IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 20, 2018, accepted June 26, 2018, date of publication August 10, 2018, date of current version September 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2861909

Multi-Objective Resource Allocation in Density-Aware Design of C-RAN in 5G

MINA BAGHANI[1], SAEEDEH PARSAEEFARD[1], (Member, IEEE), AND THO LE-NGOC[2], (Fellow, IEEE)
[1]Communication Technologies and Department, ITRC, Tehran 14155-3961, Iran
[2]Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9, Canada

Corresponding author: Saeedeh Parsaeefard (saeede.parsaeefard@gmail.com)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada Grant and in part by the Iran Telecommunication Research Center under Project 907950200.

ABSTRACT In this paper, a multi-objective resource allocation algorithm in a novel density-aware design of virtualized software-defined cloud radio access network (C-RAN) is proposed. We consider two design modes based on the average density of users: 1) high-density mode when a large number of low-cost remote radio heads (RRHs) without baseband processing capability are controlled by one single base station and 2) low-density mode when a small number of RRHs with baseband processing capability are deployed. In high-density mode, the challenge of front-haul capacity limitation is tackled via separating control plane and data plane in a heterogeneous structure. Besides, the fully centralized processing and management, and energy-efficient use of infrastructure in low traffic time by turning off RRHs are achieved. In the low-density mode, the transmission delay due to the large distance between the sparse RRHs and cloud unit, is more critical. This practical issue is handled by sharing the baseband processing and resource management among these units in a hierarchical structure. This resulting heterogeneous /hierarchical virtualized software-defined cloud-RAN (HVSD-CRAN) offers various tradeoffs in resource management objectives such as throughput and delay versus power and cost. Consequently, we resort to multi-objective optimization theory to propose a resource allocation framework in HVSD-CRAN.

INDEX TERMS Density-aware RAN design, function splitting, multi-objective resource management.

I. INTRODUCTION

Fifth generation wireless networks (5G) objects to support three different types of services: mobile broadband (MBB), mission-critical machine communication (MCC) and massive machine-type communication (mMTC) [1] where ultra high throughput, considerably low delay and massive connection of users are their three distinguishable goals, respectively. To attain them, all parts of 5G architecture and specially a radio access network (RAN) should be revolutionized where 5G leverages Cloud-RAN (C-RAN), a software-defined and virtualized structure to reach this objective.

C-RAN is a promising structure for 5G RAN proposed to decrease capital expenditure (CAPEX) an operating expenses (OPEX) and achieve higher quality of service (QoS) for users [2]. In C-RAN, all base-station (BS) baseband processing functions are deployed in a centralized baseband unit (BBU) and connected to the remote radio-heads (RRHs) through a front-haul link. On the other hand, to support the expansion of new applications and different utilization patterns of users, 5G needs a flexible structure adapted to the different users' requirements and services. A software-defined structure is a promising approach to design the flexible RAN [3]. High infrastructure utilization is another important design object in 5G from the network efficiency perspective, which can be addressed via a virtualization concept where the infrastructure of one operator can be shared between different service providers (SPs) with the aim to decrease costs of new 5G services [4]. In a virtualized wireless networks, the cloud, transmission and infrastructure resources, are sliced into multiple SPs. However, the preserving isolation between slices, i.e., the activity of users of one slice does not affect the QoS of users of the other slices, is necessary for virtualization. Consequently, the structure of new RAN in 5G can be considered as a combination of these three techniques [5]–[11]. However, disadvantages of each approach should be concurred to reach an appropriate structure.

5G is obligated to provide the various services for a large number of users with diverse QoS requirements, e.g., minimum required rate and maximum tolerable delay. On the

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

other hand, to reach the appropriate profit of SPs, 5G should be deployed with the minimum cost while considering the maximum energy efficiency and network utilization. To reach these conflicting goals in 5G RAN design, the density of users should be considered as one of the important parameters here. In this paper, we propose a density-aware C-RAN design, called the heterogeneous/hierarchical virtualized software-defined cloud RAN (HVSD-CRAN) for 5G where there exist a mixture of high-density and low-density regions.

In a high-density region with many RRHs, due to existence of high interference between RRHs, the resource management should be deployed in a centralized manner. On the other hand, due to the high data traffic, the front-haul capacity limitation is more critical, which can be tackled by decoupling the control-traffic and considering one control BS and a set of Data RRHs (D-RRHs). By this topology, in the low data traffic time, a subset of D-RRHs can be turned off to decrease energy consumption without disturbing the coverage. Also, the resource management is deployed in centralized manner via information gathered in the control BS. This topology is heterogeneous due to the use of control BS and D-RRHs.

In low-density region where RRHs are deployed in a sparse manner, using a centralized BBU cloud leads to high transmission delay due to the large distance between RRHs and cloud. To overcome this challenge, the function splitting between cloud and RRHs is proposed where RRHs should have processing capability called remote radio systems (RRSs) [12], [13]. However, this approach increases the cost of RRHs, which should be considered as one of the criteria to design RAN in 5G. Also, in this case, the fully centralized resource management is not an appropriate approach due to the large delay to report the channel state information (CSI) to the BBU cloud. However, due to the existence of virtualization isolation constraints for each slice, the fully distributed resource allocation cannot be attained. To hold this type of constraints while reaching to the distributed resource allocation algorithm to the maximum extend, we apply the Lagrange dual decomposition algorithm and propose the semi-distributed resource management for low-density region. As a result, RANs of low-density region have a hierarchical processing and management structure.

Another important aspect of RAN design in 5G is that there exists a large set of trade-off parameters and objectives related to the three mentioned categories of services. Therefore, the resource allocation problems are dealing with multi conflicting objective functions [14], [15]. Thus, a framework of the multi-objective resource allocation (MORA) problem is investigated in this paper to propose MORA algorithm in HVSD-CRAN. In a high-density mode, supporting all users by higher throughput and lower power consumption is critical. Therefore, maximizing throughput and minimizing number of active RRHs to reduce the consumed power are selected as two objective functions of MORA problem and a centralized D-RRHs, sub-carrier and power assignment algorithm is proposed. In a low-density mode, the delay and the cost of baseband processing at RRSs in the hierarchical processing topology are considered as the objective functions of MORA problem. As mentioned earlier, delay is a very limiting criteria, which is overcome by hierarchically processing in low-density mode of proposed system model. However, the processing at RRSs is costly due to their limited available resources e.g., processing units and power supply. As a result, these two objectives are conflicting. In this case, a semi-distributed resource allocation algorithm is proposed in which determining the level of splitting baseband processing of each RRS, sub-carrier and power allocation between users are performed. The weighted sum method is used to solve the multi-objective optimization (MOO) problems for both cases [16]. By changing the weights of objective functions or equivalently their priorities, a set of Pareto optimal solutions is derived.

This paper is organized as follows. In Section II, the previous works in 5G RAN design and MORA problems are surveyed. Section III describes the proposed HVSD-CRAN model. Two MORA problems related to the two modes of HVSD-CRAN are represented in Section IV. In Section V, via simulation results, the performance of the proposed resource allocation algorithms is studied. Concluding remarks are given in Section VI.

II. C-RAN & MULTI-OBJECTIVE OPTIMIZATION PROBLEMS: A BRIEF LITERATURE REVIEW

Since we propose density-aware C-RAN structure in 5G, we initialize our discussion with reviewing proposed C-RAN structures in this context. Since we develop MORA algorithms for two modes of HVSD-CRAN, the MOO problems in the literatures are reviewed in the second subsection II-B.

A. OVERVIEW OF RAN STRUCTURE IN 5G

There exists a surge of research to revolutionize RAN structure in order to attain QoS requirements of 5G services. Although it is not easy to overview all these proposed structures and categorize them [17], it is obvious that for RAN structure three main promising proposed techniques are C-RAN, software-defined and virtualized structures. Fig. 1 presents a time-line of proposed RAN structures.

The concept of C-RAN is first proposed in 2011 [2] where each BS consists of three parts: 1) RRH which only receives and sends RF signals to users without any baseband processing ability, 2) cloud unit which is responsible to process the baseband functions through RF signals, and 3) Front-haul link which is an interface between RRH and cloud unit. One of the major implementation challenges in C-RAN is to develop the high capacity and low-delay front-haul links [17]. In later years, different RAN structures are proposed to overcome this challenge, e.g., [7], [9].

Next, the software-defined structure for RAN is proposed in [3]. Basically, the flexible programable structure can be achieved by the software-defined approach in which the control plane and data plane are decoupled [18]. This concept is borrowed from core network to provide a centralized management layer in the software-defined network controller entity

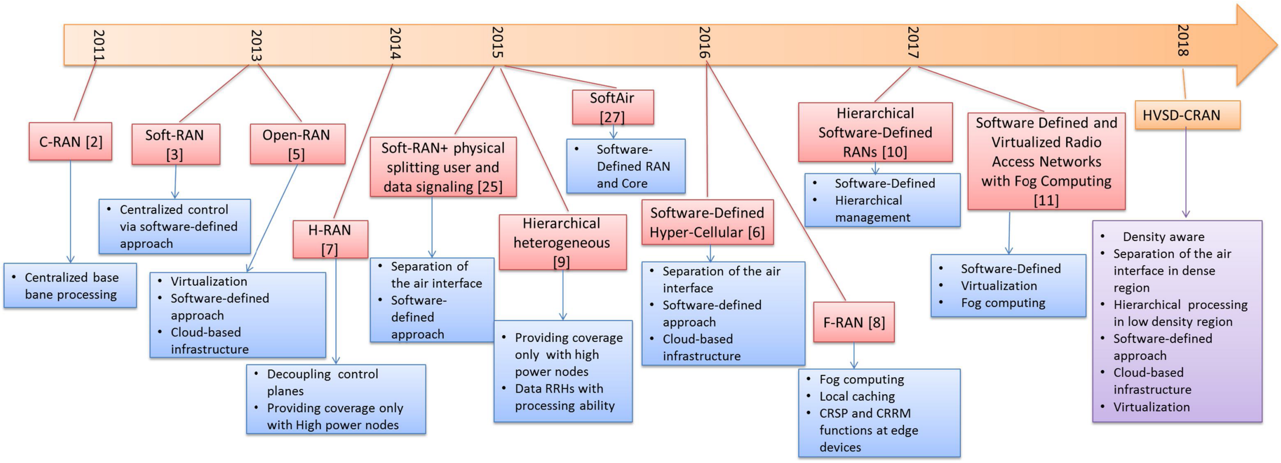

2011
2013
2014
2015
2016
2017
2018
C-RAN [2]
Centralized base bane processing
Soft-RAN [3]
Centralized control via software-defined approach
Open-RAN [5]
Virtualization
Software-defined approach
Cloud-based infrastructure
H-RAN [7]
Decoupling control planes
Providing coverage only with High power nodes
Soft-RAN+ physical splitting user and data signaling [25]
Separation of the air interface
Software-defined approach
SoftAir [27]
Software-Defined RAN and Core
Hierarchical heterogeneous [9]
Providing coverage only with high power nodes
Data RRHs with processing ability
Software-Defined Hyper-Cellular [6]
Separation of the air interface
Software-defined approach
Cloud-based infrastructure
F-RAN [8]
Fog computing
Local caching
CRSP and CRRM functions at edge devices
Hierarchical Software-Defined RANs [10]
Software-Defined
Hierarchical management
Software Defined and Virtualized Radio Access Networks with Fog Computing [11]
Software-Defined
Virtualization
Fog computing
HVSD-CRAN
Density aware
Separation of the air interface in dense region
Hierarchical processing in low density region
Software-defined approach
Cloud-based infrastructure
Virtualization

FIGURE 1. Time line of proposed RAN structures.

in order to utilize the network resources in a more efficient manner. Consequently, the OpenRAN is proposed in which the software-defined structure and vitualization (borrows from network function virtualization [19]) are considered in the cloud infrastructure [5].

One of the approaches to overcome the front-haul capacity limitation of C-RAN is the idea of *"physical decoupling"* which is a separation of the signals required for full coverage from those needed to support high data rate transmission with control-traffic decoupled air interface. This idea is first proposed in [20] and [21] for the forth generation wireless networks to overcome disadvantage of local resource management. However, due to the high interconnection between data and control signals, implementation of this approach is not straightforward [22], [23]. Decreasing the overhead signal is another advantage of physical decoupling [24] which can decrease the front-haul link data traffic. In [7], this approach is applied to overcome the front-haul capacity limitation of C-RAN where one BS has processing capabilities and only receives the control signal of all users. This BS is refereed to a control BS. Other simple and low financial cost RRHs only receive the data signal of assigned users and send data to the cloud for processing. This structure is called heterogeneous-RAN (H-RAN) since there exists one control BS and RRHs in RAN to provide the coverage. Afterwards, this idea is used in the software-defined RAN to achieve the flexible control for system [25].

Another approach to overcome the front-haul link capacity limitation is *"function splitting"* which is splitting the baseband processing tasks between RRHs and cloud BBU. This idea leads to the lower data volume to transmit in the front-haul link [12]. In [26], the required rates of the front-haul link for different levels of splitting are studied. The RRH with the processing ability is called RRS [13]. This method can overcome the additional transmission delay of the front-haul link specially where the distance between RRH and cloud center is large, e.g., low-density region where RRHs are implemented sparsely. However, the disadvantage of this solution is RRS financial cost increment since each RRS should have the processing capability.

In [9], the two approaches to overcome the front-haul capacity limitation are combined. In fact, it is assumed that one control BS is responsible to the coverage and D-RRHs have capability of function splitting and process part of the required processing functions. Although the front-haul rate limitation is completely removed in this system model, the full use of benefits of centralized processing cannot be achieved which is very important in high-density region due to the high interference of RRHs to each other. On the other hand, assuming the processing capability for D-RRHs in dense region contradicts by idea of C-RAN which is using a large number of very simple and low financial cost RRHs in the system.

Another RAN structure is SoftAir [27] where the software approach is used in both RAN and core of communication systems. Accordingly, the first open, flexible and programmable platform for the software-defined RAN is proposed in [28].

Afterwards, in [6], a cloud-based software-defined RAN with physical decoupling is proposed. Via this approach, the load of the front-haul link decreases and RAN becomes efficient from the energy consumption perspective due to the ability of turning off a set of RRHs in the low traffic time. Note that considering one control BS and D-RRHs cannot be implemented in low-density region since deploying less number of RRHs is sufficient to provide the coverage and capacity for users. Thus, the effects of RRHs on each other are negligible in this situation and a central high power BS to manage all resources is not necessary. In another RAN, proposed in 2016, the concept of fog in C-RAN called fog RAN (F-RAN) [8] is proposed in which RRHs are equipped with caching capability to decrease the latency of popular contents.

The hierarchical software-defined RAN is another RAN structure in which the management is split between the central unit and RRHs [10]. This RAN is not suitable for

TABLE 1. Advantages and disadvantages of different proposed RAN.

RAN	Advantage	Disadvantages
C-RAN [2]	Centralized processing/Centralized management Low CAPEX and OPEX	Limitation of front-haul link capacity /High delay Inflexibility/ Without network efficiency Without energy efficiency Not density-aware: Design for peak traffic load
SoftRAN [3]	Centralized management Flexibility	Distributed processing/High delay Without energy efficiency/ High CAPEX and OPEX Without network efficiency Not density-aware: Design for peak traffic load
OpenRAN [5]	Low CAPEX and OPEX/Centralized management Centralized processing/Network efficient	High delay/ Without energy efficiency Not density-aware: Design for peak traffic load
H-RAN [7]	Centralized management/Centralized processing Low CAPEX and OPEX /Energy efficient Lower front-haul rate limitation	High delay in low-density region Without network efficiency
SoftRAN with physical splitting of control and data plane [25]	Centralized management/Flexibility Energy efficient	Distributed processing/High delay in low-density region Without network efficiency/High CAPEX and OPEX
Hierarchical heterogeneous [9]	Centralized management/Low delay Energy efficient	Semi-distributed processing/ High CAPEX and OPEX Without network efficiency/Inflexibility Not density-aware
SoftAir [27]	Centralized management/Flexibility	Distributed processing/High delay High CAPEX and OPEX/ Without network efficiency Not density-aware: Design for peak traffic load
Software-defined hyper cellular [6]	Flexibility/Energy efficient Low CAPEX and OPEX/Centralized management Centralized processing	High delay in low-density region Without network efficiency Redundant infrastructure for low-density region
F-RAN [8]	Low redundant data in video traffic Lower front-haul rate limitation Energy efficient	Without network efficiency/ Semi-distributed management Semi-distributed processing/High delay for unpopular contents in low-density region
Hierarchical software-defined RAN [10]	Lower delay/Flexibility	Semi-distrusted management/ Semi-distributed processing High CAPEX and OPEX in dense region Without energy efficiency especially in dense region Not density-aware: Design for peak traffic load
Virualized software-defined RAN with fog computing [11]	Flexibility/Low delay for papular contents	Semi-distributed management/ Semi-distributed processing Without energy efficiency Not density-aware: Design for peak traffic load
HVSD-CRAN	Centralized management and processing management in high-density mode Energy efficient/Low delay Flexibility/Network benefit Low CAPEX and OPEX	Not responsive in rare condition of high users traffic in low-density region

dense region due to the high financial cost of RRHs with the processing ability. Another topology for RAN is a software-defined virtualized RAN with fog computing [11].

Fig. 1 affirms that all proposed RANs in 5G are developed based on six main techniques. As mentioned earlier, the first three main technologies are software-defined, cloud and virtualized structures. Three complementary techniques are physical decoupling, function splitting, and applying caching at RRHs. The advantages and disadvantages of the proposed RANs based on these six techniques are summarized in Table 1. At first, the software-defined structure leading to the centralized management increases delay. By virtualization, the network efficiency can be earned in the network infrastructure units. The cloud structure leads to the central processing and management. Again, this method suffers from delay due to the centralized processing and needs the high-capacity front-haul links. The physical decoupling leads to the ability of more efficient usage of energy by turning off the subset of D-RRHs. The function splitting can overcome the challenge of delay and front-haul link capacity limitations. However, here, RRHs should be capable of processing which increases their costs. Finally, considering caching for RRHs leads to decrease of the front-haul link load and delay for popular contents. However, using caching for all RRHs increments their costs. Obviously, non of the previous works considers the density of users to design RAN in 5G to overcome the mentioned drawbacks. In this paper, to cover this point, we propose our HVSD-CRAN and compare our system model with the existing proposed approaches.

B. MULTI-OBJECTIVE RESOURCE ALLOCATION PROBLEMS

Optimal dynamic resource allocation is critical in wireless networks to attain high utilization while satisfy QoS of users and other network considerations [29]. During the past three decades, there exists a surge of research in this context, e.g., [30], and obviously, it is not easy to provide the comprehensive survey and unified taxonomy here [30]. However, one important point is that in the resource allocation problems before 5G, usually, one performance metric is considered as an objective function of optimization problem called a single-objective optimization problem [31]. The single-objective resource allocation problem has one global optimal objective function value which earns by at least one solution in the feasible region. However, achieving this point is not guaranteed due to the high computational complexity. As a result, sub-optimal solutions with reasonable complexity are proposed in many papers which can be implemented in practical systems. The robustness and scalability are other important issues to be studied in the resource allocation strategies [29]. Another important aspect of single-objective resource allocation problem is how to solve the optimization problem in either the centralized or distributed manner. In the centralized approach, although the comprehensive optimization problem is solved, the extra load of the exchanged signals between the centralized unit and BSs is a practical implementation issue. The distributed approach is scalable and reduces this extra message passing between nodes, while the global optimal solution cannot be guaranteed here.

Furthermore, in 5G, the resource allocation encounters new perspective: Diverse services of 5G lead to the different objectives for system design, e.g., throughput, latency, and number of connections. There are trade-offs between these objectives and their priories in addition of dynamic behavior of users in three mentioned categories of services should be highly considered. Therefore, single-objective resource allocation is not sufficient and 5G should resort to more mature framework such as a MORA.

For instance, the trade-off between the spectral efficiency (SE) and energy efficiency (EE) is studied in [14], [15], and [32]–[36]. Also, MORA with coverage besides SE and EE is investigated in [15] for Internet of Things (IoT) services. However, there exist other conflicting objectives in 5G such as delay and cost which are very critical in 5G compared to the previous generation of wireless networks. To highlight this point, we propose two MOO problems in the proposed system model as examples.

For dense region, to provide high throughput and full coverage, there exist a large number of RRHs. Supporting all users in a dense region by very high throughput is one of the requirements of 5G in this region. Consequently, the throughput is an important object for MOO here. On the other hand, by considering one control BS in topology of the proposed system model for dense region, there is an ability to turn off the subset of D-RRHs to minimize the power consumption in low data traffic times [36]. Therefore, there exists a capability to manage the power consumption of 5G in more appropriate manner which is highly desirable to the efficient use of energy. Therefore, maximizing total network throughput and minimizing its power consumption by turning off the subset of D-RRHs are suitable objectives of the resource allocation problem in a dense region, which have an inherent conflict for resource allocation problem.

In a low-density region, we consider other two objective functions, e.g., delay and the cost of processing in RRSs. Here, we define the cost of processing as a function related to the amount of baseband functions that RRSs should process. We consider this function as a new utility for MORA since RRSs have a limited resources and when the processing capabilities are increased, the cost of using limited resources of RRSs are increased accordingly. These two objects have conflicting goals, since to decrease the delay, we need RRSs via more processing capabilities and vice versa. To cover this point, in this paper, we propose the new model for cost of RRSs based on their capabilities of function processing.[1]

Up to our knowledge, there exists no other related work that considers the resource allocation problem with MOO in 5G based on our objective functions. Therefore, definitions of objectives and sets of variables in MORA of this paper, are novelties of resource allocation in this work.

III. HVSD-CRAN SYSTEM MODEL

HVSD-CRAN is a proposed density-aware RAN in 5G in which cloud, virtualization and software-defined structure are considered. For considering density in the system model, we categorize the coverage area of 5G into the high and low-density regions. As a result, first, we explain the challenges in two types of regions and the solutions to overcome them. Next, the proposed HVSD-CRAN structure is explained.

A. HIGH-DENSITY REGION ASPECTS

To provide high data transmission rate for users in high-density region, densification is a promising approach where large number of RRHs are deployed in one specific area. On the other hand, due to the high interference of each RRH on the neighboring ones, the resource management should be deployed in one central unit, e.g., BBU cloud in our setup. However, for decreasing the front-haul load of C-RAN and energy efficiency in low traffic times, a high power BS is used as the control BS and receives control signals of users and allocates resources of the network. Other simple and low financial cost RRHs receive the data signals. Although the design of system is according to the peak traffic time, the D-RRHs can be turned off because the control BS is responsible for coverage in the system [36].

B. LOW-DENSITY REGION ASPECTS

In low-density region, the RRSs are implemented in a sparse manner, and distances between RRSs and central BBU cloud are high which lead to the high latency in this link. To overcome this challenge, the function splitting between RRSs and BBU cloud is a suitable approach [12]. Here we propose that the level of function splitting can be changed according to the instantaneous conditions and system requirements which is enabled via the software-defined structure of HVSD-CRAN. When more functions are processed by RRSs, the delay decreases accordingly while RRS's cost is increased since more processing capabilities should be used in each RRS. On the other hand, in low-density region, the interference between RRSs is low, and thus, the distributed resource allocation is more appropriate to reduce the extra message passing between RRSs and the BBU cloud. However, due to the virtualization in 5G, isolation between SPs should be guaranteed which requires the centralized management.

[1] Note that, in this paper, only transmission delay of the front-haul link is considered due to the large distance between RRSs and BBU cloud in low-density region and processing delay is neglected due to high capacity BBUs in the cloud.

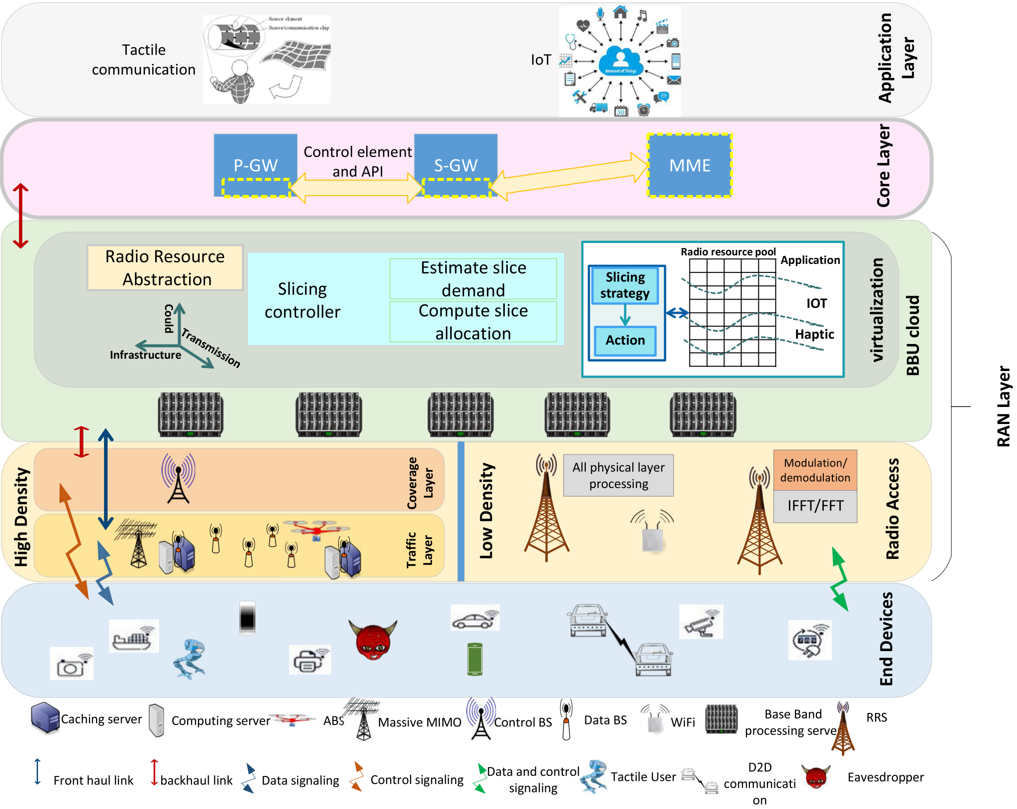


FIGURE 2. HVSD-CRAN system model.

C. HVSD-CRAN STRUCTURE AND RESOURCE MANAGEMENT

To cover different aspects of low and high-density regions, we propose a new system model depicted in Fig. 2 where there exist two modes:

- *High-density mode:* where there exist two types of BSs, i.e., control BS and D-RRHs in radio access of RAN layer in Fig. 2. The heterogeneity of this mode comes from these two types of BSs. In this topology, the end devices in Fig. 2 should have two connections for control and data messages due to the physical decoupling. The coverage is guaranteed by control BS, called coverage layer. However, there is a capability to turn off the subset of D-RRHs in traffic layer of Fig. 2 in low traffic time. The data signal of end device is sent to D-RRHs and all their baseband processing functions are deployed in a BBU cloud of RAN layer. According to the virtualoization feature of our system model, the radio resources can be sliced to use by different SPs to support various services for users.
- *Low-density mode:* where there exists hierarchical processing between cloud and RRSs in RAN layer of Fig. 2. In this mode, the baseband function splitting is used to overcome the high delay of large distance of front-haul link. In this topology, each RRS processes part of the baseband processing functions which can be dynamically determined to improve desired performance according to the instantaneous conditions of system. According to the virtualization characteristics, the radio resources are sliced in BBU cloud of RAN layer to have view of whole network. As a result, the isolation constrains between slices should be guaranteed in a centralized manner.

For both modes, the flexibility for designed system is achieved by software-defined structure. The resource slicing, for both modes, can be categorized as [37]

- *Transmission resources* e.g., power and bandwidth
- *Cloud resources* e.g., processing and memory units of BBUs
- *Infrastructure resources* e.g., RRHs, front-haul link, switches

The resource management structure of HVSD-CRAN is illustrated in Fig. 3. In this topology, two resource management units are considered. In centralized radio management (CRM), a subset of parameters, e.g., power and spectrum, that should be allocated by view of whole network, are assigned. Also, by considering virtualization, all the radio resources are sliced between SPs. When the effects of resource allocation in one RRS on the other RRSs are negligible, it is more convenient to implement local radio management (LRM) in radio access of RAN layer in Fig. 3. The resource management policy of these two modes of HVSD-CRAN are

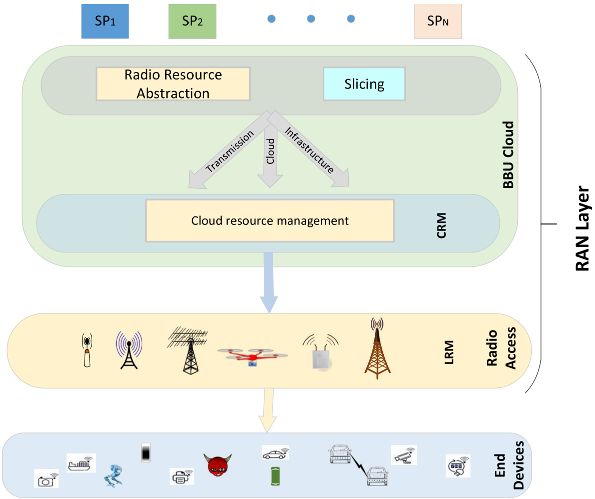


FIGURE 3. Management structure of HVSD-CRAN system model.

- *High-density mode:* In this mode due to the high effects of dense D-RRHs on each other and the lack of processing ability in D-RRHs, all resources are allocated in CRM in BBU cloud in Fig. 3. The resource management in this mode is fully centralized.
- *Low-density mode:* The parameters which are related to isolation requirements of virtulization as guaranteing QoS for SPs, are managed in CRM. Other parameters are allocated in LRM in Fig. 3 since in this mode, the effects of RRS's parameters on each other can be neglected due to the large distance of RRSs. Thus, the resource management is semi-distributed in this mode.

IV. MULTI-OBJECTIVE RESOURCE MANAGEMENT DESIGN IN HVSD-CRAN

In this section, two examples of resource allocation problem for the high and low-density modes with different objective functions are proposed. In each problem, two objective functions are applied for resource allocation optimization problem based on the network conditions.

Assume that the coverage area of specific region is provided via a set of $\mathcal{M} = \{1, \cdots M\}$ RRHs which are either D-RRHs in high-density mode or RRSs in low-density mode. The total bandwidth is divided into a set of $\mathcal{N} = \{1, \cdots N\}$. Also, there are Z SPs where the users of SP z are in a set $\mathcal{Z}_z$. The minimum required rate of SP z is represented by R_z. The variance of additive white Gaussian noise is N_0. The RRH m has the maximum power limit $p_m^{\max}$. The channel gain from RRH m to the user k on sub-carrier n is represented by $h_{n,k,m}$ and $h_{n,k}^m$ in high and low-density modes, respectively. Also, $p_{n,k,m}$ or $p_{n,k}^m$ is the allocated power to user k on sub-carrier n by RRH m in high and low-density modes, respectively.

In this paper, for high and low-density modes, we apply two models of subscripts. Since the interference between RRHs in high-density mode is high, we consider the RRH assignment parameter for this case, and consequently, index m is in the subscript of parameters for this case. However, for low-density mode, users are assigned to RRS based on their received signal strength due to the sparsity of RRSs. Therefore, we deploy m as a superscript of each parameter for this case.

A. HIGH-DENSITY MODE RESOURCE MANAGEMENT

Assume a network with K users in a set $\mathcal{K}$. Due to the dense deployment of D-RRHs and assuming their capabilities to be turned off, the assignment of users to the M D-RRHs is optimized in the resource allocation problem. Also, the sub-carrier assignment and transmit power allocation are the other variables which should be determined. Note that due to the having dense region, the resource should be centrally allocated by considering the interference of D-RRHs on each other. The rate of user $k \in \mathcal{K}$ on sub-carrier $n \in \mathcal{N}$ by assigning to the D-RRH $m \in \mathcal{M}$ is

$$R_{n.k,m} = x_{n,k,m} \log(1 + \frac{p_{n,k,m} h_{n,k,m}}{N_0 + I_{n,k,m}}), \tag{1}$$

where $x_{n,k,m}$ is equal to one if sub-carrier n of D-RRH m is assigned to the user k and otherwise is zero, and $I_{n,k,m} = \sum_{\substack{m' \in \mathcal{M} \\ m' \neq m}} \sum_{k' \in \mathcal{K}} p_{n,k',m'} h_{n,k,m'} x_{n,k',m'}$.

As explained before, in high-density mode there exists an ability of turning off D-RRH m expressed by b_m. The b_m is equal to zero where D-RRH m is turned off and otherwise is one. When at least one user is assigned to one D-RRH, this D-RRH is active and the variable b_m should be equal to one. This fact can be considered in the resource allocation problem as

$$\mathrm{C}_1^{\mathrm{H}}: \; x_{k,n,m} \leq b_m, \;\; \forall k, n, m.$$

By considering Z number of SPs in the system, the quality of them should be satisfied in the resource allocation. We assume that the total rate of users of each SP should be higher than a defined threshold R_z. Thus, we have

$$\mathrm{C}_2^{\mathrm{H}}: \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{Z}_z} \sum_{n \in \mathcal{N}} x_{n,k,m} \log(1 + \frac{p_{n,k,m} h_{n,k,m}}{N_0 + I_{n,k,m}}) \geq R_z, \quad \forall z.$$

Also, each D-RRH has the maximum transmit power limitation as

$$\mathrm{C}_3^{\mathrm{H}}: \; \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}} x_{n,k,m} p_{n,k,m} \leqslant P_{\max}^m, \quad \forall m.$$

Each sub-carrier in each D-RRH should be assigned to only one user which can be mathematically represented by

$$\mathrm{C}_4^{\mathrm{H}}: \; \sum_{k \in \mathcal{K}} x_{n,k,m} \leq 1, \quad \forall n, m.$$

In 5G, one of design criteria is maximization of total rate of network which is considered as one of the objective functions in our MORA problem. According to (1), the total rate of network can be defined as

$$R_T = \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}} x_{n,k,m} \log(1 + \frac{p_{n,k,m} h_{n,k,m}}{N_0 + I_{n,k,m}}). \tag{2}$$

On the other hand, the consumed power of each D-RRH is composed of two elements which are transmission and circuit power. If no user is assigned to one D-RRH, it is turned off and circuit power should be set to zero, otherwise its circuit power is P_C. Therefore, the total power consumption of network is

$$P_T = \sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} x_{n,k,m}p_{n,k,m} + \sum_{m\in\mathcal{M}} P_C\left(b_m\right), \tag{3}$$

where P_C is a constant value equal to the circuit power consumption of each D-RRH in active state.

Finally, the MORA problem is

$$\begin{aligned} &\max_{X,P,b} \{R_T, -P_T\} \\ &\text{subject to}: \ \mathrm{C}_1^{\mathrm{H}} - \mathrm{C}_4^{\mathrm{H}}. \end{aligned} \tag{4}$$

According to the Appendix, these two objective functions can be encapsuled in one single-objective function by weighted sum method as

$$U^{\mathrm{H}}(X, P, b) = \alpha \frac{R_T}{R_T^{\max}} - (1-\alpha)\frac{P_T}{P_T^{\min}}, \tag{5}$$

where α, $R_T^{\max}$ and $P_T^{\min}$ are parameters of weighted sum rate method for MOO problem explained in the Appendix. Also, X is a three dimensions matrix of D-RRH and sub-carrier assignments of users, P is a matrix of power allocation between users and b is a vector of binary values which indicate the ON and OFF states of D-RRHs.

Thus, the final MORA problem in high-density mode may be

$$\begin{aligned} &\max_{X,P,b} U^{\mathrm{H}}(X, P, b) \\ &\text{subject to}: \ \mathrm{C}_1^{\mathrm{H}} - \mathrm{C}_4^{\mathrm{H}}. \end{aligned} \tag{6}$$

The proposed problem is nonconvex by considering all variables jointly. To reach an efficient algorithm, the iterative algorithm is proposed that contains two steps: 1) when it is assumed that P is known and solved (6) for finding the optimal user and sub-carrier assigning (X) and D-RRHs activity(b), 2) according to the derived result of the first step assignment of this iteration, the power allocation problem is solved. The problem of each step is nonconvex due to the interference term in the rate formula. To overcome this challenge, we introduce the predefined threshold for tolerate interference and add *a* new constraint for the optimization problem as

$$\mathrm{C}_0^{\mathrm{H}}: I_{n,k,m} = \sum_{\substack{m'\in\mathcal{M} \\ m'\neq m}}\sum_{k'\in\mathcal{K}} p_{n,k',m'}h_{n,k,m'}x_{n,k',m'} \leq I_{\mathrm{th}}, \quad \forall n, m, k,$$

where I_{th} is the tolerable level of interference of all users in the network. Then, I_{th} is used in the first objective function and the first constraint instead of $I_{n,k,m}$ which leads to the lower limit of network rate. The problem of the first step is still nonconvex due to the integer variables. By relaxing integer variables as continuous ones, the problem in this step, is transformed to a convex one and can be solved by CVX tool. The problem of the second step is convex too and can be solved by CVX tool. The overal resource allocation procedure is presented in Algorithm 1.

Algorithm 1 Resource Allocation Algorithm for High-Density Mode

```
for $\alpha = 0 : 0.2 : 1$
Initialization: Set $t := 1$ and initialize $P^*(0) = P_{\max}/N$
    Repeat
        Step 1: Derive $X^*(t)$ and $b^*(t)$ to maximize
        (5) with constraints $C_0^H, C_1^H, C_2^H, C_4^H$ by
        considering fixed value of $P^*(t-1)$
         Step 2: For fixed value $b^*(t)$ and $X^*(t-1)$,
        find $P^*(t)$ to maximize (5) with constraints
        $C_0^H, C_2^H, C_3^H$
        Step 3: if $|P^*(t) - P^*(t-1)| \leq \varepsilon$
                Set $P^\alpha = P^*(t)$, $X^\alpha = X^*(t)$, and
                $b^\alpha = b^*(t)$
                Stop repeat.
            else
        set $t := t + 1$ and go to Step 1.
end
```

B. LOW-DENSITY MODE RESOURCE MANAGEMENT

Assume a network with M RRSs. In low-density mode, the RRSs are sparsely distributed. Thus, the distance of each user to one RRS is much lower than the distance to the other RRSs. Therefore, the users are assigned to RRSs by considering the minimum distance to the RRSs where $\mathcal{K}_m$, $m \in \{1, \ldots, \mathrm{M}\}$ defined sets of user assigned to RRS m, respectively. Each RRS uses all sub-carriers. The user of SP z are distributed in the whole coverage region of M RRSs. The total transmission rate of RRS $m \in \mathcal{M}$ in downlink is equal to

$$R_m = \sum_{k\in\mathcal{K}_{\mathrm{m}}}\sum_{n\in\mathcal{N}} x_{n,k}^m \log(1 + \frac{p_{n,k}^m h_{n,k}^m}{N_0 + \sigma_I^2}), \tag{7}$$

where σ_I^2 is an interference variance on each RRS from other RRSs which is low value due to the large distance between RRSs and it is in order of additive wight Gaussian noise variance N_0. Thus, the summation of these two values is expressed by $N_{0I} = N_0 + \sigma_I^2$. Other parameters are defined before.

Assume a set $\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}}$ containing the users of SP z assigned to the RRS m where $\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}} \subset \mathcal{Z}_{\mathrm{z}}$ and $\bigcup_{m\in\mathcal{M}} \mathcal{Z}_{\mathrm{z}}^{\mathrm{m}} = \mathcal{Z}_{\mathrm{z}}$. By considering virtualization, due to the slice isolation constraint, the minimum sum rate of all users belonging to each SP, should be guaranteed which is represented as

$$\mathrm{C}_1^{\mathrm{L}}: \sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}}}\sum_{n\in\mathcal{N}} x_{n,k}^m \log(1 + \frac{p_{n,k}^m h_{n,k}^m}{N_{0I}}) \geq R_z, \quad \forall z.$$

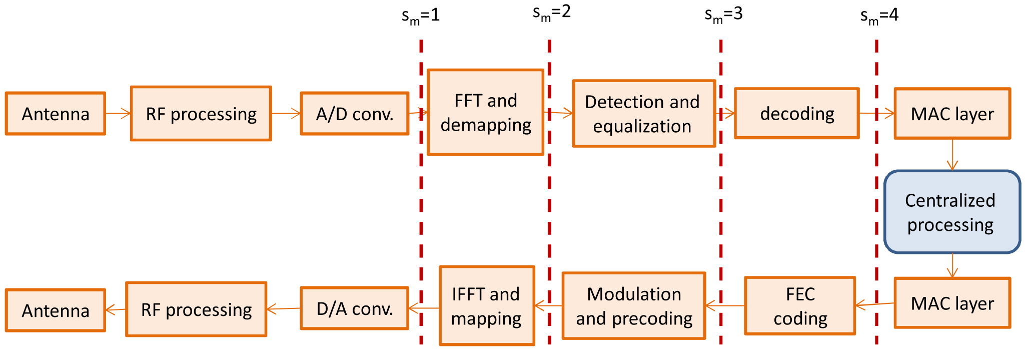


FIGURE 4. Level of function splitting in C-RAN [26].

One practical implementation is that each sub-carrier of each RRS should be allocated to only one user as

$$\mathrm{C}_2^{\mathrm{L}}: \sum_{k\in\mathcal{K}_{\mathrm{m}}} x_{n,k}^m \leq 1, \quad \forall n, m.$$

Each RRS has a maximum transmission power limitation as

$$\mathrm{C}_3^{\mathrm{L}}: \sum_{k\in\mathcal{K}_{\mathrm{m}}} \sum_{n\in\mathcal{N}} x_{n,k}^m p_{n,k}^m \leqslant P_{\max}^m \quad \forall m.$$

The transmitted data for all users of RRS m in the access link (RRS to users) is proportional to the R_m. Note that this transmitted data is the received data at RRS m from cloud unit via front-haul link. Thus, the transmitted data of RRS m in front-haul link is proportional to R_m. In the literatures, the baseband processing functions can be split between RRS and BBU cloud in four levels which are demonstrated in Fig. 4 [26]. By applying the less baseband processing functions on data in cloud unit, the required rate for transmission in front-haul link decreases [26] or equivalently, the transmitted data load decreases. As a result, if the processing splitting indicator of RRS m (s_m) becomes high (low baseband processing in cloud unit), the transmitted data in front-haul link decreases. For modeling this inverse relation between transmitted data load and function splitting level, one can assume that the transmitted data in front-haul link is equal to the transmitted data from RRS in access link which is divided by function splitting level. As a result, the transmitted data on front-haul link by considering function splitting level of s_m is

$$D_m^F = \frac{R_m}{s_m}, \quad \forall m, \tag{8}$$

which is a transmitted data load on the front-haul link m. When the capacity of front-haul link between RRS m and cloud is limited to U_m bits per second, there is a limitation on the transmitted data load on this link per second. Therefore, the constraint of front-haul data transmission is written as

$$\mathrm{C}_4^{\mathrm{L}}: \sum_{k\in\mathcal{K}_{\mathrm{m}}} \sum_{n\in\mathcal{N}} \frac{x_{n,k}^m \log(1+\frac{p_{n,k}^m h_{n,k}^m}{N_{0I}})}{s_m} \leq U_m, \ \forall m.$$

By considering maximum capacity limitation of front-haul link U_m, when there exists D_m bit data load, the transmission delay is $\tau_m = \frac{D_m}{U_m}$. Now, via (8), the transmission delay of RRS m (τ_m) is

$$\tau_m = \frac{D_m^F}{U_m}, \quad \forall m. \tag{9}$$

As delay is one of the critical measurements in 5G, delay of front-haul link is an appropriate choice for objective function in resource allocation problem. Thus, to consider front-haul transmission delay of all RRSs by assuming same priority for them, $\tau_{\mathrm{T}} = \sum_{m\in\mathcal{M}} \tau_m$ should be minimized.

Although delay of system decreases by increasing the processing level of RRSs, the cost of processing at RRSs increases due to the limited resources (power or processing units) in it. Thus, we can assume the relation between data splitting level and cost of processing is monotonic increasing function which is expressed by function $F(s_m)$. Therefore, the total cost of network is a summation of cost of all RRSs as

$$G = \sum_{m\in\mathcal{M}} F(s_m). \tag{10}$$

For simplicity, we consider $F(s_m) = C \times s_m$ where C is a cost coefficient. Therefore, we have

$$G = C \sum_{m\in\mathcal{M}} s_m. \tag{11}$$

These two objectives τ and G are conflicting with each other. In fact, although increasing the value of s_m increases the cost function, it decreases the delay according to (8), accordingly. As a result, the MORA problem is

$$\begin{aligned} &\min_{X,P,s} \{\tau, G\} \\ &\text{subject to}: \ \mathrm{C}_1^{\mathrm{L}} - \mathrm{C}_4^{\mathrm{L}}. \end{aligned} \tag{12}$$

Again, the weighted sum rate is deployed to formulate multi-objective optimization as a single-objective problem as

$$\min_{X,P,s} \alpha \frac{\tau}{\tau^{\min}} + (1-\alpha)\frac{G}{G^{\min}}$$
$$\text{subject to: } \mathrm{C}_1^L - \mathrm{C}_4^L, \tag{13}$$

where α, $\tau^{\min}$ and $G^{\min}$ are the parameters of weighted sum approach for MOO problem explained in the Appendix. Also, X and P are the matrices of sub-carrier and power allocation, respectively, and, s is a vector of function splitting levels of RRSs.

In distributed resource allocation approach, there is not any data exchanging between centralized manager and local elements and it is scalable. However, in the new structure of 5G, there are a subset of requirements which should be managed in the centralized unit and guaranteed in whole system like virtulization constraints. Therefor, a subset of the variables should be allocated in the centralized manner. By considering the mentioned points, the semi-distributed management which is the compound of distributed and centralized methods, is the best choice for this setup. In semi-distributed approach, the subset of the variables are centrally assigned and the other ones are allocated in the distributed manner. Also, in semi-distributed resource allocation approach the data exchange between a centralized manager and local elements significantly decreases compared to the fully centralized manner. On the other word, the semi-distributed resource allocation method is one approach to overcome the need of high capacity requirement of front-haul link. This method is a suitable choice in low-density mode due to the low effects of sparse RRSs on each other.

One of the methods for proposing semi-distributed resource allocation is decomposition methods [38]. Note that, in our resource allocation problem, only C_1 is the constraint that couples the optimization problems of all RRSs. Thus, by omitting it, the problem can be decomposed in to m different problems. For this purpose, by relaxing this constraint via the Lagrangian method [38], we transform (13) into

$$\min_{x,p,s,\boldsymbol{\lambda}} \alpha \frac{\tau}{\tau^{\min}} + (1-\alpha)\frac{G}{G^{\min}}$$
$$+ \sum_{z\in\mathcal{Z}} \lambda_z \left(R_z - \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}}} \sum_{n\in\mathcal{N}} x_{n,k}^m \log(1+\frac{p_{n,k}^m h_{n,k}^m}{N_{0I}}) \right), \tag{14}$$

where $\boldsymbol{\lambda}$ is a vector of λ_z for all $z \in \{1, \ldots, Z\}$ which are Lagrangian multipliers. This problem can be separated into two level optimization problems in which all problems of low level are solved in distributed manner, and high level problem is solved in centralized manner. The centralized problem in high level is

$$\min_{\boldsymbol{\lambda}} \sum_{z\in\mathcal{Z}} \lambda_z$$
$$\left(R_z - \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}}} \sum_{n\in\mathcal{N}} {x_{n,k}^m}^{t-1} \log(1+\frac{{p_{n,k}^m}^{t-1} h_{n,k}^m}{N_{0I}}) \right). \tag{15}$$

This problem can be solved by subgradient method as

$$\lambda_z(t) = \Bigg[R_z - \lambda_z(t-1)$$
$$-\beta \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}}} \sum_{n\in\mathcal{N}} {x_{n,k}^m}^{t-1} \log(1+\frac{{p_{n,k}^m}^{t-1} h_{n,k}^m}{N_{0I}}) \Bigg]^+ \forall z. \tag{16}$$

The new value of vector $\boldsymbol{\lambda}$ is transmitted to the RRSs.

On the lower level, the optimization problem of RRS m is

$$\min_{X,P,s} F_D = \alpha \frac{\sum_{k\in\mathcal{K}_{\mathrm{m}}} \sum_{n\in\mathcal{N}} \frac{x_{n,k}^m \log(1+\frac{p_{n,k}^m h_{n,k}^m}{N_{0I}})}{s_m}}{U_m} + (1-\alpha) C s_m$$
$$+ \sum_{z\in\mathcal{Z}} \lambda_z^t \left(- \sum_{k\in\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}}} \sum_{n\in\mathcal{N}} x_{n,k}^m \log(1+\frac{p_{n,k}^m h_{n,k}^m}{N_{0I}}) \right)$$
$$\text{subject to : } \mathrm{C}_1^{\mathrm{L}} - \mathrm{C}_4^{\mathrm{L}}. \tag{17}$$

This problem is nonconvex. Thus, we propose iterative algorithm to solve it. In each iteration, three types of variables (X, P, s) are solved separately by assuming fixed values for other two types of variables. The problem of finding X by given values of P, s of RRS m is

$$\min_X F_D$$
$$\text{subject to: } \mathrm{C}_1^L - \mathrm{C}_4^L. \tag{18}$$

This problem is convex and can be solved by CVX tool. The second problem is power allocation problem which is

$$\min_P F_D$$
$$\text{subject to: } \mathrm{C}_1^{\mathrm{L}}, \mathrm{C}_3^{\mathrm{L}}, \mathrm{C}_4^{\mathrm{L}}. \tag{19}$$

The objective function is nonconvex. Also, C_4^{L} is nonconvex too. Thus, we use the successive convex approximation to convert the problem to the convex one. As can be seen, in (17) two logarithmic functions of p are subtracted. Thus, our problem is difference of convex function. To convert the problem to the convex one, the difference of convex (DC) approximation method [39] is applied and the problem is solved iteratively. Also, this approximation is used to convert the nonconvex constraint (C_4^{L}). By considering parameter t_1 as the iteration index of finding the optimum power allocation, we have

$$\min_p \alpha \Bigg(\sum_{k\in\mathcal{K}_{\mathrm{m}}} \sum_{n\in\mathcal{N}} \frac{x_{n,k}^m \log(1+\frac{{p_{n,k}^m}^{t_1-1} h_{n,k}^m}{N_{0I}})}{U_m s_m}$$
$$+ \sum_{k\in\mathcal{K}_{\mathrm{m}}} \sum_{n\in\mathcal{N}} \frac{x_{n,k}^m \frac{h_{n,k}^m}{N_{0I}+{p_{n,k}^m}^{t_1-1} h_{n,k}^m}}{U_m s_m} (p_{n,k}^m - {p_{n,k}^m}^{t_1-1}) \Bigg)$$
$$+ \sum_{z\in\mathcal{Z}} \lambda_z \left(- \sum_{k\in\mathcal{Z}_{\mathrm{z}}^{\mathrm{m}}} \sum_{n\in\mathcal{N}} x_{n,k}^m \log(1+\frac{p_{n,k}^m h_{n,k}^m}{N_{0I}}) \right)$$
$$\text{subject to: } C_1^{\mathrm{L}}, C_3^{\mathrm{L}},$$

$$\sum_{k\in\mathcal{K}_{\mathrm{m}}}\sum_{n\in\mathcal{N}}\frac{x_{n,k}^{m}\log(1+\frac{{p_{n,k}^{m}}^{t_1-1}h_{n,k}^{m}}{N_{0I}})}{s_m}$$
$$+\sum_{k\in\mathcal{K}_{\mathrm{m}}}\sum_{n\in\mathcal{N}}\frac{x_{n,k}^{m}\frac{h_{n,k}^{m}}{N_{0I}+{p_{n,k}^{m}}^{t_1-1}h_{n,k}^{m}}}{s_m}(p_{n,k}^{m}-{p_{n,k}^{m}}^{t_1-1})) \leqslant U_m. \tag{20}$$

Finally, by relaxing the value of s to the real value in [0 4], the convex problem for assigning s_m is

$$\min_{s} F_D$$
$$\text{subject to: } \mathrm{C}_4^{\mathrm{L}}. \tag{21}$$

Similarly, this problem can be solved by CVX. These three optimization problems are solved iteratively until the difference of derived optimum value of variables of two iterations is lower than defined threshold ϵ. The overal resource allocation procedure is presented in Algorithm 2.

C. CONVERGENCE AND COMPLEXITY ANALYSIS

The proposed resource allocation algorithms are based on block coordinate descent (BCD) method where one group of variables is optimized by assuming fix value for other groups of variables. In [40], it is proved that the convergence of BCD is guaranteed when the optimization problem of each step is convex. Thus, the convergence of algorithms to the local optimum, not necessarily the global optimum, are guaranteed.

As the interior point method is used in CVX tool for solving convex problems, the number of required iterations is $\frac{\log(c/(t\kappa))}{\log(\varphi)}$ [41] where c, t and κ are the number of constraints of convex optimization problem, the initial point to approximate the accuracy of interior point method, and the stopping criterion for interior point method, respectively. Also, φ is used for updating the accuracy of interior point method. Thus, the complexity of each resource allocation algorithm is the summation of the complexity of all steps. Note that, the convergence of DC algorithm is achieved by complexity of $O(\log(1/\epsilon))$ where ϵ is the stopping criterion [42]. By these two complexity formulas, the complexity of each of the proposed resource allocation algorithm can be achieved.

V. SIMULATION RESULTS

As mentioned earlier, the solution of MOO problem is a set of Pareto optimal solutions. As a result, in this section, the outputs of Pareto optimal set (Pareto frontier set) of the proposed resource allocation algorithms are demonstrated. These points are derived by changing the priority scale of objective functions (α, $1-\alpha$) in weighted sum method. The results of two different resource allocation problems based on the two proposed modes are investigated in the two following subsections.

A. HIGH-DENSITY MODE

For this mode, we consider a network of $M = 4$ D-RRHs and $K = 80$ users with $N = 20$ sub-carriers. The users belong to two different SPs. The maximum power $P_{\max}^{m}$, maximum interference tolerance I_{th} and additive wight Gaussian noise variance N_0 are 50 W, 10 W and 1 W, respectively. The users are uniformly distributed in 2×2 square region. By dividing the whole region into 4 same square, the D-RRHs are in the center of these small squares. The channel gain is modeled as $h_{n,k}^{m} = (d_k^m)^{-\phi}\beta_{n,k}^{m}$ where d_k^m is a distance between user k and D-RRH m, a path loss exponent (ϕ) is 3 and the small scale fading gain $\beta_{n,k}^{m}$ has the exponential distribution with unit variance.

The Pareto frontier sets by assuming different values for P_c and R_z are derived via Algorithm 1, and the results are depicted in Fig. 5. By moving from left to right of the curves, the value of α in the weighted multi-objective problem increases by step 0.2 from zero to one. As expected, by increasing α and decreasing the priority of power consumption objective, more D-RRHs becomes active, which leads to the more power consumption. On the other hand,

Algorithm 2 Resource Allocation Algorithm for Low-Density Mode

```
for $\alpha = 0 : 0.2 : 1$
Initialization: Set $t := 1$ and initialize
$P^*(0) = P_{\max}/N$, $S^*(0) = 1$, and $\boldsymbol{\lambda}^*(0) = 1$
    Repeat
        For each RRS
        Step 1: Derive $X^*(t)$ to maximize (18)
            considering fixed value of $P^*(t-1)$,
             $s^*(t-1)$, and $\boldsymbol{\lambda}^*(t-1)$
         Step 2: For fixed value $X^*(t)$, $s^*(t-1)$ and
            $\boldsymbol{\lambda}^*(t-1)$,
            Set $t_1 := 1$, $P(t_1 - 1) = P^*(t-1)$
            Repeat
                Step I:Solve (20) to find P($t_1$)
                if $|P^*(t_1) - P^*(t_1 - 1)| \leq \varepsilon$
                  Set $P(t) = P^*(t_1)$ and
                  Stop Repeat
              else
                 set $t_1 := t_1 + 1$ and go to Step I
        Step 3:Derive $s^*(t)$ to maximize (21)
            considering fixed value of $P^*(t)$, $X^*(t)$,
             and $\boldsymbol{\lambda}^*(t-1)$
        End For each RRS
        Step 4: Update $\boldsymbol{\lambda}^*(t)$ according to (16)
            considering fixed value of $P^*(t)$, $X^*(t)$
             $s^*(t)$
         Step 5: if $|\boldsymbol{\lambda}^*(t) - \boldsymbol{\lambda}^*(t-1)| \leq \varepsilon$ and
              $|P^*(t) - P^*(t-1)| \leq \varepsilon$
              Set $\mathrm{X}^{\alpha} = X(t)$, $\mathrm{P}^{\alpha} = P(t)$, and
              $\mathrm{s}^{\alpha} = s(t)$
              Stop repeat
            else
                set $t := t + 1$ and go to Step 1
end
```

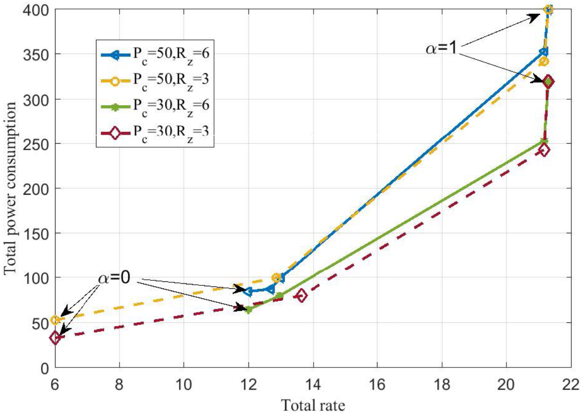


FIGURE 5. Power versus rate via Algorithm 1 for various values of $\alpha \in [0\ 1]$, P_C and R_Z.

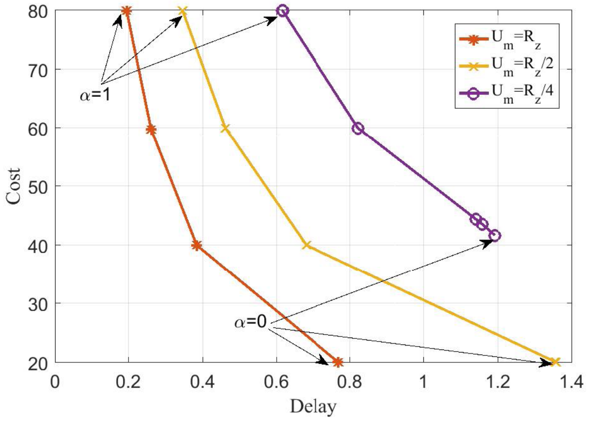


FIGURE 6. Cost versus delay via Algorithm 2 for $\alpha \in [0\ 1]$ and different front-haul link capacity limitations.

the priority of rate objective increases and higher rate are achieved with high value of α. When $\alpha = 0$, the rate in the objective function is not considered and the minimum required rate of each SP is guaranteed, which leads to the total of 6 and 12 for minimum rate bound of 3 and 6, respectively. On the other hand, when $\alpha = 1$, all D-RRHs are active to increase the rate since their power consumption is not important in MOO problem.

Choosing α can be considered as a planning design factor in resource allocation problem. For instance, in high traffic time, when providing higher throughput is more critical than decrement of power consumption, in HVSD-CRAN, we can set $\alpha > 0.5$. However, in low traffic time, the power consumption has more priority compared to the throughput which leads to $\alpha < 0.5$. As a result, according to the dynamic behavior of traffic, one of the Pareto optimal solutions can be selected by adjusting the value of α in the weighted sum method. For example, IoT services are usually based on sensors with limited battery life-time with low traffic rate. While virtual and augmented reality services and 3D and ultra-HD videos require high data rate. By considering the traffic variations of these types of users per each day, different Pareto optimal solutions can be selected for the resource allocation problem in HVSD-CRAN in 5G.

B. LOW-DENSITY MODE

For this mode, again we consider 4 RRSs where each of them has a 2×2 square coverage area. Each RRS has 20 users in its coverage area. The same channel model used in the previous subsection is applied. The minimum required rate of each SP is set to 150 bps. The imposed cost of processing in RRS is $C = 20$. The maximum capacity of front-haul link, i.e., U_m is considered as a ratio of R_z.

The Pareto frontier set derived via Algorithm 2 is demonstrated in Fig. 6. In this figure, the value of α increases from 0 to 1 with step of 0.2 by moving from right to left of the curves. As expected, by increasing α, the priority of cost function decreases and priority of delay increases. Therefore, the level of splitting increases to reduce the delay for $\alpha = 1$. In this case, the solution reaches $s_m = 4$ for all curves (cost= 80).

Also, by decreasing the maximum front-haul link capacity (U_m), the level of function splitting increases. As can be seen, for $U_m = R_s/4$ in $\alpha = 0$ higher average function splitting level achieved and at least s_m is equal to 2 (cost= 40). The reason is that higher order of s_m should be used in the system to satisfy the front-haul link capacity constraint in low U_m.

For this scenario, α can be considered as a planning factor to distinguish users of different services in 5G. For instance, when there exists a set of users belonging to autonomous driving and MCC, e.g., for tactile services, it is better to set $\alpha > 0.5$. By increasing the traffic of these users in the system, the priority of delay function (α) should be increased. Accordingly, more baseband processing functions should be done at RRSs to reduce the delay for these types of users.

VI. CONCLUSION

The purpose of this study is twofold. In the first part, new density-aware RAN structure in 5G is proposed. The virtualized software-defined structure of this system model is different in low and high-density modes. In dense mode, heterogeneous RRHs which are one control BS and many D-RRHs, are implemented. However, in low-density mode, hierarchical signal processing is considered in which the baseband processing function is split between RRSs and BBU cloud. The resource management of these two cases are also implemented in fully centralized and semi-distributed manners, respectively. In the second fold, the multi-objective resource allocation framework is proposed for these two cases. In high-density mode, the rate and power consumption are considered as two major conflicting objective functions. The sum of the delay and cost of processing in RRSs are minimized in low-density mode where the resources of access link and level of splitting are variables of optimization problem. The weighted sum method is used to solve multi-objective resource allocation problems for two scenarios. The Pareto optimal solution sets of two problems are derived by varying the weight of objective functions. The simulation results demonstrate how Pareto optimal sets are varied under different system settings.

APPENDIX

In MOO problems, comparing the objective functions are not easy as single-objective problem because there are a vector of functions. As a result, in multi-objective optimization problem, there is not only one solution like single-objective problem. In these problems, there is a set of solutions which is called Pareto optimal solutions. The Pareto optimal solution refers to a solution, around which there is no way of improving any objective without decreasing at least one other objectives.

The general MOO problem is [16]

$$\begin{aligned} &\min_x [f_1(x), f_2(x), \ldots, f_O(x)],\\ &\text{subject to: } g_i(x) \leqslant 0, \quad i = 1, \ldots, M,\\ &\qquad\qquad\quad h_j(x) = 0, \quad j = 1, \ldots, K, \end{aligned} \tag{22}$$

where O, M, and K are the number of objective functions, inequality constraints and equality constraints, respectively. Also x is a vector of N independent variables of optimization problem (22). The feasible region is a set of x in which all constraints of problem are satisfied. The comparison of two points from feasible region, in order to find the solution of optimization, is not very simple in MOO problems because there is a vector of objective functions. Therefor, to compare two feasible solutions $(\boldsymbol{x_A}, \boldsymbol{x_B})$, it is said that $\boldsymbol{x_A}$ is dominate to $\boldsymbol{x_B}$ when we have

$$\begin{aligned} &f_i(\boldsymbol{x_A}) \leq f_i(\boldsymbol{x_B}),\ i \in \{1, \ldots, O\},\\ &f_j(\boldsymbol{x_A}) < f_j(\boldsymbol{x_B})\ \exists j \in \{1, \ldots, O\}. \end{aligned} \tag{23}$$

A feasible solution x is called "strongly non-dominated" if there is no solution dominated it. The set of non-dominated solutions is called Pareto optimal set. This set should be derived via MOO problem. The set of Pareto optimal outcomes is often called the Pareto frontier.

MOO solution techniques can be categorized in three general categories which are scalarization, meta-heuristic and game theory method [16]. In scalarization based methods, the objective functions are combined to form single-objective optimization problem. The simplest and common method among scalarization based techniques, is weighted sum in which all objective functions are summed together by considering different weights for each of them as

$$\begin{aligned} &\min_x \sum_{q=1}^{O} w_q \frac{f_q(x)}{N_q}\\ &\text{subject to: } g_i(x) < 0, \quad i = 1, \ldots, M,\\ &\qquad\qquad\quad h_j(x) = 0, \quad j = 1, \ldots, K, \end{aligned}$$

where $0 \leq w_q \leq 1$ and $\sum_{q=1}^{O} w_q = 1$. In other word, the values of w_q, $q = 1, .., O$ specify the priority of objective functions relative to each other. N_q is the normalizing factor due to the fact that the different objective functions have different units (e.g. bit per second for throughput and joule for energy efficiency) and values of objective functions are not in the same range. N_q can be equal to minimum or maximum value of $f_q(x)$ [43]. It can be derived by solving optimization problem by considering only $f_q(x)$ as an objective function. In this method, the Pareto optimal solution can be derived by considering different w_q for objective functions. This method is computationally efficient compared to the other scalarization methods and do not change the original optimization problem by not adding any new constraint [16].

REFERENCES

[1] K. Mallinson, "The path to 5G: As much evolution as revolution," 3GPP Article, May 2016.

[2] China Mobile, "C-RAN the road towards green RAN," China Mobile Res. Inst., White Paper ver. 2.5, 2011, pp. 1–10.

[3] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *Proc. 2nd Workshop Hot Topics Softw. Defined Netw.*, 2013, pp. 25–30.

[4] H. Wen, P. K. Tiwary, and T. Le-Ngoc, *Wireless Virtualization*. Springer, 2013.

[5] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "OpenRAN: A software-defined RAN architecture via virtualization," *ACM SIGCOMM Comput. Commun.*, vol. 43, no. 4, pp. 549–550, 2013.

[6] S. Zhou, T. Zhao, Z. Niu, and S. Zhou, "Software-defined hyper-cellular architecture for green and elastic wireless access," *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 12–19, Jan. 2016.

[7] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[8] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul./Aug. 2016.

[9] J. Liu, S. Xu, S. Zhou, and Z. Niu, "Redesigning fronthaul for next-generation networks: Beyond baseband samples and point-to-point links," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 90–97, Oct. 2015.

[10] X. Chen, Z. Han, Z. Chang, G. Xue, H. Zhang, and M. Bennis, "Adapting downlink power in fronthaul-constrained hierarchical software-defined RANs," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[11] K. Liang, L. Zhao, X. Chu, and H.-H. Chen, "An integrated architecture for software defined and virtualized radio access networks with fog computing," *IEEE Netw.*, vol. 31, no. 1, pp. 80–87, Jan./Feb. 2017.

[12] U. Dötsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Tech. J.*, vol. 18, no. 1, pp. 105–128, Jun. 2013.

[13] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "Control-data separation across edge and cloud for uplink communications in C-RAN," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[14] Y. Hao, B. 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.

[15] M. S. Omar *et al.*, "Multiobjective optimization in 5G hybrid networks," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1588–1597, Jun. 2018.

[16] J.-H. Cho, Y. Wang, I.-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.

[17] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, 3rd Quart., 2016.

[18] N. McKeown *et al.*, "OpenFlow: Enabling innovation in campus networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, 2008.

[19] *Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV*, document ETSI GS NFV 003, 2014.

[20] Z. Niu, S. Zhou, S. Zhou, X. Zhong, and J. Wang, "Energy efficiency and resource optimized hyper-cellular mobile communication system architecture and its technical challenges," *Sci. China Inf. Sci.*, vol. 42, no. 10, pp. 1191–1203, 2012.

[21] H. Ishii, Y. Kishiyama, and H. Takahashi, "A novel architecture for LTE-B: C-plane/U-plane split and phantom cell concept," in *Proc. IEEE GLOBECOM Workshop*, Dec. 2012, pp. 624–630.

[22] *Study on Small Cell Enhancements for E-UTRA and E-UTRAN–Higher Layer Aspects*, document 3GPP TR 36.842, 2013.

[23] X. Xu, G. He, S. Zhang, Y. Chen, and S. Xu, "On functionality separation for green mobile networks: Concept study over LTE," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 82–90, May 2013.

[24] A. Mohamed, O. Onireti, M. Imran, A. Imrany, and R. Tafazolli, "Correlation-based adaptive pilot pattern in control/data separation architecture," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 2233–2238.

[25] Z. Zaidi, V. Friderikos, and M. A. Imran, "Future RAN architecture: SD-RAN through a general-purpose processing platform," *IEEE Veh. Technol. Mag.*, vol. 10, no. 1, pp. 52–60, Mar. 2015.

[26] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis, "Fronthaul and backhaul requirements of flexibly centralized radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 105–111, Oct. 2015.

[27] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," *Comput. Netw.*, vol. 85, pp. 1–8, Jul. 2015.

[28] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A flexible and programmable platform for software-defined radio access networks," in *Proc. 12th Int. Conf. Emerg. Netw. Exp. Technol.*, 2016, pp. 427–441.

[29] S. Parsaeefard, A. R. Sharafat, and N. Mokari, *Robust Resource Allocation in Future Wireless Networks*. New York, NY, USA: Springer, 2017.

[30] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1656–1686, 3rd Quart., 2016.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[32] 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.

[33] L. Deng, Y. Rui, P. Cheng, J. Zhang, Q. T. Zhang, and M. Li, "A unified energy efficiency and spectral efficiency tradeoff metric in wireless networks," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 55–58, Jan. 2013.

[34] Q.-V. Pham and W.-J. Hwang, "Fairness-aware spectral and energy efficiency in spectrum-sharing wireless networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10207–10219, Nov. 2017.

[35] 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.

[36] P. Luong, F. Gagnon, C. Despins, and L.-N. 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, Nov. 2017.

[37] M. Derakhshani, S. Parsaeefard, T. Le-Ngoc, and A. Leon-Garcia, "Leveraging synergy of SDWN and multi-layer resource management for 5G networks," *IET Netw.*, Feb. 2018, doi: 10.1049/iet-net.2017.0004.

[38] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

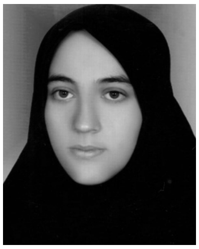
[39] J. Duchi, S. Boyd, and J. Mattingley, "Sequential convex programming," Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, Tech. Rep. EE364b, 2007.

[40] M. Razaviyayn, M. Hong, and Z.-Q. Lue, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

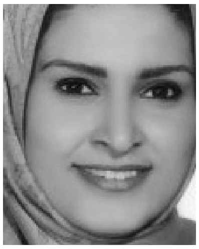
[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[42] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, 2003.

[43] R. T. Marler and J. S. Arora, "Function-transformation methods for multi-objective optimization," *Eng. Optim.*, vol. 37, no. 6, pp. 551–570, 2005.



MINA BAGHANI received the B.Sc. and M.Sc. degrees from Shahed University, Tehran, Iran, in 2008 and 2011, respectively, and the Ph.D. degree in electrical engineering from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, in 2017. Her current research interests include multilayer coding, nonlinear optimization methods, cognitive radio networks, and resource allocation in wireless networks.



SAEEDEH PARSAEEFARD (S'09–M'14) received the B.Sc. and M.Sc. degrees from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2003 and 2006, respectively, and the Ph.D. degree in electrical and computer engineering from Tarbiat Modares University, Tehran, in 2012. From 2010 to 2011, she was a Visiting Ph.D. Student with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA, USA. She was a Post-Doctoral Research Fellow of the Telecommunication and Signal Processing Laboratory, Department of Electrical and Computer Engineering, McGill University, Canada, in 2013. Her current research interests include the applications of robust optimization theory and game theory on the resource allocation and management in wireless networks. She received the IEEE Women in Engineering Award in Iran in 2018.

THO LE-NGOC (F'97) received the B.Eng. degree (Hons.) in electrical engineering and the M.Eng. degree from McGill University, Montreal, QC, Canada, in 1976 and 1978, respectively, and the Ph.D. degree in digital communications from the University of Ottawa, Canada, in 1983. From 1977 to 1982, he was with Spar Aerospace Ltd., Sainte-Anne-de-Bellevue, Canada, where he was involved in the development and design of satellite communications systems. From 1982 to 1985, he was with SR Telecom Inc., Saint-Laurent, QC, Canada, where he developed the new point-to-multipoint DA-TDMA/TDM subscriber radio system SR500. From 1985 to 2000, he was a Professor with the Department of Electrical and Computer Engineering, Concordia University, Montreal. Since 2000, he has been with the Department of Electrical and Computer Engineering, McGill University. His research interest is in the area of broadband digital communications. He is a fellow of The Engineering Institute of Canada, The Canadian Academy of Engineering, and the Royal Society of Canada. He was a recipient of the 2004 Canadian Award in telecommunications research and the IEEE Canada Fessenden Award in 2005. He is currently the Canada Research Chair (Tier I) in Broadband Access Communications.