IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 12 March 2023, accepted 28 March 2023, date of publication 3 April 2023, date of current version 6 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3263951

RESEARCH ARTICLE

Multi-Objective Energy Efficient Resource Allocation in Massive Multiple Input Multiple Output-Aided Heterogeneous Cloud Radio Access Networks

NAHID AMANI[1], SAEEDEH PARSAEEFARD[2], (Senior Member, IEEE), AND HALIM YANIKOMEROGLU[3], (Fellow, IEEE)

[1]Department of Communication Technology, ICT Research Institute, Tehran 1439955471, Iran
[2]Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S, Canada
[3]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

Corresponding author: Nahid Amani (n_amani@itrc.ac.ir)

The work of Halim Yanikomeroglu was supported by an NSERC (Natural Sciences and Engineering Research Council of Canada) Discovery Grant.

ABSTRACT In this work, a novel energy efficient multi-objective resource allocation algorithm for heterogeneous cloud radio access networks (H-CRANs) is proposed where the trade-off between increasing throughput and decreasing operation cost is considered. H-CRANs serve groups of users through femto-cell access points (FAPs) and remote radio heads (RRHs) equipped with massive multiple input multiple output (MIMO) connected to the base-band unit (BBU) pool via front-haul links with limited capacity. We formulate an energy-efficient multi-objective optimization (MOO) problem with a novel utility function. Our proposed utility function simultaneously improves two conflicting goals as total system throughput and operation cost. With this MOO, we jointly assign the sub-carrier, transmit power, access point (AP)(RRH/FAP), RRH, front-haul link, and BBU. To address the conflicting objectives, we convert the MOO problem into a single-object optimization problem using an elastic-constraint scalarization method. With this approach, we flexibly adjust trade-off parameters to choose between two objective functions. To propose an efficient algorithm, we deploy successive convex approximation (SCA) and complementary geometric programming (CGP) approaches. Finally, via simulation results we discuss how to select the values of trade-off parameters, and we study their effects on conflicting objective functions (i.e., throughput and operation cost in MOO problem). Simulation results also show that our proposed approach can offload traffic from C-RANs to FAPs with low transmit power and thereby reduce operation costs by switching off the under-utilized RRHs and BBUs. It can be observed from the simulation results that the proposed approach outperforms the traditional approach in which each user is associated to the AP (RRHs/FAPs) with the largest average value of signal strength. The proposed approach reduces operation costs by 30% and increases throughput index by 25% which in turn leads to greater energy efficiency (EE).

INDEX TERMS 5G, multi-objective optimization problem, elastic-constraints method.

I. INTRODUCTION

A. MOTIVATIONS

To meet the demand of expanding services in fifth-generation (5G) wireless networks (e.g., services from vertical industries, enterprises, and Internet companies), networks with high data rates and enhanced quality-of-service (QOS) are increasingly expected. Indeed, with the proliferation of massive number of online heterogeneous devices, such as tablets, sensors for home security, and wearable health monitors, the energy consumption of future networks could result in serious restrictions, which must be considered [1]. This tremendous increase in mobile data traffic with high data rates raises new challenges in terms of increasing energy efficiency (EE) and providing high-quality services for various traffic types in 5G networks [1], [2], [3]. Cloud radio access networks

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

(C-RANs), massive multiple input multiple output (MIMO) and heterogeneous networks (HetNets) are three key technologies suggested for 5G that can significantly enhance EE [1], [2], [3].

Massive MIMO systems provide opportunities for increasing spectral efficiency (SE) and simultaneously improving EE [4], [5], [6]. In addition, the deployment of small-cell variations, such as low-power femto access points (FAPs), in HetNets is a promising approach for handling massive heterogeneous traffic. FAPs also enhance SE and EE in 5G. By applying traffic offloading, FAPs can provide data distribution to release the pressure of macro access points and improve the QoS of users, particularly for cell-edge users, which may lead to switching off of under-utilized high-power macro access points [7], [8].

Moreover, C-RANs are a promising architecture for significantly enhancing EE in 5G. enhance EE [9], [10], [11]. In C-RANs, by separating the remote radio head (RRH) from the baseband units (BBUs), all baseband processing functions are carried out through a centralized cloud called a BBU pool. BBUs are connected to the RRHs by limited capacity front-haul links. With their fully centralized processing and management design, C-RANs are able to achieve cooperative gains, such as interference management and load balancing. Thus by switching off RRHs and BBUs, EE can be improved [11], [12].

To leverage these three technologies, heterogeneous cloud radio access networks (H-CRANs) have been introduced, which include a large number of RRHs equipped with massive MIMO in order to enhance EE [13]. Combining them offers a large set of parameters and flexibility in system design, but at the expense of highly complex multi-objective resource allocation (MORA) problems [14], [15]. In high density H-CRANs, handling all users by high throughput and low energy consumption cost is important. By deploying a large number of RRHs in a cell, SE and the throughput of the whole network will be improved, which may lead to under-utilized RRHs and BBUs, and consequently increased energy consumption costs and decreased EE [16]. Therefore, to reduce energy consumption costs, simultaneously maximizing the throughput and minimizing the number of active RRHs and BBUs are considered two conflicting objective functions for improving EE [17]. Hence, 5G should follow other frameworks, such as a MORA, where conflict objectives can be integrated in a multi-objective optimization (MOO) problem [14], [18].

To address these challenges, the main contributions of this paper are summarized as follows:

1) We formulate a centralized EE MORA optimization problem in MIMO-aided H-CRAN. In so doing, we propose a novel utility function and a new set of constraints that aim to maximize the total system throughput while minimizing operation costs in order to jointly assign cloud assignment parameters (CAPs), including sub-carrier, access points (APs)(RRH/FAP) assignment, RRH and front-haul link to active BBUs, and transmit power allocation to each user. Our proposed resource allocation problem in this paper highlights the new aspects of H-CRAN with a combination of three technologies (e.g., C-RANs, MIMO and HetNets), a new utility function, and a novel set of assigned resource parameters.
2) We discuss how to select the values of trade-off parameters, and we study their effects on conflicting objective functions (i.e., throughput and operation cost in MOO problem). The simulation results show that by adjusting the value of trade-off parameters and by considering energy consumption cost of RRHs, BBUs and total transmit power of all users as operation costs in utility function, the traffic of users associated with under-utilized RRHs can be offloaded to neighboring FAPs with low transmit power. Consequently, under-utilized RRHs and BBUs and their corresponding front-haul links can be switched off for greater EE. Simultaneous reduction of these three costs is the novelty of the MORA algorithm in this work.
3) The conflicting objectives and highly complex relation between various optimization variables and their effects on each other make the formulated MOO problem much more difficult to solve. To tackle this issue, we apply an elastic-constraint scalarization method [18], [19] to convert the MOO problem into a single-objective optimization (SOO) problem with low computational complexity, which allows for trade-off parameters and a flexible choice between increasing throughput and decreasing operation cost functions for different preferences.
4) Due to interference among users from different APs and the existence of binary variables, the proposed optimization problem is inherently non-convex, NP-hard, and suffers from high computational complexity. By applying successive convex approximation (SCA) and complementary geometric programming (CGP) techniques to various transformations and convexification approaches [20], [21], [22], such as arithmetic-geometric mean approximation (AGMA), we convert the SOO problem into a convex problem, which can be solved with available software (e.g., CVX) [23].
5) The simulation results showed that our proposed approach can offload traffic from C-RANs to FAPs with low transmit power, and reduce operation costs by switching off under-utilized RRHs and BBUs. Also, the simulation results illustrate that Pareto optimal solutions are different under diverse sets of system parameters.

B. RELATED WORKS

Many systems or applications have been developed for distributed environments with the goal of attaining multiple objectives in the face of environmental challenges such as

TABLE 1. Summary of related work.

Ref.	System model	Objective function	Optimization variables	Solution
[8]	HetNets	Max EE	Sub-carrier and transmit power	Markov approximation and noncooperative gam/near-optimal
[13]	H-CRAN	Max EE	Sub-carrier and transmit power	Lagrange dual decomposition/global optimal
[15]	OFDMA networks	Max EE MOO	sub-carrier and power	Sum weighted method/local optimal
[16]	C-RAN	Max EE	AP assignment	Heuristic algorithm/local optimal
[19]	Permutation flowshop scheduling	Min energy consumption and flowtime MOO	Processing time and energy consumption of jobs	Augmented epsilon-constraint/pareto optimal
[26]	OFDMA networks	Max EE	AP, transmit power	Heuristic algorithms/local optimal
[27]	OFDMA networks	max-min EE	transmit power	Heuristic algorithms/local optimal
[28]	C-RAN	Max EE	Sub-carrier and transmit power	Dinkelbach method/local optimal
[29]	MIMO-enabled HetNet	Max EE and SE MOO	AP assignment and transmit power	weighted sum method/local optimal
[30]	MIMO-enabled HetNet	Max EE MOO	AP assignment,sub-carrier and transmit power	weighted Tchebycheff method/Pareto optimal
[31]	C-RAN	Max sum rate and min sum power MOO	RRH selection, RRH user association and transmit beamforming	branch-and-reduce-and-bound/global optimal
[32]	MISO-NOMA	Max EE MOO	Sub-carrier and transmit beamforming	SCA and weighted sum method/Pareto optimal
[33]	mmWave NOMA	Max EE MOO	transmit power, active and passive precoding	weighted sum method/suboptimal
This work	MIMO-aided H-CRAN	Max EE MOO with novel utility function	AP assignment, Sub-carrier and transmit power allocation and RRH, fronthaul allocation to active BBUs	CGP, SCA and elastic-constraint scalarization method/Pareto optimal

high dynamics/hostility, or severe resource constraints (e.g., energy or communications bandwidth). Often the multiple objectives are conflicting with each other, requiring optimal tradeoff analyses between the objectives [18]. In this context, MOO problems are considered when balancing the trade-off among two or more objectives [24], [25]. There are common solution approaches for MOO problems, such as weighting method, goal programming, and elastic-constraint scalarization method [19].

This work focuses on the intersection of two main areas in H-CRAN resource allocation problems: EE and MOO. In the area of EE, there has been a surge of research (e.g., in [4], [8], [13], [16], [26], [27], and [28]). In these works, the EE MOO problem is not considered. For example, in order to maximize EE in HetNets, the authors in [8] and [28] formulated a joint power and sub-carrier assignment as a noncooperative game. Reference [13] used a Lagrange dual decomposition method to minimize the power consumption of users while allocating access points (APs) and power to each user in a H-CRAN.

Reference [15] formulated the MOO resource allocation problem for maximizing the achievable rate/SE and minimizing the total power consumption by using the sum weighted method. The authors applied the generalized framework of the resource allocation for the EE-SE trade-off to optimally allocate the subcarriers' power for OFDMA with imperfect channel estimation. Reference [19] studied two contradictory objectives, namely, total flowtime and total energy consumption (TEC) in a green permutation flowshop environment. To address the conflicting objectives of minimizing TEC and total flowtime, the augmented epsilon-constraint approach was employed to obtain Pareto-optimal solutions. In [29] and [30], the authors formulated a joint dynamic radio resource allocation MOO problem for massive MIMO-enabled HetNets with the aim of maximizing EE and SE simultaneously. The studies in [29] and [30] employed the weighted sum method and weighted Tchebycheff method, respectively, to transform the MOO problem into an SOO problem. In [31], the researchers formulated a joint design for RRH selection, RRH user association, and transmit beamforming, where a branch-and-reduce-and-bound algorithm was applied to simultaneously optimize the achievable sum rate and total power consumption using the MOO concept in C-RANs. The authors in [17] formulated an EE MOO problem for uplink multi-cell networks using a joint design for sub-channel assignment, power control, and antenna selection, where the weighted Tchebycheff method was deployed. Reference [32] formulated a joint SE-EE based design as a MOO problem to achieve a good trade-off between SE and EE. This work exploited a priori articulation scheme combined with the weighted sum approach to transform the original MOO problem as a conventional SOO problem and used SCA technique to solve the non-convex

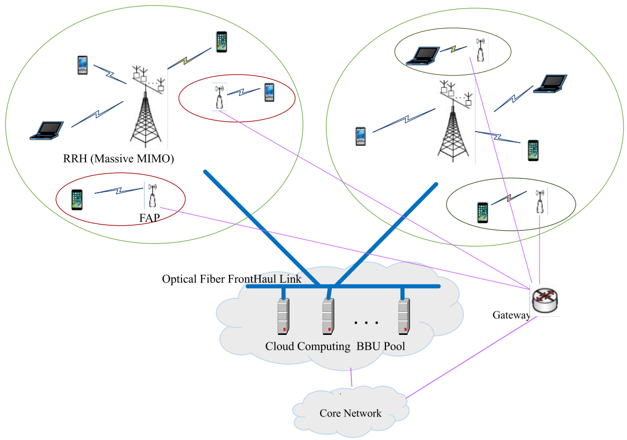

FIGURE 1. C-RAN architecture with cloud computing BBU pool and massive MIMO RRHs.

SOO problem. In [33] the EE-SE tradeoff problem modeled as a MOO through jointly optimizing power allocation, active precoding at the BS, and passive precoding at the in intelligent reflecting surface (IRS) millimeter wave (mmWave) non-orthogonal multiple access (NOMA) systems. Then, the authors used the weighted-sum method to transform the MOO problem into a SOO problem. The above mentioned works are summarized in Table 1. However, to the best of our knowledge, no other works have considered the EE MORA problem in H-CRANs. Moreover, none of the aforementioned works has considered user association jointly with dynamic radio resource allocation and C-RAN limitations (e.g., maximum BBU capacity and front-haul capacity) with a view to reducing RRHs and BBUs energy consumption costs, particularly by switching off under-utilized RRHs and BBUs in MIMO-aided H-CRAN. This work aims to fill this gap.

In what follows, Section II describes the system model and formulation of the MORA problem. Section III introduces the proposed two-step iterative MORA algorithm, followed by an analysis of its computational complexity. Section IV presents the simulation results. Section V concludes this paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a down-link transmission in a two-tier orthogonal frequency division multiple access (OFDMA) based H-CRAN serving a set of $\mathcal{N} = \{1, \ldots, N\}$ single-antenna users. H-CRAN covers all of the users by a set of single-antenna $\mathcal{F} = \{1, \ldots, F\}$ FAPs and a set of $\mathcal{R} = \{1, \ldots, R\}$ RRHs, as shown in Fig. 1. As see in Table 2, $\mathcal{M} = \mathcal{R} \bigcup \mathcal{F}$ denote the set of all APs in this region. Each RRH $r \in \mathcal{R}$ is equipped with $J_r \gg 1$ antennas (i.e., massive MIMO) and is connected to the BBU pool by a limited capacity front-haul link. FAPs are connected to the core network through back-haul links. A BBU pool consists of a set of $\mathcal{B} = \{1, \ldots, B\}$ BBUs to process the received baseband signals from all RRHs. The total bandwidth W is divided into a set of sub-carriers $\mathcal{S} =$

TABLE 2. Table of notations.

Variable	Description
$\beta_{m,n}$	AP m is assigned to n^{th} user
$\alpha^s_{m,n}$	Ap m allocates sub-carrier s to n^{th} user
$h^s_{m,n}$	Channel gain of user n to AP m on sub-carrier s
$p^s_{m,n}$	Transmit power of user n to AP m on sub-carrier s
$z_{r,b} \in \{0,1\}$	If $z_{r,b} = 1$, RRH r is associated to BBU b
$x_b \in \{0,1\}$	If $x_b = 1$, BBU b is in state on
$\tau_r \in \{0,1\}$	If $\tau_r = 1$, RRH r is in state on
Parameter	**Description**
J_r	Number of antennas of RRH r
N_m	Total number of users assigned to AP m
σ^2	Noise power
$I^s_{m,n}$	Interference to user n in AP m on sub-carrier s
I^{th}_s	Maximum interference for sub-carrier s
$R^s_{m,n}$	Throughput of user n to AP m on sub-carrier s
R^{rsv}_n	Minimum reserved rate of user n
$\boldsymbol{\tau}$	Matrix of τ_r
$T^{\max}_b$	Maximum load handled by each BBU $b \in \mathcal{B}$
$F^{\max}_{r,b}$	Front-haul capacity limitation
$p^{\max}_m$	Transmit power limitation of each AP m
μ_a	Energy consumption cost of each antenna
μ_b	Energy consumption cost of each BBU
μ_p	Cost of the total transmit power
Set	**Description**
$\mathcal{F}$	Set of femto access points (FAPs)
$\mathcal{R}$	Set of RRHs
$\mathcal{B}$	Set of BBUs
$\mathcal{M} = \mathcal{R} \bigcup \mathcal{F}$	Set of all APs
$\mathcal{N}$	Set of users
$\mathcal{S}$	Set of sub-carriers
$\mathbf{P}$	Matrix of transmit powers
$\boldsymbol{\beta}$	Matrix of $\beta_{m,n}$
$\boldsymbol{\alpha}$	Matrix of $\alpha^s_{m,n}$
$\mathbf{Z}$	Matrix of $z_{r,b}$
$\mathbf{x}$	Vector of all BBUs
Function	**Description**
$U_1(\boldsymbol{\tau})$	Energy consumption cost of RRHs
$U_2(\mathbf{X})$	Energy consumption cost of BBUs
$U_3(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P})$	Total transmit power of all users
$U_4(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P})$	Total throughput
U	Network utility function

$\{1, \ldots, S\}$. We use $\beta_{m,n}$ as an AP association indicator for user $n \in \mathcal{N}$ of AP $m \in \mathcal{M}$ where

$$\beta_{m,n} = \begin{cases} 1, & \text{if AP } m \text{ is assigned to } n^{th} \text{ user,} \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, the total number of users associated to the AP m is $N_m = \sum_{n \in \mathcal{N}} \beta_{m,n}, \quad \forall m \in \mathcal{M}$. Also, we define a binary variable $\alpha^s_{m,n}$ as the sub-carrier allocation indicator, where

$$\alpha^s_{m,n} = \begin{cases} 1, & \text{if AP } m \text{ allocates sub-carrier } s \text{ to the } n^{th} \text{ user,} \\ 0, & \text{otherwise.} \end{cases}$$

Let $p^s_{m,n}$ and $h^s_{m,n}$ represent the transmit power and channel gain of user $n \in \mathcal{N}$ to AP $m \in \mathcal{M}$ on sub-carrier $s \in \mathcal{S}$, respectively. We consider the number of simultaneously served users by a RRH r, be smaller than the number of transmit antennas as $J_r \gg N_r$. The achievable throughput of user $n \in \mathcal{N}$ over sub-carrier $s \in \mathcal{S}$

in AP $m \in \mathcal{M}$ is as [30], [34]

$$R^s_{m,n}(\mathbf{P}, \boldsymbol{\beta}) = \begin{cases} \log_2(1 + (\frac{J_m - N_m + 1}{N_m} \frac{p^s_{m,n} h^s_{m,n}}{\sigma^2 + I^s_{m,n}})), & \text{if } m \in \mathcal{R}, \\ \log_2(1 + \frac{p^s_{m,n} h^s_{m,n}}{\sigma^2 + I^s_{m,n}}), & \text{if } m \in \mathcal{F}, \end{cases} \tag{1}$$

where $I^s_{m,n} = \sum_{\forall m' \in \mathcal{M}, m' \neq m} \sum_{\forall n' \neq n} p^s_{m',n'} h^s_{m,n'}$ is the interference to user $n \in \mathcal{N}$ in AP $m \in \mathcal{M}$ and sub-carrier $s \in \mathcal{S}$, and σ^2 is the noise power. Moreover, $\mathbf{P}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ are matrices of all $p^s_{m,n}$, $\alpha^s_{m,n}$ and $\beta_{m,n}$, respectively, for all $n \in \mathcal{N}, \forall m \in \mathcal{M}$ and $s \in \mathcal{S}$. Due to the dense deployment of RRHs and the capability of turning them on or off, we consider a vector of all RRHs as $\boldsymbol{\tau} = [\tau_r]_{1 \times R}$, where τ_r denotes the on and off states of RRH $r \in \mathcal{R}$ as

$$\tau_r = \begin{cases} 1, & \text{if RRH } r \text{ is in state on,} \\ 0, & \text{otherwise.} \end{cases}$$

The RRHs and BBUs are connected by a limited capacity front-haul link. Therefore, the binary variable $z_{r,b}$ is introduced to assign RRH r to BBU b, and stipulates front-haul link between them as

$$z_{r,b} = \begin{cases} 1, & \text{If the RRH } r \text{ is associated to the BBU } b, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, $\mathbf{Z} = [z_{r,b}]_{R \times B}$ is defined as a matrix of associated RRHs to BBUs. The binary variable matrix $\mathbf{X} = [x_b]_{1 \times B}$ is introduced to describe the on and off states of BBU $b \in \mathcal{B}$ as

$$x_b = \begin{cases} 1, & \text{if BBU } b \text{ is in state On,} \\ 0, & \text{otherwise.} \end{cases}$$

We consider $\mathfrak{S} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P}, \boldsymbol{\tau}, \mathbf{Z}, \mathbf{X}\}$, and with the aim of decreasing network operation cost, we define a novel network utility function as

$$U(\mathfrak{S}) = \underbrace{\sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} \alpha^s_{m,n} \beta_{m,n} R^s_{m,n}(\mathbf{P}, \boldsymbol{\beta})}_{U_4(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{P})} - [\mu_a \underbrace{\sum_{r \in \mathcal{R}} \tau_r J_r}_{U_1(\boldsymbol{\tau})} + \underbrace{\sum_{b \in \mathcal{B}} \mu_b \times x_b}_{U_2(\mathbf{X})} + \mu_p \underbrace{\sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} \alpha^s_{m,n} \beta_{m,n} p^s_{m,n}}_{U_3(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{P})}], \tag{2}$$

which is the total throughput ($U_4(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P})$) minus the total operation cost function. We define operation cost function as the sum of the energy consumption cost of active RRHs, active BBUs, and total transmit power of all users, which are denoted by $U_1(\boldsymbol{\tau})$, $U_2(\mathbf{X})$ and $U_3(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{P})$, respectively. Simultaneous reduction of these three costs is the novelty of the MORA algorithm in this work. In (2), we consider the energy consumption cost of RRHs; in other words, $U_1(\mathbf{X})$ is proportional to the total rate transmitted rate by the number of allocated antennas to active RRHs. Hence, μ_a is defined as the energy consumption cost of each antenna, which is static and proportional to the maximum transmission rate of each antenna, and its unit is bps/Hz. Furthermore, μ_b is defined as the energy consumption cost of each BBU $b \in \mathcal{B}$, which is static and proportional to its maximum load capacity [35]. Hence, the unit of $U_2(\mathbf{X})$ will be bps/Hz. In (2), due to the fact that different objective functions have different units (e.g. bps/Hz for throughput and Watt for transmit power), and the values of objective functions are not in the same range [8], μ_p is considered to be a dimension regulation factor and a cost of the total transmit power of all users, and accordingly its unit is bps/Hz/Watt. This means that the unit of utility function (2) is bps/Hz, and, based on (2), the MORA problem to maximize EE can be written as

$$\begin{aligned}
&\max_{\mathfrak{S}} U(\mathfrak{S}),\\
&\text{subject to}: \text{C1}: \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} p^s_{m,n} \leq p^{\max}_m, \ \forall m \in \mathcal{M},\\
&\text{C2}: \sum_{\forall m \in \mathcal{M}} \sum_{s \in \mathcal{S}} \alpha^s_{m,n} \beta_{m,n} R^s_{m,n}(\mathbf{P}, \boldsymbol{\beta}) \geq R^{\text{rsv}}_n, \ \forall n \in \mathcal{N},\\
&\text{C3}: \sum_{m \in M} \beta_{m,n} \leq 1, \ \forall n \in \mathcal{N},\\
&\text{C4}: \sum_{\forall n \in \mathcal{N}} \alpha^s_{m,n} \leq 1, \ \forall m \in \mathcal{M}, \forall s \in \mathcal{S},\\
&\text{C5}: \alpha^s_{m,n} \leq \beta_{m,n}, \ \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \forall s \in \mathcal{S},\\
&\text{C6}: \sum_{b \in \mathcal{B}} z_{r,b} \leq 1, \ \forall r \in \mathcal{R},\\
&\text{C7}: \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} z_{r,b} \alpha^s_{r,n} \beta_{r,n} R^s_{r,n}(\mathbf{P}, \boldsymbol{\beta}) \leq F^{\max}_{r,b}, \ \forall r \in \mathcal{R},\\
&\text{C8}: \sum_{r \in \mathcal{R}} \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} z_{r,b} \alpha^s_{r,n} \beta_{r,n} R^s_{r,n}(\mathbf{P}, \boldsymbol{\beta}) \leq T^{\max}_b \times x_b,\\
&\text{C9}: z_{r,b} - x_b \leq 0, \ \forall r \in \mathcal{R}, \forall b \in \mathcal{B},\\
&\text{C10}: \tau_r - \sum_{b \in \mathcal{B}} z_{r,b} \leq 0, \ \forall r \in \mathcal{R},\\
&\text{C11}: \sum_{n \in \mathcal{N}} \beta_{r,n} \leq \tau_r \times \varrho, \ \forall r \in \mathcal{R},
\end{aligned} \tag{3}$$

$\alpha^s_{m,n} \in \{0, 1\}, \beta_{m,n} \in \{0, 1\}, \tau_r \in \{0, 1\}, z_{r,b} \in \{0, 1\}, x_{r,b} \in \{0, 1\}, \ \forall m, n, s, r, b.$

In (3), C1 denotes the transmit power of each AP $m \in \mathcal{M}$ limited by $p^{\max}_m$. The required minimum reserved rate (i.e., R^{rsv}_n) of each user can be indicated by C2. Due to OFDMA constraints, C3 and C4 stipulate that each user $n \in \mathcal{N}$ can only be served by at most one AP, and each sub-carrier must be associated to at most one user within each AP, respectively. C5 stipulates that AP m can allocate sub-carrier s to user n when Ap m is assigned to user n. C6 stipulates that each RRH $r \in \mathcal{R}$ can be assigned to at most one BBU. Based on C7 and C8, the total allocated load to the front-haul link $z_{r,b}$ and the

TABLE 3. Two-step iterative MORA algorithm.

Input: Set $t = t_1 = 0$, $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 = 10^{-3}$,
$\mathbf{P}(t = 0) = P^{\max}/S$.
Repeat: set $t = t + 1$
Step 1: Computes the CAP parametrs, i.e., $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}, \mathbf{X}$ in order to jointly assign sub-carrier, access point to each user and RRH, front-haul link to active BBUs.
Input of Step 1: set $t_1 = 0, \omega(t_1) = \omega(t), \mathbf{P}(t_1) = \mathbf{P}(t)$.
Repeat: set $t_1 = t_1 + 1$.
Step 1.1: Update CGP variables according to (15)-(34), and solve (9) for $\boldsymbol{\alpha}(t_1), \boldsymbol{\beta}(t_1), \boldsymbol{\tau}(t_1), \mathbf{Z}(t_1), \mathbf{X}(t_1)$ using CVX,
Until $\| \boldsymbol{\beta}^*(t_1) - \boldsymbol{\beta}^*(t_1 - 1) \| \leq \kappa_1, \| \mathbf{Z}^*(t_1) - \mathbf{Z}^*(t_1 - 1) \| \leq \kappa_2$ and $\| \mathbf{X}^*(t_1) - \mathbf{X}^*(t_1 - 1) \| \leq \kappa_3$.
Output: $\boldsymbol{\alpha}^*(t_1), \boldsymbol{\beta}^*(t_1), \boldsymbol{\tau}^*(t_1), \mathbf{Z}^*(t_1)$, and $\mathbf{X}^*(t_1)$.
set $\boldsymbol{\alpha}^*(t) = \boldsymbol{\alpha}^*(t_1), \boldsymbol{\beta}^*(t) = \boldsymbol{\beta}^*(t_1), \boldsymbol{\tau}^*(t) = \boldsymbol{\tau}^*(t_1)$,
$\mathbf{Z}^*(t) = \mathbf{Z}^*(t_1)$, and $\mathbf{X}^*(t) = \mathbf{X}^*(t_1)$
Step 2: Transmit Power Allocation:
Input of Step 2: set $t_2 = 0$, $\boldsymbol{\alpha}^*(t_2) = \boldsymbol{\alpha}^*(t)$,
$\boldsymbol{\beta}^*(t_2) = \boldsymbol{\beta}^*(t), \boldsymbol{\tau}^*(t_2) = \boldsymbol{\tau}^*(t), \mathbf{Z}^*(t_2) = \mathbf{Z}^*(t)$,
and $\mathbf{X}^*(t_2) = \mathbf{X}^*(t)$.
Repeat: set $t_2 = t_2 + 1$.
Step 2.1: Update CGP variables according to (38)-(46)
Step 2.2: Solve (37) for $\mathbf{P}(t_2)$ using CVX ,
Until $\| \mathbf{P}^*(t_2) - \mathbf{P}^*(t_2 - 1) \| \leq \kappa_4$.
Output: $\mathbf{P}^*(t_2)$.
set $\mathbf{P}^*(t) = \mathbf{P}^*(t_2)$.
Until $\| \boldsymbol{\beta}^*(t) - \boldsymbol{\beta}^*(t-1) \| \leq \kappa_1, \| \mathbf{Z}^*(t) - \mathbf{Z}^*(t-1) \| \leq \kappa_2, \| \mathbf{X}^*(t) - \mathbf{X}^*(t-1) \| \leq \kappa_3$ and $\| \mathbf{P}^*(t) - \mathbf{P}^*(t-1) \| \leq \kappa_4$.
Output: $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\tau}^*, \mathbf{Z}^*, \mathbf{X}^*$, and $\mathbf{P}^*(t)$

BBU b is bounded by $F_{r,b}^{\max}$ and $T_b^{\max}$, respectively. A front-haul link can be activated between RRH $r \in \mathcal{R}$ and BBU $b \in \mathcal{B}$ when BBU b is switched on. Consequently, RRH $r \in \mathcal{R}$ can be activated when at least one corresponding front-haul link is switched on (i.e., the variable τ_r should be equal to one). To mathematically represent these two practical considerations, we have C9 and C10. Finally, C11 stipulates that each user n can be associated to RRH r when RRH r is active and $\varrho \gg 1$ is a constant value.

(3) represents a MOO problem because it consists of conflicting objectives U_1, U_2, U_3, and U_4. For example, with increasing $\mathbf{P}$, U_3, U_4 will increase while U_3 decreases (3). Hence, there is a trade-off between the increasing total throughput and the decreasing total operation cost. Therefore, finding optimal solutions for (3) is an over-constrained problem [18]. To tackle this computational complexity, we apply the elastic-constraints method [18] (See Appendix I), which allows us to generate a single objective function by selecting one of the multiple objective functions (i.e., U_4) as the primary objective function and to consider the remaining objective functions as constraints. To apply the elastic-constraints method, we rewrite (3) as

$$\begin{aligned} &\min_{\mathfrak{S}} [-U_4 + \sum_{i=1}^{3} U_i], \\ &\text{subject to : C1} - \text{C11}. \end{aligned} \tag{4}$$

Then, based on the elastic-constraints method, we consider the total throughput as the primary objective function (i.e., U_4) and objectives U_1, U_2 and U_3 as new constraints. Therefore, (4) can be reformulated as an SOO problem accordingly;

$$\begin{aligned} &\min_{\mathfrak{S}, L_i, s_i'} [-\sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} \alpha_{m,n}^s \beta_{m,n} R_{m,n}^s(\mathbf{P}, \boldsymbol{\beta}) + \sum_{i=1}^{3} \pi_i s_i'], \\ &\text{subject to : C1} - \text{C11}, s_i', L_i \geq 0, \\ &\qquad \text{C12} : \mu_{\text{a}} \sum_{r \in R} \tau_r J_r + L_1 - s_1' = \varepsilon_1, \\ &\qquad \text{C13} : \sum_{b \in \mathcal{B}} \mu_b \times x_b + L_2 - s_2' = \varepsilon_2, \\ &\qquad \text{C14} : \mu_{\text{p}} \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{s \in \mathcal{S}} \alpha_{m,n}^s \beta_{m,n} p_{m,n}^s \\ &\qquad\qquad + L_3 - s_3' = \varepsilon_3. \end{aligned} \tag{5}$$

In (5), slack variables L_i and surplus variables s_i' are utilized to convert the upper bound ε_i on the objective value U_i into an equality constraint, and π_i is the penalty coefficient for a given objective U_i. We consider ε_1, ε_2, and ε_3 as the upper bounds of the energy consumption costs of RRHs, BBUs, and the total transmit power of all users, respectively. In this context, by changing the values of ε_i for a given objective U_i, a set of Pareto optimal solutions is derived [36]. Then, according to network conditions, one of the Pareto optimal solutions can be selected. In Section IV, we will investigate the effect of the variation value of ε_is as trade-off parameters on the network performance through simulation results. Now, we first focus on solving (5), which is a non-convex and NP-hard resource allocation problem with high computational complexity [37].

III. TWO-STEP ITERATIVE ALGORITHM FOR DYNAMIC RESOURCE ALLOCATION

To solve (3), we propose an efficient two-step iterative algorithm as summarized in Table 3. Each step is also iterative. At each iteration t, Step 1 with a given (fixed) power allocation vector computes the Pareto optimal solution for the CAPs, i.e., $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\mathbf{Z}$, $\mathbf{X}$. Then, in Step 2, the transmit power is allocated among users based on a given fixed set of CAP parameters derived in Step 1. The whole process can be explained as

$$\underbrace{\underbrace{\boldsymbol{\alpha}(0), \boldsymbol{\beta}(0), \boldsymbol{\tau}(0), \mathbf{Z}(0), \mathbf{X}(0)}_{\text{output of Step 1}} \rightarrow \underbrace{\mathbf{P}(0)}_{\text{output of Step 2}}}_{\text{Initialization}} \rightarrow \dots \rightarrow$$

$$\underbrace{\underbrace{\boldsymbol{\alpha}^*(t), \boldsymbol{\beta}^*(t), \boldsymbol{\tau}^*(t), \mathbf{Z}^*(t), \mathbf{X}^*(t)}_{\text{output of Step 1}} \rightarrow \underbrace{\mathbf{P}^*(t)}_{\text{output of Step 2}}}_{\text{Iteration } t} \rightarrow$$

$$\underbrace{\underbrace{\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\tau}^*, \mathbf{Z}^*, \mathbf{X}^*}_{\text{output of Step 1}} \rightarrow \underbrace{\mathbf{P}^*}_{\text{output of Step 2}}}_{\text{Optimal solution}},$$

where $t \geq 0$ is the iteration index in each step. Also, $\boldsymbol{\alpha}^*(t)$, $\boldsymbol{\beta}^*(t)$,$\boldsymbol{\tau}^*(t)$, $\mathbf{Z}^*(t)$, $\mathbf{X}^*(t)$, and $\mathbf{P}^*(t)$ are optimal values obtained at iteration t. The iterative procedure stops when the convergence criteria are met. That is, when

$$\| \boldsymbol{\beta}^*(t) - \boldsymbol{\beta}^*(t-1) \| \le \kappa_1, \| \mathbf{Z}^*(t) - \mathbf{Z}^*(t-1) \| \le \kappa_2,$$
$$\| \mathbf{X}^*(t) - \mathbf{X}^*(t-1) \| \le \kappa_3 \text{ and } \| \mathbf{P}^*(t) - \mathbf{P}^*(t-1) \| \le \kappa_4,$$

where $0 < \kappa_1, \kappa_2, \kappa_3, \kappa_4 \ll 1$. However, the optimization problems of Steps 1 and 2 are still non-convex and encounter high computational complexity. To solve them, by using CGP [22] along with various transformations and convexification approaches, we convert non-convex problems into the equivalent lower-bound GP problems. See Section III-A in [37] for more information about the preliminaries of CGP.

A. STEP 1: CAP ALLOCATION ALGORITHM

Assuming a fixed value of $\mathbf{P}(t)$ in a high signal-to-interference-plus-noise ratio (SINR) scenario, we have

$$\widetilde{R}^s_{m,n}(\mathbf{P}, \boldsymbol{\beta}) \approx \begin{cases} \log_2(\frac{J_m - N_m + 1}{N_m}\gamma^s_{m,n}(t)), & \text{if } m \in \mathcal{R}, \\ \log_2(1 + \gamma^s_{m,n}(t)), & \text{if } m \in \mathcal{F}, \end{cases} \tag{6}$$

where $\gamma^s_{m,n}(t) = \frac{p^s_{m,n}(t)h^s_{m,n}}{\sigma^2 + \sum_{m' \neq m}\sum_{n' \neq n} p^s_{m',n}(t)h^s_{m,n'}}$ is the SINR of user $n \in \mathcal{N}$ at AP $m \in \mathcal{M}$ on sub-carrier $s \in \mathcal{S}$ and it has a fixed value for this step. Therefore, at iteration t_1, (5) is converted to

$$\min_{\Im, L_i, s'_i}\Big(- \sum_{m\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}} \alpha^s_{m,n}\beta_{m,n}\widetilde{R}^s_{m,n}(\mathbf{P}(t), \boldsymbol{\beta})$$
$$- \sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}} \alpha^s_{m,n}\beta_{m,n}\widetilde{R}^s_{m,n}(\mathbf{P}(t)) + \sum_{i=1}^{3}\pi_i s'_i\Big),$$
$$\text{subject to: C2} - \text{C14}. \tag{7}$$

In (7) the optimization variables are $\Im$, L_i, and s'_i, and (7) has less computational complexity than (5). However, due to the binary variables and the non-linear function of the throughput, (7) is still a non-convex optimization problem. To overcome these issues, we aim to convert (7) into the standard form of GP. Since, in (7), N_m is a function of $\beta_{m,n}$. Therefore, converting (7) into the standard form of GP suffers from high computational complexity. Hence, we first relax the binary variables as $\alpha^s_{m,n} \in [0, 1]$, $\beta_{m,n} \in [0, 1]$, $\tau_r \in [0, 1]$, $z_{r,b} \in [0, 1]$ and $x_{r,b} \in [0, 1]$. Then, by applying various transformations and DC approximation, we try to develop the analytical framework to transform the non-convex optimization problems into the equivalent lower-bound standard form of GP. Hence, at iteration t_1, (7) is converted into (see Appendix II)

$$\min_{\Im, L_i, s'_i}\Big(- \sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t_1)\beta_{m,n}(t_1)\widetilde{R}^s_{m,n}(\mathbf{P}(t), \boldsymbol{\beta})$$
$$- \sum_{m\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t_1)\beta_{m,n}(t_1)$$
$$\times[\log_2(J_m\gamma^s_{m,n}(t)) - \Gamma_{m,n}(t_1)$$
$$+\Gamma_{m,n}(t_1-1) - \log(N_m(t_1-1))] + \sum_{i=1}^{3}\pi_i s'_i\Big), \tag{8}$$

where $\Gamma_{m,n}(t_1) = \sum_{n\in N} \frac{\beta_{m,n}(t_1)}{\sum_{n\in N}\beta_{m,n}(t_1-1)}$ and $\Gamma_{m,n}(t_1-1) = \sum_{n\in N} \frac{\beta_{m,n}(t_1-1)}{\sum_{n\in N}\beta_{m,n}(t_1-1)}$. Now, based on (8), AGMA, and Proposition 1, we derive the GP approximation of (7) for each iteration. We assume t_1 as the index of iterations in Step 1.

Proposition 1: Consider the positive auxiliary variable $\varpi_0(t_1) > 0$ and $\Lambda_1 \gg 1$. Also, consider $Y(t) = \alpha^s_{m,n}(t)\beta_{m,n}(t)$. Then, the GP approximation of (7) is

$$\min_{\omega(t_1)} \varpi_0(t_1)$$
subject to:
C3, C4, C6, C11,

$$\widetilde{\text{C}}00 : (\Lambda_1 + I_1(t_1) + \sum_{i=1}^{3}\pi_i s'_i) \times (\frac{\varpi_0(t_1)}{\eta_0(t_1)})^{-\eta_0(t_1)}$$
$$\times \prod_{\substack{m\in\mathcal{R}\\ n\in\mathcal{N}\\ s\in\mathcal{S}}} \left(\frac{Y(t_1)(\log_2(J_m\gamma^s_{m,n}(t)) + \Gamma_{m,n}(t_1-1))}{\Upsilon^s_{m,n}(t_1)}\right)^{-\Upsilon^s_{m,n}(t_1)}$$
$$\prod_{\substack{m\in\mathcal{F}\\ n\in\mathcal{N}\\ s\in\mathcal{S}}} \left(\frac{Y(t_1)\widetilde{R}^s_{m,n}(\mathbf{P}(t), \boldsymbol{\beta})}{\psi^s_{m,n}(t_1)}\right)^{-\psi^s_{m,n}(t_1)} \le 1,$$

$$\widetilde{\text{C}}2.1 : R^{\text{rsv}}_n \times \prod_{\substack{m\in\mathcal{F}\\ s\in\mathcal{S}}} \left[\frac{Y(t_1)\widetilde{R}^s_{m,n}(\mathbf{P}(t), \boldsymbol{\beta})}{\theta^s_{m,n}(t_1)}\right]^{-\theta^s_{m,n}(t_1)}$$
$$\le 1, \forall n \in \mathcal{N},$$

$$\widetilde{\text{C}}2.2 : [R^{\text{rsv}}_n + \sum_{m\in\mathcal{R}}\sum_{s\in\mathcal{S}} Y(t_1)[\log_2(N_m(t_1-1))$$
$$+ \sum_{n\in\mathcal{N}} \frac{\beta_{m,n}(t_1)}{} \sum_{n\in\mathcal{N}} \beta_{m,n}(t_1)]]$$
$$\prod_{\substack{m\in\mathcal{R}\\ s\in\mathcal{S}}} \left(\frac{Y(t_1)\log_2(J_m\gamma^s_{m,n}(t)}{\varsigma_{m,s}(t_1)}\right)^{-\varsigma_{m,s}(t_1)}$$
$$\prod_{\substack{m\in\mathcal{R}\\ s\in\mathcal{S}}} \left(\frac{Y(t_1)\Gamma_{m,n}(t_1-1)}{\upsilon_{m,s}(t_1)}\right)^{-\upsilon_{m,s}(t_1)} \le 1, \forall n \in \mathcal{N},$$

$$\widetilde{\text{C}}5 : (\alpha^s_{m,n}(t_1) + 1) \times (\frac{1}{\lambda(t_1)})^{-\lambda(t_1)} \times (\frac{\beta_{m,n}(t_1)}{\xi(t_1)})^{-\xi(t_1)}$$
$$\le 1, \quad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \forall s \in \mathcal{S},$$

$$\widetilde{\text{C}}7 : \sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}} z_{r,b}Y(t_1)[\log_2(J_r\gamma^s_{r,n}(t) + \Gamma_{r,n}(t_1-1)]$$
$$\left(\frac{F^{\max}_{r,b}}{\chi_1(t_1)}\right)^{-\chi_1(t_1)} \prod_{\substack{n\in\mathcal{N}\\ s\in\mathcal{S}}} \left(\frac{z_{r,b}(t_1)Y(t_1)\log_2(N_r(t_1-1))}{\chi_2(t_1)}\right)^{-\chi_2(t_1)}$$
$$\prod_{\substack{n\in\mathcal{N}\\ s\in\mathcal{S}}} \left(\frac{z_{r,b}(t_1)Y(t_1)\Gamma_{r,n}(t_1-1)}{\chi_3(t_1)}\right)^{\chi_3(t_1)} \le 1, \forall r \in \mathcal{R},$$

$$\widetilde{\text{C}}8 : x_b(t_1)^{-1} \times \sum_{r\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}} z_{r,b}Y(t_1)$$

$$\log_2(J_r\gamma^s_{r,n}(t)+\Gamma_{r,n}(t_1-1)]\left(\frac{T_b^{\max}}{\chi_4(t_1)}\right)^{-\chi_4(t_1)}$$
$$\prod_{\substack{r\in\mathcal{R}\\ n\in\mathcal{N}\\ s\in\mathcal{S}}}\left(\frac{z_{r,b}(t_1)Y(t_1)\log_2(N_r(t_1-1))}{\chi_5(t_1)}\right)^{-\chi_5(t_1)}$$
$$\prod_{\substack{r\in\mathcal{R}\\ n\in\mathcal{N}\\ s\in\mathcal{S}}}\left(\frac{z_{r,b}(t_1)Y(t_1)\Gamma_{r,n}(t_1-1)}{\chi_6(t_1)}\right)^{\chi_6(t_1)}\leq 1,\forall b\in\mathcal{B},$$
$$\widetilde{\mathrm{C}}9:(z_{r,b}(t_1)+1)\times(\frac{1}{\phi(t_1)})^{-\phi(t_1)}\times(\frac{x_b(t_1)}{\delta(t_1)})^{-\delta(t_1)}$$
$$\leq 1,\quad \forall r\in\mathcal{R},\forall b\in\mathcal{B},$$
$$\widetilde{\mathrm{C}}10:(\tau_r(t_1)+1)\times(\frac{1}{\rho(t_1)})^{-\rho(t_1)}\prod_{\substack{r\in\mathcal{R}\\ b\in\mathcal{B}}}\times(\frac{z_{r,b}(t_1)}{\varphi(t_1)})^{-\varphi(t_1)}$$
$$\leq 1,\quad \forall r\in\mathcal{R},\forall b\in\mathcal{B}, \tag{9}$$

and for i = [1], [2], [3], we have

$$\widetilde{\mathrm{C}}12.1-\widetilde{\mathrm{C}}14.1:\ q_i^{-1}(t_1)U_i(t_1)$$
$$+q_i^{-1}(t_1)L_i(t_1)\leq 1,$$
$$\widetilde{\mathrm{C}}12.2-\widetilde{\mathrm{C}}14.2:q_i(t_1)\times(\frac{\varepsilon_i}{e_i(t_1)})^{-e_i(t_1)}\times(\frac{s'_i(t_1)}{d_i(t_1)})^{-d_i(t_1)}$$
$$\leq 1,$$

where

$$\omega(t_1)=\varpi_0(t_1),\mathfrak{S}(t_1),s'_i(t_1),L_i(t_1)\}, \tag{10}$$
$$I_1(t_1)=\sum_{m\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}Y(t_1)\log_2(N_m(t_1-1))$$
$$+\sum_{m\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}Y(t_1)(t_1)\Gamma_{m,n}(t_1), \tag{11}$$
$$I_2(t_1)=\sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}Y(t_1)\widetilde{R}^s_{m,n}(\mathbf{P}(t),\boldsymbol{\beta})$$
$$+\sum_{m\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}Y(t_1)(\log_2(J_m\gamma^s_{m,n}(t))$$
$$+\Gamma_{m,n}(t_1-1)), \tag{12}$$
$$I_3(t_1)=F^{\max}_{r,b}$$
$$+\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}z_{m,b}(t_1)Y(t_1)(\log_2(N_r(t_1-1))$$
$$+\Gamma_{m,n}(t_1)), \tag{13}$$
$$I_4(t_1)=T^{\max}_b$$
$$+\sum_{r\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}z_{m,b}(t_1)Y(t_1)(\log_2(N_r(t_1-1))$$
$$+\Gamma_{m,n}(t_1)), \tag{14}$$
$$\theta^s_{m,n}(t_1)=\frac{Y(t_1-1)\widetilde{R}^s_{m,n}(\mathbf{P}(t),\boldsymbol{\beta})}{\sum\limits_{m\in\mathcal{F}}\sum\limits_{s\in\mathcal{S}}\alpha^s_{m,n}(t_1-1)\beta^s_{m,n}(t_1-1)\widetilde{R}^s_{m,n}(\mathbf{P}(t),\boldsymbol{\beta})}, \tag{15}$$
$$\eta_0(t_1)=\frac{\varpi_0(t_1-1)}{\varpi_0(t_1-1)+I_2(t_1-1)}, \tag{16}$$
$$\psi^s_{m,n}(t_1)=\frac{Y(t_1-1)\widetilde{R}^s_{m,n}(\mathbf{P}(t),\boldsymbol{\beta})}{\varpi_0(t_1-1)+I_2(t_1-1)}, \tag{17}$$
$$\Upsilon^s_{m,n}(t_1)=\frac{Y(t_1-1)(\log_2(J_m\gamma^s_{m,n}(t))+\Gamma_{m,n}(t_1-1))}{\varpi_0(t_1-1)+I_2(t_1-1)}, \tag{18}$$
$$\varsigma_{m,s}(t_1)=\frac{Y(t_1-1)\log_2(J_m\gamma^s_{m,n}(t))}{\sum\limits_{m\in\mathcal{R}}\sum\limits_{s\in\mathcal{S}}Y(t_1-1)\big(\log_2(J_m\gamma^s_{m,n}(t))+\Gamma_{m,n}(t_1-1)\big)}, \tag{19}$$
$$\upsilon_{m,s}(t_1)=\frac{Y(t_1-1)\Gamma_{m,n}(t_1-1)}{\sum\limits_{m\in\mathcal{R}}\sum\limits_{s\in\mathcal{S}}Y(t_1-1)\big(\log_2(J_m\gamma^s_{m,n}(t))+\Gamma_{m,n}(t_1-1)\big)}, \tag{20}$$
$$\lambda(t_1)=\frac{1}{1+\beta_{m,n}(t_1-1)}, \tag{21}$$
$$\xi(t_1)=\frac{\beta_{m,n}(t_1-1)}{1+\beta_{m,n}(t_1-1)}, \tag{22}$$
$$\chi_1(t_1)=\frac{F^{\max}_{r,b}}{I_3(t-1)}, \tag{23}$$
$$\chi_2(t_1)=\frac{z_{m,b}(t_1-1)Y(t_1-1)\log_2(N_r(t_1-1))}{I_3(t-1)}, \tag{24}$$
$$\chi_3(t_1)=\frac{z_{m,b}(t_1-1)Y(t_1-1)\Gamma_{m,n}(t_1-1)}{I_3(t-1)}, \tag{25}$$
$$\chi_4(t_1)=\frac{T^{\max}_b}{I_4(t-1)}, \tag{26}$$
$$\chi_5(t_1)=\frac{z_{m,b}(t_1-1)Y(t_1-1)\log_2(N_r(t_1-1))}{I_4(t-1)}, \tag{27}$$
$$\chi_6(t_1)=\frac{z_{m,b}(t_1-1)Y(t_1-1)\Gamma_{m,n}(t_1-1)}{I_4(t-1)}, \tag{28}$$
$$\phi(t_1)=\frac{1}{1+x_b(t_1-1)}, \tag{29}$$
$$\delta(t_1)=\frac{x_b(t_1-1)}{1+x_b(t_1-1)}, \tag{30}$$
$$\rho(t_1)=\frac{1}{1+\sum\limits_{b\in\mathcal{B}}z_{r,b}(t_1-1)}, \tag{31}$$
$$\varphi(t_1)=\frac{z_{r,b}(t_1-1)}{1+\sum\limits_{b\in\mathcal{B}}z_{r,b}(t_1-1)}, \tag{32}$$

and for i = [1], [2], [3], we have

$$e_i(t_1)=\frac{\varepsilon_i}{\varepsilon_i+s'_i(t_1-1)}, \tag{33}$$
$$d_i(t_1)=\frac{s'_i(t_1-1)}{\varepsilon_i+s'_i(t_1-1)}. \tag{34}$$

Proof: See Appendix III. ■

In (9), by applying AGMA, we get the monomial approximation $\widetilde{\mathrm{C}}2.1$, $\widetilde{\mathrm{C}}2.2$, $\widetilde{\mathrm{C}}5$, $\widetilde{\mathrm{C}}9$, $\widetilde{\mathrm{C}}10$ and $\widetilde{\mathrm{C}}12-\widetilde{\mathrm{C}}14$ for C2, C5, C9, C10, C12 − C14. Now, (9) is iteratively solved by CVX [23] at each iteration. The iterative algorithm will stop when the optimal solutions of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\mathbf{Z}$, $\mathbf{X}$ are derived and the convergence criteria are met.

B. STEP 2: POWER ALLOCATION ALGORITHM

From the values obtained of $\boldsymbol{\alpha}^*(t)$, $\boldsymbol{\beta}^*(t)$, $\boldsymbol{\tau}^*(t)$, $\mathbf{Z}^*(t)$, and $\mathbf{X}^*(t)$ obtained from Step 1, the optimization problem for power allocation in Step 2 is

$$\begin{aligned}&\min_{\mathbf{P}}[-\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^s_{m,n}(t)\beta_{m,n}(t)\widetilde{R}^s_{m,n}(\mathbf{P}(t_2))\\&\quad+\mu_{\mathrm{a}}\sum_{r\in\mathcal{R}}\tau_r J_r(t)+\sum_{b\in\mathcal{B}}\mu_b\times x_b(t)\\&\quad+\mu_{\mathrm{p}}\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^s_{m,n}(t)\beta_{m,n}(t)p^s_{m,n}(t_2)],\\&\text{subject to: C1, C2, C7, C8,}\end{aligned}\tag{35}$$

where t_2 is the index of iterations in Step 2. Note that in (35), the only optimization variable is $\mathbf{P}$. Therefore, (35) has less computational complexity than (3). Since, $\boldsymbol{\tau}$ and $\mathbf{X}$ have fixed values, the objectives U_1 and U_2 are constant and do not affect the problem solution of this step. Hence, similar to (5), by considering the total throughput (i.e., U_4) as the primary objective function and the total transmit power of all users (i.e., U_3) as constraint C0, we can transform (35) into a single objective optimization problem as follows:

$$\begin{aligned}&\min_{\mathbf{P},s'_4}[-\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^s_{m,n}(t)\beta_{m,n}(t)\widetilde{R}^s_{m,n}(\mathbf{P}(t_2,\boldsymbol{\beta}))\\&\quad+\mu_{\mathrm{a}}\sum_{r\in\mathcal{R}}\tau_r J_r(t)+\sum_{b\in\mathcal{B}}\mu_b\times x_b(t)+\pi_4 s'_4(t_2)],\\&\text{subject to: C1, C2, C7, C8,}\\&\qquad\mathrm{C0}:\mu_{\mathrm{p}}\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^s_{m,n}\beta_{m,n}p^s_{k,n}(t_2)\\&\qquad\quad+L_4(t_2)-s'_4(t_2)=\varepsilon_3,\end{aligned}\tag{36}$$

where the slack and surplus variables $L_4(t_2)$ and $s'_4(t_2)$ are associated with the bound ε_3 on objective U_3. Also, π_4 is the penalty coefficient. Since $\widetilde{R}^s_{m,n}(\mathbf{P}(t_2))$ is a non-linear function, (36) is a non-convex optimization problem. To tackle this computational complexity, we first apply DC approximation of $\widetilde{R}_{k,s,n}(\mathbf{P})$ at iteration t_2, and then by using AGMA, we will convert (36) into a GP approximation as shown in Proposition 2.

Proposition 2: Consider the positive auxiliary variable $\varpi_1(t_2) > 0$ and $\Lambda_2 \gg 1$. The GP-based reformulation of (36) for each iteration t_2 is

$$\begin{aligned}&\min_{\mathbf{P},L_4,s'_4}\ \varpi_1(t_2)\\&\text{subject to: C1,}\\&\widetilde{\mathrm{C}}01:[\Lambda_2+\pi_4 s'_4(t_2)+\mu_{\mathrm{a}}\sum_{r\in\mathcal{R}}\tau_r J_r(t)+\sum_{b\in\mathcal{B}}\mu_b\times x_b(t)\\&+\sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\frac{h^s_{m,n}}{\sigma^2+\sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}p^s_{m,n}(t_2-1)h^s_{m,n}}\\&\times\frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)})]\\&(\frac{\varpi_1(t_2)}{a_0(t_2)})^{-a_0(t_2)}\prod_{\substack{m\in\mathcal{F}\\n\in\mathcal{N},s\in\mathcal{S}}}\left[\frac{\log_2(1+\frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)})}{b_0(t_2)}\right]^{-b_0(t_2)}\\&\prod_{\substack{m\in\mathcal{F}\\n\in\mathcal{N},s\in\mathcal{S}}}(\frac{\frac{h^s_{m,n}}{\sigma^2+\sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}p^s_{m,n}(t_2-1)h^s_{m,n}}\times\frac{p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2)}}{d_0(t_2)})^{-d_0(t_2)}\\&\times\prod_{\substack{m\in\mathcal{R}\\n\in\mathcal{N},s\in\mathcal{S}}}\left[\frac{\log_2(\frac{J_m}{N_m(t)}\frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)})}{b_1(t_2)}\right]^{-b_1(t_2)}\\&\prod_{\substack{m\in\mathcal{R}\\n\in\mathcal{N},s\in\mathcal{S}}}(\frac{\frac{1}{p^s_{m,n}(t_2-1)}\times\frac{J_m}{N_m(t)}\frac{p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2)}}{d_1(t_2)})^{-d_1(t_2)}\le 1,\\&\widetilde{\mathrm{C}}0.1:\ q_p(t_2)\times(\frac{\kappa_3}{e(t_2)})^{-e(t_2)}\times(\frac{s'_4(t_2)}{d(t_2)})^{-d(t_2)}\le 1,\\&\widetilde{\mathrm{C}}0.2:\ q_p^{-1}(t_2)\mu_{\mathrm{p}}\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^s_{m,n}(t)\beta_{m,n}(t)p^s_{m,n}(t_2)\\&+q_p^{-1}(t_2)L_4(t_2)\le 1,\\&\widetilde{\mathrm{C}}2.1:\prod_{\substack{m\in\mathcal{F}\\s\in\mathcal{S}}}\alpha^s_{m,n}(t)\beta_{m,n}(t)(\sigma^2+I^s_{m,n}(t_2))\times\left[\frac{\sigma^2}{\lambda_0(t_2)}\right]^{-\lambda_0(t_2)}\\&\prod_{\substack{m\in\mathcal{F}\\s\in\mathcal{S}}}\left[\frac{p^s_{m,n}(t_2)h^s_{m,n}}{\lambda_{m,s}(t_2)}\right]^{-\lambda_{m,s}(t_2)}\le 2^{-R^{\mathrm{rsv}}_n},\ \forall n\in\mathcal{N},\\&\widetilde{\mathrm{C}}2.2:\prod_{\substack{\forall m\in\mathcal{R}\\\forall s\in\mathcal{S}}}\alpha^s_{m,n}(t)\beta_{m,n}(t)\left(\frac{\sigma^2+I^s_{m,n}(t_2)}{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2)h^s_{m,n}}\right)\le 2^{-R^{\mathrm{rsv}}_n},\\&\widetilde{\mathrm{C}}7:\prod_{\substack{n\in\mathcal{N}\\s\in\mathcal{S}}}z_{r,b}(t)\alpha^s_{r,n}(t)\beta_{r,n}(t)(\frac{\sigma^2+I^{th}_s+\frac{J_m}{N_m(t)}p^s_{r,n}(t_2)h^s_{r,n}}{\sigma^2+I^{th}_s})\\&\le 2^{F^{\max}_{r,b}},\ \forall r\in\mathcal{R},\forall b\in\mathcal{B},\\&\widetilde{\mathrm{C}}8:\prod_{\substack{r\in\mathcal{R}\\n\in\mathcal{N}\\s\in\mathcal{S}}}z_{r,b}(t)\alpha^s_{r,n}(t)\beta_{r,n}(t)(\frac{\sigma^2+I^{th}_s+\frac{J_m}{N_m(t)}p^s_{r,n}(t_2)h^s_{r,n}}{\sigma^2+I^{th}_s})\\&\le 2^{T^{\max}_b\times x_b(t)},\forall b\in B,\end{aligned}\tag{37}$$

where

$$G=\sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}[\log_2(1+\frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)})$$

$$+\frac{h^s_{m,n}}{\sigma^2 + \sum\limits_{m\in\mathcal{M}}\sum\limits_{n\in\mathcal{N}} p^s_{m,n}(t_2-1)h^s_{m,n}}$$

$$\times\frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)}]$$

$$+\varpi_1(t_2-1)\sum_{m\in\mathcal{R}}$$

$$\times\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}[\log_2(\frac{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)})$$

$$+\frac{1}{p^s_{m,n}(t_2-1)}\times\frac{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)}],$$

$$a_0(t_2) = \frac{\varpi_1(t_2-1)}{G}, \tag{38}$$

$$b_0(t_2) = \frac{\log_2(1 + \frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)})}{G}, \tag{39}$$

$$d_0(t_2) = \frac{\frac{h^s_{m,n}}{\sigma^2+\sum\limits_{m\in\mathcal{M}}\sum\limits_{n\in\mathcal{N}} p^s_{m,n}(t_2-1)h^s_{m,n}} \times \frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)}}{G}, \tag{40}$$

$$b_1(t_2) = \frac{\log_2(\frac{J_m}{N_m(t)}\frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)})}{G}, \tag{41}$$

$$d_1(t_2) = \frac{\frac{1}{p^s_{m,n}(t_2-1)}\times(\frac{J_m}{N_m(t)})\frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2+I^s_{m,n}(t_2-1)}}{G}, \tag{42}$$

$$(t_2) = \frac{\varepsilon_3}{\varepsilon_3 + s'_4(t_2-1)}, \tag{43}$$

$$d(t_2) = \frac{s'_4(t_2-1)}{s'_4(t_2-1) + s'_4(t_2-1)}, \tag{44}$$

$$\lambda_0(t_2) = \frac{\sigma^2}{\sigma^2 + \sum\limits_{\substack{m\in\mathcal{F}\\ s\in\mathcal{S}}} p^s_{m,n}(t_2-1)h^s_{m,n}}, \tag{45}$$

and

$$\lambda_{m,s}(t_2) = \frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + \sum\limits_{\substack{m\in\mathcal{F}\\ s\in\mathcal{S}}} p^s_{m,n}(t_2-1)h^s_{m,n}}. \tag{46}$$

Proof: See Appendix IV. ■

To reduce the computational complexity of the constraints C7 and C8, we suppose that interference for each sub-carrier $s \in \mathcal{S}$ is bounded to the maximum aggregated value of I^{th}_s [38], [39], [40]. By applying AGMA, monomial approximation of C01, C0, C2, C7 and C8 are $\widetilde{\text{C}}01$, $\widetilde{\text{C}}0.1$, $\widetilde{\text{C}}0.2$, $\widetilde{\text{C}}2.1$, $\widetilde{\text{C}}2.2$, $\widetilde{\text{C}}7$, and $\widetilde{\text{C}}8$, respectively. The optimization problem (37) is iteratively solved until the convergence criteria $\| P^*(t_2) - \mathbf{P}^*(t_2-1) \| \leq \kappa_4$ are met.

TABLE 4. Simulation parameters.

Parameter	Value	Parameter	Value
Cell radius	500 m	$\mathcal{M} = \mathcal{R}\bigcup\mathcal{F}$	12
σ^2	[1,2] Watt	I^{th}_s	[10,15] Watt
$p^{\max}_m$	20 Watt	J_r	[100,250]
$F^{\max}_{r,b}$	[10,50] bps/Hz	$T^{\max}_b$	[20,100] bps/Hz
μ_b	[20,100] bps/Hz	μ_a	[0.1,3] bps/Hz
ε_1	[50,200] bps/Hz	ε_2	[20,100] bps/Hz
ε_3	[5,30] bps/Hz		

C. CONVERGENCE AND COMPUTATIONAL COMPLEXITY

Based on [40], our proposed resource allocation algorithm belongs to block SCA algorithm. It was shown in [41] that with AGMA approximation, the SCA method converges to a locally optimal solution that satisfies the Karush-Kuhn-Tucker (KKT) conditions. Thus, by applying AGMA approximation, the convergence of (9) and (37) to a local optimal solution are guaranteed.

CVX uses interior point method for solving GP sub-problems in Steps 1 and 2. According to [42], using this method, the required number of iterations to solve this is $\frac{\log(c/(\nu t^0))}{\log\zeta}$, where c, ν, and t^0 are the total number of constraints, the stopping criterion, and the initial point to approximate the accuracy, respectively. Also, ζ is used for updating the accuracy of the method. The numbers of constraints in (9) and (37) are $c_1 = MNS + MS + 3RB + B + 2R + 3N + 7$ for Step 1 and $c_2 = 2N + RB + B + M + 3$ for Step 2. Furthermore, for each iteration in Steps 1 and 2, $i_1 = 2FNS + 3RNS + MNS + 2RB + 6$ and $i_2 = 2NSRB + 3FNS + 3RNS + 2$ are the number of calculations required to transform the non-convex problems using AGMA into the GP approximations, respectively. Consequently, the total number of calculations for Steps 1 and 2 of our proposed algorithm is $i_1 \times \frac{\log(c_1/(\nu_1 t_1{}^0))}{\log\zeta_1}$ and $i_2 \times \frac{\log(c_2/(\nu_2 t_2{}^0))}{\log\zeta_2}$, respectively. According to this analysis, the computational complexity of Step 1 and 2 become logarithm functions $O((NS(R+F+M)+RB)\log(MNS+RB))$ and $O(NS(F+RB)\log(N+M+RB))$, respectively, with polynomial complexity and not exponential complexity (as summarized in Table 5). The computational complexity of Steps 1 and 2 are sensitive to the number of users (i.e., N). Fig. 2 illustrates that with increasing N, the number of iterations required for convergence of both Step 1 and 2 will be increased.

IV. SIMULATION RESULTS

To evaluate the performance of our approach, we consider $N \in [20, 300]$ users uniformly distributed inside a region served by three RRHs ($R = 3$), two BBUs ($B = 2$), and nine FAPs ($F = 9$). The channel gain between user $n \in \mathcal{N}$ and AP $m \in \mathcal{R}$ and $m \in \mathcal{F}$ are modeled as $h^s_{m,n} = \frac{1}{1+(d_{m,n})^4}$ and $h^s_{m,n} = \wp_{m,s,n} d^{-\iota}_{m,n}$, respectively, where $d_{m,n} > 0$ is the distance of user $n \in \mathcal{N}$ to AP $m \in \mathcal{M}$, $\iota = 3$ is the path loss exponent, and $\wp_{m,s,n} \sim \text{Exp}(1)$ [43]. We set $\mu_P = 1$ and $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 = 10^{-3}$ for all of the simulations. Furthermore, we assume the ratio for the

TABLE 5. Computational complexity.

	Step 1	Step 2
No. of constraints	$MNS + MS + 3RB + B + 2R + 3N + 7$	$2N + RB + B + M + 3$
No. of transforms	$2FNS + 3RNS + MNS + 2RB + 6$	$2NSRB + 3FNS + 3RNS + 2$
Computational complexity	$O((NS(R + F + M) + RB)\log(MNS + RB))$	$O(NS(F + RB)\log(N + M + RB))$

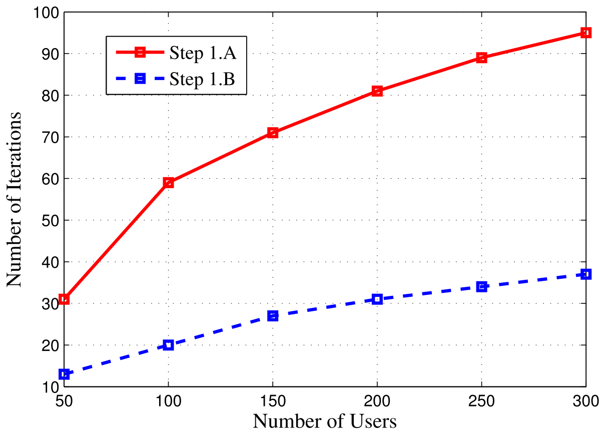

FIGURE 2. Number of required iterations versus number of users.

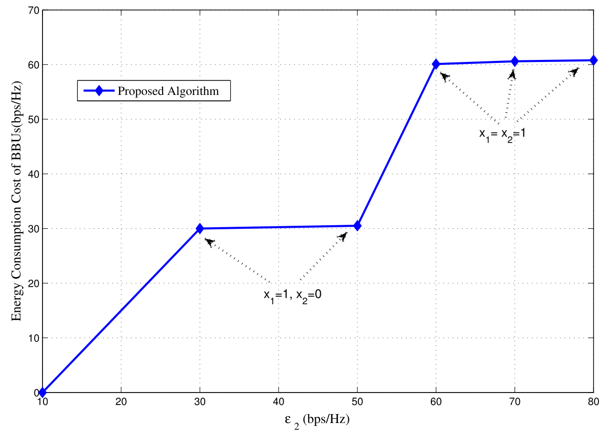

FIGURE 4. The energy consumption cost of BBUs relative to ε_2.

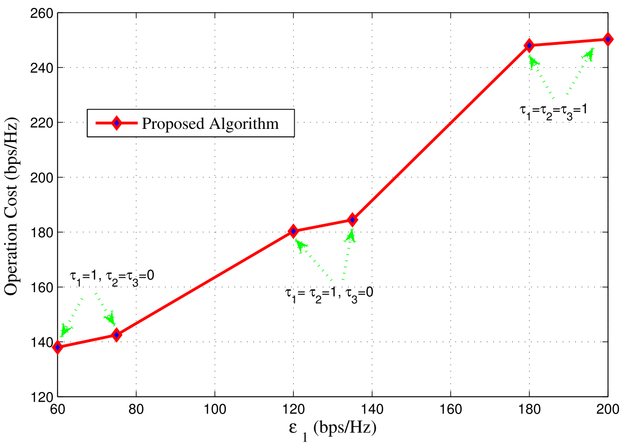

FIGURE 3. Network operation cost relative to ε_1 and $J_1 = J_2 = J_3 = 200$.

maximum transmit power of RRHs to FAPs to be equal to 0.5. The values of the maximum BBU load, front-haul link capacity, energy consumption cost of active BBUs and each antenna, and the number of antennas mounted on the RRH $r \in \mathcal{R}$ are randomly chosen, as shown in Table 4.

In the following, we investigate the effects of different trade-off parameters (e.g., ε_1, ε_2, and ε_3) on the Pareto optimal sets derived by the proposed MORA algorithm. Based on C0 and C1 in (36), the total transmit power of all users and the transmit power of each AP are bounded to ε_3 and $p_m^{\max}$, respectively. Therefore, C1 significantly affects and limits C0, while choosing different values of ε_3 does not affect the Pareto optimal sets derived by the proposed MORA algorithm. But simulation results reveal that ε_1 and ε_2 significantly impact the Pareto optimal sets. Since ε_1 and ε_2 are randomly chosen from a predetermined rang, the appropriate selection of these parameters can significantly enhance network performance in terms of H-CRAN operation cost, outage probability, total throughput, and EE. Therefore, we evaluate their effects on two major conflicting objective functions, namely throughput and operation cost. Also, to study the performance of the proposed algorithm in terms of coverage, we evaluate traffic offloading and outage probability for different values of ε_1. In Fig. 3, the effect of ε_1 on the network operational cost is demonstrated. As we can see, the operation cost increases in function of ε_1. This is because, based on C12 in (5), total energy consumption cost of RRHs is bounded to ε_1. This constraint affects the feasibility region of (5). Therefore, by increasing ε_1, the upper bound allowable energy consumption cost of RRHs will increase, which leads to more active RRHs. For instance, in Fig. 3 when $\varepsilon_1 = 60$, only one RRH is in state on (e.g., $\tau_1 = 1, \tau_2 = \tau_3 = 0$) and total operation cost is 142.5 bps/Hz, which is lower than 180.3 bps/Hz when $\varepsilon_1 = 120$ and two RRHs are in state on (e.g., $\tau_1 = \tau_2 = 1, \tau_3 = 0$). Similarly, in Fig. 4, with increasing ε_2, more BBUs become active, which leads to a higher energy consumption cost of BBUs.

Besides, by decreasing ε_1, some under-utilized RRHs will be switched off; therefore, the chance to choose sub-carriers and APs, and assign transmit power to each user will be reduced. To tackle this issue, in the proposed approach, joint radio resource assignment (e.g., sub-carrier, transmit power and AP allocation to each user) manages the inter-tier interference between different APs. Thus, the traffic of users associated to the under-utilized RRHs can be offloaded to neighboring low-power FAPs. Consequently, under-utilized

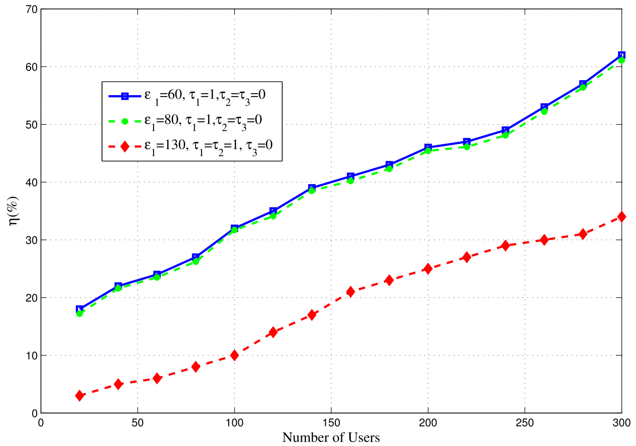

FIGURE 5. Traffic offloading from RRHs to FAPs relative to the number of users and $J_1 = J_2 = J_3 = 200$.

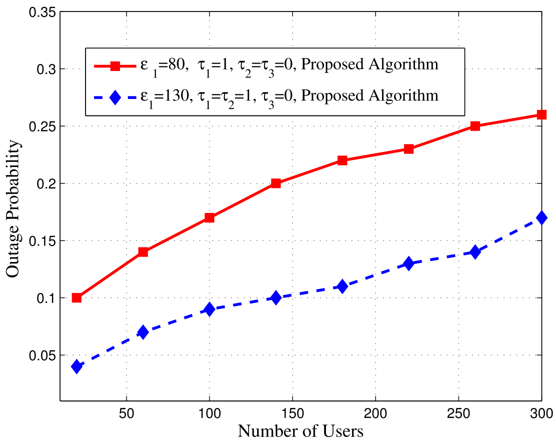

FIGURE 6. Outage probability relative to the number of users and $J_1 = J_2 = J_3 = 200$.

RRHs and BBUs can be switched off, which leads to greater EE. We define the ratio of network traffic offloading (η) as

$$\eta = \frac{\text{Total number of users moved from RRHs to FAPs}}{\text{Total number of users}}.$$

Fig. 5 shows that η increases with decreasing ε_1 and increasing the number of users to satisfy the minimum guaranteed rate of each user. Vs with increasing ε_1, the more RRHs will be switched on, which leads to improved coverage for users. Therefore, users who are located near RRHs can obtain a higher SINR with less transmit power and interference, which leads to a decrease of η. For instance, in Fig. 5, when $\varepsilon_1 = 60$, only one RRH is switched on and traffic offloading is more than that of when $\varepsilon_1 = 130$ and two RRHs are switched on. Additionally, Fig. 5 indicates that for $\varepsilon_1 = 60$ and $\varepsilon_1 = 80$, the results are close to each other. This because in both cases, due to the minimal difference in ε_1 values, only one RRH will be switched on, leading to a similar feasibility region of resource allocation and traffic offloading. Consequently, by adjusting the value of ε_1, traffic can be effectively offloaded from the C-RAN to low-power

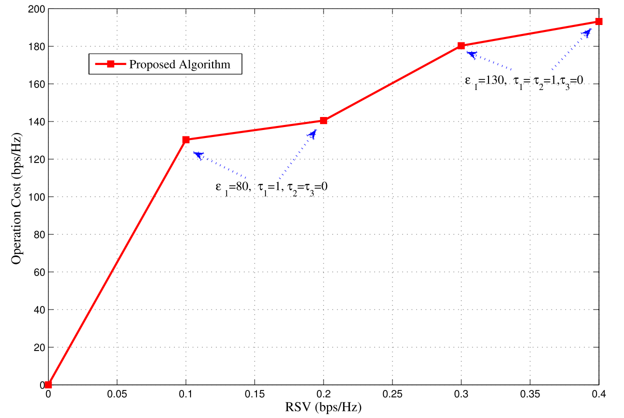

FIGURE 7. Network operation cost relative to R_n^{rsv} and $J_1 = J_2 = J_3 = 200$.

FAPs and reduce the energy consumption cost by switching off the under-utilized RRHs, BBUs, and their corresponding front-haul links, which leads to improved EE. On the other hand, according to the dynamic behavior of traffic, one of the Pareto optimal solutions can be selected by adjusting the value of ε_1 in the elastic-constraints method.

Consider the following outage probability of C1 for user $n \in \mathcal{N}$

$$\begin{aligned} &\Pr(\text{outage}) \\ &= \Pr\{\sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}} \alpha^s_{m,n}\beta_{m,n}R^s_{m,n}(\mathbf{P},\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &< R_n^{\mathrm{rsv}}\}. \end{aligned}$$

Fig. 6 shows the outage probability relative to the total number of users. As we can see, the outage probability increases as the number of users increases. We can also see that increasing ε_1 in (5) from $\varepsilon_1 = 80$ to $\varepsilon_1 = 130$ leads to more RRHs being switched on; therefore, the feasibility region of resource allocation will improve to meet the minimum guaranteed rate of each user. Consequently, it can provide better coverage and a more achievable rate of all users, which leads to a decrease in the outage probability and infeasibility of (5). However, the operation cost will increase as more RRHs are switched on.

In Fig. 7, the effect of minimum required rate of each user, e.g., R_n^{rsv}, on the operation cost is illustrated. As we can see, as there in an increase in R_n^{rsv} to meet C2, the operation cost increases, which leads to an increase in ε_1 and the number of active RRHs and energy consumption cost of RRHs. This figure also shows that immediately after $R_n^{\mathrm{rsv}} = 0.2$ bps/Hz, $\tau_1 = \tau_2 = 1$, e.g., RRH 1 and RRH 2 are switched on. Therefore, the operation cost linearly increases as R_n^{rsv} increases. For $R_n^{\mathrm{rsv}} \in [0.1, 0.2]$, this operation cost with increasing R_n^{rsv} does not have a significant increment. This is because RRH 2 and RRH 3 are switched off, while only RRH 1 is switched on.

For different values of ε_1, namely $\varepsilon_1 = 80$ and $\varepsilon_1 = 130$, the performance of our approach with a traditional wireless network scenario in terms of the total throughput.

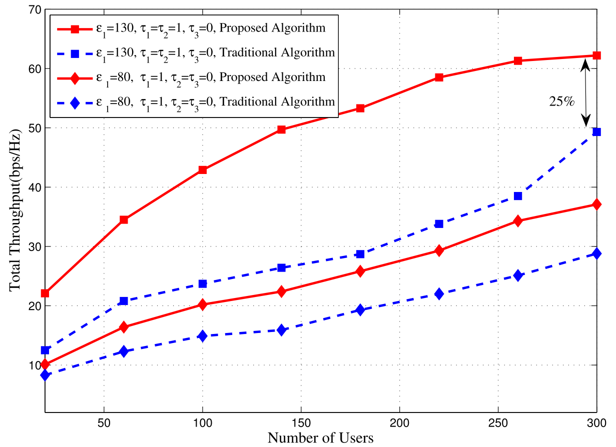

FIGURE 8. Total throughput relative to the number of users with $\varepsilon_3 = 4$ bps/Hz and $J_1 = J_2 = J_3 = 200$.

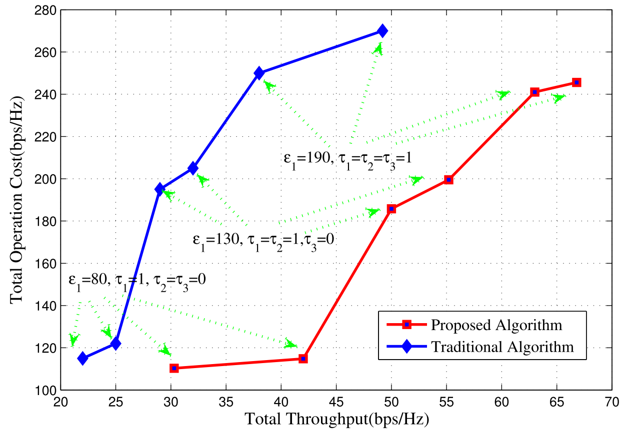

FIGURE 9. Total throughput relative to operation cost and $J_1 = J_2 = J_3 = 200$.

In the traditional approach, each user is associated to the AP (RRHs/FAPs) based on the largest value of received SINR [44]. In Fig. 8, due to multi-user diversity gain [37], the total throughput increases as the number of users increases. In both scenarios, Fig. 8 shows that when $\varepsilon_1 = 80$, the total throughput is less than that of $\varepsilon_1 = 130$. This is mainly because in the case of $\varepsilon_1 = 130$, there are more RRHs switched on. Therefore, users are assigned to the RRHs with the higher channel gain, which is due to the close distance. Hence, with less transmit power, users can obtain high data rate which leads to an improvement in the total throughput.

Moreover, as we can see in Fig. 8, the proposed approach outperforms the traditional algorithm in both cases (i.e., $\varepsilon_1 = 80$ and $\varepsilon_1 = 130$). For instance, Fig. 8 shows that when $N = 300$ and $\varepsilon_1 = 130$ in the proposed approach, the total throughput achieved is more than 25% higher than the traditional algorithm. This is because in the proposed approach effective control of inter-tier interference between APs, traffic offloading from the C-RAN to FAPs, and spectrum reuse by FAPs increase the throughput. Compared to the traditional algorithm, the AP assignment is predetermined and there is no traffic offloading. Users are assigned to RRHs, regardless of the maximum load capacity of the front-haul links, BBUs, and the upper bound limitations of allowable energy consumption cost of RRHs and BBUs. Therefore, some of users may not be connected to the network.

The Pareto optimal sets achieved by the two scenarios are compared in Fig. 9. In both scenarios, as the curve moves from left to right, the value of ε_1 increases at each step. In both scenarios, as ε_1 increases, the priority of the operation cost objective decreases; therefore, more RRHs becomes active, which leads to higher operation costs. By contrast, at the same time as ε_1 increases, the priority of the throughput objective increases, which leads to higher data rates. For instance, when data traffic is low (e.g., $N = 60$), Fig. 9 shows that by setting $\varepsilon_1 = 80$ bps/Hz, the minimum required rate of users can be satisfied with only one active RRH ($\tau_1 = 1$). Yet by increasing the data traffic (e.g., $N = 300$), the priority of the throughput objective increases compared to the operation cost objective. Hence, the value of ε_1 should increase. Therefore, by setting $\varepsilon_1 = 190$ bps/Hz, all three RRHs are activated to satisfy the minimum required rate of all users since their energy consumption cost is not important.

To compare the two scenarios in terms of the total throughput and the total operation cost, Fig. 9 shows that when the achieved total throughput is equal to 50 bps/Hz, the proposed approach has a lower total operation cost of 185.8 bps/Hz, compared to 268.3 bps/Hz with the traditional algorithm which is 30% less than it. In this case, with our proposed approach, τ_1, τ_2 and x_1 are equal to one. This means that RRH 1, RRH 2, and only one BBU with the lowest operation cost (μ_b) are switched on. By contrast, in the traditional algorithm, all three RRHs and two BBUs are switched on. This is because our proposed approach can offload traffic from the C-RAN to FAPs with low transmit power. Therefore, under-utilized RRHs and high cost BBUs can be switched off. Consequently, the total throughput and operation cost in the proposed approach will be enhanced compared to that of the traditional approach, which leads to improved EE.

All simulation results are based on randomly chosen ε_1 and ε_2 from a predetermined rang. An appropriate selection of these parameters can considerably improve network performance in terms of EE, total throughput, operation cost of H-CRANs, and outage probability. For instance, in high data traffic, when it more critical to provide higher throughput than it is to decrease energy consumption costs, we can choose higher values of ε_1. Conversely, in low data traffic, where energy consumption costs are more important than the throughput, ε_1 can be decreased. On the other hand, ε_1 is a trade-off parameter that adjusts the priority of each utility function according to the application type, QoS requirements, and traffic variations. For this reason, one of the Pareto optimal solutions can be selected by adjusting the value of trade-off parameters (e.g., ε_1 and ε_2) in the elastic-constraints method. Consequently, choosing ε_1 and ε_2 can be considered a planning design factor in resource allocation problems. Investigating the optimal value of these parameters relative to network conditions is a topic we will examine in future research.

V. CONCLUSION

In this paper, we proposed an EE MORA framework for H-CRANs to reduce operation costs, which involved introducing a novel utility function to minimize the energy consumption costs of RRHs, BBUs, and total transmit power of users. We formulated the problem as a joint AP, sub-carrier, RRH-BBU, and front-haul link assignment and power allocation optimization problem in MIMO-aided H-CRANs while ensuring that the minimum rate of each user was met. In high density H-CRANs, throughput and operation cost are considered to be two conflicting objective functions. To tackle this issue, we converted the MOO problem into an SOO problem using an elastic-constraint scalarization method, which allows for trade-off parameters and a flexible choice between increasing throughput and decreasing operation cost functions for different preferences. choosing trade-off parameters can be considered a planning design factor in resource allocation problems. Investigating the optimal value of these parameters relative to network conditions is a topic we will examine in future research. The simulation results showed that our proposed approach can offload traffic from C-RANs to FAPs with low transmit power, and reduce operation costs by switching off under-utilized RRHs and BBUs. Also, the simulation results illustrate that Pareto optimal solutions are different under diverse sets of system parameters.

APPENDIX I.

The general MOO problem is defined as

$$\begin{aligned} &\min_{x\in A}[f_1(x), f_2(x), \ldots, f_L(x)],\\ &\text{subject to}: \ g_i(x) \leq 0,\ i = 1, \ldots, O,\\ &\qquad\qquad h_j = 0,\ j = 1, \ldots, Y, \end{aligned} \tag{47}$$

where L, O, and Y represent the number of objective functions, inequality constraints, and equality constraints, respectively. The elastic-constraints method is a class of scalarization techniques, which is a generalization of both the weighted sum method and the ε-constraint method which generates all Pareto optimal solutions for MOO problems [11], [18]. Via this method, the objective functions are combined form a single objective optimization problem; therefore, (47) is formulated as

$$\begin{aligned} &\min_{x\in A}[f_e(x) + \sum_{k\neq e} t_k s_k],\\ &\text{subject to}: g_i(x) \leq 0,\ i = 1, \ldots, O,\\ &\qquad h_j = 0,\ j = 1, \ldots, Y,\\ &\qquad f_k(x) + L_k - s_k = \varepsilon_k,\ e, k \in \{1, \ldots, L\}, k \neq e,\\ &\qquad L_k, s_k \geq 0,\ x \in A, \end{aligned}$$

where $f_e(x)$ and $f_k(x)$, $k \neq e$ are objective functions in the primary MOO problem (47). By applying the elastic-constraints method, $f_e(x)$ and $f_k(x)$ are considered to be the objective function and new equality constraint, respectively. ε_k is an upper bound on violated objective value, which is used to penalize the constraint violation, and t_k is the penalty coefficient for a given objective k [45], [46]. The elastic-constraints method uses two sets of variables, including slack variables, L_k, and surplus variables, s_k, in order to transform the upper bounds on objective values into equality constraints for any $x \in A$ (i.e., a set of feasible solutions) based on an appropriate selection of s_k and L_k. In the other words, the values of ε_k specify the priority of objective functions relative to each other, and the Pareto optimal solutions can be derived by considering different ε_k for the objective functions.

APPENDIX II.

From the assumption of $J_m \gg N_m(t_1)$, we will have $\frac{J_m - N_m(t_1)+1}{N_m(t_1)} \approx \frac{J_m}{N_m(t_1)}$ [35]. Then, we can rewrite (7) as

$$\begin{aligned} &\min_{\mathfrak{S}, L_i, s'_i}[-\sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^s_{m,n}\beta_{m,n}R^s_{m,n}(\mathbf{P}(t), \boldsymbol{\beta})\\ &\quad \sum_{m\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^s_{m,n}\beta_{m,n}\log_2(\frac{J_m}{N_m(t_1)}\gamma^s_{m,n}(t)) + \sum_{i=1}^{3}\pi_i s'_i],\\ &\text{subject to: C2} - \text{C14}. \end{aligned} \tag{48}$$

Let us rewrite $\log_2(\frac{J_m}{N_m(t_1)}\gamma^s_{m,n}(t_1)) \approx \log_2(J_m\gamma^s_{m,n}(t_1)) - \log_2(N_m(t_1))$ and substitutes it in (48). (48) is not in a GP standard form because throughput is a logarithm function, which is a non-linear function. We apply DC approximation and obtain a linear approximation of $\log_2(N_m(t_1))$ as

$$\begin{aligned} &\log_2(N_m(t_1))\\ &\approx \log_2(N_m(t_1-1))\\ &\quad + \nabla\log_2(N_m(t_1-1))(N_m(t_1) - N_m(t_1-1)), \end{aligned} \tag{49}$$

where $N_m = \sum_{n\in\mathcal{N}} \beta_{m,n}$. Further simplifying (49), we have

$$\begin{aligned} &\log_2(N_m(t_1))\\ &\approx \log(N_m(t_1-1))\\ &+ \sum_{n\in\mathcal{N}}\frac{\beta_{m,n}(t_1)}{\sum_{n\in\mathcal{N}}\beta_{m,n}(t_1-1)} - \sum_{n\in\mathcal{N}}\frac{\beta_{m,n}(t_1-1)}{\sum_{n\in N}\beta_{m,n}(t_1-1)}, \end{aligned} \tag{50}$$

where via substituting (50) into (48), we will have (8).

APPENDIX III.

We have four steps: 1) C2 transforms to $\widetilde{\text{C}}2.1$ and $\widetilde{\text{C}}2.2$, 2) C5, C9 and C10 transform to $\widetilde{\text{C}}5$, $\widetilde{\text{C}}9$ and $\widetilde{\text{C}}10$ 3) C12 − C14 transform to $\widetilde{\text{C}}12.1$ − $\widetilde{\text{C}}14.1$ and $\widetilde{\text{C}}12.2$ − $\widetilde{\text{C}}14.2$ and 4) objective function (8) is converted into (9).

Step 1: For this step we have two cases:

- When $m \in \mathcal{F}$, we can rewrite C2 as

$$\sum_{m\in\mathcal{F}}\sum_{s\in\mathcal{S}}\alpha^{s}_{m,n}(t_1)\beta_{m,n}(t_1)\widetilde{R}^{s}_{m,n}(\mathbf{P}(t),\boldsymbol{\beta}) \geq R^{\mathrm{rsv}}_{n},$$

By applying AGMA, we come to $\widetilde{\mathrm{C}}2.1$ in (9).

- Based on (6), when $m \in \mathcal{R}$, C2 is not in a GP standard form because throughput is a non-linear function. Therefore, we apply DC approximation and by substituting (50) into C2, we have

$$\sum_{m\in\mathcal{R}}\sum_{s\in\mathcal{S}}\alpha^{s}_{m,n}(t_1)\beta_{m,n}(t_1)[\log_2(J_m\gamma^{s}_{m,n}(t) - \log_2(N_m(t_1-1)) - \sum_{n\in\mathcal{N}}\frac{\beta_{m,n}(t_1)}{\sum\limits_{n\in\mathcal{N}}\beta_{m,n}(t_1-1)} + \sum_{n\in\mathcal{N}}\frac{\beta_{m,n}(t_1-1)}{\sum\limits_{n\in\mathcal{N}}\beta_{m,n}(t_1-1)}] \geq R^{\mathrm{rsv}}_{n}, \quad \forall n\in\mathcal{N}. \tag{51}$$

Now, by applying AGMA, we come to $\widetilde{\mathrm{C}}2.2$ in (9).

Step 2: Due to the negative terms in C5, C9, and C10, they do not satisfy the properties of posynomials in GP formulations. Therefore, by adding 1 to both the left and right hand sides of C5, C9, C10, we have C5 : $\alpha^{s}_{m,n}+1 \leq \beta_{m,n}+1$, C9 : $z_{r,b}+1 \leq x_b+1$, C10 : $\tau_r+1 \leq \sum_{b\in\mathcal{B}} z_{r,b}+1$. Now, by using AGMA, we get the monomial approximation for C5, C9, and C10 as $\widetilde{\mathrm{C}}5$, $\widetilde{\mathrm{C}}9$, and $\widetilde{\mathrm{C}}10$, respectively in (9) [11].

Step 3: At iteration t_1, we can rewrite C12 − C14 as $U_i + L_i = s'_i + \varepsilon_i$ for ı=[1,2,3], which are not monomial functions. Therefore, we use an auxiliary variable $q_i \geq 0$ to relax and convert C12 − C14 into the posynomial inequalities as

$$U_i + L_i \leq q_i(t_1) \leq s'_i + \varepsilon_i, \quad \text{for i=[1,2,3]}. \tag{52}$$

(52) can be rewritten as $\frac{U_i+L_i}{q_i(t_1)} \leq 1$ and $\frac{q_i(t_1)}{s'_i+\varepsilon_i} \leq 1$. Since the above constraints do not satisfy the properties of posynomial functions, we approximate them to posynomial functions by using AGMA for i = [1], [2], [3] as follows:

$$\widetilde{\mathrm{C}}12.1 - \widetilde{\mathrm{C}}14.1 : q_i^{-1}(t_1)U_i(t_1) + q_i^{-1}(t_1)L_i(t_1) \leq 1,$$

and

$$\widetilde{\mathrm{C}}12.2 - \widetilde{\mathrm{C}}14.2 : q_i(t_1) \times (\frac{\varepsilon_i}{e_i(t_1)})^{-e_i(t_1)} \times (\frac{s'_i(t_1)}{d_i(t_1)})^{-d_i(t_1)} \leq 1,$$

where $e_i(t_1)$ and $d_i(t_1)$ are introduced in (33) and (34), respectively.

Step 4: To obtain the positive conditions of objective function (8) in GP, we use the positive auxiliary variable $\varpi_0(t_1)$ and $\Lambda_1 \gg 1$ to define the following constraint

$$\mathrm{C00} : \Lambda_1 - \sum_{m\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^{s}_{m,n}(t_1)\beta_{m,n}(t_1)\widetilde{R}^{s}_{m,n}(\mathbf{P}(t),\boldsymbol{\beta}) - \sum_{m\in\mathcal{R}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^{s}_{m,n}(t_1)\beta_{m,n}(t_1)[\log_2(J_m\gamma^{s}_{m,n}(t)) - \Gamma_{m,n}(t_1) + \Gamma_{m,n}(t_1-1) - \log_2(N_m(t_1-1))] + \sum_{i=1}^{3}\pi_i s'_i) \leq \varpi_0.$$

This can be rewritten as

$$\frac{\Lambda_1 + I_1(t_1) + \sum\limits_{i=1}^{3}\pi_i s'_i}{\varpi_0(t_1) + I_2(t_1)} \leq 1, \tag{53}$$

where $I_1(t_1)$ and $I_2(t_1)$ are introduced in (11) and (12), respectively. Now, (53) is always positive. Finally, the equivalent optimization problem becomes

$$\min_{\omega(t_1)} \varpi_0\,(t_1)$$
$$\text{subject to: C00, C2} - \text{C14}, \tag{54}$$

where $\omega(t_1)$ is defined in (10). Since C00, C2, C5, C9, C10 and C12 − C14 do not have standard form of GP formulations, we approximate them by using AGMA to arrive at (9).

APPENDIX IV.

We have three steps: 1) C2 converts to $\widetilde{\mathrm{C}}2.1$, and $\widetilde{\mathrm{C}}2.2$, 2) C0 is transformed into $\widetilde{\mathrm{C}}0.1$ and $\widetilde{\mathrm{C}}0.2$, and 3) objective function (36) is transformed into (37).

Step 1: For this step we have two cases:

- when $m \in \mathcal{F}$, we can rewrite C2 as: $\log_2(\prod_{\substack{m\in\mathcal{F}\\ s\in\mathcal{S}}}(1+\frac{p^{s}_{m,n}(t_2)h^{s}_{m,n}}{\sigma^2+I^{s}_{m,n}(t_2)})^{-1}) \leq -R^{\mathrm{rsv}}_{n}$, which can be mathematically represented as

$$\prod_{\substack{m\in\mathcal{F}\\ s\in\mathcal{S}}}(\frac{\sigma^2+I^{s}_{m,n}(t_2)}{\sigma^2+I^{s}_{m,n}(t_2)+p^{s}_{m,n}(t_2)h^{s}_{m,n}}) \leq 2^{-R^{\mathrm{rsv}}_{n}}.$$

By using AGMA, we come to $\widetilde{\mathrm{C}}2.1$ in (37).

- Based on (6), when $m \in \mathcal{R}$, by assuming $J_m \gg N_m(t)$, we will have $\frac{J_m-N_m(t)+1}{N_m(t)} \approx \frac{J_m}{N_m(t)}$ and we can rewrite C2 as:

$$\sum_{m\in\mathcal{R}}\sum_{s\in\mathcal{S}}\alpha^{s}_{m,n}(t)\beta_{m,n}(t)\left(\log_2(\frac{\frac{J_m}{N_m(t)}p^{s}_{m,n}(t_2)h^{s}_{m,n}}{\sigma^2+I^{s}_{m,n}(t_2)})\right) \geq R^{\mathrm{rsv}}_{n}.$$

This can be mathematically represented as

$$\log_2\prod_{\substack{m\in\mathcal{R}\\ s\in\mathcal{S}}}\alpha^{s}_{m,n}(t)\beta_{m,n}(t)(\frac{\frac{J_m}{N_m(t)}p^{s}_{m,n}(t_2)h^{s}_{m,n}}{\sigma^2+I^{s}_{m,n}(t_2)}) \geq R^{\mathrm{rsv}}_{n}.$$

Therefore, we arrive at $\widetilde{\mathrm{C}}2.2$ in (37).

Step 2: Similar to Step 3 in Appendix III, at iteration t_2, we can rewrite C0 as

$$\mu_{\mathrm{p}}\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\sum_{s\in\mathcal{S}}\alpha^{s}_{m,n}(t)\beta_{m,n}(t)p^{s}_{m,n}(t_2)+L_4(t_2)=s'_4(t_2)+\epsilon_3.$$

This is not monomial function. Hence, we apply an auxiliary variable $q_p \geq 0$ to transform C0 into the posynomial inequalities as

$$\mu_{\mathrm{p}} \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t)p^s_{m,n}(t_2) + L_4(t_2) \leq q_p(t_2) \leq s'_4(t_2) + \epsilon_3, \tag{55}$$

(55) can be rewritten as $\frac{\mu_{\mathrm{p}} \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t)p^s_{m,n}(t_2)+L_4(t_2)}{q_p(t_2)} \leq 1$ and $\frac{q_p(t_2)}{s'_4(t_2)+\epsilon_3} \leq 1$. Now, by using AGMA, we convert them into monomial functions as

$$\widetilde{\mathrm{C}}0.1: \; q_p(t_2) \times (\frac{\epsilon_3}{e(t_2)})^{-e(t_2)} \times (\frac{s'_4(t_2)}{d(t_2)})^{-d(t_2)} \leq 1,$$

$$\widetilde{\mathrm{C}}0.2: \; q_p^{-1}(t_2)\mu_{\mathrm{p}} \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t)p^s_{m,n}(t_2) + q_p^{-1}(t_2)L_4(t_2) \leq 1,$$

where $e(t_2)$ and $d(t_2)$ are introduced in (43) and (44), respectively.

Step 3: we can rewrite objective function in (36) as

$$\min_{\mathbf{P}}[-\sum_{m\in\mathcal{R}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t) \log_2(\frac{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2)}) - \sum_{m\in\mathcal{F}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t)\log_2(1 + \frac{p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2)}) + \mu_{\mathrm{a}} \sum_{r\in\mathcal{R}} \tau_r J_r(t) + \sum_{b\in\mathcal{B}} \mu_b \times x_b(t) + \pi_4 s'_4(t_2)]. \tag{56}$$

To obtain a standard GP formulation, the objective function in (56) must be transformed into a positive term. Hence, we use the positive auxiliary variable $\varpi_1(t_2)$ and $\Lambda_2 \gg 1$, and rewrite (56) as

$$\mathrm{C}01: \; \Lambda_2 - \sum_{m\in\mathcal{R}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t) \times \log_2(\frac{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2)}) - \sum_{m\in\mathcal{F}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t)\log_2(1 + \frac{p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2)}) + \mu_{\mathrm{a}} \sum_{r\in\mathcal{R}} \tau_r J_r(t) + \sum_{b\in\mathcal{B}} \mu_b \times x_b(t) + \pi_4 s'_4(t_2) \leq \varpi_1(t_2).$$

Since throughput is a non-linear function, C01 is a non-posynomial constraint. We use DC approximation of logarithmic functions to solve this:

$$\begin{aligned}
\mathrm{C}01: \; & \Lambda_2 + \mu_{\mathrm{a}} \sum_{r\in\mathcal{R}} \tau_r J_r(t) + \sum_{b\in\mathcal{B}} \mu_b \times x_b(t) + \pi_4 s'_4(t_2) \\
& + \sum_{m\in\mathcal{F}} \sum_{n\in\mathcal{N}} \sum_{s\in\mathcal{S}} [\alpha^s_{m,n}(t)\beta_{m,n}(t) \\
& \times (\frac{h^s_{m,n}}{\sigma^2 + \sum_{m\in\mathcal{F}} \sum_{n\in\mathcal{N}} p^s_{m,n}(t_2-1)h^s_{m,n}} \\
& \times \frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)} - \frac{h^s_{m,n}}{\sigma^2 + \sum_{m\in\mathcal{F}} \sum_{n\in\mathcal{N}} p^s_{m,n}(t_2-1)h^s_{m,n}} \\
& \times \frac{p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2)} - \log_2(1 + \frac{p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)})] \\
& + \sum_{m\in\mathcal{R}} \sum_{n\in\mathcal{N}} \\
& \times \sum_{s\in\mathcal{S}} \alpha^s_{m,n}(t)\beta_{m,n}(t)[-\log_2(\frac{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)}) \\
& - \frac{1}{p^s_{m,n}(t_2-1)} \times \frac{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)} + \frac{1}{p^s_{m,n}(t_2-1)} \\
& \times \frac{\frac{J_m}{N_m(t)}p^s_{m,n}(t_2-1)h^s_{m,n}}{\sigma^2 + I^s_{m,n}(t_2-1)}] \leq \varpi_1(t_2).
\end{aligned}$$

Now, by applying AGMA approximations, C01 can be transformed into a posynomial constraint. Finally, we come to the following equivalent optimization problem

$$\begin{aligned} & \min_{\mathbf{P}} \varpi_1(t_2) \\ & \text{subject to}: \; \mathrm{C}0, \mathrm{C}01, \mathrm{C}1, \mathrm{C}2, \mathrm{C}7, \mathrm{C}8. \end{aligned} \tag{57}$$

Since constraints C0, C01, C2, C7, C8 are not standard form of GP, we use AGMA approximations for them to reach (37).

REFERENCES

[1] M. Usama and M. Erol-Kantarci, "A survey on recent trends and open issues in energy efficiency of 5G," *Sensors*, vol. 19, pp. 1–23, Jul. 2019.

[2] S. Malathy, P. Jayarajan, H. Ojukwu, F. Qamar, M. N. Hindia, K. Dimyati, K. A. Noordin, and I. S. Amiri, "A review on energy management issues for future 5G and beyond network," *Wireless Netw.*, vol. 27, no. 4, pp. 2691–2718, Apr. 2021.

[3] I. B. Sofi and A. Gupta, "A survey on energy efficient 5G green network with a planned multi-tier architecture," *J. Netw. Comput. Appl.*, vol. 118, pp. 1–28, Sep. 2018.

[4] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Feb. 2013.

[5] L. Wang, K.-K. Wong, M. Elkashlan, A. Nallanathan, and S. Lambotharan, "Secrecy and energy efficiency in massive MIMO aided heterogeneous C-RAN: A new look at interference," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 8, pp. 1375–1389, Dec. 2016.

[6] K. Wang, Y. Zhou, J. Li, L. Shi, W. Chen, and L. Hanzo, "Energy-efficient task offloading in massive MIMO-aided multi-pair fog-computing networks," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2123–2137, Apr. 2021.

[7] H. R. Chi and A. Radwan, "Multi-objective optimization of green small cell allocation for IoT applications in smart city," *IEEE Access*, vol. 8, pp. 101903–101914, 2020.

[8] T. Z. Oo, N. H. Tran, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Offloading in HetNet: A coordination of interference mitigation, user association and resource allocation," *IEEE Trans. Mobile Comput.*, vol. 16, no. 8, pp. 2276–2291, Aug. 2017.

[9] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[10] N. Yu, Z. Song, H. Du, H. Huang, and J. Xiaohua, "Dynamic resource provisioning for energy efficient cloud radio access networks," *IEEE Trans. Cloud Comput.*, vol. 7, no. 4, pp. 964–974, Dec. 2019.

[11] J. Tang, R. Wen, T. Q. S. Quek, and M. Peng, "Fully exploiting cloud computing to achieve a green and flexible C-RAN," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 40–46, Nov. 2017.

[12] H. Safi, A. M. Montazeri, J. Rostampoor, and S. Parsaeefard, "Spectrum sensing and resource allocation for 5G heterogeneous cloud radio access networks," *IET Commun.*, vol. 16, pp. 358–384, Feb. 2022.

[13] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[14] M. Baghani, S. Parsaeefard, and Le-Ngoc, "Multi-objective resource allocation in density aware designed of C-RAN in 5G," *IEEE Access*, vol. 6, pp. 45177–45190, 2018.

[15] O. Amin, E. Bedeer, M. H. Ahmed, and O. A. Dobre, "Energy efficiency–spectral efficiency tradeoff: A multiobjective optimization approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 1975–1981, Apr. 2016.

[16] A. Jahid, M. A. Alsharif, P. Uthansakul, J. Nebhen, and A. Aly, "Energy efficient throughput aware traffic load balancing in green cellular networks," *IEEE Access*, vol. 9, pp. 90587–90602, 2021.

[17] A. Khalili, M. R. Mili, M. Rasti, and S. Parsaeefard, "Antenna selection strategy for energy efficiency maximization in uplink OFDMA networks: A multi-objective approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 595–609, Jan. 2020.

[18] J.-H. Cho, Y. Wang, R. Chen, K. S. Chan, and A. Swami, "A survey on modeling and optimizing multi-objective systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1867–1901, 3rd Quart., 2017.

[19] H. öztop, M. F. Tasgetiren, D. Eliiyi, Q. Pan, and L. Kandiller, "An energy-efficient permutation flowshop scheduling problem," *Expert Syst. Appl.*, vol. 150, Feb. 2020, Art. no. 113279.

[20] M. Chiang, "Geometric programming for communication systems," *Found. Trends Commun. Inf. Theory*, vol. 2, nos. 1–2, pp. 1–154, Aug. 2005.

[21] M. Chiang, C. W. Tan, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, Jul. 2007.

[22] G. Xu, "Global optimization of signomial geometric programming problems," *Eur. J. Oper. Res.*, vol. 233, no. 3, pp. 500–510, 2014.

[23] *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: http://cvxr.com/cvx

[24] D. D. Olatinwo, A. M. Abu-Mahfouz, and G. P. Hancke, "A hybrid multi-class MAC protocol for IoT-enabled WBAN systems," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6761–6774, Mar. 2021.

[25] H. Liu, X. Cao, Q. Deng, D. Pei, and J. Wang, "Multi-objective optimization of resource allocation for uplink transmission in two-tier heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Smart Internet Things (SmartIoT)*, Aug. 2019, pp. 275–282.

[26] Y. Li, M. Sheng, C. Yang, and X. Wang, "Energy efficiency and spectral efficiency tradeoff in interference-limited wireless networks," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1924–1927, Aug. 2013.

[27] Y. Li, M. Sheng, X. Wang, Y. Zhang, and J. Wen, "Max–min energy-efficient power allocation in interference-limited wireless networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 9, pp. 4321–4326, Sep. 2015.

[28] Z. Zhou, M. Dong, K. Ota, G. Wang, and L. T. Yang, "Energy-efficient resource allocation for D2D communications underlaying cloud-RAN-based LTE-A networks," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 223–234, Jun. 2016.

[29] Y. Hao, Q. Ni, H. Li, and S. Hou, "Energy and spectral efficiency tradeoff with user association and power coordination in massive MIMO enabled HetNets," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2091–2094, Oct. 2016.

[30] Y. Hao, Q. Ni, H. Li, and S. Hou, "On the energy and spectral efficiency tradeoff in massive MIMO-enabled HetNets with capacity-constrained backhaul links," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4720–4733, Nov. 2017.

[31] P. Luong, F. Gagnon, C. Despins, and L. Tran, "Optimal joint remote radio head selection and beamforming design for limited fronthaul C-RAN," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5605–5620, Aug. 2017.

[32] H. M. Al-Obiedollah, K. Cumanan, J. Thiyagalingam, J. Tang, A. G. Burr, Z. Ding, and O. A. Dobre, "Spectral-energy efficiency trade-off-based beamforming design for MISO non-orthogonal multiple access systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 6593–6606, Oct. 2020.

[33] F. Xu and H. Zhang, "Energy efficiency and spectral efficiency tradeoff in IRS-assisted downlink mmWave NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1433–1437, Jul. 2022.

[34] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "User association and load balancing for cellular massive MIMO," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2014, pp. 1–10.

[35] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband processing units virtualization for cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 189–192, Apr. 2015.

[36] V. Pareto, *The Rise and Fall of the Elites: An Application of Theoretical Sociology*. New York, NY, USA: Bedminster Press, vol. 48, no. 1, 1969.

[37] 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.

[38] A. Abdelnasser, E. Hossain, and D. I. Kim, "Tier-aware resource allocation in OFDMA macrocell-small cell networks," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 695–710, Mar. 2015.

[39] M. Hasan and E. Hossain, "Distributed resource allocation in D2D-enabled multi-tier cellular networks: An auction approach," in *Proc. IEEE ICC*, London, U.K., Jun. 2015, pp. 2949–2954.

[40] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Ph.D. dissertation, Dept. Elect. Eng., Univ. Minnesota, Minneapolis, MN, USA, 2014.

[41] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, "Joint subchannel assignment and power allocation for OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 342–355, Jan. 2014.

[42] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[43] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[44] S. Parsaeefard, V. Jumba, A. D. Shoaei, M. Derakhshani, and T. Le-Ngoc, "User association in cloud RANs with massive MIMO," *IEEE Trans. Cloud Comput.*, vol. 9, no. 2, pp. 821–833, Apr. 2021.

[45] M. Ehrgott, "A discussion of scalarization techniques for multiple objective integer programming," *Ann. Oper. Res.*, vol. 147, no. 1, pp. 343–360, Aug. 2006.

[46] K. Klamroth and T. Jørgen, "Constrained optimization using multiple objective programming," *J. Global Optim.*, vol. 37, no. 3, pp. 325–355, Mar. 2007.

NAHID AMANI received the B.S. degree in computer engineering from the Amirkabir University of Technology, in 2000, and the M.S. degree in computer engineering from the Iran University of Science and Technology, Tehran, Iran, in 2003. She is currently pursuing the Ph.D. degree in computer engineering with the Amirkabir University of Technology. She is a Faculty Member with the Iran Telecommunication Research Center. Her current research interest includes resource allocation in wireless networks.

SAEEDEH PARSAEEFARD (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Tarbiat Modares University, Tehran, Iran, in 2012. From November 2010 to October 2011, she was a visiting Ph.D. student with the University of California at Los Angeles, Los Angeles, CA, USA. She was also a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada, from 2013 to 2015. She is currently a Research Scientist and a Visiting Faculty Member with the University of Toronto, Toronto, ON, Canada. She has published more than 50 journal articles and conference papers all in the best venue of IEEE Transactions and ComSoc flag conferences. Her research has been cited more than 1000 times with H-index of 16 and I10- index of 27. She is the coauthor of two books in *Wireless Network Virtualization* and *Robust Resource Allocation in Future Wireless Networks* (Springer). Her research interests include applying optimization theory, game theory, machine learning tools for better understanding and analyzing interactions of heterogonous, non-cooperative, and distributed multi-agent systems in uncertain environments, in particular wireless networks (5G and 6G), the IoT, the IIoT, and URRLC applications. She was a recipient of the IEEE Women in Engineering Award (in region eight), in 2018.

HALIM YANIKOMEROGLU (Fellow, IEEE) received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 1990, and the M.A.Sc. degree in electrical engineering (now ECE) and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Canada, in 1992 and 1998, respectively.

Since 1998, he has been with the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, where he is currently a Full Professor. He has given more than 110 invited seminars, keynotes, panel talks, and tutorials in the last five years. He has supervised or hosted over 150 postgraduate researchers in his laboratory with Carleton University. His extensive collaborative research with industry resulted in 39 granted patents. His research interests include wireless communications and networks, with a special emphasis on non-terrestrial networks (NTNs).

Dr. Yanikomeroglu is a Fellow of the Engineering Institute of Canada (EIC) and the Canadian Academy of Engineering (CAE). He is also a member of the IEEE ComSoc Governance Council, IEEE ComSoc GIMS, IEEE ComSoc Conference Council, and IEEE PIMRC Steering Committee. He is also a Distinguished Speaker of the IEEE Communications Society and the IEEE Vehicular Technology Society and an Expert Panelist of the Council of Canadian Academies (CCA|CAC). He received several awards for his research, teaching, and service, including the IEEE ComSoc Fred W. Ellersick Prize, in 2021, IEEE VTS Stuart Meyer Memorial Award, in 2020, and IEEE ComSoc Wireless Communications TC Recognition Award, in 2018. He received the Best Paper Awards at IEEE Competition on Non-Terrestrial Networks for B5G and 6G (grand prize), in 2022, IEEE ICC 2021, and IEEE WISEE 2021 and 2022. He is currently serving as the Chair for the Steering Committee of IEEE's Flagship Wireless Event and Wireless Communications and Networking Conference (WCNC). He served as the general chair and the technical program chair for several IEEE conferences. He has also served on the editorial boards for various IEEE periodicals.