IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 19, 2020, accepted February 4, 2020, date of publication February 17, 2020, date of current version March 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974287

Joint Radio Resource Allocation and Content Caching in Heterogeneous Virtualized Wireless Networks

YAN KYAW TUN[1], ANSELME NDIKUMANA[1,2], SHASHI RAJ PANDEY[1], ZHU HAN[1,3], (Fellow, IEEE), AND CHOONG SEON HONG[1], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Kyung Hee University, Yongin-Si 17104, South Korea
[2]Faculty of Computing and Information Sciences, University of Lay Adventists of Kigali, Kigali, Rwanda
[3]Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004, USA

Corresponding author: Choong Seon Hong (cshong@khu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant NRF-2017R1A2A2A05000995, in part by the Ministry of Science and ICT (MSIT), South Korea, and in part by the Grand Information Technology Research Center Support Program supervised by the Institute for Information and Communications Technology Promotion (IITP) under Grant IITP-2019-2015-0-00742.

ABSTRACT An efficient content caching policy at the edge of the mobile cellular network can improve the quality of services of the mobile users and reduces network congestion at the backhaul. On the other hand, the wireless network virtualization emerges as a cutting-edge technique to address the limited network capacity problem due to the exponential growth of the mobile data traffic. In addition, the wireless network virtualization can bring huge benefits such as reducing the capital expenditure (CAPEX) and the operational expenditure (OPEX) as well as improving the network capacity. In this regard, one of the key requirements to recognize the benefits of the above mentioned two technologies is to have an applicable resource allocation framework, that enables the deployment of the content caching schemes in the virtualized wireless network. In this study, we investigate a novel joint radio resource allocation and content caching problem to efficiently utilize the radio resource blocks, the transmit power, and the available cache storage at the base stations (BSs). The goal of the formulated problem presented in this paper focuses on minimizing the delays experienced by the end mobile users of mobile virtual network operators (MVNOs). We show that the formulated problem is a non-convex, mixed integer non-linear problem (MINLP), which is NP-hard and simply intractable. Therefore, we deploy the block upper-bound minimization (BSUM) algorithm to solve the formulated problem. Numerical results show that our method outperforms the existing baseline resource allocation schemes, with up to 19% performance gain in terms of network delay.

INDEX TERMS Wireless network virtualization, resource allocation, content caching, block successive upper-bound minimization (BSUM).

I. INTRODUCTION

Wireless Network Virtualization (WNV) has been seen as one of the exciting innovations for solving unprecedented challenges in the future wireless networks (5G) due to massive growth in mobile devices, mobile video traffic and data-oriented mobile applications [1]. Wireless network virtualization enables multiple mobile network operators (MVNOs) to share network resources such as spectrum, power, and physical infrastructure owned by the infrastructure provider (InP). Thus, WNV logically disconnects the existing cellular network into two entities: infrastructure provider (InP) and mobile virtual network operators (MVNOs). Moreover, WNV facilitates to achieve higher peak rate, lower delay, lower network deploying cost, and improves network capacity [2]–[4]. In spite of the fact that network virtualization is an auspicious technology for the future wireless network, there are several challenges to be addressed. One of such is *how to efficiently share network resources (i.e., wireless subchannels, power) among different MVNOs* while ensuring the two main requirements: *inter-isolation* (i.e., no interference among different MVNOs) and *intra-isolation*

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

(i.e., no interference among the users of the same MVNO) [5]. In this regard, there are two kinds of resource sharing framework in WNV: 1) InP is the central player, and directly shares/allocates the network resources to the users of MVNOs [6], [7], and 2) InP at first shares network resources to MVNOs, and then each MVNO manages and allocates network resources to its mobile users [8]–[10]. In this work, we choose the first *resource sharing framework* in the wireless network virtualization.

Additionally, 75% of the global mobile traffic in the future wireless network will be mobile video traffic by 2020 [11]. This massive surge in network traffic is going to put burden on radio access network (RAN) as well as on the backhaul network that is coupling RAN with the core network (CN). Hence, efficient resource utilization, i.e., proper usage of radio resources (e.g., wireless channels, power) in the radio access network, and relieving congestion at the backhaul are becoming critical affairs. Deploying content storage capacity and caching popular contents at each base station (BS), i.e., at the edge of the radio access network *reduces end-to-end latency* experienced by the mobile users while downloading contents, and further improves network performances due to the reduction of traffic congestion at the backhaul link [12]–[14].

A. RESEARCH CONTRIBUTIONS

In this work, different from the previous works, we consider the joint deployment of wireless network virtualization (WNV) and content caching at the edge of the access network to fulfill traffic demand of the future wireless network. This approach will reduce traffic congestion at the backhaul link during high-traffic hours while meeting QoS requirements of the users of MVNOs, with minimum end-to-end delay. Nonetheless, in WNV, multiple MVNOs coexist on the same infrastructure and share the limited storage capacity of InP at each BS. Therefore, the network performance is severely affected by the duplication of cached contents at each BS, specially, when the InP simply divides the limited storage capacity of each BS into multiple slices, and then allocates each storage slice to each MVNO. Because of this, *an efficient storage sharing and contents caching framework between MVNOs is required, jointly, with an efficient wireless resources (i.e., channels and power) allocation* to upgrade the performance of the network.

In order to address the aforementioned challenges of content caching jointly with the radio resource allocation in the virtualized wireless network, we propose a unique framework for content storage sharing, content caching, subchannels and power allocation in WNV. The summary of the main contributions of this paper is given as follows:

- Firstly, we formulate a joint content caching, channels and power allocation problem in the virtualized wireless network that minimizes the network delay experienced by the users of all MVNOs while downloading contents, subject to the QoS requirement of users of all MVNOs, cache storage capacity of BSs, and subchannels and power constraints of BSs. The formulated problem is a non-convex, mixed-integer nonlinear problem (MINLP), which is NP-hard due to the combinatorial properties.
- To address the proposed problem, we firstly relax the binary variables into a continuous form. Then, we transform the relaxed problem into the proximal upper bound minimization problem and apply BSUM algorithm [15], [16]. BSUM is a new powerful approach for solving non-convex, non-smooth optimization problems, where the original complex problem is decomposed into smaller sub-problems.
- In the simulation section, firstly we present the total network delay experienced by the users of all MVNOs under the proposed solution approach based on BSUM method. Then, the performance of our proposed algorithm is compared with other baseline schemes: Equal cache storage (ECS) allocation, Equal power EP) allocation, Equal channel (EC) allocation, and Equal channel + Equal power + Equal cache storage (EC+EP+ECS) allocation. We observe that the proposed algorithm attains a significant performance gain: likely 19.75%, 10.06%, 5.35%, and 4.28% in comparison with (EC+EP+ECS), ECS, EP, and EC allocation scheme, respectively. Lastly, we show that our proposed algorithm converges quickly to the optimal.

The remainder of this paper is arranged according to the following: In Section II, we summarize related works. The system model and content caching in the heterogeneous virtualized wireless networks framework are presented in Section III. Section IV demonstrates the proposed joint resource allocation and content caching problem in the heterogeneous virtualized wireless networks. The overview of BSUM algorithm and the solution approach based on the BSUM method to solve the proposed problem are introduced in Sections V and VI, respectively. Section VII discusses the numerical results of our proposed solution approach. Finally, Section VIII concludes the paper.

II. RELATED WORKS

We can categorize the existing works into two groups: (i) communication resources (i.e., subchannels, power) allocation, and (ii) content caching in virtualized wireless network.

(i) Communication resources allocation in virtualized wireless network: In [17], the authors introduced a joint subchannels and power allocation in WNV. The work of [18] considered user clustering and non-orthogonal multiple access (NOMA) based joint channels and power allocation in WNV. A Lyapunov optimization based hierarchical wireless resource management in virtualized wireless networks was presented in [19]. However, in the above mentioned works, the authors considered a single cell scenario that undermines inter-tier (inter-cell) interference and that can make it easy to solve the optimization problem. In [20], the authors proposed

a stochastic optimization based cells selection and resource allocation framework in virtualized wireless networks. The work of [21] introduced an efficient power allocation algorithm and interference management in multi-cell virtualized wireless network. Nonetheless, in this work, the authors have omitted subchannels allocation. A joint user association and allocation of resources in WNV and a stochastic game-based wireless spectrum allocation was proposed in [22] and [23]. Furthermore, some of the research works studied spectrum sharing from an InP to the MVNOs. As an example, in [24], the resource allocation in WNV was studied by using an opportunistic spectrum sharing framework. A bankruptcy game-based dynamic spectrum sharing among MVNOs was proposed in [25]. However, in all of these research works, individual users were not considered for spectrum sharing among MVNOs.

Moreover, a few existing works proposed resource allocation in WNV from an economic perspective rather than considering the technological aspects. In our previous work [8], we introduced Generalized Kelly Mechanism (GKM) based hierarchical resource allocation in the virtualized wireless networks. In this work, an InP firstly allocates its wireless resources among MVNOs based on an auction mechanism, and then each MVNO efficiently shares the obtained wireless resource from the InP to its mobile users. Similarly, VCG (Vickrey-Clarke-Groves) auction mechanism based power allocation and spectrum sharing by using a sequential auction in the wireless network virtualization were proposed in [26] and [27]. In [28], the authors proposed stochastic resource allocation in WNV. In their work, they tried to maximize the profit of the mobile network operator (i.e., the profit of the infrastructure provider (InP)). Furthermore, a cyclic three-sided matching game based resource allocation in the WVN was introduced in [29]. All of these works aimed at maximizing the profits of the MVNOs.

(ii) Content caching in the virtualized wireless network: Although there are several existing works which focus on improving the performance of content caching, i.e., to diminish the transmission delay, and increase the cache hit ratio at the edge of the wireless network, only a few works considered efficient content caching in virtualized wireless network [30]–[32]. However, signal-to-noise ratio (SNR) based content placing at the base station will not be efficient because content caching decision at each BS will be long-term scale, but SNR is changing over a short period. In [33], the authors have proposed a deep learning based cache buying strategies for the MVNOs where each MVNO buys cache storage from the InP in order to maximize their own users' experience. The authors in [34] proposed software-defined networking based resource allocation and content caching problem in the virtualized network where they tried to maximize the utility of the mobile virtual network operators (MVNOs). In [35], the authors considered the optimal caching strategy in the virtualized wireless networks, and the work in [36] introduced the joint resource allocation in content caching for wireless network virtualization with the objective of minimizing the content request rejection rate. However, the authors did not consider the frequency reuse between the BSs, i.e., no consideration for inter-tier interference, backhaul delay, and power allocation.

In this work, we consider the significant requirements that were mostly overlooked in the prior works, and jointly combine the communication and caching schemes in the heterogeneous virtualized wireless network to formulate an optimization problem. Moreover, in order to solve our proposed optimization problem, we introduce the BSUM algorithm. BSUM is a powerful framework for solving big data, non-convex and non-smooth optimization problem. Furthermore, BSUM allows to effectively decompose the initial optimization problem into multiple smaller-subproblems, which can be solved disjointly and in parallel.

III. SYSTEM MODEL

We consider a WNV scenario where an infrastructure provider (InP) installs a set of base stations (BSs) $\mathcal{B}$, comprising a macro base station (MBS) and $(|\mathcal{B}|-1)$ small-cell base stations (SBSs), as illustrated in Fig. 1. Each BS $b \in \mathcal{B}$ is operating on a set of channels $\mathcal{N} = \{1, 2, \ldots, N\}$, where each wireless channel has total bandwidth ω. Moreover, each BS is connected to the core network (CN) with the high-capacity wired backhaul link. This set of BSs installed by the InP serves a set of mobile virtual network operators (MVNOs) $\mathcal{M} = \{1, 2, \ldots, M\}$, where each MVNO is providing a specific services to a set of users $\mathcal{U} = \{1, 2, \ldots, U\}$.

In this work, we assume that the mobile users of MVNOs are associated to the specific BS in prior, and orthogonal frequency division multiple access (OFDMA) based wireless access protocol is used among the users of a BS to avoid intra-cell interference. We define $y_{mu}^{bn} \in \{0, 1\}$ as a channel assignment variable which indicates whether or not channel n of BS b is assigned to user u of MVNO m.

$$y_{mu}^{bn} = \begin{cases} 1, & \text{if channel } n \text{ of MBS } b \text{ is assigned} \\ & \text{to user } u \text{ in MVNO } m\,, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Then, the received signal to interference plus noise ratio (SINR) of user u of MVNO m on channel n of BS b can be expressed as follows:

$$\gamma_{mu}^{bn} = \frac{p_{mu}^{bn} h_{mu}^{bn}}{I_{mu}^{bn} + \sigma^2}, \tag{2}$$

where h_{mu}^{bn}, p_{mu}^{bn} are the channel gain and the allocated power to the user u of MVNO m who is assigned to channel n of BS b, respectively, and

$$I_{mu}^{bn} = \sum_{b' \in \mathcal{B}, b' \neq b} \sum_{m \in \mathcal{M}} \sum_{u' \in \mathcal{U}, u' \neq u} p_{mu'}^{b'n} h_{mu'}^{b'n},$$

is the interference on mobile user u of MVNO m who is assigned to channel n of BS b, and σ^2 is the noise power spectral. Therefore, we can calculate the data rate of user u

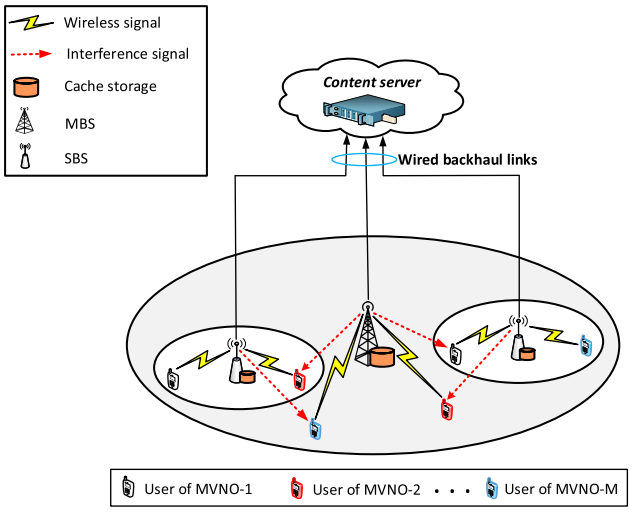

FIGURE 1. System model.

of MVNO m on channel n of BS b as follows:

$$R_{mu}^{bn} = \omega \log_2(1 + \gamma_{mu}^{bn}). \tag{3}$$

To calculate the data rate of users of all MVNOs at each BS, we consider the following constraints:

- Transmit Power Constraint: The transmit power allocated to the users of all MVNOs at each base station must be less than the maximum transmit power of that base station. Therefore, we can write the transmit power constraint as follows:

$$\text{C1}: \sum_{m\in\mathcal{M}} \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} p_{mu}^{bn} \le P_b^{\max}, \quad \forall b \in \mathcal{B}, \tag{4}$$

  where $P_b^{\max}$ is the maximum transmit power of BS $b \in \mathcal{B}$.
- QoS Constraint: The InP guarantees pre-agreement with MVNOs by the minimum data rate requirement constraint as:

$$\text{C2}: \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} y_{mu}^{bn} \omega \log_2(1+\gamma_{mu}^{bn}) \ge R_m^{\min}, \quad \forall m \in \mathcal{M}, \tag{5}$$

  where $R_m^{\min}$ is the minimum rate requirement of MVNO $m \in \mathcal{M}$.
- Resource Scheduling Constraint: According to OFDMA technique for channel access, a channel is assigned to at most one mobile user in each BS. Therefore, we have the following constraint:

$$\text{C3}: \sum_{m\in\mathcal{M}} \sum_{u\in\mathcal{U}} y_{mu}^{bn} \le 1, \quad \forall b \in \mathcal{B}, \ \forall n \in \mathcal{N}. \tag{6}$$

On the other hand, the communication cost (i.e., delay) is very high to access contents from the content server located at the core network (CN). In order to reduce the communication delay and save the backhaul link capacity, in this work, we assumed that the InP deploys and manages the

TABLE 1. Summary of key notations.

Notation	Definition
$\mathcal{B}$	Set of base stations (BSs), $\lvert\mathcal{B}\rvert = B$
$\mathcal{N}$	Set of channels, $\lvert\mathcal{N}\rvert = N$
ω	Total bandwidth of each channel
$\mathcal{M}$	Set of MVNOs, $\lvert\mathcal{M}\rvert = M$
$\mathcal{U}$	Set of mobile users of MVNO $m \in \mathcal{M}$, $\lvert\mathcal{U}\rvert = U$
y_{mu}^{bn}	Channel assignment variable
p_{mu}^{bn}	Allocated power to user u of MVNO m who is assigned to channel n of BS b
h_{mh}^{bn}	Achievable channel gain of user u of MVNO m
γ_{mu}^{bn}	Received signal to interference plus noise ratio (SINR)
σ^2	Noise power spectral
I_{mu}^{bn}	Interference experienced by user $u \in \mathcal{U}$ of MVNO $m \in \mathcal{M}$ that is associated with the base station $b \in \mathcal{B}$
$P_b^{\max}$	Total transmit power of base station $b \in \mathcal{B}$
R_{mu}^{bn}	Data rate achieved by the user $u \in \mathcal{U}$ of MVNO $m \in \mathcal{M}$
$R_m^{\min}$	Minimum requirement for data rate of MVNO $m \in \mathcal{M}$
z_{mu}^{bs}	Cache decision variable
C_b	Total cache capacity of the base station $b \in \mathcal{B}$
$\mathcal{S}$	Set of contents, $\lvert\mathcal{S}\rvert = S$
L_s	Size of content $s \in \mathcal{S}$
$H_{m,u}^{b}$	Backhaul delay incurred by user $u \in \mathcal{U}$ of MVNO $m \in \mathcal{M}$ associated with the BS b when downloading content s from the content server
$T_{m,n,u}^{b,s}$	Transmission delay incurred by the user $u \in \mathcal{U}$ of MVNO $m \in \mathcal{M}$ when downloading content s from the base station b
$D_{b,n,m,u}^{s}$	Total delay experienced by the user $u \in \mathcal{U}$ of MVNO $m \in \mathcal{M}$ that is associated with the BS $b \in \mathcal{B}$ when downloading content s
$\mathcal{Y}$	Feasible set of channel assignment variable
$\mathcal{P}$	Feasible set of power control variable
$\mathcal{Z}$	Feasible set of cache decision variable
$\mathcal{X}$	Set of index
i	Iteration number
μ_x	Positive penalty parameter

cache storage at each BS with proactive caching [37], where cache storage is shared and used by all MVNOs. Let C_b denote the maximum storage capacity of the cache repository deployed at BS $b \in \mathcal{B}$. Without loss of generality, different BSs have different storage capacity. Here, the mobile users of MVNOs are willing to download contents in the following content catalog $\mathcal{S} = \{1, 2, \ldots, S\}$. We consider that all of the contents in the content list have same size (i.e., $L_1 = \ldots = L_s = \ldots = L_S$). This is feasible using advanced coding technique, by applying which all of the contents can be divided into the same block length [1]. In the coverage area of BS $b \in \mathcal{B}$, each content has a different value of popularity which depends on the probability of it being requested by the users of MVNOs. In this work, we consider the content requests by the users of MVNOs (i.e., content popularity) follow the Zipf-like distribution [38], where the probability of requesting content $s \in \{1, 2, \ldots, S\}$ by the users is given by:

$$Z(s, \lambda) = \frac{\Theta}{s^{\lambda}}, \tag{7}$$

where $\Theta = \left(\sum_{i=1}^{S} \frac{1}{i^{\lambda}}\right)^{-1}$, and $0 \le \lambda \le 1$ is the exponent characterizing the Zipf distribution. When $\lambda = 0$, all of the contents have the same popularity. Otherwise, when $\lambda \ne 0$,

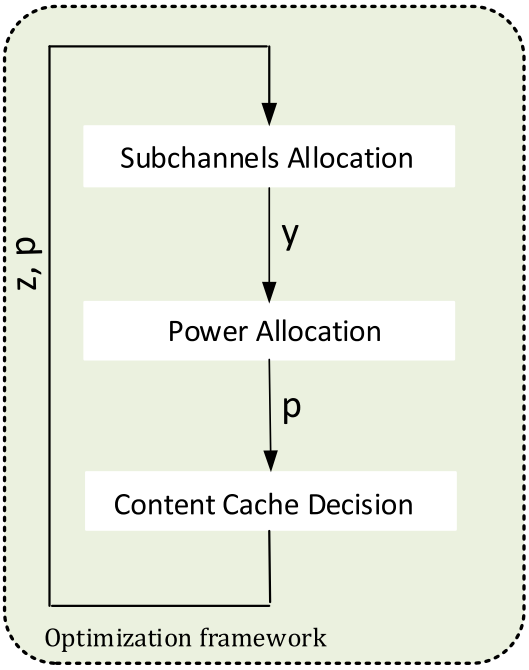

FIGURE 2. Optimization framework.

the contents have different popularity. Furthermore, the contents are ranked in order of their popularity. Here, we consider that BSs cache the contents in the order of their popularity starting from the highest popularity. Now, we define another new variable, $q_{mu}^{bs} \in \{0, 1\}$ which represents whether user $u \in \mathcal{U}$ of MVNO $m \in \mathcal{M}$ requests the content $s \in \mathcal{S}$ at BS b or not, i.e.,

$$q_{mu}^{bs} = \begin{cases} 1, & \text{if content } s \text{ is requested by user } u \text{ of} \\ & \text{MVNO } m \text{ at BS } b, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Then, let the new variable z_m^{bs} be the cache decision variable, which represents whether or not content s is cached at BS b and shared to MVNO m as

$$z_m^{bs} = \begin{cases} 1, & \text{if the content } s \text{ is cached at the BS } b \text{ and shared} \\ & \text{to MVNO } m, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

In particular, each BS caches the popular contents among the mobile users of MVNOs in advance for possible future requests [2], [3]. Moreover, mobile users of different MVNOs can share the cached contents at each BS's content storage, so that the same content will not be cached at each BS to serve the mobile users from different MVNOs. However, user mobility analysis in joint radio resource allocation and content caching is outside the scope of this paper and will be subjected to our future research. Therefore, we can express the above constraint in mathematical form as follows:

$$\sum_{m \in \mathcal{M}} z_m^{bs} \leq 1, \quad \forall s \in \mathcal{S}, \ \forall b \in \mathcal{B}. \tag{10}$$

Moreover, the maximum content storage capacity constraint at each BS m can be expressed as follows:

$$\sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{S}} z_m^{bs} L_s \leq C_b, \quad \forall b \in \mathcal{B}. \tag{11}$$

IV. PROBLEM FORMULATION

Next, we derive the joint radio resource allocation and content caching problem in the heterogeneous virtualized wireless networks which aims to minimize the overall network delay. The network delay consists of two main terms: the wireless transmission delay, and backhaul delay. Firstly, we introduce wireless transmission delay of user $u \in \mathcal{U}$ in MVNO $m \in \mathcal{M}$ to download content $s \in \mathcal{S}$ when associating with BS $b \in \mathcal{B}$. The wireless transmission delay of user u of MVNO m on channel n of BS b to download content s can be expressed as follows:

$$T_{m,n,u}^{b,s} = \frac{L_s}{\omega \log_2(1 + \gamma_{mu}^{bn})}, \tag{12}$$

where L_s is the size of content s. From (12), we observe that the transmission delay experienced by user u of MVNO m to download the content from the base station b depends on the size of the content, and signal to interference plus noise ratio (SINR).

The second component of the network delay is the backhaul delay. We add the backhaul delay experienced by the user u in MVNO m associated with the BS b as $H_{m,u}^b$ when downloading the content from the content server. In case of wired backhaul, the characterizing components of the backhaul delay are: a) the average link distance, b) the average traffic load, and c) the average number of SBSs associated with a single small cell gateway interface. In this work, we follow the backhaul delay model as in [39].

Therefore, the network delay experienced by user u of MVNO m on channel n of MBS b to download content s can be expressed as follows:

$$D_{m,u,n}^{b,s} = \left[(1 - z_m^{bs}) \left(H_{m,u}^b + \frac{L_s}{\omega \log_2(1 + \gamma_{mu}^{bn})} \right) \right] + \frac{z_m^{bs} L_s}{\omega \log_2(1 + \gamma_{mu}^{bn})}. \tag{13}$$

Considering the limitation of bandwidth efficiency, transmission power restriction, and the storage capacity constraint, the joint resource allocation and content caching in multi-cells virtualized wireless networks problem to minimize the total delay experienced by all MVNOs users can be formulated as follows:

$$\min_{\boldsymbol{y},\boldsymbol{p},\boldsymbol{z}} \sum_{b \in \mathcal{B}} \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} \sum_{u \in \mathcal{U}} y_{mu}^{bn} \sum_{s \in \mathcal{S}} q_{mu}^{bs} D_{m,u,n}^{b,s}, \tag{14}$$

$$\text{s.t.} \sum_{m \in \mathcal{M}} \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} p_{mu}^{bn} \leq P_b^{\max}, \quad \forall b \in \mathcal{B}, \tag{15}$$

$$\sum_{b \in \mathcal{B}} \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} y_{mu}^{bn} \omega \log_2(1 + \gamma_{mu}^{bn}) \geq R_m^{\min}, \quad \forall m, \tag{16}$$

$$\sum_{m \in \mathcal{M}} \sum_{u \in \mathcal{U}} y_{mu}^{bn} \leq 1, \quad \forall b \in \mathcal{B}, \ \forall n \in \mathcal{N}, \tag{17}$$

$$\sum_{m\in\mathcal{M}} z_m^{bs} \le 1, \quad \forall s \in \mathcal{S},\ \forall b \in \mathcal{B}, \tag{18}$$

$$\sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}} z_m^{bs} L_s \le C_b, \quad \forall b \in \mathcal{B}, \tag{19}$$

$$p_{m,u}^{b,n} \ge 0, \quad \forall b \in \mathcal{B},\ \forall n \in \mathcal{N},\ \forall m \in \mathcal{M}, \forall u \in \mathcal{U}, \tag{20}$$

$$y_{m,u}^{b,n} \in \{0,1\}, \quad \forall b \in \mathcal{B},\ \forall n \in \mathcal{N},\ \forall m \in \mathcal{M},\ \forall u \in \mathcal{U}, \tag{21}$$

$$z_m^{bs} \in \{0,1\}, \forall b \in \mathcal{B}, \quad \forall m \in \mathcal{M},\ \forall s \in \mathcal{S}, \tag{22}$$

where the objective function described in (14) is a non-convex function because of the inter-cell interference, and the decision variables (i.e., $\boldsymbol{y}, \boldsymbol{p}$ and z) are coupling in the objective function. Moreover, the constraints mentioned in (15)-(22) are non-linear and which are the mixture of continuous and binary variables, i.e., $\boldsymbol{y}, \boldsymbol{p}, z$. As a conclusion, the above optimization problem is a non-convex mixed integer nonlinear problem (MINLP), which is NP-hard and takes exponential time complexity to find the optimal solution. Therefore, we deploy BSUM algorithm to solve our proposed problem.

V. OVERVIEW OF BSUM ALGORITHM

BSUM, a distributed algorithm, can be used for parallel computing. The use of BSUM in centralized algorithms results in advantages for both solution speed and problem decomposability. Hence, the following function is considered as a block-structured optimization problem for the introduction of the standard form of BSUM [40]:

$$\begin{aligned} &\min_{\boldsymbol{a}} f(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_x), \\ &\text{s.t. } \boldsymbol{a}_x \in \mathcal{D}_x, \\ &\quad \forall x \in \mathcal{X},\ x = 1, 2, \ldots, X, \end{aligned} \tag{23}$$

where $\mathcal{D} := \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_X, f(\cdot)$ is the continuous function, and $\mathcal{X}$ denotes the set of index. When $x = 1, 2, \ldots, X$, $\mathcal{D}_x$ is considered as a close convex set and a_x as a block of variables. If we apply block coordinate descent (BCD) in every iteration i, it will optimize a single block of variable by solving the following problem:

$$a_x^i \in \min_{\boldsymbol{a}_x \in \mathcal{D}_x} f(\boldsymbol{a}_x, \boldsymbol{a}_{-x}^{i-1}), \tag{24}$$

where $\boldsymbol{a}_{-x}^{i-1} := (\boldsymbol{a}_1^{i-1}, \ldots, \boldsymbol{a}_{x-1}^{i-1}, \boldsymbol{a}_{x+1}^{i-1}, \ldots, \boldsymbol{a}_x^{i-1})$, $a_k^i = a_k^{i-1}$ for $x \neq k$.

However, both (23) and (24) are challenging to solve. The hardest condition will happen when the function in (23) is non-convex and BCD cannot guarantee convergence for every time. So, by applying BSUM at a given feasible point $\boldsymbol{e} \in \mathcal{D}$, the proximal upper-bound function $g(\boldsymbol{a}_x, \boldsymbol{e})$ of $f(\boldsymbol{a}_x, \boldsymbol{e}_{-x})$ can be introduced. There are three common schemes used for choosing the proximal upper-bound function: linear upper-bound, quadratic upper-bound, and Jensen's upper-bound. Here, the selected proximal upper-bound function $g(\boldsymbol{a}_x, \boldsymbol{e})$ has to satisfy the following assumption.

Assumption 1:
(*i*) $g(\boldsymbol{a}_x, \boldsymbol{e}) = f(\boldsymbol{e})$,
(*ii*) $g(\boldsymbol{a}_x, \boldsymbol{e}) > f(\boldsymbol{a}_x, \boldsymbol{e}_{-x})$,
(*iii*) $g'(\boldsymbol{a}_x, \boldsymbol{e}; \boldsymbol{r}_x)|_{\boldsymbol{a}_x = \boldsymbol{e}_x} = f'(\boldsymbol{e}; \boldsymbol{r}), \boldsymbol{e}_x + \boldsymbol{r}_x \in \mathcal{D}_x$.

Assumption 1(*i*) and 1(*ii*) guarantee that the objective function f is globally upper-bounded by the proximal upper-bound function g. Then, Assumption 1(*iii*) ensures the existence of first-order derivative satisfying the fact that $g(\boldsymbol{a}_x, \boldsymbol{e})$ takes negative steps towards the gradient of the objective function $f(\boldsymbol{a}_x, \boldsymbol{e}_{-x})$ in the direction of $\boldsymbol{r}$. For ease of presentation, the following proximal upper bound function is defined by adding quadratic penalty term to the objective function:

$$g(\boldsymbol{a}_x, \boldsymbol{e}) = f(\boldsymbol{a}_x, \boldsymbol{e}_{-x}) + \frac{\theta}{2}||(\boldsymbol{a}_x - \boldsymbol{e}_x)||^2, \tag{25}$$

where θ is the positive parameter, i.e., $\theta > 0$. The BSUM algorithm addresses the proximal upper-bound function in (25) at each iteration i along with the following update:

$$\begin{cases} a_x^i \in \min\limits_{\boldsymbol{a}_x \in \mathcal{D}_x} g(\boldsymbol{a}_x, \boldsymbol{a}_x^{i-1}), & \forall x \in \mathcal{X}, \\ \boldsymbol{a}_x^{(i)} = \boldsymbol{a}_k^{i-1}, & \forall k \notin \mathcal{X}. \end{cases} \tag{26}$$

The detailed procedures of the BSUM algorithm is presented in Algorithm 1. The BSUM can be viewed as the generalized form of BCD which successively updates the primary variable blocks to maximize the upper-bound function of the original objective function. It can be used to solve separable convex optimization problems with linear coupling constraints that are smooth or non-smooth. Specifically, each block of variables is iteratively updated to minimize the upper-bound proximal function before convergence to both a coordinate-wise minimum and a stationary solution. If a block of variables reaches the minimum point $\boldsymbol{a}_x^* = \boldsymbol{a}_x^{(i+1)}$, a stationary solution is assumed to be a coordinate-wise minimum. In other words, the whole point vector cannot find a better minimum direction at stationary points [40], [41]. Based on [40], we have the following remark:

Remark 1 (Convergence of BSUM Algorithm): In order to converge to the ϵ-optimal solution, BSUM algorithm takes at most $\mathcal{O}(\log(1/\epsilon))$ iterations. In other words, it is a sub-linear convergence.

The ϵ-optimal solution can be determined as $\boldsymbol{a}_x^\epsilon \in \{\boldsymbol{a}_x | \boldsymbol{a}_x \in \mathcal{D}_x, g(\boldsymbol{a}_x, \boldsymbol{a}^i, \boldsymbol{e}^i) - g(\boldsymbol{a}_x^*, \boldsymbol{a}^i, \boldsymbol{e}^i)\} \le \epsilon$, where $g(\boldsymbol{a}_x^*, \boldsymbol{a}^i, \boldsymbol{e}^i)$ is the optimal of $g(\boldsymbol{a}_x, \boldsymbol{e})$ w.r.t $\boldsymbol{a}_x$.

Algorithm 1 Standard BSUM Algorithm

1: **Initialization:** Set $i = 0, \epsilon > 0$, and find initial feasible solution $\boldsymbol{a}^0$;
2: **repeat**
3: Choose index set $\mathcal{X}$;
4: Let $\boldsymbol{a}_x^{(i+1)} \in \min g(\boldsymbol{a}_x, \boldsymbol{a}_{-x}^i), \forall x \in \mathcal{X}$;
5: Set $\boldsymbol{a}_x^{(i+1)} = \boldsymbol{a}_k^i, \forall k \notin \mathcal{X}$;
6: $i = i + 1$;
7: **until** $\| \frac{g_x^{(i)} - g_x^{(i+1)}}{g_x^{(i)}} \| \le \epsilon$;
8: Then, set $\boldsymbol{a}_x^{(i+1)}$ as the desired solution.

For analyzing the complexity of the BSUM Algorithm 1, we use the complexity analysis described in [42] and make the following remark:

Remark 2 (Complexity of BSUM Algorithm): The Algorithm 1, which is the standard BSUM algorithm, uses proximal upper-bound minimization technique, where each block of variables is iteratively updated to minimize the upper-bound proximal function until it converges to the coordinate-wise minimum point which is a stationary point. Based on Assumption 1 (i, ii) and Remark 1, with flexible update rules, BSUM algorithm has sub-linear rate of convergence. Therefore, as described and proved in [42], the Algorithm 1 has sub-linear iteration complexity $\mathcal{O}(1/i)$, where i is the index of iteration.

VI. SOLUTION APPROACH FOR JOINT RADIO RESOURCE ALLOCATION AND CONTENT CACHING

To address our proposed problem, firstly, we relax the channel assignment variable $y_{m,u}^{b,n}$ and the cache decision variable z_m^{bs} in constraints (21) and (22) into the continuous form, i.e., $0 \leq y_{m,u}^{b,n} \leq 1$ and $0 \leq z_m^{bs} \leq 1$. Then, the above optimization problem (14) can be rewritten into the following form:

$$\min_{\boldsymbol{y},\boldsymbol{p},z} \sum_{b\in\mathcal{B}}\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}}\sum_{u\in\mathcal{U}} y_{mu}^{bn}\sum_{s\in\mathcal{S}} q_{mu}^{bs} D_{m,u,n}^{b,s} \tag{27}$$

$$\text{s.t. (15)-(20)}, \tag{28}$$

$$y_{m,u}^{b,n} \in [0,1], \quad \forall b \in \mathcal{B},\ \forall n \in \mathcal{N},\ \forall m \in \mathcal{M}, \forall u \in \mathcal{U}, \tag{29}$$

$$z_m^{bs} \in [0,1], \quad \forall b \in \mathcal{B},\ \forall m \in \mathcal{M},\ \forall s \in \mathcal{S}. \tag{30}$$

Likewise, we can rewrite the optimization problem in (27) more concisely as

$$\min_{y\in\mathcal{Y},p\in\mathcal{P},z\in\mathcal{Z}} \mathcal{O}(\boldsymbol{y},\boldsymbol{p},z), \tag{31}$$

where $\mathcal{O}(\boldsymbol{y},\boldsymbol{p},\boldsymbol{z}) \triangleq \sum_{b\in\mathcal{B}}\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}}\sum_{u\in\mathcal{U}} y_{mu}^{bn}\sum_{s\in\mathcal{S}} q_{mu}^{bs} D_{m,u,n}^{b,s}$ is the objective function. Moreover, $\mathcal{Y} \triangleq \{\boldsymbol{y} : \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{n\in\mathcal{N}} y_{mu}^{bn}\omega\log_2(1+\gamma_{mu}^{bn}) \geq R_m^{\min}, \sum_{m\in\mathcal{M}}\sum_{u\in\mathcal{U}} y_{mu}^{bn} \leq 1, y_{mu}^{bn} \in [0,1], \forall b,n,m,u\}$, $\mathcal{P} \triangleq \{\boldsymbol{p} : \sum_{m\in\mathcal{M}}\sum_{u\in\mathcal{U}}\sum_{n\in\mathcal{N}} p_{mu}^{bn} \leq P_b^{\max}, p_{m,u}^{b,n} \geq 0, \forall b,n,m,u\}$, and $\mathcal{Z} \triangleq \{\boldsymbol{z} : \sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}} z_m^{bs} L_s \leq C_b, \sum_{m\in\mathcal{M}} z_m^{bs} \leq 1, z_m^{bs} \in [0,1], \forall b,m,s\}$ are the feasible sets of $\boldsymbol{y},\boldsymbol{p}$ and $\boldsymbol{z}$, respectively.

The problem in (27), equivalent to (31), is non-convex after relaxing the binary variables into the continuous form, and in general, is challenging to solve. Even though, as an example, the subproblem of optimizing the variable $\boldsymbol{y}$ while the other variables such as $\boldsymbol{p}$ and $\boldsymbol{z}$ are fixed is still non-convex because of intercell (or inter-tier) interference. Therefore, we cannot directly use the traditional block coordinated descent (BCD) algorithm [43] because it cannot guarantee the convergence when the problem is non-convex. Consequently, we use the BSUM algorithm, which is a different method from the exact BCD. The BSUM algorithm guarantees convergence to the set of stationary points of the non-convex problem [41].

At each iteration i, $\forall x \in \mathcal{X}^i$, where $\mathcal{X}$ is the indexes set, we define the proximal upper-bound function $\mathcal{Q}_x$ of the objective function described in (31). At this point, we add the quadratic penalty term to the objective function in (31) to keep the proximal upper-bound function convex as

$$\mathcal{Q}_x(\boldsymbol{y}_x;\boldsymbol{y}^i,\boldsymbol{p}^i,\boldsymbol{z}^i) = \mathcal{Q}(\boldsymbol{y}_x;\tilde{\boldsymbol{y}},\tilde{\boldsymbol{p}},\tilde{\boldsymbol{z}}) + \frac{\mu_x}{2} \parallel (\boldsymbol{y}_x - \tilde{\boldsymbol{y}}) \parallel^2, \tag{32}$$

where μ_x is the positive penalty parameter, i.e., $\mu_x > 0$.

(32) is the proximal upper-bound function of the objective function in (31) for the vector of variable $\tilde{\boldsymbol{y}}$, and it can be deployed for other vectors for variables $\tilde{\boldsymbol{p}}$, and $\tilde{\boldsymbol{z}}$, respectively. Additionally, the proximal upper-bound function is convex as a result of its penalty term, $\frac{\mu_x}{2} \parallel (\boldsymbol{y}_x - \tilde{\boldsymbol{y}}) \parallel^2$. Strictly speaking, with respect to $\boldsymbol{y}_x$, $\boldsymbol{p}_x$, and $\boldsymbol{z}_x$, the proximal upper-bound function has the unique minimizers vector $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{p}}$, and $\tilde{\boldsymbol{z}}$ at each iteration i, which are examined as the solution of the previous iteration $(i-1)$. Then, the solution in each iteration $(i+1)$ can be updated by solving the following subproblems:

$$\boldsymbol{y}_x^{(i+1)} \in \min_{\boldsymbol{y}_x\in\mathcal{Y}} \mathcal{Q}_x\left(\boldsymbol{y}_x;\boldsymbol{y}^{(i)},\boldsymbol{p}^{(i)},\boldsymbol{z}^{(i)}\right), \tag{33}$$

$$\boldsymbol{p}_x^{(i+1)} \in \min_{\boldsymbol{p}_x\in\mathcal{P}} \mathcal{Q}_x\left(\boldsymbol{p}_x;\boldsymbol{p}^{(i)},\boldsymbol{y}^{(i+1)},\boldsymbol{z}^{(i)}\right), \tag{34}$$

$$\boldsymbol{z}_x^{(i+1)} \in \min_{\boldsymbol{z}_x\in\mathcal{Z}} \mathcal{Q}_x\left(\boldsymbol{z}_x;\boldsymbol{z}^{(i)},\boldsymbol{y}^{(i+1)},\boldsymbol{p}^{(i+1)}\right). \tag{35}$$

Moreover, the subproblems in (33)-(35) can be solved by using our proposed block successive upper-bound minimization (BSUM) algorithm for joint resource allocation and content caching in WNV as shown in Algorithm 2.

VII. SIMULATION RESULTS

In this section, we assess the performance of our proposed BSUM algorithm based joint resource allocation and content caching in the heterogeneous virtualized wireless networks.

Algorithm 2 BSUM Algorithm for Joint Radio Resource Allocation and Content Caching in WNV

1: **Initialization:** Set $i = 0$, $\epsilon > 0$, and find initial feasible solutions $(\boldsymbol{y}^{(0)},\boldsymbol{p}^{(0)},\boldsymbol{z}^{(0)})$;
2: **repeat**
3: Choose index set $\mathcal{X}$;
4: Find the solution for channel allocation problem at each BS, $\boldsymbol{y}_x^{(i+1)}$, by solving the following subproblem;
5: $\boldsymbol{y}_x^{(i+1)} \in \min_{\boldsymbol{y}_x\in\mathcal{Y}} \mathcal{Q}_x\left(\boldsymbol{y}_x;\boldsymbol{y}^{(i)},\boldsymbol{p}^{(i)},\boldsymbol{z}^{(i)}\right)$;
6: Set $\boldsymbol{y}_x^{(i+1)} = \boldsymbol{y}_k^i, \forall k \notin \mathcal{X}$;
7: Find the solution for power allocation problem at each BS, $\boldsymbol{p}_x^{(i+1)}$, by solving the subproblem in (34);
8: Find the solution for cache decision problem at each BS, $\boldsymbol{z}_x^{(i+1)}$, by solving the subproblem in (35);
9: $i = i + 1$;
10: **until** $\parallel \frac{\mathcal{Q}_x^{(i)} - \mathcal{Q}_x^{(i+1)}}{\mathcal{Q}_x^{(i)}} \parallel \leq \epsilon$;
11: Then, set $\left(\boldsymbol{y}_x^{(i+1)},\boldsymbol{p}_x^{(i+1)},\boldsymbol{z}_x^{(i+1)}\right)$ as the desired solutions.

TABLE 2. Summary of simulation parameters.

Simulation Parameters	Values
Number of MBS	1
Coverage radius of MBS	500 m
Number of SBSs	3
Number of MVNOs	3
Number of users of each MVNO	[5,10]
Carrier frequency	2 GHz
Frame structure	FDD
System bandwidth	10 MHz
Number of subchannels	50
System bandwidth of each subchannel	150 KHz
Total transmit power of MBS	48 dBm
Total transmit power of SBS	38 dBm
Thermal noise density	-174 dBm/Hz
Fading model	Rayleigh fading
Number of contents	100
Size of each content	15 KB
Storage capacity of each BS	can be varied from 20% 50% of total contents size
Convergence threshold (ϵ)	10^{-4}

A. SIMULATION SETUP

In this network scenario, the InP deploys a single macro base station (MBS) having the coverage radius of 500 m. Then, 3 SBSs with the radius of 100 m are randomly deployed by the InP within the coverage area of the MBS. Moreover, the InP installs the cache storage at the base stations, where the maximum cache storage capacity of each base station can be varied from 15% to 50% of the total content size. From above, it is clear that the total cache storage capacity of the BS is always smaller than the total content size. In this network setup, 3 MVNOs are sharing the infrastructure provided by the InP, where each MVNO is subscribed by the number of users $U \in [5, 10]$. Here, every user of each MVNO is associated with BS that has the strongest wireless signal strength. Then, mobile users of MVNOs access the list of 1000 contents (i.e., $S = 1000$), where the size of each content is 15 KB. In our simulation, we also consider that each BS is operating at the maximum system bandwidth 10 MHz. Therefore, according to LTE standard, the maximum available subchannels at each BS is 50, where each subchannel has the system bandwidth 15 kHz. Furthermore, the maximum transmit power of the MBS and each SBS are 48 dBm and 38 dBm, respectively, and the thermal noise density of the system is -174 dBm/Hz. Moreover, the small scale fading model, i.e., Rayleigh fading model is considered. The details of the parameters used in our simulation are presented in Table 2.

B. NUMERICAL RESULTS

In this subsection, we mainly focus on the performance gain of the proposed algorithm. Moreover, we compare the performance of our proposed solution approach with the existing baseline schemes. In this work, we consider the following 4 baseline schemes to compare our work with.

- Equal channel + Equal power + Equal cache storage (EC+EP+ECS) allocation: In this scheme, the channels and the transmit power of BSs are equally allocated to the users of all MVNOs. Moreover, the total cache storage capacity of each BS is also equally partitioned amongst MVNOs and the cache decision is independently made for each MVNO. Furthermore, the contents will not be shared amongst different MVNOs to serve their mobile users.
- Equal cache storage (ECS) allocation: Under equal cache storage (ECS) allocation scheme, the cache storage of each BS is equally shared amongst the associated MVNOs. Furthermore, contents sharing amongst MVNOs to serve their mobile users is also not possible. Then, channels and the transmit power of each BS are allocated to the users of MVNOs by using our proposed algorithm.
- Equal power (EP) allocation: In EP allocation scheme, the transmit power of each BS is equally allocated to its associated users of MVNOs.
- Equal channel (EC) allocation: The channels of each BS are equally allocated to its associated users of all MVNOs.

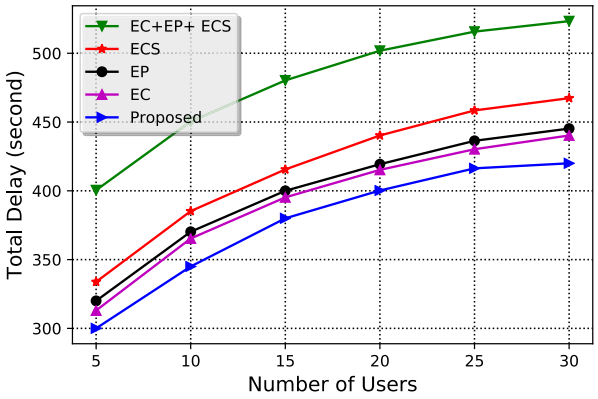

FIGURE 3. Total network delay versus number of users of MVNOs.

Fig. 3 demonstrates the total delay of the network under different number of users. Here, all BSs are operating at the maximum bandwidth 10 MHz (i.e., maximum available channels at each BS is 50) and the maximum content storage capacity of each BS is 50% of the total contents size. From Fig.3, we observe that the total delay increases with the increase in the number of users of MVNOs. Furthermore, we observe that the total network delay is 523.344 seconds (EC+EP+ECS), 467.345 seconds (ECS), 445.23 seconds (EP), 440 seconds (EC), and 420 seconds (Proposed), for 30 users in the network. We clearly observe that our proposed algorithm outperforms other baseline schemes. From there, we can conclude that equal sharing of radio resources and cache storage of the BSs among the users of MVNOs is not efficient.

Fig. 4 demonstrates the network delay experienced by the users of all MVNOs under different number of available channels (i.e., $N \in [50, 100, 150, 200]$). Concurrently, the cache storage capacity of each BS is 50% of the total contents size. As an example, we recognize from Fig. 4 that

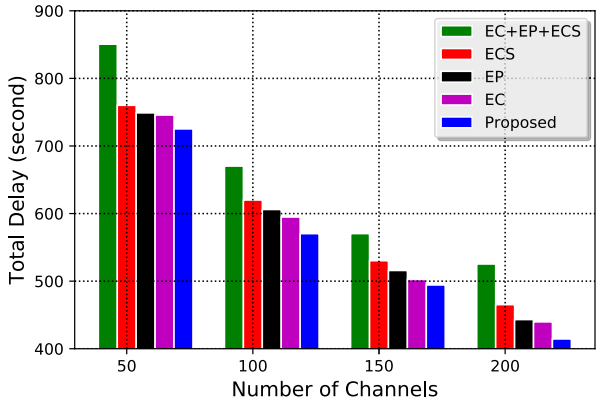

FIGURE 4. Total network delay versus number of channels.

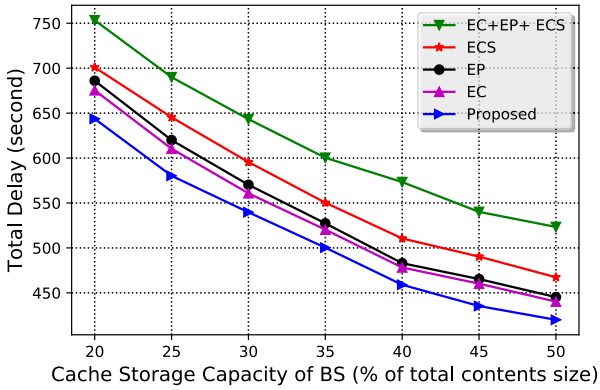

FIGURE 5. Total network delay versus cache storage capacity of BS.

the total delay when the number of available channels is 50 is 850.57 seconds (EC+EP+ECS), 760 seconds (ECS), 748 seconds (EP), 745 seconds (EC), and 726.23 seconds (Proposed). Thus, our proposed algorithm achieves the lowest network delay compared with other baseline schemes. Moreover, we present the delay experienced by the users of all MVNOs for different cache storage capacities of the BSs (i.e., $C_b \in [20\%, 50\%]$) in Fig. 5. Meanwhile, the maximum available bandwidth at each BS is 10 MHz (i.e., the total available channel at each BS is 50). In Fig. 5, we observe that the delay is the lowest under our proposed scheme. Therefore, similar to the previous results, our proposed algorithm outperforms other baseline schemes.

Fig. 6 demonstrates the number of channels assigned to each MVNO's users under our proposed algorithm. From Fig. 6, we observe that the number of channels allocated to the users of MVNO-1 is 70, MVNO-2 is 87, and MVNO-3 is 42, respectively. Here, the number of channels assigned to each MVNO's users depends on the pre-agreement between the InP and MVNO, as characterized in (5). It is clear that the BSs need to allocate more channels to the users of MVNO with the higher QoS requirements. Similar to the scenario of channels allocation, the transmit power allocated to the users of MVNOs is shown in Fig. 7. From Fig. 7, we observe that the transmit power of the BSs allocated to the users of MVNO-1, MVNO-2 and MVNO-3 are 50.34 dBm, 73.14 dBm, and

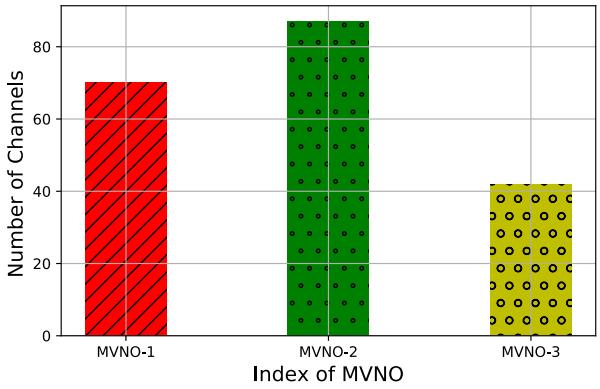

FIGURE 6. Number of channels allocated to each MVNO.

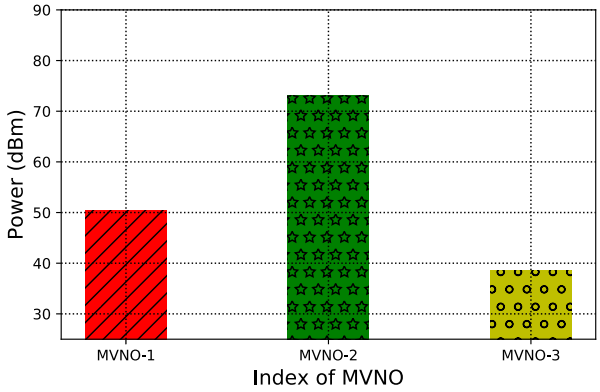

FIGURE 7. Transmit power allocated to each MVNO.

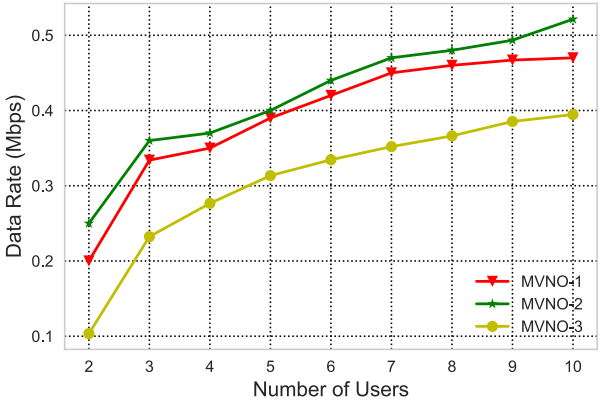

FIGURE 8. Achievable data rate versus the number of users.

38.52 dBm, respectively. Therefore, it is clear that higher transmit power is allocated to the users of MVNO with the higher QoS requirements.

As illustrated in Fig. 8, with the number of users, the achievable data rate of each MVNO increases. It is clearly seen that the achievable data rate of MVNO-2 is the highest among all MVNOs. Statistically, the data rates of MVNO-1, MVNO-2 and MVNO-3 are 4.7 Mbps, 5.2 Mbps and 3.9 Mbps, respectively. The reason is that each BS allocates more channels and power to the users of MVNO-2 following the initial agreement between InP and MVNOs. In Fig. 9, we compare the performance of the proposed algorithm

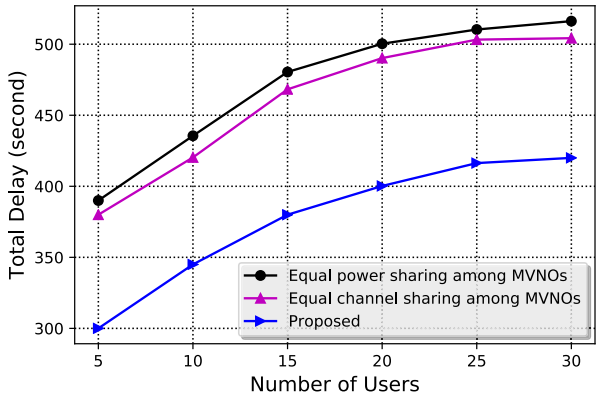

FIGURE 9. Performance comparison of our proposed algorithm with equal channels and power sharing schemes.

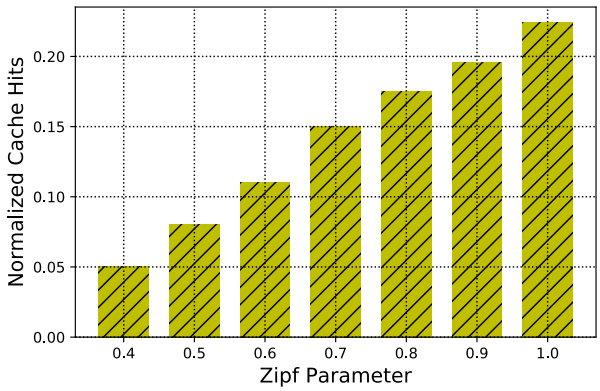

FIGURE 10. Normalized cache hits versus zipf parameter.

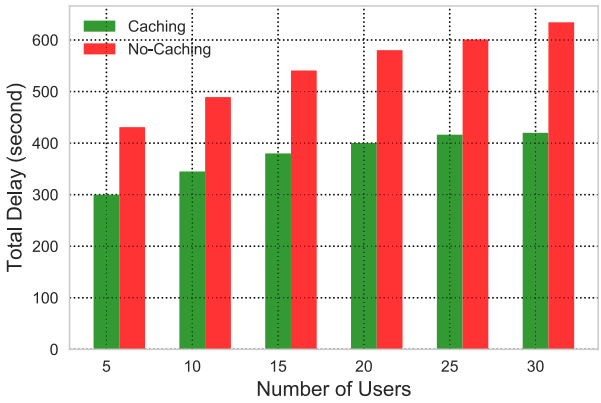

FIGURE 11. Total delay in network versus number of users.

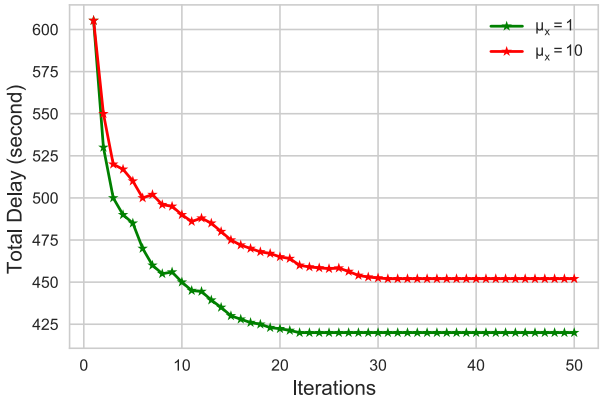

FIGURE 12. Convergence rate of the proposed algorithm with different penalty parameter values.

with the other two schemes: 1) Equal power sharing among MVNOs, and 2) Equal channel sharing among MVNOs. In the equal power sharing among MVNOs, the transmit power the BS is equally shared among MVNOs without considering the pre-agreement (i.e., without considering the QoS requirements of the MVNO), as in (5). It is same as the equal channel sharing scheme, where the channels are equally allocated among MVNOs. In the above aforementioned schemes, the MVNO with the large number of users will be allocated equal proportion of channels and power of the BS compared with the MVNO who has the few number of users. In other words, fairness is not guarantee among MVNOs. To this end, from Fig. 9, we observe that the total delay when the number of users in the network is 30 is 516 seconds (Equal power sharing among MVNOs), 504 seconds (Equal channel sharing among MVNOs), and 420 seconds (Proposed). Therefore, when compared with other existing schemes, our proposed algorithm achieves the lowest delay.

In Fig. 10, we present the normalized cache hits at the BSs based on the content requests made by the users of MVNOs. The requested contents, if not cached at the BS (cache misses), need to be retrieved from the content server located at the core network (CN). We observe that the normalized cache hits increase with the Zipf parameter. This is intuitive as well, as the algorithm caches the contents having high user demand, the cache hit rates will increase correspondingly.

In Fig. 11, we demonstrates the delay experienced by the users of MVNOs under caching (i.e., the popular contents are cached at the storage repository of the BSs to serve users of all MVNOs) and no-caching scenarios. In the no-caching scenario, an InP does not deploy content storage capacity at the BSs. Therefore, the contents requested by the users of MVNOs are downloaded directly from the content server located at the core network. Therefore, the backhaul delay will be present for each user of MVNOs when the contents are downloaded from the content server. As an example, in Fig. 11, the total network delay experienced by the users of all MVNOs is 634.31 seconds (not caching), and 420 seconds (caching) for 30 users in the network. According to the result, we can conclude that deploying the cache storage and caching the popular contents at the BSs can reduce the network delay experienced by the users of MVNOs.

Finally, in Fig. 12, the convergence rate of the proposed algorithm is shown. We also compare the proposed algorithm's convergence rate with different values of penalty parameter (i.e., different μ_x). When $\mu_x = 1$, the algorithm converges within 30 iterations. However, the proposed algorithm converges in fewer than 40 iterations when we set the value of the parameter as 10 (i.e., $\mu_x = 10$). According to the above results, we observe that the value of the penalty parameter can effect on both the convergence rate and the value of the function in (31).

VIII. CONCLUSION

In this paper, we have formulated a problem of joint radio resource allocation and content caching in the heterogeneous virtualized wireless networks. We have shownthat the problem formulated is a non-convex. Therefore, by adding a proximal term, we transformed the proposed problem into the convex form and then solved the transformed problem using a successive block upper-bound minimization algorithm (BSUM). Simulation results have reflected that the total content downloading delay experienced by the users of MVNOs with our solution approach is lesser than that provided by the other baseline schemes. Finally, we introduced the convergence of our proposed algorithm, where it is clear that our proposed algorithm converges within a few iterations. In future, we will consider joint communication and computation resource allocation in the virtualized wireless networks.

REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tutr.*, vol. 17, no. 1, pp. 358–380, 1st Quart., 2015.

[3] A. Haider, R. Potter, and A. Nakao, "Challenges in resource allocation in network virtualization," in *Proc. 20th ITC Specialist Seminar*, vol. 18, May 2009, pp. 1–9.

[4] H. Wen, P. K. Tiwary, and T. Le-Ngoc, "Current trends and perspectives in wireless virtualization," in *Proc. Int. Conf. Sel.Topics Mobile Wireless Netw. (MoWNeT)*, Aug. 2013, pp. 62–67.

[5] 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.

[6] D. H. N. Nguyen, Y. Zhang, and Z. Han, "Contract-based spectrum allocation for wireless virtualized networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7222–7235, Nov. 2018.

[7] R. A. Banez, H. Xu, N. H. Tran, J. B. Song, C. S. Hong, and Z. Han, "Network virtualization resource allocation and economics based on prey-predator food chain model," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4738–4752, Oct. 2018.

[8] Y. K. Tun, N. H. Tran, D. T. Ngo, S. R. Pandey, Z. Han, and C. S. Hong, "Wireless network slicing: Generalized kelly mechanism-based resource allocation," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1794–1807, Aug. 2019.

[9] T. M. Ho, N. H. Tran, L. B. Le, Z. Han, S. M. A. Kazmi, and C. S. Hong, "Network virtualization with energy efficiency optimization for wireless heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 18, no. 10, pp. 2386–2400, Oct. 2019.

[10] L. Li, N. Deng, W. Ren, B. Kou, W. Zhou, and S. Yu, "Multi-service resource allocation in future network with wireless virtualization," *IEEE Access*, vol. 6, pp. 53854–53868, 2018.

[11] *The 1000x Mobile Data Challenge*. Accessed: Nov. 2013. [Online]. Available: https://www.qualcomm.com/invention/5g

[12] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.

[13] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[14] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[15] Z. Han, M. Hong, and D. Wang, *Signal Processing and Networking for Big Data Applications*. Cambridge, U.K.: Cambridge Univ. Press, Apr. 2017.

[16] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo, "A block successive upper bound minimization method of multipliers for linearly constrained convex optimization," 2014, *arXiv:1401.7079*. [Online]. Available: http://arxiv.org/abs/1401.7079

[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.

[18] T. M. Ho, N. H. Tran, S. M. A. Kazmi, Z. Han, and C. S. Hong, "Wireless network virtualization with non-orthogonal multiple access," in *Proc. NOMS - IEEE/IFIP Netw. Operations Manage. Symp.*, Taipei, Taiwan, Apr. 2018, pp. 1–9.

[19] H. Jiang, T. Wang, and S. Wang, "Multi-scale hierarchical resource management for wireless network virtualization," *IEEE Trans. Cognit. Commun.and Netw.*, vol. 4, no. 4, pp. 919–928, Dec. 2018.

[20] K. Teague, M. J. Abdel-Rahman, and A. B. MacKenzie, "Joint base station selection and adaptive slicing in virtualized wireless networks: A stochastic optimization framework," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Honolulu, HI, USA, Feb. 2019, pp. 859–863.

[21] Y. K. Tun, C. W. Zaw, and C. S. Hong, "DownlinK power allocation in virtualized wireless networks," in *Proc. 19th Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Seoul, South Korea, Sep. 2017, pp. 346–349.

[22] S. Parsaeefard, R. Dawadi, M. Derakhshani, and T. Le-Ngoc, "Joint user-association and resource-allocation in virtualized wireless networks," *IEEE Access*, vol. 4, pp. 2738–2750, 2016.

[23] F. Fu and U. C. Kozat, "Stochastic game for wireless network virtualization," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 84–97, Feb. 2013.

[24] M. Yang, Y. Li, D. Jin, J. Yuan, L. Su, and L. Zeng, "Opportunistic spectrum sharing based resource allocation for wireless virtualization," in *Proc. 7th Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput.*, Taichung, Taiwan, Jul. 2013, pp. 51–58.

[25] B. Liu and H. Tian, "A bankruptcy game-based resource allocation approach among virtual mobile operators," *IEEE Commun. Lett.*, vol. 17, no. 7, pp. 1420–1423, Jul. 2013.

[26] B. Fan, H. Tian, and B. Liu, "Game theory based power allocation in LTE air interface virtualization," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, USA, Mar. 2015, pp. 972–976.

[27] F. Fu and U. C. Kozat, "Wireless network virtualization as a sequential auction game," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–9.

[28] M. M. Gomez, S. Chatterjee, M. J. Abdel-Rahman, A. B. MacKenzie, M. B. H. Weiss, and L. DaSilva, "Market-driven stochastic resource allocation framework for wireless network virtualization," *IEEE Syst. J.*, to be published.

[29] N. Raveendran, Y. Gu, C. Jiang, N. H. Tran, M. Pan, L. Song, and Z. Han, "Cyclic three-sided matching game inspired wireless network virtualization," *IEEE Trans. Mobile Comput.*, to be published.

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

[31] Q. Chen, F. R. Yu, T. Huang, R. Xie, J. Liu, and Y. Liu, "Joint resource allocation for software-defined networking, caching, and computing," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 274–287, Feb. 2018.

[32] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5g mobile wireless networks," *IEEE Netw.*, vol. 29, no. 3, pp. 68–74, May 2015.

[33] K. Thar, T. Z. Oo, Y. K. Tun, D. H. Kim, K. T. Kim, and C. S. Hong, "A deep learning model generation framework for virtualized multi-access edge cache management," *IEEE Access*, vol. 7, pp. 62734–62749, 2019.

[34] K. Wang, H. Li, F. Richard Yu, and W. Wei, "Virtual resource allocation in software-defined information-centric cellular networks with Device-to-Device communications and imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10011–10021, Dec. 2016.

[35] X. Li, X. Wang, C. Zhu, W. Cai, and V. C. M. Leung, "Caching-as-a-service: Virtual caching framework in the cloud-based mobile networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2015, pp. 372–377.

[36] 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.

[37] E. Bastug, M. Bennis, E. Zeydan, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data meets telcos: A proactive caching perspective," *J. Commun. Netw.*, vol. 17, no. 6, pp. 549–557, Dec. 2015.

[38] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM . Conf. Comput. Commun. 18th Annu. Joint Conf. IEEE Comput. Commun. Soc.*, New York, NY, USA, Mar. 1999, pp. 126–134.

[39] D. C. Chen, T. Q. S. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3194–3206, Jun. 2015.
[40] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.
[41] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, Jan. 2013.
[42] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo, "Iteration complexity analysis of block coordinate descent methods," *Math. Program.*, vol. 163, nos. 1–2, pp. 85–114, Aug. 2016.
[43] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, Jun. 2001.

YAN KYAW TUN received the B.E. degree in marine electrical systems and electronics engineering from Myanmar Maritime University, Thanlyin, Myanmar, in 2014. He is currently pursuing the Ph.D. degree in computer science and engineering with Kyung Hee University, South Korea, for which he received a scholarship, in 2015. His research interests include network economics, game theory, network optimization, wireless communication, wireless network virtualization, mobile edge computing, and wireless resource slicing for 5G.

ANSELME NDIKUMANA received the B.S. degree in computer science from the National University of Rwanda, in 2007, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in August 2019. Since September 2019, he has been with the Faculty of Computing and Information Sciences, University of Lay Adventists of Kigali, Rwanda, where he is currently a Lecturer, and also with the Department of Computer Science and Engineering, Kyung Hee University, Yongin, South Korea. His professional experience includes a Chief of information systems, a System Analyst, and a Database Administrator with Rwanda Utilities Regulatory Authority, from 2008 to 2014. His research interests include deep learning, multiaccess edge computing, information centric networking, and in-network caching.

SHASHI RAJ PANDEY received the B.E. degree in electrical and electronics with specialization in communication from Kathmandu University, Nepal, in 2013. He is currently pursuing the Ph.D. degree in computer science and engineering with Kyung Hee University, South Korea. After graduation, he served as a Network Engineer with Huawei Technologies Nepal Co., Pvt., Ltd, Nepal, from 2013 to 2016. His research interests include network economics, game theory, wireless communications and networking, edge computing, and machine learning.

ZHU HAN (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was a Research and Development Engineer with JDSU, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate with the University of Maryland. From 2006 to 2008, he was an Assistant Professor with Boise State University, ID, USA. He is currently a John and Rebecca Moores Professor with the Department of Electrical and Computer Engineering as well as the Department of Computer Science, University of Houston, TX, USA. He is also with the Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, South Korea. He is also a Chair Professor with National Chiao Tung University, China. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grids. He has received the NSF Career Award, in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society, in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing*, in 2015, the IEEE Leonard G. Abraham Prize in the field of communications systems (best paper award for the IEEE JSAC), in 2016, and several best paper awards in IEEE conferences. He was an IEEE Communications Society Distinguished Lecturer, from 2015 to 2018, and he has been an AAAS Fellow, since 2019, and an ACM Distinguished Member, since 2019. He has been a 1% highly-cited researcher according to the Web of Science, since 2017.

CHOONG SEON HONG (Senior Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Kyung Hee University, Seoul, South Korea, in 1983 and 1985, respectively, and the Ph.D. degree from Keio University, in 1997. In 1988, he joined KT, where he was involved in broadband networks, as a Member of Technical Staff. In 1993, he joined Keio University, Japan. He was with the Telecommunications Network Laboratory, KT, as a Senior Member of Technical Staff and the Director of the Networking Research Team, until 1999. Since 1999, he has been a Professor with the Department of Computer Engineering, Kyung Hee University. His research interests include future internet, ad hoc networks, network management, and network security. He is a member of the ACM, the IEICE, the IPSJ, the KIISE, the KICS, the KIPS, and the OSIA. He has served as the General Chair, the TPC Chair/Member, or an Organizing Committee Member of international conferences such as NOMS, IM, APNOMS, E2EMON, CCNC, ADSN, ICPP, DIM, WISA, BcN, TINA, SAINT, and ICOIN. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, the *International Journal of Network Management*, and the IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS, and an Associate Technical Editor of the *IEEE Communications Magazine*.