0}\left\{a_k\,\omega - \log G(\omega)\right\}. \tag{19}$$

If it is known that the probability distribution of Δ_k, i.e. $\Delta_k \sim \begin{pmatrix} -D_k, \cdots, 0, 1, \cdots, A_k \\ \pi_k^{-D_k}, \cdots, \pi_k^0, \pi_k^1, \cdots, \pi_k^{A_k}\end{pmatrix}$, $\log G(\omega)$ can be caculated as:

$$\log G(\omega) = \log\left\{\sum_{d=-D_k}^{A_k}\pi_k^d\,e^{d\omega}\right\}. \tag{20}$$

According to equation (18), for sufficiently large T, the queue overflow can approximately be derived as follows:

$$P_{overflow}^k(t+T) \approx e^{-Tf(a_k)}. \tag{21}$$

Since the estimated queue overflow probability decreases exponentially with T, it is necessary to select an appropriate T to obtain the accurate queue overflow probability. The value of T will be discussed in the following simulation section.

Although the estimated value of queue overflow probability can be obtained according to equation (21), it is impossible to acquire the moment generating function $G(w)$ because the probability distribution of Δ_k is unknown. Therefore, a sliding window based method is adopted to estimate π_k^d online.

Supposing that the size of sliding window is T_w, the observation vector of virtual network k in current period t can be represented as: $W_k(t) = [\Delta_k(t-1), \cdots, \Delta_k(t-T_w)]$. Hence, the estimation of m_k is:

$$\hat{m}_k = \frac{\sum_{d=t-T_w}^{t-1}\Delta_k(d)}{T_w}. \tag{22}$$

R_k^d is defined as the number of times $\Delta_k(j) = d$ occurs in the sliding window, and the probability of its occurrence can be calculated by $\tilde{\pi}_k^d = R_k^d / T_w$. If $\tilde{\pi}_k^d$ is employed directly as the estimation of π_k^d, it might cause large fluctuation of π_k^d in different short periods. To this end, in the light of literature [33], exponential smoothing method is introduced to soft the estimation, which is specifically described as:

$$\widehat{\pi}_k^d(t) = \eta\,\widehat{\pi}_k^d(t-1) + (1-\eta)\,\tilde{\pi}_k^d(t), \tag{23}$$

where $\eta \in [0, 1]$ is to measure the impact of both current estimations and previous information on parameter estimations. As η approaches 0, it means that $\widehat{\pi}_k^d(t)$ prefers the current estimation $\tilde{\pi}_k^d(t)$. As η approaches 1, it means that $\widehat{\pi}_k^d(t)$ is largely affected by previous estimations.

2) DYNAMIC TIME-FREQUENCY RESOURCE SCHEDULING STRATEGY

Concerning the estimation model of the queue overflow probability above, we allocate appropriate time-frequency RBs for each virtual network according to its priority, so as to meet QoS requirements.

As $\hat{m}_k \geq a_k$, it means that the expected average queue growth rate of virtual network k in next T short periods is higher than the acceptable one. If the current service rate is kept constant, the virtual network k is highly likely to encounter queue overflow after T short periods. As $\hat{m}_k < a_k$, despite the expected average queue growth rate is lower than the receivable one, it is still unable to conclude that queue overflow will not happen in virtual network k during short period from t to $t+T$. However, the virtual network in $\hat{m}_k \geq a_k$ is more urgent than that in $\hat{m}_k < a_k$, therefore, priority should be given to the former time-frequency RBs scheduling problem. For any of virtual networks, we will calculate the overflow residual time T_k in $\hat{m}_k \geq a_k$ and the queue overflow probability $P_{overflow}^k(t+T)$ in $\hat{m}_k < a_k$ respectively. The overflow residual time T_k of virtual network k can be approximated as:

$$T_k = \frac{B_k - Q_k(t)}{E\left[\Delta_k(t)\right]} \approx \frac{a_k}{\hat{m}_k}. \tag{24}$$

The smaller the value of equation (24), the higher the priority of virtual network k, and the smaller the difference between the queue overflow probability and the threshold $P_{overflow}^k(t+T) - \varepsilon_k$, thus the lower the priority of virtual network k.

To this end, the scheduling scheme of time-frequency RBs based on the determined priorities of virtual networks is listed as follows:

1. Construct a set K_1 for virtual networks in the case of $\hat{m}_k \geq a_k$, and select the virtual network $k = \arg\min_{k\in K_1}\{T_k\}$ during short period t.
2. Allocate time-frequency RBs for virtual network k. r represents the service rate provided by a single time-frequency RB, and it is assumed to be same. Hence the service rate of virtual network k can be calculated as:

$$C_k(t) = mr. \tag{25}$$

By increasing the number of time-frequency RBs m until meet $A_k(t) < C_k(t)$ in virtual network k, accordingly in order to ensure that the queue of virtual network k no longer augments at least in current short period to alleviate the pressure on queue growth in future.

3. Supposing that $K_1 = K_1 \setminus \{k\}$ and $N = N - m$, then select next virtual network $k' = \arg\min_{k' \in K_1} \{T_{k'}\}$, and repeat step 1 and step 2 till $K_1 = \emptyset$.
4. Construct a set K_2 for virtual networks which is in the case of $\hat{m}_k < a_k$, and select virtual network $k^* = \arg\max_{k^* \in K_2} \left\{P^{k^*}_{overflow}(t+T) - \varepsilon_{k^*}\right\}$, then repeat step 2 to configure the time-frequency RBs, while update the number of virtual network sets and the RBs by analogy to step 3 up to meet $K_2 = \emptyset$.
5. If $K_1 = \emptyset$, $K_2 = \emptyset$ and $N \neq 0$, then construct a set K_3 for virtual networks which satisfies $C_k(t) < Q_k(t) + A_k(t)$. And select a virtual network $k'' = \arg\min_{k'' \in K_3} \{Q_{k''}(t) + A_{k''}(t) - C_{k''}(t)\}$ in this set to allocate the time-frequency RBs m, till meet $Q_{k''}(t) + A_{k''}(t) - C_{k''}(t) < \bar{C}_{k''}(t)$.
6. Supposing that $K_3 = K_3 \setminus \{k''\}$ and $N = N - m$, then repeat step 5 up to $N = 0$ or $K_3 = \emptyset$.

The overall process is presented in Algorithm 1. The complexity of scheduling time-frequency RBs in short period mainly derives from estimating the queue overflow probability $P^k_{overflow}(t+T)$ of each virtual network. Since $\log G(\omega)$ is a convex function, the approximation of the queue overflow probability of each virtual network can be obtained by adjusting parameters ω to maximize the rate function $f(a_k)$. In order to accelerate convergence of this algorithm, we adopt the golden section search algorithm to acquire the optimal $f(a_k)$. Supposing that J represents the maximum number of iterations of the search algorithm, the complexity of equation (21) is $O(T_w J)$. Assuming that $K = |K|$ is the amount of virtual networks in this system, therefore, the algorithm complexity mentioned in this paper is $O(T_w JK)$ in the worst case.

IV. SIMULATION RESULTS AND DISCUSSIONS

In this section, we demonstrate the performance improvements of our proposed algorithm virtual resource allocation based on load forecasting (RALF). In the simulation, the settings of specific simulation parameters are as shown in Table 1. Since RALF proposed in this paper is composed of cache space reservation strategy and time-frequency resource scheduling strategy, thus in order to better reflect the performance of RALF, we respectively compare with two basic strategies.

1) Cache space comparison scheme: As previously elaborated in Section III.A, we proposed an ARMA prediction of cache space reservation, which has a capability to dynamically adjust cache space reservation according to the prediction results under long time-scales. That is to say, it is a dynamic reservation scheme. Therefore, we utilize two static cache space reservation schemes as comparisons, i.e., static conservative cache allocation (SCCA) and static abundant cache allocation (SACA), which reserve relatively few and abundant cache space for each virtual network respectively. And neither scheme vary with the service and simulation time.

Algorithm 1 The Online Dynamic Time-Frequency RBs Scheduling Algorithm

```
 1: Observe the current state $Q_k(t)$ of each virtual network queue and the size of reserved cache space $B_k$ in short period;
 2: for $k = 1; k < K; k++$ do
 3:     Calculate $a_k$,and estimate $\hat{m}_k$ according to equation (22);
 4:     if $\hat{m}_k \geq a_k$ then
 5:         Add to the set of virtual network $K_1$, and estimate the overflow residual time $T_k$ according to equation (24);
 6:     else
 7:         Add to the set of virtual network $K_2$, and execute golden section search algorithm to estimate $P^k_{overflow}(t+T)$;
 8:     end if
 9: end for
10: while $K_1 \neq \emptyset$ do
11:     Supposing that $m = 1$, then select a virtual network $k = \arg\min_{k \in K_1} \{T_k\}$;
12:     while $A_k(t) > C_k(t)$ do
13:         $m \leftarrow m + 1, C_k(t) \leftarrow mr, N \leftarrow N - 1$;
14:     end while
15:     $K_1 = K_1 \setminus \{k\}$;
16: end while
17: while $K_2 \neq \emptyset$ do
18:     Supposing that $m = 1$, then select a virtual network $k^* = \arg\max_{k^* \in K_2} \left\{P^{k^*}_{overflow}(t+T) - \varepsilon_{k^*}\right\}$;
19:     Repeat step 12-14;
20:     $K_2 = K_2 \setminus \{k^*\}$;
21: end while
22: if $N \neq 0$ then
23:     for $k = 1; k < K; k++$ do
24:         if $C_k(t) < Q_k(t) + A_k(t)$ then
25:             Add to the set of virtual network $K_3$;
26:         end if
27:     end for
28:     while $K_3 \neq \emptyset$ and $N \neq 0$ do
29:         Supposing that $m = 1$, then select a virtual network $k'' = \arg\min_{k'' \in K_3} \{Q_{k''}(t) + A_{k''}(t) - C_{k''}(t)\}$;
30:         while $Q_{k''}(t) + A_{k''}(t) > \bar{C}_{k''}(t) + C_{k''}(t)$ do
31:             $m \leftarrow m + 1, \bar{C}_{k''}(t) \leftarrow mr, N \leftarrow N - 1$;
32:         end while
33:         $K_3 = K_3 \setminus \{k''\}$.
34:     end while
35: end if
```

2) Time-frequency resource comparison scheme: As described earlier in Section III.B, a time-frequency resource scheduling strategy with QoS constrains is proposed.

TABLE 1. Simulation parameters.

Simulation Parameters	Value
Virtual network numbers	2,3,4,5,6
System bandwidth	10MHz(50RBs)
Short period	1ms
Long period	300ms
Load arrival process	Poisson distribution
Bit arrival rate	$\lambda = 58.7kbit/ms$
Price per unit of RB α	1.2,2.1,1.5 unit/RB
Price per unit of cache space ρ	8,6,4 unit/kbit
Queue overflow probability ε	0.13,0.05,0.12
The service rate of single RB	$r = 3024bit/ms$
The size of sliding window T_ω	60ms
Smoothness index η	0.7
Simulation time	6600ms

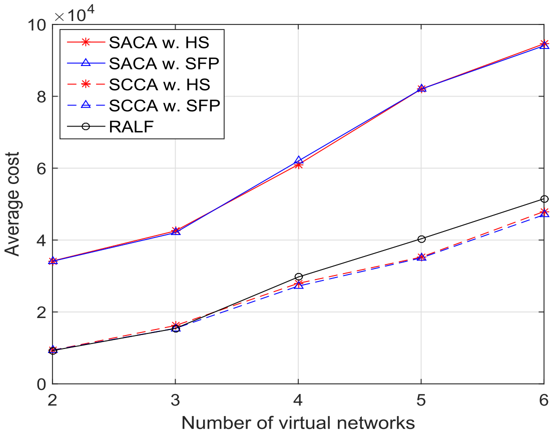


FIGURE 4. The cache space reservation policy based on ARMA prediction in long period.

In this scheme, we set priorities of virtual networks based on queue overflow probability estimation model instead of just based on queue length. Then we dynamically allocate appropriate time-frequency RBs for each virtual network according to its priority. Hence, hard slice (HS) and slice with fixed prioritization (SFP) are compared. The HS scheme provides a fixed number of time-frequency RBs for each virtual network, that is allocating RBs equally for each network. While SFP scheme is similar to the literature [6], which dynamically allocates time-frequency RBs for each virtual network in terms of a fixed prioritization, i.e., the longer the queue length is, the more RBs for the virtual network.

Then the algorithms presented above are evaluated from three aspects: total overheads of leasing resources, average utilization and bit loss rate. Fig.4 depicts a comparison of total resource leasing cost of different scenarios and different virtual networks. As can be seen, it illustrates that the cost of RALF grows slowly when the number of virtual networks is small, and increases significantly only in the case of the augmentation of virtual networks. However, the other four schemes consistently maintain similar growth rates. It is mainly because that the RALF algorithm mentioned adopts the cache space reservation mechanism based on load prediction. To be more specific, on the one hand, compared with SACA which always allocates abundant cache space for each virtual network, the average cost of RALF is much lower. On the other hand, the average cost of RALF is near to SCCA if the number of virtual networks is smaller than 4. It is because when the number of virtual networks is small, according to the reservation policy based on ARMA prediction, the cache space needs to be reserved is relatively fewer just like SCCA. However, with the number of virtual networks increasing, the cache space needs to be reserved will be larger, while SCCA still reserve the fixed few cache space for each virtual network, so the average cost of RALF will be larger than the SCCA. Although the cost of RALF is not the smallest, large number of data overflows occur because of insufficient cache space in SCCA, thus the bit loss rate of RALF is much lower than SCCA as shown in Fig 6. As a consequence, the proposed algorithm predicts the possible load variations of each virtual network by leveraging the load information in historical periods, then executes cache space configuration in advance to avoid resource wastes, while the other four resource allocation schemes ignore the dynamics of service, which may cause mismatch between resource allocation and actual needs to some extent.

Fig.5 and Fig.6 reveal the evaluations of resource utilization and bit loss rate respectively of different schemes with different number of virtual networks. When the number of virtual networks is small, the time-frequency RBs in system can cope leisurely with the transmission demand, hence the bit loss rate of each scheme is at an ideal level. But with the number of virtual networks increasing, they begin to compete with the limited time-frequency RBs, meanwhile the average utilization rate rises continuously, however the lack of resources might cause more serious bit loss. Although the bit loss rate of SACA scheme is lower, it can be seen from Fig.5 that it has the worst performance in average utilization. The average utilization rate of SCCA is comparatively high, but because of insufficient cache space configuration, a large number of data overflows occur, and the performance on bit loss rate as presented in Fig.6 is dissatisfactory. Moreover, compared to SACA scheme, the average bit loss rate of RALF is slightly larger after the number of virtual networks is not smaller than 4, it is because we also take the resource leasing cost into consideration. With the number of virtual networks increasing, the things we do are not only allocating more cache space and RBs, but also trying to minimize the cost. Just as depicted in Fig.4, the average cost of SACA is significantly larger than RALF. In addition, we can also see from Fig.6 that HS and SFP have relatively similar bit loss rate under the circumstance of fewer virtual networks, but as depicted in Fig.5 the average utilization rate of HS is lower than that of SFP. It is because that HS sets a fixed number of time-frequency RBs for each virtual network but ignores the real

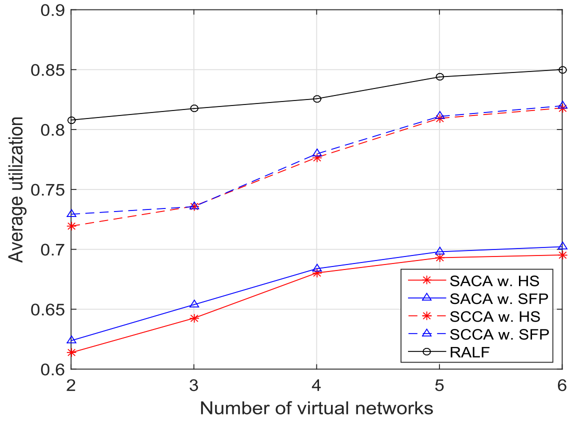


FIGURE 5. Comparison of average utilization of different schemes.

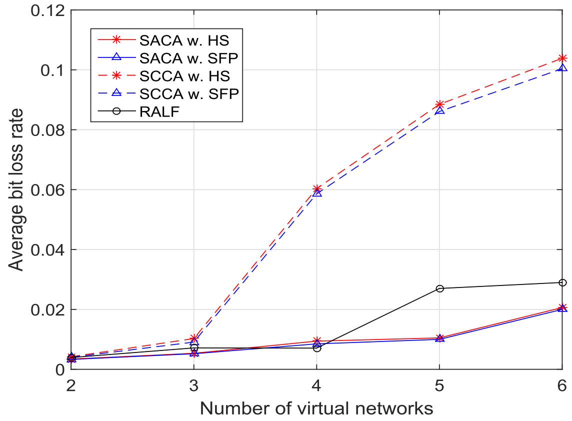


FIGURE 6. Comparison of average bit loss rate of different schemes.

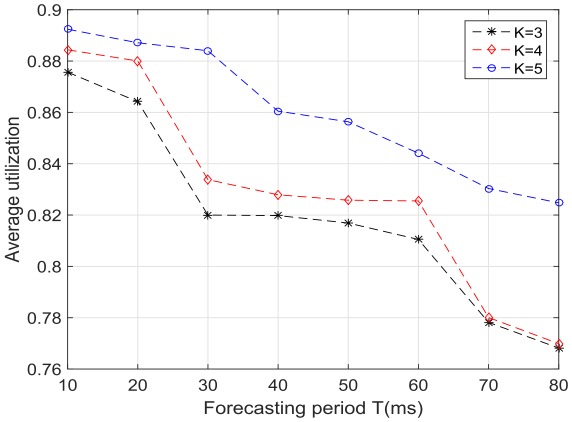


FIGURE 7. Average utilization rate of different T.

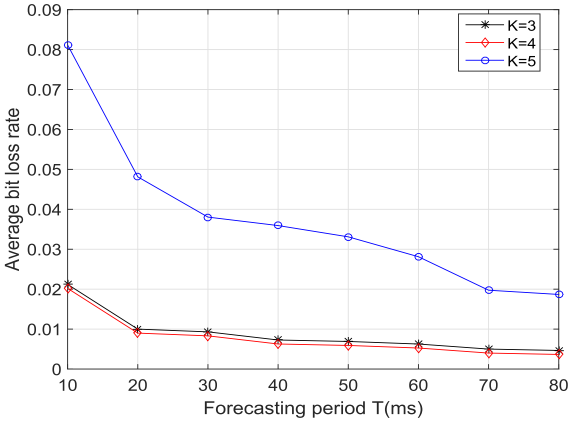


FIGURE 8. Average bit loss rate of different T.

needs, which finally results in some physical resources being idle. Compared with other strategies, the RALF algorithm proposed in this paper has an ideal comprehensive effect on average cost, average utilization rate as well as bit loss rate. To put it simply, the advantage of RALF algorithm lies in its adaptive adjustment ability. Cache space can be reserved in advance for future load variations, simultaneously priority of each virtual network is dynamically adjusted according to the queue overflow probability, and the transmission service of virtual network with high overflow probability is given a priority and dynamic adjustment is employed in both long and short periods. On the contrary, the other four schemes are trapped by their fixed periodic allocation mechanism, which leads to excessive or insufficient resource allocation.

Furthermore, in this section, we design a simulation experiment to study the effect of the forecasting period T on system performance in short period. Simulation parameters are set as: the number of virtual networks is $K = 3, 4, 5$, and value of the estimated period is $T = 20, 30, 40, 50, 60, 70, 80$. Regardless of the number of virtual networks, the average utilization rate in Fig.7 and average bit loss rate of the cache space and time-frequency RBs in Fig.8 both decrease to some extent with the increasing of forecasting period T. And especially in the cases of $K = 3$ and $K = 4$, the two curves are very close. But as illustrated in Fig.7, if $K = 5$, the average utilization rate rises prominently, and the average bit loss rate as displayed in Fig.8 also increases as well. Obviously the results of different K as shown in Fig.7 and Fig.8 are consistent with the results of comparison of average utilization rate and average bit loss rate of different schemes as shown in Fig.5 and Fig.6. Also equation (21) implies that the larger the forecasting period is, the more accurate the estimation of queue overflow probability will be. Therefore in order to optimally guarantee the queue overflow probability constraint of each virtual network, the larger T should be selected in theory. Nevertheless, as presented in Fig.7, as the forecasting period T increases from 10 to 80, the average utilization of system obviously decreases. Accordingly so as to ensure that all aspects of the system performance are in good condition, it is critical to select a reasonable value of forecasting period T.

Finally, we need to validate the forecasting ability of the average load state of the proposed strategy. In this case, the number of virtual networks is set as 3, and the x-axis is the time index of long period, and the y-axis is the average load value after the smooth processing. Since as with ARMA, both the Markov prediction model and grey-prediction model do not require a large number of labeled data, we adopt the Grey-Markov (GM) prediction model [34] as a comparison. It combines the grey model and Markov model so as to get

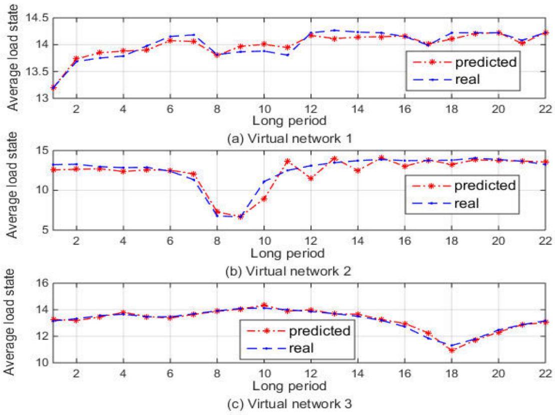


FIGURE 9. Comparison between the actual and the predicted values of average load based on ARMA prediction in three different virtual networks during 22 consecutive long periods.

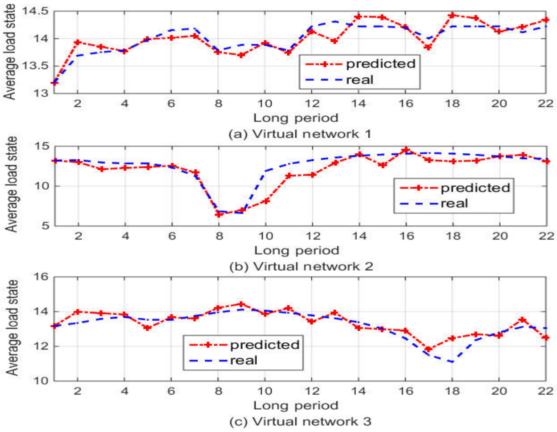


FIGURE 10. Comparison between the actual and the predicted values of average load based on GM prediction in three different virtual networks during 22 consecutive long periods.

more accurate prediction than using these two models separately to some extent. Fig.9 and Fig.10 show the comparison of the ARMA predicted value with the true value of average load state and the GM predicted value with the true value. Fig.9(a), 9(b) and 9(c) respectively depict the imitative effect of the actual average load value and the predicted value in three different virtual networks during 22 consecutive long periods. It can be observed that the actual average load states of each virtual network are basically consistent with the predicted average load fluctuations in 22 long periods. Furthermore, by comparing Fig.9 with Fig.10, although the GM model can overcome some shortcomings in grey model and Markov model as previously mentioned, the ARMA model still has a better performance on prediction accuracy than the GM model.

V. CONCLUSION

A dynamic virtual resource allocation algorithm based on load forecasting was proposed to solve the problem of unreasonable allocation of virtual resources which caused by the service uncertainty and the feedback delay in wireless virtualized network. Considering the differentiating features of different resources, cache space and time-frequency RBs are deemed as virtualized carriers, and a multiple time-scales hybrid scheduling mechanism is proposed. Aiming at minimizing the cost of resources leasing, the dynamic scheduling strategy of cache space and time-frequency RBs is executed in long and short period respectively. Simulation results indicate that in our proposed algorithm the overhead of leasing resources can effectively be reduced and meanwhile the utilization rate of physical resources can be improved. However, the cache space reservation policy designed is relatively simple. In order to better adapt to the caching requirements of different virtualized application scenarios, in the future, we need to further study more precise model for cache space scheduling.

REFERENCES

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016. doi: 10.1109/COMST.2016.2532458.

[2] X. Costa-Pèrez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013. doi: 10.1109/MCOM.2013.6553675.

[3] C. Liang and F. Yu, "Wireless virtualization for next generation mobile cellular networks," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 61–69, Feb. 2015. doi: 10.1109/MWC.2015.7054720.

[4] X. Lu, Q. Ni, D. Zhao, W. Cheng, and H. Zhang, "Resource virtualization for customized delay- bounded QoS provisioning in uplink VMIMO-SC-FDMA systems," *IEEE Trans. Commun.*, vol. 67, no. 4, pp. 2951–2967, Apr. 2019. doi: 10.1109/TCOMM.2018.2886337.

[5] *UK Strategy and Plan for 5G & Digitisation—Driving Economic Growth and Productivity*, Future Commun. Challenge Group, London, U.K., Jan. 2017, pp. 1–52. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/582640/FCCG_Interim_Report.pdf

[6] M. Kalil, A. Al-Dweik, M. F. A. Sharkh, A. Shami, and A. Refaey, "A framework for joint wireless network virtualization and cloud radio access networks for next generation wireless networks," *IEEE Access*, vol. 5, pp. 20814–20827, 2017. doi: 10.1109/ACCESS.2017.2746666.

[7] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. M. Leung, "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, Aug. 2017. doi: 10.1109/MCOM.2017.1600940.

[8] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 462–476, Sep. 2016. doi: 10.1109/TNSM.2016.2597295.

[9] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, Mar. 2015. doi: 10.1109/COMST.2014.2352118.

[10] M. M. Rahman, C. Despins, and S. Affes, "Design optimization of wireless access virtualization based on cost & QoS trade-off utility maximization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6146–6162, Sep. 2016. doi: 10.1109/TWC.2016.2580505.

[11] O. Sallent, J. Pérez-Romero, R. Ferrús, and R. Agustí, "On radio access network slicing from a radio resource management perspective," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 166–174, Oct. 2017. doi: 10.1109/MWC.2017.1600220WC.

[12] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management & prioritization in 5G mobile systems," in *Proc. 22th Eur. Wireless Conf.*, Oulu, Finland, May 2016, pp. 1–6.

[13] Q. Zhu and X. Zhang, "Game-theory based buffer-space and transmission-rate allocations for optimal energy-efficiency over wireless virtual networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[14] T. D. Tran and L. B. Le, "Joint resource allocation and content caching in virtualized content-centric wireless networks," *IEEE Access*, vol. 6, pp. 11329–11341, 2018. doi: 10.1109/ACCESS.2018.2804902.

[15] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proc. IEEE Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9. doi: 10.1109/INFOCOM.2017.8057230.

[16] L. Tang, X. Yang, X. Wu, T. Cui, and Q. Chen, "Queue stability-based virtual resource allocation for virtualized wireless networks with self-backhauls," *IEEE Access*, vol. 6, pp. 13604–13616, 2018. doi: 10.1109/ACCESS.2018.2797088.

[17] S. Parsaeefard, V. Jumba, M. Derakhshani, and T. Le-Ngoc, "Joint resource provisioning and admission control in wireless virtualized networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, USA, Mar. 2015, pp. 2020–2025. doi: 10.1109/WCNC.2015.7127778.

[18] S. Parsaeefard, R. Dawadi, M. Derakhshani, and T. Le-Ngoc, "Joint user-association and resource-allocation in virtualized wireless networks," *IEEE Access*, vol. 4, no, pp. 2738–2750, 2016. doi: 10.1109/ACCESS.2016.2560218.

[19] X. Lu, K. Yang, and H. Zhang, "An elastic sub-carrier and power allocation algorithm enabling wireless network virtualization," *Wireless Pers. Commun.*, vol. 75, no. 4, pp. 1827–1849, Apr. 2013.

[20] G. Liu, F. R. Yu, H. Ji, and V. C. M. Leung, "Virtual resource management in green cellular networks with shared full-duplex relaying and wireless virtualization: A game-based approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7529–7542, Sep. 2016. doi: 10.1109/TVT.2015.2497360.

[21] S. M. A. Kazmi, N. H. Tran, T. M. Ho, and C. S. Hong, "Hierarchical matching game for service selection and resource purchasing in wireless network virtualization," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 121–124, Jan. 2018. doi: 10.1109/LCOMM.2017.2701803.

[22] S. Gu, Z. Li, C. Wu, and H. Zhang, "Virtualized resource sharing in cloud radio access networks through truthful mechanisms," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1105–1118, Mar. 2017. doi: 10.1109/TCOMM.2016.2637900.

[23] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Resource provisioning in wireless virtualized networks via massive-MIMO," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 237–240, Jun. 2015. doi: 10.1109/LWC.2015.2402126.

[24] H. Ahmadi, I. Macaluso, I. Gomez, L. Doyle, and L. A. DaSilva, "Substitutability of spectrum and cloud-based antennas in virtualized wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 114–120, Apr. 2017. doi: 10.1109/MWC.2016.1500303WC.

[25] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, "Resource allocation for information-centric virtualized heterogeneous networks with in-network caching and mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11339–11351, Dec. 2017. doi: 10.1109/TVT.2017.2737028.

[26] T. LeAnh, N. H. Tran, D. T. Ngo, and C. S. Hong, "Resource allocation for virtualized wireless networks with backhaul constraints," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 148–151, Jan. 2017. doi: 10.1109/LCOMM.2016.2617307.

[27] Y.-M. Chu, N.-F. Huang, and S.-H. Lin, "Quality of service provision in cloud-based storage system for multimedia delivery," *IEEE Syst. J.*, vol. 8, no. 1, pp. 292–303, Mar. 2014. doi: 10.1109/JSYST.2013.2257338.

[28] M. Amiri and L. Mohammad-Khanli, "Survey on prediction models of applications for resources provisioning in cloud," *J. Netw. Comput. Appl.*, vol. 82, pp. 93–113, Mar. 2017.

[29] J. Li, X. Liu, and Z. Han, "Research on the ARMA-based traffic prediction algorithm for wireless sensor network," *J. Electron. Inf. Technol.*, vol. 29, no. 5, pp. 1224–1227, 2007.

[30] M. Mandhes, *Large Deviations for Gaussian Queues: Modelling Communication Networks*. Chichester, U.K.: Wiley, 2007, pp. 55–60.

[31] A. Dembo and O. Zeitouni, "Large deviations techniques and applications," *J. Amer. Stat. Assoc.*, vol. 95, no. 452, pp. 303–304, 2010.

[32] J. Yang, Y. Ran, S. Chen, W. Li, and L. Hanzo, "Online source rate control for adaptive video streaming over HSPA and LTE-style variable bit rate downlink channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 643–657, Feb. 2016. doi: 10.1109/TVT.2015.2398515.

[33] E. S. Gardner, Jr., "Exponential smoothing: The state of the art—Part II," *Int. J. Forecasting*, vol. 4, no. 1, pp. 637–666, 1985.

[34] Y. Zhou, B. Qi, and B. Zhang, "Online prediction of electrical load for distributed management of PEV based on Grey-Markov model," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, Chongqing, China, May 2017, pp. 6911–6916.

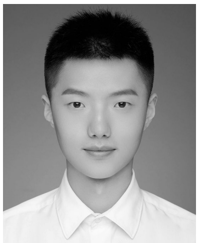
LUN TANG received the Ph.D. degree in communication and information system from Chongqing University, Chongqing, China.

He is currently a Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications. His current research interests include 5G cellular networks, interference management, and small cell networks.

XIAOYU HE received the B.S. degree in electronic information engineering from the Southwest University of Science and Technology, China, in 2017. She is currently pursuing the M.S. degree with the Key Laboratory of Mobile Communication Technology, Chongqing University of Posts and Telecommunications (CQUPT), China. Her research interests include network function virtualization (NFV), resource allocation in 5G C-RAN, and applications of reinforcement learning technique in mobile networks.

XIXI YANG received the B.S. degree and the M.S. degree in communication and information systems from the Chongqing University of Posts and Telecommunications (CQUPT), China, in 2015. Her current research interests include radio resource allocation, network virtualization, and wireless self-backhaul small cell networks.

YANNAN WEI received the B.S. degree in telecommunication engineering from the Chongqing University of Posts and Telecommunications (CQUPT), China, in 2017, where he is currently pursuing the M.S. degree with the Key Laboratory of Mobile Communication Technology. His research interests include resource allocation in heterogeneous cellular networks, wireless backhaul networks, and applications of Lyapunov optimization technique in mobile networks.



XIAO WANG received the B.S. degree in communication engineering science from the Chongqing University of Posts and Telecommunications (CQUPT), China, in 2017, where he is currently pursuing the M.S. degree with the Key Laboratory of Mobile Communication Technology. His research interests include 5G network slicing, resource allocation of network function virtualization (NFV), and applications of machine learning in 5G mobile networks.

QIANBIN CHEN (M'03–SM'14) received the Ph.D. degree in communication and information system from the University of Electronic Science and Technology of China, Chengdu, China, in 2002.

He is currently a Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, and the Director of the Chongqing Key Laboratory of Mobile Communication Technology. He has authored or coauthored more than 100 articles in journals and peer-reviewed conference proceedings, and coauthored seven books. He holds 47 granted national patents.