Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks

Ying Loong Lee, *Member, IEEE,* Jonathan Loo, Teong Chee Chuah, and Li-Chun Wang, *Fellow, IEEE*

***Abstract*—Multitenant cellular network slicing has been gaining huge interest recently. However, it is not well-explored under the heterogeneous cloud radio access network (H-CRAN) architecture. This paper proposes a dynamic network slicing scheme for multitenant H-CRANs, which takes into account tenants' priority, baseband resources, fronthaul and backhaul capacities, quality of service (QoS) and interference. The framework of the network slicing scheme consists of an upper-level, which manages admission control, user association and baseband resource allocation; and a lower-level, which performs radio resource allocation among users. Simulation results show that the proposed scheme can achieve a higher network throughput, fairness and QoS performance compared to several baseline schemes.**

***Index Terms*—Multitenancy, H-CRAN, network slicing, resource management**

I. Introduction

THE demand for broadband multimedia applications has been rising exponentially, leading to a dramatic increase in network capital and operating expenditures. Network sharing has been introduced as a promising solution to the problem by allowing multiple operators to share the network infrastructure and resources [2]. For 5G systems, network virtualization emerges as the key enabling technology of network sharing [3]–[6]. It allows network sharing to be performed in the form of multitenancy, whereby multiple virtual networks can be created on physical network hardware, each being known as a *network slice*. The owner of each network slice is known as the *tenant* or virtual network operator (VNO) [7]. The advantage of network slicing is its flexibility to meet different quality of service (QoS) requirements of multiple VNOs, which is one of the key design goals of 5G networks [6]. Also, network slicing allows infrastructure providers to provide "Network-Slice-as-a-Service" [8]–[10] to service providers, specifically those who do not own physical network infrastructures or those with insufficient network coverage. There are two types of network slicing: core and radio access network slicing [10]. In this paper, we focus on the latter where baseband resources, base stations (BSs), radio resources, and transmission power are shared among the tenants to serve their associated users.

The studies in [7], [11]–[21] have investigated network sharing and slicing based on traditional cellular architectures where baseband processing is decentralized to BSs. However, this architecture is energy- and cost-inefficient [22]. Thus, cloud radio access networks (C-RANs) [22] have emerged as a new architecture which centralizes baseband resources to a single baseband unit (BBU) pool that connects to several radio transceiver units, known as remote radio heads (RRHs), via optical fronthaul links. The fact that BBU pools can be virtualized at cloud data centers facilitates network slicing implementation in C-RANs. Researchers have begun to explore network slicing in multitenant C-RANs [1], [23], [24].

Heterogeneous cellular networks, which consist of macrocells overlaid with small-cells, have been intensively studied due to their potential for extending network coverage and improving capacity [25]–[27]. Multitenancy has also been studied under this architecture in [28]. Combining multitenancy with C-RAN is considered in [29], resulting in the heterogeneous C-RAN (H-CRAN) architecture, where both their advantages are inherited. In [1], we have investigated network slicing for multitenant H-CRANs. However, the study only focused on the small-cell tier. In this paper, we extend the network slicing framework for multitenant H-CRANs to both small-cell and macrocell tiers.

The objective of this study is to develop a dynamic network slicing scheme for downlink multitenant H-CRANs. The contributions of this study are summarized as follows:

1) A network slicing model for multitenant H-CRANs is introduced. Here, network slicing is performed as a process of allocating network resources to users associated with different tenants, where each admitted user receives a set of resources (i.e., associated RRH and BBU with certain capacity, subchannels and transmission power levels).
2) A network slicing problem is formulated to maximize the weighted network throughput of the multitenant H-CRAN, subject to the interference, QoS, fronthaul and backhaul capacity, and BBU pool capacity constraints.
3) A two-level dynamic resource management framework is proposed with the upper-level managing user admission, user association and BBU capacity allocation, and the lower-level managing radio resource allocation.
4) For the upper-level, the admission control problem is solved by a dynamic programming approach whose complexity can be tuned; the user association problem

This work was supported in part by the Ministry of Higher Education Malaysia under FRGS grant MMUE/140082 and TM Research & Development under Grant MMUE/160065. A preliminary version of this work appeared in the proceedings of the International Conference on Advances in Electrical, Electronic and Systems Engineering, Putrajaya, Malaysia, Nov. 2016 [1].

Y. L. Lee and T. C. Chuah are with the Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia (e-mail: yingloonglee@gmail.com; tcchuah@mmu.edu.my).

J. Loo is with the School of Computing and Engineering, University of West London, Ealing W5 5RF, UK (e-mail: jonathan.loo@uwl.ac.uk).

L.-C. Wang is with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: lichun@g2.nctu.edu.tw).

TABLE I
LIST OF NOTATION

Notation	Description	Notation	Description
$\mathcal{N}$	Set of tenants: $\{1, \dots, \lvert\mathcal{N}\rvert\}$	$\mathcal{U}$	Set of UEs: $\{1, \dots, \lvert\mathcal{U}\rvert\}$
$\mathcal{K}$	Set of PRBs: $\{1, \dots, \lvert\mathcal{K}\rvert\}$	$\mathcal{U}_n$	Set of UEs belonging to tenant n
$\mathcal{S}$	Set of RRHs: $\{0, 1, \dots, \lvert\mathcal{S}\rvert\}$ where $s = 0$ indicates the M-RRH	a_u	$a_u = 1$ if UE u is admitted to the network; else $a_u = 0$
b_{su}	$b_{su} = 1$ if UE u associates with RRH s; else $b_{su} = 0$	ω_{sku}	$\omega_{sku} = 1$ if PRB k is allocated to UE u that is associated with RRH s, else $\omega_{sku} = 0$
p_{sku}	The power of RRH s on PRB k for UE u	v_n	Priority weight of tenant n
w_u	Priority weight of UE u	$c_{\mathrm{vB},u}$	Virtual BBU capacity of UE u
C_{cBUP}	BBU pool capacity	$R_{\mathrm{min},u}$	Required minimum data rate of UE u
$R_{\mathrm{max},u}$	Required maximum data rate of UE u	g_{sku}	Channel gain of PRB k between UE u and RRH s
Γ_{sku}	SINR experienced by UE u from RRH s on PRB k	P_{AWGN}	AWGN power
B	Bandwidth of a PRB (i.e., 180 kHz)	$I_{\mathrm{max},sku}$	Interference threshold of PRB k for UE u associated with RRH s
C_{fh}	Fronthaul capacity	C_{bh}	Backhaul capacity
$C_{\mathrm{xh},s}$	$C_{\mathrm{xh},s} = C_{\mathrm{bh}}$ if s = 0; else $C_{\mathrm{xh},s} = C_{\mathrm{fh}}$	W_u	$W_u = v_n w_u$ if $u \in \mathcal{U}_n$; else $W_u = 0$
$P_{\mathrm{max},s}$	Maximum allowable transmission power of RRH s	r_{sku}	Capacity achieved by UE u associated with RRH s on PRB k
U_{a}	Set of UEs admitted to the network	$U_{\mathrm{RRH},s}$	Set of admitted UEs that are associated with RRH s
$\bar{g}_{su}$	Wideband channel gain between UE u and RRH s	$\Gamma_{\mathrm{wb},su}$	Wideband SINR experienced by UE u from RRH s
$\Gamma^{\mathrm{LB}}_{sku}$	Lower bound SINR experienced by UE u from RRH s on PRB k	r^{LB}_{sku}	Lower bound capacity achieved by UE u associated with RRH s on PRB k
η^{LB}_{sku}	Lower bound spectral efficiency achieved by UE u associated with RRH s on PRB k		

is solved using a suboptimal low-complexity greedy algorithm; the BBU capacity allocation problem is solved by linear programming.

5) For the lower-level, the resource allocation problem is formulated as a nonconvex mixed-integer programming problem and solved by the Lagrangian dual method.

The rest of this paper is organized as follows: Section II reviews the related studies. Section III presents the system model of a multitenant H-CRAN and formulates the network slicing problem. Section IV presents the proposed network slicing scheme and proposes algorithms for implementation. Section V compares the performance of the proposed scheme with several baseline schemes. Concluding remarks are made in Section VI.

II. RELATED WORK

Several studies have investigated network slicing and sharing for multitenant cellular networks [1], [7], [8], [11]–[21], [23], [24], [28]. In [8], the impact of network slicing to radio access networks in terms of resource utilization, tenants' protection, traffic differentiation, etc was discussed. In [11], resource allocation among tenants was studied using an auction game where tenants sequentially compete for network resources. In [12], an entity named hypervisor was introduced to manage network slicing among tenants using virtualization technology. In [13], a network virtualization substrate was proposed, which runs a slice scheduler that allocates radio resources to the tenants, and a flow scheduler that schedules data transmission for the users in each network slice. The authors in [14] suggested to reserve radio resources for each tenant and to admit users based on resource availability of the associated network slice. Physical and virtual resource sharing for multitenant cellular networks were compared in [15]. In [7], the authors proposed to associate users with the best cell sector in order to balance the loads among network slices and cell sectors. A resource sharing scheme was proposed in [16] to pool and share the radio resources of all network slices among tenants. In [17], resources are assigned to each tenant as per the service-level agreement (SLA) and resource allocation among tenants is performed based on throughput maximization with some degree of fairness. The authors in [18] proposed to assign minimum amounts of radio resources as per the SLA among tenants while device-to-device users and cellular users can share radio resources if they belong to the same tenant. In [19], a heuristic-based admission control and resource allocation scheme was proposed to perform network slicing according to the priority of the tenants. A resource sharing scheme based on software-defined networking was designed in [20] to share network resources among operators under scenarios with multiple radio access technologies (e.g., cellular, WiFi). In [21], the authors devised a fair resource slicing scheme for multitenant networks through game formulation based on the α-fairness utility. A resource negotiation scheme was introduced in [28] for multitenant heterogeneous cellular networks to allow BSs with insufficient radio resources to access the spare resources of their neighboring BSs. In [23], a virtualized multitenant C-RAN architecture was designed to perform capacity allocation among tenants as per the SLA. The authors in [24] suggested an auction approach for network slicing in C-RANs where computational, storage and radio resources are competed for. In [1], a network slicing scheme for multitenant H-CRANs was proposed as a process of assigning BBU capacity, RRHs and radio resources to the users.

Despite the numerous related studies carried out such as [7], [11]–[21], they only focused on traditional cellular architectures which are less energy- and cost-efficient compared to the C-RAN architecture [22]. Although the studies in [1], [23] and [24] considered the C-RAN architecture, the work in [23] and [24] have not jointly considered the fronthaul capacity constraints and baseband resource allocation in the BBU pool whereas the study in [1] only considered network slicing in the small-cell tier. Moreover, most of the related studies in [7],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2017.2789294, IEEE Transactions on Wireless Communications

[11]–[14], [16]–[18], [28] do not consider the diverse multiuser channel quality in the network slicing process for achieving different throughput-fairness tradeoffs in multitenant networks. Therefore, the current work aims to fill the aforementioned gaps by extending the work of [1].

III. System Model and Problem Formulation

A Long Term Evolution (LTE)-based multitenant H-CRAN model consisting of a macrocell and several small-cells is considered in this study. In this model, we assume that each tenant has its own core network (or core network slice) that connects to the H-CRAN. The macrocell RRH (M-RRH) and the small-cell RRHs (S-RRHs) are connected to a cloud BBU pool via the backhaul and fronthaul links, respectively. Also, C/U splitting is assumed in the system model, whereby the control and data planes are separated such that the control plane is managed by the M-RRH in the network [29]. We follow the LTE specifications in which the channel bandwidth is divided into physical resource blocks (PRBs) with each occupying 180 kHz in the frequency domain and 0.5 ms in the time domain [30]. In LTE, a minimum of two PRBs (in the time domain) can be allocated to a UE as resource allocation is performed every transmission time interval (TTI) of 1 ms. Hereafter, we assume that 1) each PRB experiences flat and slow fading; 2) the network is perfectly synchronized; and 3) shared spectrum mode, whereby both M-RRH and S-RRHs can access the entire channel bandwidth, is adopted. For system modeling, we employ a set of notation as given in Table I. It is noteworthy that $\mathcal{U} = \bigcup_{n\in\mathcal{N}} \mathcal{U}_n$ and $\bigcap_{n\in\mathcal{N}} \mathcal{U}_n = \emptyset$.

In this study, we assume that the BBUs are virtualized in the BBU pool and adopt the BBU model in [31] where each virtual BBU associates with one UE and has a certain amount of computational capacity quantified in the form of user data rate. This model has also been considered in [32] and [33] for resource allocation in C-RANs where the computational capacity[1] of the virtual BBU is modeled as the service rate of a processing queue by abstracting the baseband data processing of the BBU. Here, we denote the computational capacity of a virtual BBU required by each UE u as $c_{\mathrm{vB},u}$. We treat the computational capacity of a virtual BBU as its supported maximum user data rate and thus the capacity allocation depends on the QoS requirements of the traffic services provided by the UE's associated tenant. We also assume that the amount of computational resources of the BBU pool is limited by the user data capacity, C_{cBUP}.

Since tenants may impose different priorities over the network due to the nature of their provisioned services, we characterize the priority of each tenant n by $v_n \in [0, 1]$ as in [19], where $\sum_{n\in\mathcal{N}} v_n = 1$. Also, we characterize the priority of UE u by $w_u \in [0, 1]$ to achieve various notions of fairness. Each UE is imposed with a set of QoS requirements depending on the type of services provided by the associated tenant. Here, we model the QoS requirements of each UE u with minimum and maximum data rates, $R_{\mathrm{min},u}$ and $R_{\mathrm{max},u}$, respectively [19]. The minimum data rate requirement is compliant with LTE and it is defined as the guaranteed bit rate (GBR) for inelastic traffic [34], [35]. Nonetheless, the minimum QoS model is still applicable to elastic traffic such as best-effort by imposing zero minimum data rate. On the other hand, the maximum data rate requirement corresponds to the maximum bit rate (MBR) and aggregate-MBR (AMBR) defined in LTE to achieve congestion control for inelastic and elastic traffic, respectively [34], [35]. Besides that, maximum QoS requirements have been used in [36] to limit interference by reducing the number of PRBs allocated to UEs. It is noted that the virtual BBU capacity of each UE u, $c_{\mathrm{vB},u}$ needs to meet its minimum data rate requirement for QoS satisfaction.

For channel modeling, the signal-to-interference-plus-noise ratio (SINR) experienced by UE u at RRH s on PRB k is modeled as [1], [36]

$$\Gamma_{sku} = \frac{p_{sku} g_{sku}}{\sum_{i\in\mathcal{S}\setminus\{s\}} \sum_{j\in\mathcal{U}\setminus\{u\}} a_j b_{ij} \omega_{ikj} p_{ikj} g_{iku} + P_{\mathrm{AWGN}}}, \quad (1)$$

where g_{sku} is the channel gain experienced by UE u at RRH s on PRB k, and P_{AWGN} is the additive white Gaussian noise (AWGN) power. Next, the achievable upper bound capacity of UE u at RRH s on PRB k can be calculated as

$$r_{sku} = B \log_2(1 + \Gamma_{sku}), \quad (2)$$

where B is the bandwidth of a PRB. It is noteworthy that the actual capacity of each UE depends on the modulation and coding scheme (MCS) selected in LTE systems, where the throughput can be calculated as a trunked version of (2). For simplicity, we consider (2) in the problem and solution formulation as well as performance evaluation where the throughput results obtained are upper bound capacity values. To suppress cross-tier and cotier interference, the threshold $I_{\mathrm{max},sku}$ is introduced, which is the maximum allowable interference experienced by UE u associated with RRH s on PRB k. Each RRH has a maximum transmission power level, $P_{\mathrm{max},s}$. Also, each S-RRH (i.e., $s \neq 0$) is connected to the BBU pool via a fronthaul link with user data capacity C_{fh} whereas the M-RRH is connected to the cloud BBU pool via a backhaul link with user data capacity C_{bh}[2]. For notational simplicity, we denote

$$C_{\mathrm{xh},s} = \begin{cases} C_{\mathrm{bh}} & s = 0 \\ C_{\mathrm{fh}} & s \neq 0 \end{cases} \quad (3)$$

[1] The computational capacity here is not equivalent to the computational burden, where the latter refers to the amount of time complexity required by the virtual BBU to run a baseband process within a time interval. Since the computational burden of a virtual BBU depends on the number of PRBs and MCS allocated to the user, we assume that the peak computational burden allowable for a virtual BBU is scaled such that the virtual BBU can support the desired computational capacity considering that the associated user is allocated maximum number of PRBs and the highest order MCS during virtual BBU capacity allocation. In this case, the virtual BBU can still support the desired computational capacity regardless of the number of PRBs and MCS allocated to the associated user. Such scaling can be achieved by modifying the specifications of the virtual BBU, e.g., the number of processor cores, core clock speed and memory capacity.

[2] The practical capacity of an optical backhaul or fronthaul link could be up to several Gb/s, which is much higher than the user data rate, because the user data is encapsulated by the common public radio interface (CPRI) protocol, which transforms the user data into a very high-frequency waveform. For instance, a single-sector LTE network with 20 MHz bandwidth and 2×2 MIMO configuration, which can achieve up to 150 Mb/s downlink user data rate, would require a CPRI rate of 2457.6 Mb/s with 10/8 coding and 15+1 IQ sample width [22], [37]. For more information about CPRI, backhaul and fronthaul, readers may refer to [38] and [39]. In this paper, we refer both the backhaul and fronthaul capacity to as their user data capacity.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2017.2789294, IEEE Transactions on Wireless Communications

In this paper, network slicing of multitenant H-CRANs can be treated as a resource allocation process that involves 1) virtual BBU capacity allocation; 2) UE admission and association; and 3) PRB and power allocation. First of all, let

$$W_u = \begin{cases} v_n w_u & u \in \mathcal{U}_n \\ 0 & u \notin \mathcal{U}_n. \end{cases} \tag{4}$$

Then, a generic network slicing problem for multitenant H-CRANs can be formulated as

$$\max_{a_u, b_{su}, c_{\mathrm{vB},u}, \omega_{sku}, p_{sku}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}} W_u a_u b_{su} \omega_{sku} r_{sku} \tag{5}$$

subject to

$$a_u \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} b_{su} \omega_{sku} r_{sku} \geq a_u R_{\mathrm{min},u} \quad \forall u \in \mathcal{U} \tag{5a}$$

$$a_u \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} b_{su} \omega_{sku} r_{sku} \leq a_u c_{\mathrm{vB},u} \quad \forall u \in \mathcal{U} \tag{5b}$$

$$a_u \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} b_{su} \omega_{sku} r_{sku} \leq a_u R_{\mathrm{max},u} \quad \forall u \in \mathcal{U} \tag{5c}$$

$$\sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}} a_u b_{su} \omega_{sku} p_{sku} \leq P_{\mathrm{max},s} \quad \forall s \in \mathcal{S} \tag{5d}$$

$$\sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}} a_u b_{su} \omega_{sku} r_{sku} \leq C_{\mathrm{xh},s} \quad \forall s \in \mathcal{S} \tag{5e}$$

$$\sum_{i \in \mathcal{S} \backslash \{s\}} \sum_{j \in \mathcal{U} \backslash \{u\}} a_j b_{ij} \omega_{ikj} p_{ikj} g_{iku} \leq I_{\mathrm{max},sku} \qquad \forall s \in \mathcal{S}, k \in \mathcal{K}, k \in \mathcal{U} \tag{5f}$$

$$a_u R_{\mathrm{min},u} \leq a_u c_{\mathrm{vB},u} \leq a_u R_{\mathrm{max},u} \quad \forall u \in \mathcal{U} \tag{5g}$$

$$\sum_{u \in \mathcal{U}} a_u c_{\mathrm{vB},u} \leq C_{\mathrm{cBUP}} \tag{5h}$$

$$a_u \sum_{s \in \mathcal{S}} b_{su} \leq 1 \quad \forall u \in \mathcal{U} \tag{5i}$$

$$\sum_{u \in \mathcal{U}} a_u b_{su} \omega_{sku} \leq 1 \quad \forall s \in \mathcal{S}, k \in \mathcal{K} \tag{5j}$$

$$a_u, b_{su}, \omega_{sku} \in \{0, 1\} \quad \forall s \in \mathcal{S}, k \in \mathcal{K}, u \in \mathcal{U} \tag{5k}$$

$$c_{\mathrm{vB},u}, p_{sku} \geq 0 \quad \forall s \in \mathcal{S}, k \in \mathcal{K}, u \in \mathcal{U}. \tag{5l}$$

Constraint (5a) ensures that each admitted UE is guaranteed a minimum data rate. The UE data rate is bounded by its associated virtual BBU capacity and the maximum data rate in constraints (5b) and (5c) respectively. The total transmission power of each RRH is limited in constraint (5d). Constraint (5e) implies that the total UE data rate carried over the backhaul or fronthaul link is bounded by its capacity. Constraint (5f) limits the interference experienced by each UE u associated with RRH s on each PRB k. Constraint (5g) ensures that the virtual BBU capacity of each UE is allocated an amount between its required minimum and maximum data rates. Constraint (5h) guarantees that the sum capacity of all virtual BBUs does not exceed the BBU pool capacity. Constraint (5i) permits each UE to associate with only one RRH. Each PRB will not be allocated to two or more UEs associating with the same RRH, as enforced by constraint (5j). Constraint (5k) ensures binary-valued a_u, b_{su} and ω_{sku}. Constraint (5l) ensures nonnegative p_{sku} and $c_{\mathrm{vB},u}$.

Let R_u be the achievable *instantaneous rate* of UE u and $\bar{R}_u$ be the past *average rate* achieved by UE u estimated with an exponentially weighted low-pass time window. It has been shown in [40] and [41] that if the time window length is sufficiently large or the time-averaging step size is sufficiently small, a network utility maximization (NUM) problem which maximizes the total utility with respect to $\bar{R}_u$ is equivalent to a weighted sum rate maximization problem with respect to R_u where the weight corresponding to UE u is the gradient of the utility with respect to $\bar{R}_u$. Clearly, (5) is one such gradient-based weighted sum rate maximization problem with W_u being the gradient of the utility with respect to the average rates. Thus, the gradient-based framework of (5) has the following implications [40]:

1) The gradient-based framework provides better flexibility to enhance spectral efficiency due to the time window used for averaging $\bar{R}_u$ where the length of which relaxes the fairness requirement, unlike the classical NUM problem that maximizes the total utility with respect to R_u which requires achieving fairness at each time epoch;
2) The gradient-based framework reduces computational complexity because the gradient of the utility relies only on previous resource allocation (e.g., past average rate);
3) The linear objective function of the weighted sum rate maximization problem in (5) can simplify the solution algorithm and achieves faster convergence;
4) It has been shown that improvements of actual average user rate can be achieved.

As such, various notions of fairness with respect to $\bar{R}_u$ can be achieved by equating W_u of (5) with the gradient of a generic fairness utility. One such utility is the weighted α-fairness function [42], [43] where $\alpha \geq 0$ is a fairness parameter. This leads to the following proposition.

Proposition 1: The solution to (5) is weighted α-fair with respect to $\bar{R}_u$ when $w_u = 1/\bar{R}_u^{\alpha}$.

Proof: Refer to Appendix A. ■

It is noteworthy that adoption of the weighted α-fairness function as the utility for the gradient-based framework of (5) as in Proposition 1 achieves fairness with respect to the *long-term average rates*. This is different from the classical NUM problem that maximizes the same fairness utility but with respect to the *instantaneous rates*. Nevertheless, the gradient-based framework is more preferred when long-term average rates are concerned, and thus is considered in this study.

To solve (5), we analyze the complexity of the problem and obtain the following:

Proposition 2: The problem in (5) is generally NP-hard.

Proof: Refer to Appendix B. ■

As such, (5) is computationally intractable. Moreover, solving this problem directly is impractical because rapid channel variations would trigger a rapidly repeated network slicing process, which would burden higher-level functions of the network and results in increased signaling overhead. Thus, we propose a two-level hierarchical approach to the problem in the next section.

IV. Proposed Network Slicing Scheme

In this section, we propose an efficient two-level network slicing framework for a multitenant H-CRAN (cf. Fig. 1).

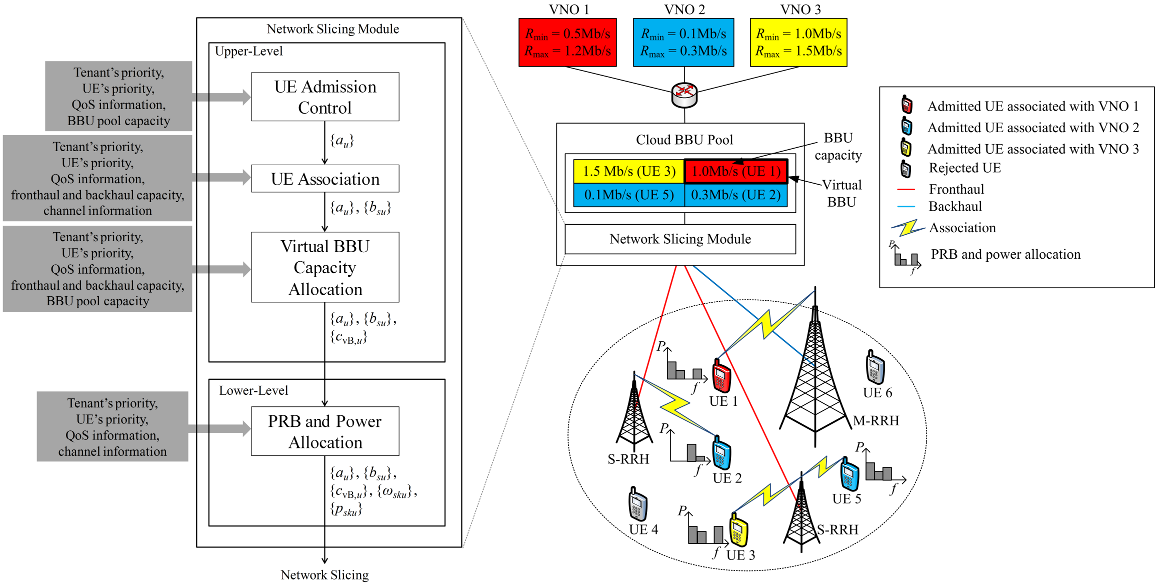


Fig. 1. Proposed network slicing framework.

A. Upper-Level

In this level, differentiated prioritization among UEs is performed. Firstly, admission control is performed to admit high-priority UEs while considering their QoS requirements and the availability of the BBU pool capacity. Then, the admitted UEs are associated, according to their priority, with the RRHs that provide them with high channel quality, subject to availability of fronthaul and backhaul capacity. Next, a virtual BBU is created for each admitted UE with a capacity that meets QoS requirements.

1) Admission Control: Here, we intend to maximize the number of UEs admitted to the network, which corresponds to maximizing the network throughput while considering their priority. Clearly, the number of admitted UEs can be increased if $c_{\mathrm{vB},u}$ is smaller. Since $c_{\mathrm{vB},u} \geq R_{\min,u}$, we can formulate the following admission control problem:

$$\max_{a_u} \sum_{u\in\mathcal{U}} W_u a_u \tag{6}$$

subject to

$$\sum_{u\in\mathcal{U}} a_u R_{\min,u} \leq R_{\mathrm{cBUP}} \tag{6a}$$

$$a_u \in \{0,1\} \quad \forall u \in \mathcal{U} \tag{6b}$$

where constraint (6a) is derived from constraint (5h) to ensure that the total minimum data rate of all admitted UEs does not exceed the BBU pool capacity.

Proposition 3: The admission control problem in (6) can be classified as an NP-complete 0-1 knapsack problem [44].

Proof: Refer to Appendix C. ■

Dynamic programming has been widely used to obtain the optimal solution for 0-1 knapsack problems. The key idea of dynamic programming is to decompose (6) into small problems and recursively calculate their objective function value based on the solutions for the even smaller problems [45]. The maximum objective function value can then be obtained by bottom-up computation using a table structure; the solution for the original problem can be therein obtained. Here, we design an admission control algorithm (cf. Algorithm 1) based on a dynamic programming approach in [46, pp. 266-272]. In steps 1-2 of Algorithm 1, we first set a number N_{w} to produce a set of discrete integer values, i.e., $\{0, 1, 2, \ldots, N_{\mathrm{w}}\}$, which corresponds to a set of BBU pool capacities $\{0, d, 2d, \ldots, N_{\mathrm{w}}d\}$ where $N_{\mathrm{w}}d = C_{\mathrm{cBUP}}$. Further, $V(u,i)$ is defined to store the maximum value of (6) for any subset of $\{1, 2, \ldots, u\}$ with the capacity being at most id. Besides that, $X(u,i)$ is defined to trace the solution obtained from solving the problem in (6). Both $V(u,i)$ and $X(u,i)$ are set to zero for all $u \in \mathcal{U}$

Algorithm 1 Dynamic programming-based UE admission control algorithm

```
1: Set $N_{\mathrm{w}}$, and set $V(u,i) = X(u,i) = 0$ and $V(0,i) = 0$ for all $u \in \mathcal{U}$ and $i \in \{0, 1, \ldots, N_{\mathrm{w}}\}$.
2: Calculate $d = \frac{C_{\mathrm{cBUP}}}{N_{\mathrm{w}}}$, set $a_u = 0$ for all $u \in \mathcal{U}$ and set $N_{\mathrm{tmp}} = N_{\mathrm{w}}$.
3: for all $u \in \mathcal{U}$ do
4:     for all $i \in \{0, 1, \ldots, N_{\mathrm{w}}\}$ do
5:         if $R_{\min,u} \leq id$ and $V(u-1,i) < W_u + V\left(u-1, i - \left\lceil \frac{R_{\min,u}}{d} \right\rceil\right)$ then
6:             Calculate $V(u,i) = W_u + V\left(u-1, i - \left\lceil \frac{R_{\min,u}}{d} \right\rceil\right)$
7:             Set $X(u,i) = 1$.
8:         else
9:             Set $V(u,i) = V(u-1,i)$.
10:        end if
11:    end for
12: end for
13: Set $u = |\mathcal{U}|$.
14: while $u > 1$ do
15:     if $X(u, N_{\mathrm{tmp}}) = 1$ then
16:         Set $a_u = 1$.
17:         Calculate $N_{\mathrm{tmp}} = N_{\mathrm{tmp}} - \left\lceil \frac{R_{\min,u}}{d} \right\rceil$.
18:         $u = u - 1$.
19:     end if
20: end while
```

and $i \in \{0, 1, 2, \ldots, N_{\mathrm{w}}\}$ initially. In addition, $V(0, i) = 0$ for all $i \in \{0, 1, 2, \ldots, N_{\mathrm{w}}\}$ is set to provide the value of (6) corresponding to the fact that no UEs are admitted. Also, $a_u = 0$ is set for all $u \in \mathcal{U}$, and $N_{\mathrm{tmp}} = N_{\mathrm{w}}$ is initialized. Next, in steps 3-12, the solution is obtained by computing $V(u, i)$, which can be obtained as follows:

Lemma 1: For $u \in \mathcal{U}$ and $i \in \{0, 1, 2, \ldots, N_{\mathrm{w}}\}$, $V(u, i)$ can be calculated as

$$V(u,i) = \begin{cases} \max\left(V(u-1,i), W_u \right. \\ \qquad \left. + V\left(u-1, i - \left\lceil \frac{R_{\mathrm{min},u}}{d} \right\rceil \right)\right) & \text{if } R_{\mathrm{min},u} \leq id \\ V(u-1,i) & \text{otherwise,} \end{cases} \tag{7}$$

Proof: Refer to Appendix D. ■

While computing $V(u, i)$, $X(u, i)$ is set to one if UE u is admitted and zero otherwise. Subsequently, in steps 13 - 20, the solution a_u is obtained based on $X(u, i)$.

The asymptotic computational complexity of Algorithm 1 can be shown to be of $O(|\mathcal{U}|N_{\mathrm{w}})$, which implies that the algorithm has a pseudo-polynomial time complexity due to N_{w} that is defined for (6). Conventionally, by setting $N_{\mathrm{w}} = C_{\mathrm{cBUP}}$, i.e., $d = 1$, the solution obtained using Algorithm 1 is optimal for (6) by definition of (6) and Lemma 1. In this case, the introduction of parameter d is redundant. However, C_{cBUP} could be too large as it would incur a prohibitively high computational complexity and the set of capacities for Algorithm 1 becomes very large, which requires large memory to store the objective function values. Notwithstanding that, our algorithm provides the option to reduce computational complexity and memory size. By setting $N_{\mathrm{w}} << C_{\mathrm{cBUP}}$, i.e., $d >> 1$, the complexity and memory size can be reduced, but possibly at the cost of solution optimality. Fig. 2 illustrates the performance tradeoff of Algorithm 1 with different N_{w} for a particular scenario. Interestingly, certain values of N_{w} can still lead to optimal solutions. In fact, those values correspond to a set of d, which are the factors of $R_{\mathrm{min},u}$. As such, the general rule of thumb for Algorithm 1 to be optimal is to select a value of N_{w} such that d is a factor of $\min\{R_{\mathrm{min},u}\}_{u \in \mathcal{U}}$. As such, Algorithm 1 can still optimally solve (6) with reasonable complexity.

2) UE Association: After solving UE admission, the rejected UEs will be omitted from the network slicing process. Next, we proceed to solve the UE association problem with the objective to associate admitted UEs with RRHs that could provide high channel quality with consideration of the UE's priority. Firstly, the wideband SINR received by each admitted UE is estimated as follows:

$$\Gamma_{\mathrm{wb},su} = \frac{P_{\mathrm{max},s}\bar{g}_{su}}{\sum_{i \in \mathcal{S}\setminus\{s\}} P_{\mathrm{max},i}\bar{g}_{iu} + P_{\mathrm{AWGN}}}, \tag{8}$$

where $\Gamma_{\mathrm{wb},su}$ and $\bar{g}_{su}$ are the wideband SINR and average gain across the channel bandwidth between RRH s and UE u, respectively. Next, we can formulate the UE association problem as

$$\max_{b_{su}} \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}_{\mathrm{a}}} b_{su} W_u \Gamma_{\mathrm{wb},su} \tag{9}$$

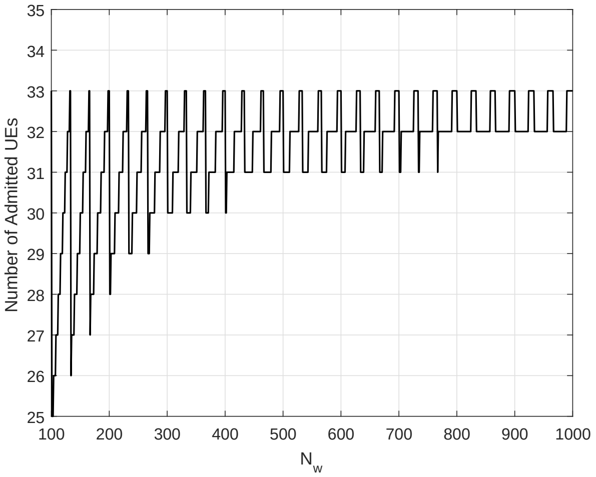


Fig. 2. Performance tradeoff of Algorithm 1 for the following particular scenario: two equal-priority tenants with each having 20 UEs, $C_{\mathrm{cBUP}} = 3000$, $R_{\mathrm{min},u} = 30$ for all UEs, and all UE have equal priority.

subject to

$$\sum_{u \in \mathcal{U}_{\mathrm{a}}} b_{su} R_{\mathrm{min},u} \leq C_{\mathrm{xh},s} \quad \forall s \in \mathcal{S} \tag{9a}$$

$$\sum_{s \in \mathcal{S}} b_{su} = 1 \quad \forall u \in \mathcal{U}_{\mathrm{a}} \tag{9b}$$

$$b_{su} \in \{0, 1\} \quad \forall s \in \mathcal{S}, u \in \mathcal{U}_{\mathrm{a}}, \tag{9c}$$

where $\mathcal{U}_{\mathrm{a}} = \{u \in \mathcal{U} | a_u = 1\}$ is the set of UEs admitted to the network. Constraint (9a) is derived from (5e) to ensure that the fronthaul (or backhaul) link can at least carry the minimum data rate required by the associated UEs, whereas (9b) is derived from (5i) to guarantee that each admitted UE associates with only one RRH.

Proposition 4: The problem in (9) can be classified as an NP-complete 0-1 multiple knapsack problem [44].

Proof: Refer to Appendix E. ■

For 0-1 multiple knapsack problems, dynamic programming techniques are inefficient as the computational complexity would be prohibitively much larger than solving a 0-1 knapsack problem. Therefore we propose a greedy algorithm for (9) to find a feasible suboptimal solution (cf. Algorithm 2). The key idea of Algorithm 2 is to associate the UE with the RRH that provides the highest weighted wideband SINR if the fronthaul (or backhaul) link of the S-RRH (or M-RRH) still has sufficient capacity to accommodate the UE. Firstly, the weighted wideband SINR of each UE is calculated for all $s \in \mathcal{S}$. Then, the RRH that provides the highest weighted wideband SINR value to the UE will be selected to evaluate (9a). If (9a) holds, the UE will associate with the RRH. Otherwise, the RRH that provides the next highest weighted wideband SINR to the UE will be selected and the same process is repeated. The asymptotic computational complexity of Algorithm 2 is of $O(|\mathcal{S}||\mathcal{U}|)$.

3) Virtual BBU Capacity Allocation: After UE association, the BBU capacity is allocated to each admitted UEs. Here,

Algorithm 2 Greedy UE Association Algorithm

```
Set $\mathcal{S}_{\text{rem}} = \mathcal{S}$, and set $C_{\text{rem},s} = C_{\text{xh},s}$, $f_s = 0$ and $b_{su} = 0$ for all $s \in \mathcal{S}$ and $u \in \mathcal{U}_{\text{a}}$.
for all $u \in \mathcal{U}_{\text{a}}$ do
    for all $s \in \mathcal{S}$ do
        Calculate $f_s = W_u \Gamma_{\text{wb},su}$.
    end for
    Find $s = \arg\max_{i \in \mathcal{S}_{\text{rem}}} f_i$.
    while $C_{\text{rem},s} \geq R_{\text{min},s}$ do
        if $C_{\text{rem},s} \geq R_{\text{min},s}$ then
            Set $b_{su} = 1$ and update $C_{\text{rem},s} = C_{\text{rem},s} - R_{\text{min},s}$.
        else
            $\mathcal{S}_{\text{rem}} = \mathcal{S}_{\text{rem}} \backslash \{s\}$.
        end if
    end while
end for
```

the following virtual BBU capacity allocation problem is introduced:

$$\max_{c_{\text{vB},u}} \sum_{u \in \mathcal{U}_{\text{a}}} W_u c_{\text{vB},u} \tag{10}$$

subject to

$$\sum_{u \in \mathcal{U}_{\text{RRH},s}} c_{\text{vB},u} \leq C_{\text{xh},s} \quad \forall s \in \mathcal{S} \tag{10a}$$

$$R_{\text{min},u} \leq c_{\text{vB},u} \leq R_{\text{max},u} \quad \forall u \in \mathcal{U}_{\text{a}} \tag{10b}$$

$$\sum_{u \in \mathcal{U}_{\text{a}}} c_{\text{vB},u} \leq C_{\text{cBUP}} \tag{10c}$$

$$c_{\text{vB},u} \geq 0 \quad \forall u \in \mathcal{U}_{\text{a}}, \tag{10d}$$

where (10a) and (10b) are derived from (5e) and (5g), respectively, and (10c) is equivalent to (5h). The objective of (10) is to maximize the virtual BBU capacity assigned to each UE with consideration of the UE's priority, fronthaul and backhaul capacities, and BBU pool capacity. Apparently, this problem can be expressed as a linear programming problem, which can be solved by the well-known *simplex method* [45].

B. Lower-Level

After the upper-level, the remaining problem is to allocate PRBs and power among admitted UEs for each RRH. From the upper-level of the proposed network slicing scheme, we obtain the solutions for UE admission, UE association and virtual BBU capacity allocation problems, which reduce the network slicing problem in (5) to

$$\max_{\omega_{sku}, p_{sku}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}_{\text{RRH},s}} W_u \omega_{sku} r_{sku} \tag{11}$$

subject to

$$\sum_{k \in \mathcal{K}} \omega_{sku} r_{sku} \geq R_{\text{min},u} \quad \forall u \in \mathcal{U}_{\text{RRH},s}, s \in \mathcal{S} \tag{11a}$$

$$\sum_{k \in \mathcal{K}} \omega_{sku} r_{sku} \leq r_{\text{vBBU},u} \quad \forall u \in \mathcal{U}_{\text{RRH},s}, s \in \mathcal{S} \tag{11b}$$

$$\sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}_{\text{RRH},s}} \omega_{sku} p_{sku} \leq P_{\text{max},s} \quad \forall s \in \mathcal{S} \tag{11c}$$

$$\sum_{i \in \mathcal{S} \backslash \{s\}} \sum_{j \in \mathcal{U}_{\text{RRH},i}} \omega_{ikj} p_{ikj} g_{iku} \leq I_{\text{max},sku} \quad \forall u \in \mathcal{U}_{\text{RRH},s}, s \in \mathcal{S}, k \in \mathcal{K} \tag{11d}$$

$$\sum_{u \in \mathcal{U}_{\text{RRH},s}} \omega_{sku} \leq 1 \quad \forall s \in \mathcal{S}, k \in \mathcal{K} \tag{11e}$$

$$\omega_{sku} \in \{0, 1\} \quad \forall s \in \mathcal{S}, k \in \mathcal{K}, u \in \mathcal{U}_{\text{RRH},s} \tag{11f}$$

$$p_{sku} \geq 0 \quad \forall s \in \mathcal{S}, k \in \mathcal{K}, u \in \mathcal{U}_{\text{RRH},s}, \tag{11g}$$

where $\mathcal{U}_{\text{RRH},s} = \{u \in \mathcal{U} | a_u = 1, b_{su} = 1\}$ is the set of admitted UEs associated with RRH s. The main objective of (11) is to allocate sufficient PRBs and power to the UEs such that their minimum data rate can be achieved by fully exploiting multiuser diversity of the channel. Although (11) is less complicated than (5), it is still generally NP-hard. Therefore, we simplify the problem into a tractable one and solve it using a conventional resource allocation approach.

Firstly, (11b) can be removed from the problem because, in practice, the maximum aggregate data rate achieved by each UE is always less than its associated BBU capacity, regardless of the amount of allocated PRBs and transmission power. As such, it is not essential to enforce (11b). Furthermore, we relax ω_{sku} into continuous values where (11f) can be transformed into

$$0 \leq \omega_{sku} \leq 1 \quad \forall u \in \mathcal{U}_{\text{RRH},s}, s \in \mathcal{S}, k \in \mathcal{K} \tag{12}$$

and express the lower bound SINR experienced by UE u associated with RRH s on PRB k as

$$\Gamma_{sku}^{\text{LB}} = \frac{p_{sku} g_{sku}}{I_{\text{max},sku} + P_{\text{AWGN}}}. \tag{13}$$

Then, the lower bound of the achievable data rate of UE u associated with RRH s on PRB k can be calculated as

$$r_{sku}^{\text{LB}} = B \log_2(1 + \Gamma_{sku}^{\text{LB}}). \tag{14}$$

Next, we can express the lower bound of (11) as

$$\max_{\omega_{sku}, p_{sku}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}_{\text{RRH},s}} W_u \omega_{sku} r_{sku}^{\text{LB}} \tag{15}$$

subject to (11c)-(11e), (11g), (12) and

$$\sum_{k \in \mathcal{K}} \omega_{sku} r_{sku}^{\text{LB}} \geq R_{\text{min},u} \quad \forall u \in \mathcal{U}_{\text{RRH},s}, s \in \mathcal{S}. \tag{15a}$$

We can further simplify the problem and rewrite (15) as

$$\max_{\omega_{sku} p_{sku}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}_{\text{RRH},s}} W_u \omega_{sku} \eta_{sku}^{\text{LB}} \tag{16}$$

subject to (11c)-(11e), (11g), (12) and

$$\sum_{k \in \mathcal{K}} \omega_{sku} \eta_{sku}^{\text{LB}} \geq \frac{R_{\text{min},u}}{B} \quad \forall u \in \mathcal{U}_{\text{RRH},s}, s \in \mathcal{S} \tag{16a}$$

where $\eta_{sku}^{\text{LB}} = r_{sku}^{\text{LB}}/B$ is the achievable lower bound spectral efficiency of UE u associated with RRH s on PRB k. We can interpret (16) as a weighted spectral efficiency maximization problem of the network.

Next, we can write the Lagrangian function of (16) as

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\omega},\mathbf{p},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\tau},\boldsymbol{\rho})\\ &=\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}} W_u\omega_{sku}\eta_{sku}^{\mathrm{LB}}\\ &+\sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\beta_{su}\left(\sum_{k\in\mathcal{K}}\omega_{sku}\eta_{sku}^{\mathrm{LB}}-\frac{R_{\mathrm{min},u}}{B}\right)\\ &+\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\lambda_{sku}\left(I_{\mathrm{max},sku}-\sum_{i\in\mathcal{S}\backslash\{s\}}\sum_{j\in\mathcal{U}_{\mathrm{RRH},i}}\omega_{ikj}p_{ikj}g_{iku}\right)\\ &+\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\mu_{sk}\left(1-\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\omega_{sku}\right)+\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\rho_{sku}p_{sku}\\ &+\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\tau_{sku}(1-\omega_{sku})+\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\nu_{sku}\omega_{sku}\\ &+\sum_{s\in\mathcal{S}}\phi_s\left(P_{\mathrm{max},s}-\sum_{k\in\mathcal{K}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\omega_{sku}p_{sku}\right),\end{aligned}\tag{17}$$

where β_{su}, ϕ_s, λ_{sku}, μ_{sk}, ν_{sku}, τ_{sku} and ρ_{sku} are the nonnegative Lagrange multipliers corresponding to constraints (16a), (11c)-(11e), (12) and (11g), respectively. Then, the Lagrangian dual function of (16) can be expressed as $\mathcal{D}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\tau},\boldsymbol{\rho}) = \max_{\omega_{sku},p_{sku}} \mathcal{L}(\boldsymbol{\omega},\mathbf{p},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\tau},\boldsymbol{\rho})$ and thus the dual optimization problem can be formulated as

$$\min_{\beta_{su},\phi_s,\lambda_{sku},\mu_{sk},\nu_{sku},\tau_{sku},\rho_{sku}} \mathcal{D}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\tau},\boldsymbol{\rho}).\tag{18}$$

Note that any dual problem (e.g., (18)) generally leads to the upper bound solution for the primal problem (e.g., (16)) due to the existence of a nonzero duality gap. However, it has been proved in [47] that the duality gap is nearly zero in a multicarrier system such as LTE if the number of PRBs is sufficiently large. Thus, solving (18) provides a near-optimal solution to (16).

Suppose that UE u has been allocated PRB k, i.e., $\omega_{sku} = 1$. Then, the derivative of (17) with respect to p_{sku} gives the following Karush-Kuhn-Tucker (KKT) conditions [48], [49]:

$$\frac{\partial\mathcal{L}}{\partial p_{sku}} = \omega_{sku}(G_{sku} - \phi_s) + \rho_s = 0\tag{19}$$

$$\omega_{sku}p_{sku}(G_{sku} - \phi_s) = 0\tag{20}$$

where

$$\begin{aligned}G_{sku} = &\frac{g_{sku}(W_u + \beta_{su})}{(I_{\mathrm{max},sku} + P_{\mathrm{AWGN}} + p_{sku}g_{sku})\ln 2}\\ &-\sum_{i\in\mathcal{S}\backslash\{s\}}\sum_{j\in\mathcal{U}_{\mathrm{RRH},i}}\lambda_{ikj}g_{skj}.\end{aligned}\tag{21}$$

From (20), the power allocation can be performed for all $u \in \mathcal{U}_{\mathrm{RRH},s}$, $s \in \mathcal{S}$ and $k \in \mathcal{K}$ as

$$\begin{aligned}p_{sku} =\Bigg[&\frac{W_u + \beta_{su}}{\left(\phi_s + \sum_{i\in\mathcal{S}\backslash\{s\}}\sum_{j\in\mathcal{U}_{\mathrm{RRH},i}}\lambda_{ikj}g_{skj}\right)\ln 2}\\ &-\frac{I_{\mathrm{max},sku} + P_{\mathrm{AWGN}}}{g_{sku}}\Bigg]^+\end{aligned}\tag{22}$$

where $[x]^+ = \max(0, x)$. Next, the derivative of (17) with respect to ω_{sku} yields the following KKT conditions:

$$\frac{\partial\mathcal{L}}{\partial\omega_{sku}} = H_{sku} - \mu_s + \nu_{sku} - \tau_{sku} = 0\tag{23}$$

$$\omega_{sku}(H_{sku} - \mu_s - \tau_{sku}) = 0\tag{24}$$

where

$$H_{sku} = (W_u + \beta_{su})\eta_{sku}^{\mathrm{LB}} - \left(\phi_s + \sum_{i\in\mathcal{S}\backslash\{s\}}\sum_{j\in\mathcal{U}_{\mathrm{RRH},i}}\lambda_{ikj}g_{skj}\right)p_{sku}.\tag{25}$$

Here, (23)-(24) yield $H_{sku} - \mu_{sk} - \tau_{sku} = 0$ if $\omega_{sku} = 1$ and $H_{sku} - \mu_{sk} - \tau_{sku} \leq 0$ if $\omega_{sku} = 0$. As such, we can deduce that the UE that corresponds to the largest value of H_{sku} among all $u \in \mathcal{U}_{\mathrm{RRH},s}$, $s \in \mathcal{S}$ and $k \in \mathcal{K}$ should be allocated the PRB. Since one PRB can only be allocated to a UE among those associated with the same RRH, which is enforced in (11e), the optimal PRB allocation can be acquired by

$$\omega_{sku} = \begin{cases}1 & u = \arg\max_{i\in\mathcal{U}_{\mathrm{RRH},s}} H_{ski}\\ 0 & \text{otherwise}\end{cases}\quad \forall s \in \mathcal{S}, k \in \mathcal{K}.\tag{26}$$

With (22) and (26), we can solve the dual problem in (17) using an iterative subgradient method [50] to update the Lagrange multipliers:

$$\beta_{su}^{(t+1)} = \left[\beta_{su}^{(t)} - \delta_1\left(\sum_{k\in\mathcal{K}}\omega_{sku}\eta_{sku}^{\mathrm{LB}} - \frac{R_{\mathrm{min},u}}{B}\right)\right]^+ \quad \forall u \in \mathcal{U}_{\mathrm{RRH},s}, s \in \mathcal{S}\tag{27}$$

$$\phi_s^{(t+1)} = \left[\phi_s^{(t)} - \delta_2\left(P_{\mathrm{max},s} - \sum_{k\in\mathcal{K}}\sum_{u\in\mathcal{U}_{\mathrm{RRH},s}}\omega_{sku}p_{sku}\right)\right]^+ \quad \forall s \in \mathcal{S}\tag{28}$$

$$\lambda_{sku}^{(t+1)} = \left[\lambda_{sku}^{(t)} - \delta_3\left(I_{\mathrm{max},sku} - \sum_{i\in\mathcal{S}\backslash\{s\}}\sum_{j\in\mathcal{U}_{\mathrm{RRH},i}}p_{ikj}\omega_{ikj}g_{iku}\right)\right]^+ \quad \forall u \in \mathcal{U}_{\mathrm{RRH},s}, s \in \mathcal{S}, k \in \mathcal{K}\tag{29}$$

where $\beta_{su}^{(t)}$, $\phi_s^{(t)}$ and $\lambda_{sku}^{(t)}$ are the values of β_{su}, ϕ_s and λ_{sku} at the t-th iteration, respectively, and δ_1, δ_2 and δ_3 are the positive step sizes that satisfy the infinite travel conditions [50]. The process of updating the PRB and power allocation, and Lagrange multipliers is repeated until convergence or a predefined maximum number of iterations, T_{max} is reached. Note that μ_{sk}, ν_{sku}, τ_{sku} and ρ_{sku} are not updated because they have been absorbed in the KKT conditions and thus do not affect the solutions in (22) and (26). As such, we can omit the terms in (17) corresponding to these Lagrange multipliers. The asymptotic computational complexity of PRB and power allocation can be shown to be of $O\left(T_{\mathrm{max}}\left((3|\mathcal{K}| + 1)\sum_{s\in\mathcal{S}}|\mathcal{U}_{\mathrm{RRH},s}| + |\mathcal{S}|\right)\right)$.

C. Implementation

The proposed network slicing scheme is implemented in the new network slicing module in the BBU pool (cf. Fig. 1). This module will first collect channel information from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2017.2789294, IEEE Transactions on Wireless Communications

the RRHs, traffic and priority information from each UE, and QoS requirements and SLA information or tenant priority from tenants via their core networks. Then, the module will execute the proposed network slicing scheme accordingly. The upper- and lower-level of the network slicing scheme are executed in coarse and fine time granularity respectively. The upper-level is intended to be executed in the order of LTE frames, that is, tens of milliseconds (as each LTE frame spans 10 ms). As we have designed the admission control, user association and virtual BBU allocation algorithms in such a way that their computational complexity is reasonably low, the execution of these algorithms in the order of LTE frames is reasonable. In fact, the admission control and user association process in LTE systems normally need tens of milliseconds to complete a user handover [51], thus our proposed scheme is compatible with LTE. The upper-level can be executed periodically, or triggered when certain conditions are met (e.g., when new or handover users are requesting admission, or the channel has varied significantly). In the latter case, the upper-level can be triggered to be executed at the start of the next LTE frame or after several LTE frames. The lower-level should be executed every 1 ms as specified by LTE for resource allocation among users and such execution is possible in such a short interval as it can be performed through parallel computation on power and PRB allocation solution updates, i.e., (22) and (26) as well as on the subgradient update, i.e., (27)-(29).

The computational burden of the proposed scheme could grow fast in a dense network if we consider that a single centralized BBU pool manages all the RRHs. Such a scenario is generally impractical because sophisticated fronthaul transport networks would be needed and the BBU processing would be computationally exhaustive [52]. Nevertheless, a multicloud radio access network architecture [52] has been proposed to address this issue, where multiple clouds are introduced with each managing a subset of nearby RRHs. Under this architecture, our proposed network slicing scheme can still be applicable and implemented in each cloud. In fact, the proposed scheme is more suitable under the multicloud architecture as the number of RRHs managed by each cloud is smaller, thus the computational complexity becomes relatively lower.

In practice, a perfectly synchronized shared spectrum network is unrealistic if rapid channel reporting is considered. Rapid channel reporting would introduce large delay in the network slicing process, especially during PRB and power allocation. As such, it is impossible to allocate PRBs every 1 ms, as specified in LTE, while channel conditions are reported as rapid as every 1 ms. Nevertheless, we can reduce the frequency of channel reporting by increasing the reporting period. In fact, this is compliant with the channel quality indicator (CQI) reporting in LTE where the reporting period can be adjusted to be multiple of TTIs. Considering that the channel is slow-fading, the network slicing solution obtained with rapid channel reporting will not diverge much from that with less rapid channel reporting. Hence, the proposed scheme can still be feasible under an imperfectly synchronized shared spectrum network.

V. Performance Evaluation

The performance of the proposed network slicing scheme is evaluated using Monte Carlo simulations. We consider a two-tenant H-CRAN that consists of one M-RRH and five S-RRHs with the first and second tenant being denoted as VNO 1 and VNO 2 respectively. The number of UEs associated with each tenant is set to 50, which is large enough for observing the numbers of admitted and rejected UEs. The coverage radius of the M-RRH is set to 500 m and the S-RRHs are randomly distributed within the network. The number of PRBs is set to 100 according to the LTE specifications. Also, we assume that $C_{\mathrm{cBUP}} = 50$ Mb/s and the M-RRH can support as much as the BBU pool, i.e., $C_{\mathrm{bh}} = 50$ Mb/s. For the S-RRH, $C_{\mathrm{fh}} = 10$ Mb/s is set as the small-cell fronthaul capacity which should be much smaller than the macrocell backhaul capacity. For the transmission power parameters, $P_{\max,0} = 43$ dBm and $P_{\max,s} = 30$ dBm for $s \neq 0$ are set following the LTE specifications. The priority weight of UEs belonging to each tenant is randomly generated. We consider an independently and identically distributed (i.i.d.) Rayleigh fading channel with zero-mean and unit-variance. Log-normal shadowing is also considered, which is i.i.d. with zero-mean and a standard deviation of 10 dB. The following path loss models are adopted: $128.1 + 37.6 \log(d_{\mathrm{m}})$ dB and $140.7 + 36.7 \log(d_{\mathrm{s}})$ where d_{m} is the distance in km between the M-RRH and the UE whereas d_{s} is that between the S-RRH and the UE [53]. The noise power spectral density and noise figure are respectively set to -174 dBm/Hz and 9 dB [53]. For the proposed scheme, δ_1, δ_2 and δ_3 are set according to the nonsummable diminishing rule [50], and $N_{\mathrm{w}} = 10000$ and $T_{\max} = 100$ are set. We adopt the following schemes for performance comparison (including the proposed scheme):

1) Fixed BBU capacity and Fixed Resource allocation (FBFR): The BBU pool capacity is predivided to the tenants according to their priority. The UEs are then admitted according to their priority based on the available BBU capacity allocated to their associated tenant. The admitted UEs are then associated with RRHs based on wideband channel quality and fronthaul and backhaul capacity availability (which is similar to the proposed scheme). After that, the admitted UEs that are associated with the same tenant receive equal BBU capacity. If the fronthaul (or backhaul) capacity of the RRH is lower than the sum of the BBU capacity of the RRH's associated UEs, these UEs will be reallocated with equal BBU capacity based on the fronthaul (or backhaul) capacity. The number of PRBs is also predivided among the tenants based on their priority, in which PRB and power allocation are performed accordingly.
2) Dynamic BBU capacity and Fixed Resource allocation (DBFR): The PRB and power allocation is performed in a similar way as in FBFR except that the BBU capacity allocation is dynamically performed as in Section IV-A.
3) Fixed BBU capacity and Dynamic Resource allocation: The BBU capacity allocation is performed in a similar way as in FBFR except that the PRB and power allocation is dynamically performed as in Section IV-B.

4) Dynamic BBU capacity and Dynamic Resource allocation (DBDR): The proposed scheme presented in Section IV that performs dynamic network slicing.

Next, we consider two traffic models that are provided by the tenants: video streaming and web browsing. The video streaming traffic model requires $R_{\text{min}} = 1$ Mb/s and $R_{\text{max}} = 1.5$ Mb/s whereas the web browsing traffic model requires $R_{\text{min}} = 100$ kb/s and $R_{\text{max}} = 300$ kb/s for each UE [19]. All simulation results are averaged over 100 runs.

A. Impact of Priority of Tenants Providing Similar Services

Here, the effect of varying the tenant priority of a two-tenant H-CRAN where both tenants provide video streaming services for both the proposed and baseline schemes is investigated.

In Fig. 3(a), DBDR achieves the best total throughput performance, which is attributed to its dynamic allocation of baseband and radio resources among the tenants. Next, we can find that dynamic BBU capacity allocation contributes slight improvements of about 1% over fixed allocation, as shown by DBDR and DBFR over FBDR and FBFR, respectively. This is because the proposed dynamic BBU capacity allocation mechanism ensures the BBU capacity allocated to each UE to meet its QoS requirement. Meanwhile, the fixed BBU capacity allocation mechanism does not take into account the QoS requirements of the admitted UE when allocating the BBU capacity to the admitted UEs. In fact, some of the admitted UEs receive insufficient BBU capacities due to insufficient fronthaul (or backhaul) capacity of the associated RRHs. However, the improvement is small because both tenants impose the same data rate requirements. Thus, the total numbers of admitted UEs as well as the BBU capacity allocated to the admitted UEs for both fixed and dynamic BBU capacity allocation are nearly identical, resulting in nearly identical throughput performance. On the other hand, dynamic PRB and power allocation results in substantial throughput improvements of about 35% over fixed PRB and power allocation, as exhibited by DBDR and FBDR over DBFR and FBFR, respectively. This is because DBFR and FBFR predivide the number of PRBs among the tenants, thus limiting the achievable throughput.

In Fig. 3(b), both DBDR and FBDR provide the best and comparable network slicing performance as the ratio of the throughput achieved by VNO 1 to that of VNO 2 is almost equivalent to the priority ratio. This thanks to the dynamic PRB and power allocation mechanism which allows full exploitation of multiuser diversity of the channel to achieve fair throughput distribution among the tenants. The throughput achieved by FBFR and DBFR for VNO 1 and VNO 2 is different when the priority ratio is 1:1 because both schemes split the number of PRBs between the two tenants in proportion to the tenants' priority and allocate PRBs to the UEs based on only the user priority. Such allocation cannot guarantee even throughput distribution between the two tenants. Besides that, although both tenants have equal numbers of PRBs when the priority ratio is 1:1, the number of PRBs received by each of the two tenants is only half of the channel bandwidth and the PRBs may not always have equal quality since the channel gain of each PRB between individual RRHs and UEs is different.

To further confirm the observations in Fig. 3(b), we calculate the normalized throughput of each UE, r_u with respect to its target minimum data rate, i.e., $r_u = \frac{R_u}{R_{\text{min},u}}$. Then, we evaluate Jain's fairness index [54] for the interslice case as $\Xi_{\text{inter}} = \frac{\left(\sum_{n \in \mathcal{N}} R_n/v_n\right)^2}{|\mathcal{N}| \sum_{n \in \mathcal{N}} (R_n/v_n)^2}$, where $R_n = \sum_{u \in \mathcal{U}_n} r_u$. Fig. 3(c) shows the fairness performance of the two-tenant H-CRAN with both VNO 1 and VNO 2 providing video streaming services. In line with Fig. 3(b), the fairness performance of DBDR and FBDR are comparable and better than the other two schemes. FBDR is superior to the proposed DBDR because the former divides the BBU pool capacity proportionally to the priority of the tenants, where both provide the same video services, and allocates virtual BBU capacity evenly to the admitted UEs. Moreover, the DBDR scheme allocates virtual BBU capacity according to the overall priority of each admitted UE, which is slightly inefficient than FBDR in this scenario. Though, the performance gap between them is negligible as both of them have achieved a high fairness index exceeding 0.9.

Fig. 3(d) illustrates the intraslice fairness performance of the H-CRAN, where Jain's fairness index is evaluated as $\Xi_{\text{intra}} = \frac{\left(\sum_{u \in \mathcal{U}_{\text{a}}} r_u\right)^2}{|\mathcal{U}_{\text{a}}| \sum_{u \in \mathcal{U}_{\text{a}}} (r_u)^2}$. In this case, both DBDR and FBDR perform as well as those in the interslice case. However, DBFR and FBFR perform poorly because fixed PRB allocation limits the number of PRBs allocated to each tenant, thus limiting the achievable throughput fairness performance.

Figs. 3(e) and 3(f) respectively plot the interslice and intraslice cumulative distribution function (CDF) of the normalized UE throughput, i.e., r_u for priority ratio 1:2. From Fig. 3(e), it is seen that compared to DBFR and FBFR, DBDR and FBDR have more UEs achieving nonzero throughput and their minimum data rates. Fig. 3(f) also shows that DBDR and FBDR achieve fairer throughput performance compared to DBFR and FBFR for both tenants. This is because the latter schemes predivide the number of PRBs among the tenants, which limit the achievable throughput and fairness performance. As the performance trends are similar for other priority ratios, we omit the corresponding CDF results.

Figs. 3(g) and 3(h) illustrate respectively the number of admitted UEs (in fractional form) with minimum QoS satisfaction (i.e., admitted UEs with their throughput larger or equal to their minimum data rate) and the number of UEs being denied admission. In Fig. 3(g), both DBDR and FBDR achieve higher numbers of UEs with minimum QoS satisfaction than DBFR and FBFR. This is because the latter schemes limit the number of PRBs available to each tenant according to the priority ratio, thus leading to inability to guarantee minimum QoS satisfaction for the UEs. In Fig. 3(h), all the schemes reject the same number of UEs across all priority ratios, except FBFR and FBDR. At the priority ratio of 1:1, the number of UEs rejected by the latter two schemes is larger than those by the other two schemes. This is due to the inefficient fixed BBU pool capacity allocation that predivides the BBU capacity to the tenants based on their priority. Although the sum of the remaining cloud BBU pool capacity of all tenants is sufficient to admit more UEs but this is not allowed by fixed cloud BBU

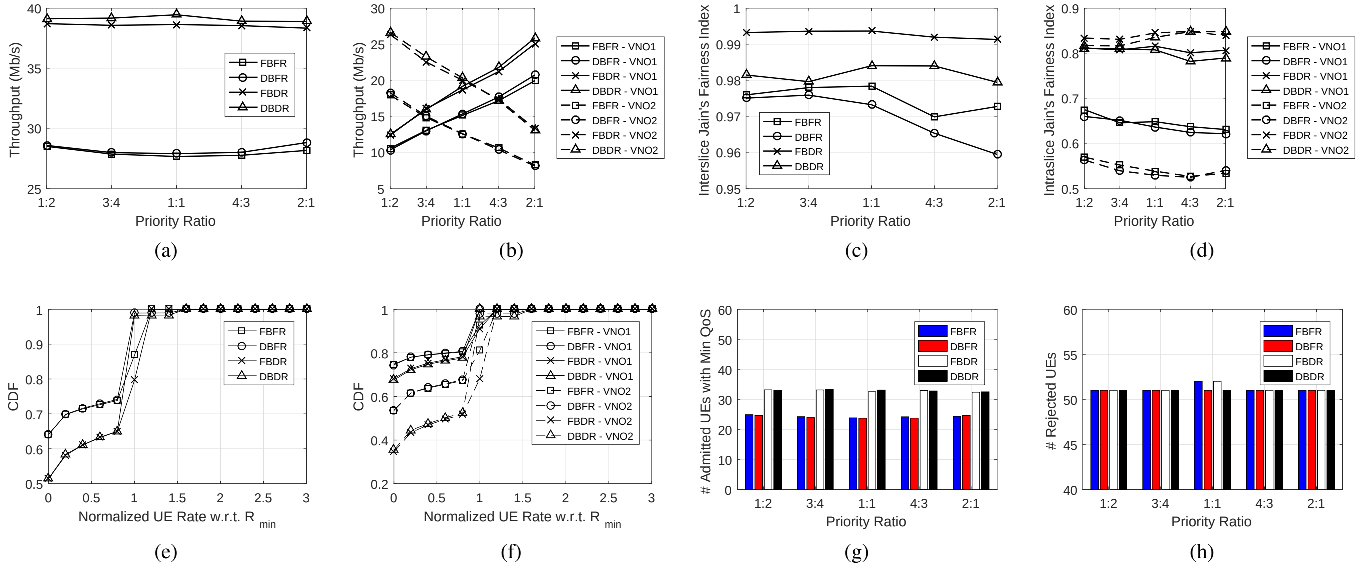


Fig. 3. (a) Total throughput, (b) throughput of VNO 1 and VNO 2, (c) interslice fairness, (d) intraslice fairness, (e) CDF of the normalized UE throughput with priority ratio 1:2, (f) CDF of the normalized UE throughput of VNO 1 and VNO 2 with priority ratio 1:2, (g) average number of admitted UEs with minimum QoS satisfaction, and (h) average number of rejected UEs of the H-CRAN where both VNO 1 and VNO 2 provide video streaming services.

pool capacity allocation.

B. Impact of Priority of Tenants Providing Different Services

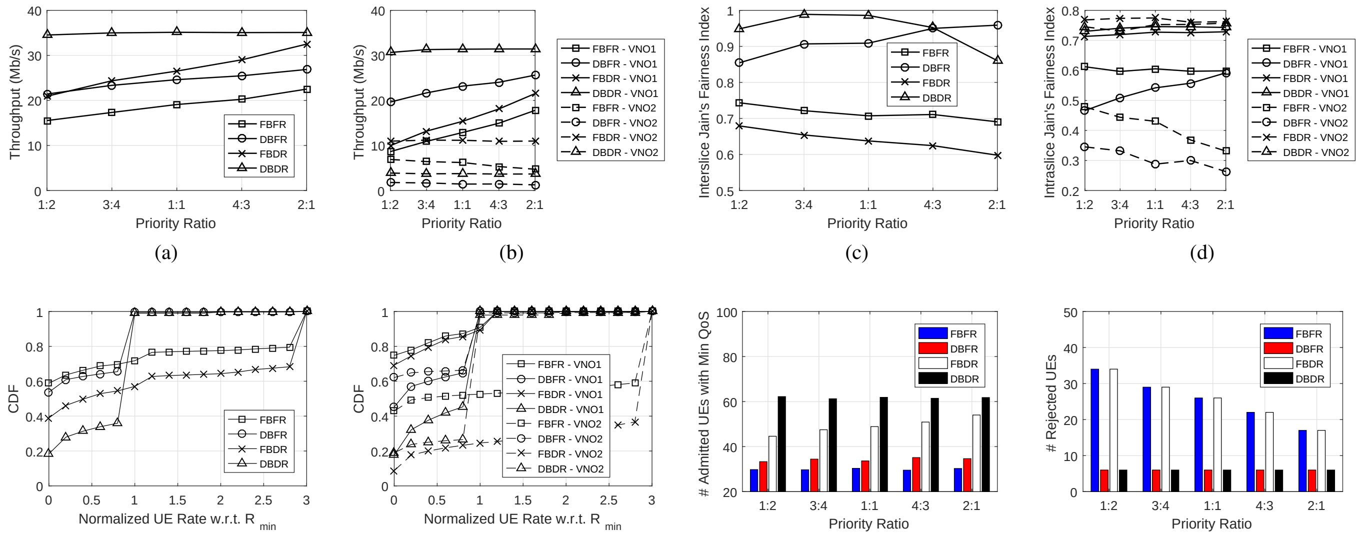
Here, we investigate the effect of varying the tenant priority of the H-CRAN where VNO 1 and VNO 2 provide video streaming and web browsing services, respectively, for all the schemes. In fact, such prioritization of these VNOs is equivalent to prioritization of services.

In Fig. 4(a), DBDR again outperforms other schemes due to its dynamic baseband and radio resource allocation. Also, Fig. 4(b) shows that the throughput achieved by VNO 1 is larger than that by VNO 2 because video streaming has a higher data rate requirements than web browsing. Notably, the difference between dynamic and fixed BBU capacity allocation is huge because the tenants imposes different data rate requirements. As fixed BBU capacity allocation splits the BBU pool capacity between tenants in proportion to their priority, the capacity allocated to VNO 2 is over-sufficient due to its much lower data rate requirements. As a result, all the UEs associated with VNO 2 are admitted and the BBU capacity allocated to these UEs is over-sufficient. Meanwhile, the number of admitted UEs associated with VNO 1 that imposes higher data rate requirements is very limited due to the inefficient fixed BBU capacity allocation. Hence, the number of admitted UEs is much lower and thus the achievable network throughput is much lower. On the other hand, the proposed dynamic BBU capacity allocation mechanism does not split the BBU pool capacity as in the fixed BBU capacity allocation mechanism but fully maximizes the number of admitted UEs according to their priority as well as their associated tenant's priority based on the availability of the whole BBU pool capacity. Therefore, it allows more UEs to be admitted to the network and hence achieving higher network throughput.

It is difficult to justify the interslice fairness performance of the schemes in Fig. 4(b) as the traffic services provided by the tenants are with different data rate requirements. Nonetheless, we can observe in Fig. 4(c) that DBDR always achieves a fairness index of above 0.8 and outperforms the other schemes across most of the priority ratios. Interestingly, FBDR and FBFR perform poorly because fixed BBU capacity allocation does not take into account heterogeneous service requirements, which results in unfair admission control and throughput performance. In Fig. 4(d), DBDR and FBDR perform equivalently as in the scenario in Section V-A due to the dynamic PRB and power allocation mechanism.

Figs. 4(e) and 4(f) respectively plot the interslice and intraslice CDFs of the normalized UE throughput, i.e., r_u for priority ratio 1:2. In Fig. 4(e), compared to other schemes, DBDR results in more UEs achieving nonzero throughput and minimum data rates. In Fig. 4(f), DBDR is overall the best performer among the schemes. Although FBDR is superior to DBDR for VNO 2 but the former is significantly inferior to DBDR as well as DBFR for VNO 1. Again, we omit the CDF results for other priority ratios due to similar trends.

Figs. 4(g) and 4(h) respectively show the number of admitted UEs (in fractional form) with minimum QoS satisfaction and the number of UEs being denied admission. It is observed in Fig. 4(g) that DBDR outperforms other schemes. This indicates that the proposed DBDR scheme is capable of providing satisfactory QoS experience for the network regardless of service types provided by the tenants, thanks to the dynamic network slicing mechanism of the proposed DBDR scheme. In Fig. 4(h), FBFR and FBDR have rejected more UEs compared to the other two schemes. Here, we can confirm that fixed BBU pool capacity allocation based on tenant priority is inefficient, especially in scenarios where the tenants are providing traffic services with different target minimum and maximum data



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. (a) Total throughput, (b) throughput of VNO 1 and VNO 2, (c) interslice fairness, (d) intraslice fairness, (e) CDF of the normalized UE throughput with priority ratio 1:2, (f) CDF of the normalized UE throughput of VNO 1 and VNO 2 with priority ratio 1:2, (g) average number of admitted UEs with minimum QoS satisfaction, and (h) average number of rejected UEs of the H-CRAN where both VNO 1 and VNO 2 provide video streaming and web browsing services, respectively.

rates. When the priority of the tenant that provides services with high target minimum and maximum data rates such as video streaming is low, the number of UEs rejected would be large if the BBU pool capacity has been allocated in advance. Only when the priority is increasing, the number of UEs admitted for the corresponding tenant can be increased, as demonstrated by the decreasing trends of the FBFR scheme and the FBDR scheme in Fig. 4(h) with increasing priority of VNO 1.

C. Impact of Fronthaul/Backhaul Capacities and Macrocell/Small-Cell Coexistence

Here, we examine the impact of C_{fh} and C_{bh} as well as the coexistence of macrocell and small-cell to the proposed network slicing scheme with the same simulation setting as in Section V-A.

Fig. 5(a) shows the throughput performance of the proposed scheme with varying C_{fh} with $C_{\text{bh}} = 50$ Mb/s. When $C_{\text{fh}} = 0$ Mb/s, all admitted UEs can only associate with the M-RRH. As such, this scenario is equivalent to that with only the macrocell, thus resulting in the most inferior throughput performance. The proposed scheme performs the best when $C_{\text{fh}} = 5$ Mb/s but the throughput gradually deteriorates with increasing C_{fh}. This is because the larger C_{fh} allows more UEs to be associated with the S-RRHs. As a result, the S-RRHs cannot satisfy the QoS requirements of the associated UEs since they have low transmission power budgets. Other performance results of the proposed scheme have similar trends as in Fig. 5(a) except that the interslice fairness and number of rejected UEs remain consistent across all C_{fh}.

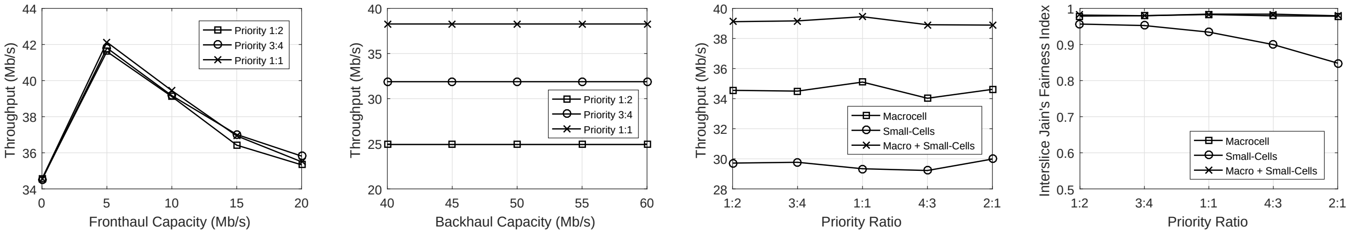
Fig. 5(b) shows that the throughput performance with varying C_{bh} with $C_{\text{fh}} = 10$ Mb/s. The throughput performance of the proposed scheme remains consistent because most of the admitted UEs associate with the S-RRHs, as the altter can provide better channel quality. Other performance results of the proposed scheme also remain consistent.

Fig. 5(c) shows that the H-CRAN with both macrocell and small-cells outperforms the macrocell-only and small-cells-only networks. This is because the former scenario has more power resources due to the coexistence of both M-RRH and S-RRHs. In addition, the UEs located near the S-RRHs can experience better channel conditions and the M-RRH can serve those UEs who are located far from the S-RRHs. The scenario with only small-cells is inferior to that with only macrocell because some UEs are far from the S-RRHs and the low-power S-RRHs need to spend more power to reach these UEs, thereby resulting in lower throughput.

The interslice fairness performance of the scenario with both macrocell and small-cells is better than that with only small-cells but is equivalent to that with only the macrocell, as shown in Fig. 5(d). This is because the M-RRH has enough power to serve all the admitted UEs and guarantee the throughput fairness among their tenants. Overall, the coexistence of both macrocells and small-cells can achieve better network slicing performance than that of only either of the two.

VI. Conclusion

This paper has proposed a new dynamic network slicing scheme for multitenant H-CRANs. The proposed scheme consists of an upper-level that manages admission control, UE association and BBU capacity allocation; and a lower-level that manages PRB and power allocation among admitted UEs. Simulation results show that the proposed scheme achieves higher throughput, fairness and QoS performance compared to the baseline schemes, especially in scenarios with tenants of heterogeneous services. Also, the simulation results infer



(a) (b) (c) (d)

Fig. 5. Total throughput performance of the H-CRAN with varying (a) fronthaul capacity and (b) backhaul capacity. (c) Total throughput and (d) interslice fairness performances of the two-tenant H-CRAN network with only a macrocell, with only small-cells, and with a macrocell and small-cells.

that compared to fixed BBU capacity allocation, the proposed dynamic BBU capacity allocation mechanism (which jointly consists of dynamic admission control, UE association and BBU capacity allocation among UEs) can more efficiently maximize the number of admitted UEs and assign BBU capacity to the latter, and achieves higher network throughput; especially in scenarios with tenants of heterogeneous services. On the other hand, the dynamic PRB and power allocation mechanism of the proposed scheme can achieve higher throughput and QoS performance compared to the baseline schemes due to efficient multiuser diversity exploitation across all PRBs. With dynamic allocation of all network resources, the proposed scheme is able to maintain a high degree of fairness in scenarios with tenants providing homogeneous and heterogeneous services.

Appendix A

Consider a gradient-based resource allocation framework where the resource allocation decision is the solution to $\sum_{u\in\mathcal{U}} F'(\bar{R}_u)R_u$, where R_u and $\bar{R}_u$ are the instantaneous and average rates respectively [40]–[42]. Clearly, (5) is a gradient-based resource allocation problem with $W_u = F'(\bar{R}_u)$. To achieve various notions of fairness, the weighted α-fairness function can be adopted as the utility [42], [43]:

$$F_\alpha(\bar{R}_u) = \begin{cases} \varpi_u(1-\alpha)^{-1}\bar{R}_u^{1-\alpha} & \alpha \geq 0, \alpha \neq 1 \\ \varpi_u \log(\bar{R}_u) & \alpha = 1 \end{cases} \tag{30}$$

where ϖ is the weighting coefficient corresponding to UE u and $\alpha \geq 0$ is a parameter that determines the notion of fairness. With (30), we yield $W_u = F'_\alpha(\bar{R}_u) = \varpi_u/\bar{R}_u^\alpha$ where $w_u = 1/\bar{R}_u^\alpha$ and $\varpi_u = v_n$ if $u \in \mathcal{U}_n$, and thus the solution to (5) is weighted α-fair. In particular, the solution corresponds to maximum throughput, proportional fairness and max-min fairness when $\alpha = 0$, $\alpha = 1$ and $\alpha \to \infty$, respectively.

Appendix B

Suppose that feasible $\{a_u\}$, $\{b_{su}\}$, $\{c_{\mathrm{vB},u}\}$ and $\{\omega_{sku}\}$ are given, (5) can be reduced to

$$\max_{p_{sku}} \sum_{s\in\mathcal{S}} \sum_{k\in\mathcal{K}} W_u a_u b_{su} \omega_{sku} r_{sku}, \tag{31}$$

subject to (5a)-(5f) and nonnegative $\{p_{sku}\}$. Clearly, (31) and (5a)-(5f) are smooth, and r_{sku} is a nonlinear logarithmic function and nonconcave with p_{sku}. Thus, (31) is a smooth, nonlinear and nonconvex programming problem with continuous variables p_{sku}, which has been proved to be generally NP-hard [55]. Moreover, (31) is a weighted sum rate utility maximization problem, which has been proved in [56] and [57] to be NP-hard, further validating the NP-hardness of (31). Then, the problem in (5) is also NP-hard as its reduced problem in (31) is already NP-hard.

Appendix C

Let us define a 0-1 knapsack problem as follows:

Definition 1: Suppose there is a set of items, $\mathcal{M}$ to be filled in a knapsack with a weight capacity of X. Each item $i \in \mathcal{M}$ has a weight of x_i and corresponds to a nonnegative profit value of y_i. A 0-1 knapsack problem is to fill the knapsack with the items in such a way that the total profit is maximized without exceeding the weight capacity of the knapsack.

By Definition 1, the following can be obtained with respect to (6): $\mathcal{U} = \mathcal{M}$, $i = u$, $X = C_{\mathrm{cBUP}}$, $x_i = R_{\min,u}$ and $y_i = W_u$.

Appendix D

There are only two options for computing $V(u,i)$ for UE u:

1) Reject UE u: In this case, the maximum objective function value for $\{1, 2, \ldots, u-1\}$ with the capacity of id is $V(u-1, i)$.
2) Admit UE u: In this case, which is only possible with $R_{\min,u} \leq id$, we gain W_u for the objective function value and add $R_{\min,u}$ to the total minimum data rate. The maximum objective function value for the set of remaining UEs $\{1, 2, \ldots, u-1\}$ with the capacity of $(id - R_{\min,u})$ is $V\left(u-1, i-\left\lceil\frac{R_{\min,u}}{d}\right\rceil\right)$. In total, $W_u + V\left(u-1, i-\left\lceil\frac{R_{\min,u}}{d}\right\rceil\right)$ is yielded.

Since (6) is a maximization problem, the larger objective function value of the above two cases is the correct one. It is noteworthy that, for $R_{\min,u} > id$, $V(u,i)$ will be computed as $V(u-1,i)$ because the capacity will be exceeded if UE u is admitted. Hence, UE u can only be rejected.

Appendix E

Let us define a 0-1 multiple knapsack problem as follows:

Definition 2: Suppose that $\mathcal{M}$ and $\mathcal{L}$ are the sets of items and knapsacks, respectively. Each item $i \in \mathcal{M}$ has a weight of x_i and gives a nonnegative profit value of y_i whereas each knapsack $j \in \mathcal{L}$ has a weight capacity of X_j. The problem is to fill all the knapsacks with the items such that the total profit of each knapsack is maximized without exceeding its weight capacity.

By mapping Definition 2 with (9), we yield $\mathcal{U}_{\text{a}} = M$, $\mathcal{L} = \mathcal{S}$, $i = u$, $j = s$, $x_i = R_{\text{min},u}$, $X_j = C_{\text{xh},s}$ and $y_i = W_u \Gamma_{\text{wb},su}$.

Acknowledgment

The authors would like to thank Prof. Geoffrey Ye Li and Dr. Guocong Song for their discussions on the NUM framework. The authors would also like to thank the editors and reviewers for their helpful comments and suggestions.

References

[1] Y. L. Lee, J. Loo, and T. C. Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *Proc. ICAEESE*, Putrajaya, Malaysia, Nov. 2016, pp. 414–420.

[2] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

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

[4] C. Liang and F. R. Yu, "Wireless virtualization for next generation mobile cellular networks," *IEEE Wireless Commun. Mag.*, vol. 22, no. 1, pp. 61–69, Feb. 2015.

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

[6] "Uk strategy and plan for 5g & digitisation - driving economic growth and productivity," Future Communications Challenge Group (FCCG), UK, pp. 1–52, Jan. 2017. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/582640/FCCG_Interim_Report.pdf

[7] P. C. Garces, X. Costa-Perez, K. Samdanis, and A. Banchs, "RMSC: A cell slicing controller for virtualized multi-tenant mobile networks," in *Proc. 81st IEEE VTC Spring*, Glasgow, UK, May 2015, pp. 1–6.

[8] I. da Silva, G. Mildh, A. Kaloxylos, P. Spapis, E. Buracchini, A. Trogolo, G. Zimmermann, and N. Bayer, "Impact of network slicing on 5G Radio Access Networks," in *Proc. EuCNC*, Athens, Greece, Jun. 2016, pp. 153–157.

[9] W. Gerhardt, C. Cordero, C. Reberger, and T. Dolan, "Mobile network as a service - a new solution architecture for mobile network operators," CISCO, pp. 1–18, Mar. 2013. [Online]. Available: http://www.cisco.com/c/dam/en_us/about/ac79/docs/sp/NaaS_White-Paper.pdf

[10] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: Enabling enterprises' own software-defined cellular networks," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 146–153, Jul. 2016.

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

[12] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-Giel, "LTE mobile network virtualization," *Mobile Netw. Appl.*, vol. 4, no. 4, pp. 424–432, Aug. 2011.

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

[14] T. Guo and R. Arnott, "Active LTE RAN sharing with partial resource reservation," in *Proc. 78th IEEE VTC Fall*, Las Vegas, Nevada, Sep. 2013, pp. 1–5.

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

[16] M. Kalil, A. Shami, and Y. Ye, "Wireless resources virtualization in LTE systems," in *Proc. IEEE INFOCOM WKSHPS*, Toronto, Ontario, May 2014, pp. 363–368.

[17] M. I. Kamel, L. B. Le, and A. Girard, "LTE wireless network virtualization: Dynamic slicing via flexible scheduling," in *Proc. 80th IEEE VTC Fall*, Vancouver, British Columbia, Sep. 2014, pp. 1–5.

[18] A. Moubayed, A. Shami, and H. Lutfiyya, "Wireless resource virtualization with device-to-device communication underlaying LTE network," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 734–740, Dec. 2015.

[19] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management & prioritization in 5g mobile systems," in *Proc. 22nd EW Conf.*, Oulu, Finland, May 2016, pp. 1–6.

[20] M. Jiang, D. Xenakis, S. Costanzo, N. Oassa, and T. Mahmoodi, "Radio resource sharing as a service in 5G: A software-defined networking approach," *Comput. Commun.*, vol. 107, pp. 13–29, Jul. 2017.

[21] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Perez, "Network slicing games: Enabling customization in multi-tenant mobile networks," in *Proc. IEEE INFOCOM*, Atlanta, GA, May 2017, pp. 1–19, to appear.

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

[23] S. Khatibi and L. M. Carreia, "A model for virtual radio resource management in virtual RANs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 68, pp. 1–12, Mar. 2015.

[24] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing in 5g: An auction-based model," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–6, to appear.

[25] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent Advances in Radio Resource Management for Heterogeneous LTE/LTE-A Networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2142–2180, Jun. 2014.

[26] Y. L. Lee, T. C. Chuah, J. Loo, and A. A. El-Saleh, "Fair Resource Allocation with Interference Mitigation and Resource Reuse for LTE/LTE-A Femtocell Networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8203–8217, Oct. 2016.

[27] ——, "Multi-Objective Resource Allocation for LTE/LTE-A Femtocell/HeNB Networks using Ant Colony Optimization," *Wireless Pers. Commun.*, vol. 92, no. 2, pp. 565–586, Jan. 2017.

[28] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable RAN virtualization in multi-tenant LTE-A heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6651–6654, Aug. 2016.

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

[30] "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation," 3GPP, Jun. 2017. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/36_series/36.211/36211-e30.zip

[31] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-Layer Resource Allocation With Elastic Service Scaling in Cloud Radio Access Network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5068–5081, Sep. 2015.

[32] K. Guo, M. Sheng, J. Tang, T. Q. S. Quek, and Z. Qiu, "Exploiting hybrid clustering and computation provisioning for green C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 12, pp. 4063–4076, Dec. 2016.

[33] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, "System cost minimization in cloud RAN with limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371–3384, May 2017.

[34] "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 14)," 3GPP, Mar. 2017. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/23_series/23.401/23401-f00.zip

[35] "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 14)," 3GPP, Mar. 2017. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/36_series/36.300/36300-e20.zip

[36] C. Koutsimanis and G. Fodor, "A dynamic resource allocation scheme for guaranteed bit rate services in ofdma networks," in *Proc. IEEE ICC*, Beijing, China, May 2008, pp. 2524–2530.

[37] K. Murphy, "Centralized RAN and fronthaul," Ericsson, pp. 1–9, May 2015. [Online]. Available: http://www.isemag.com/wp-content/uploads/2016/01/C-RAN_and_Fronthaul_White_Paper.pdf

[38] A. de la Oliva, J. A. Hernandez, D. Larabeiti, and A. Azcorra, "An overview of the CPRI specifications and its application to C-RAN-based LTE scenarios," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 152–159, Feb. 2016.

[39] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G backhaul challenges and emerging research directions: A survey," *IEEE Access*, vol. 4, pp. 1743–1766, Apr. 2016.

[40] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-Part II: Algorithm Development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[41] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," in *Proc. 2002 Allerton Conf. Communication, Control and Computing*, 2002, pp. 1–10.

[42] J. Huang, V. G. Subramaniam, R. Agrawal, and R. A. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 288–296, Jan. 2009.

[43] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[44] D. S. Johnson and M. Garey, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[45] S. S. Rao, *Engineering Optimization: Theory and Practice*. Jon Wiley & Sons, 2009.

[46] J. Kleinberg and E. Tardos, *Algorithm Design*. Pearson Education, 2006.

[47] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[48] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[49] K. Kim, Y. Han, and S.-L. Kim, "Joint subcarrier and power allocation in uplink OFDMA systems," *IEEE Commun. Lett.*, vol. 9, no. 6, pp. 526–528, Jun. 2005.

[50] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," Stanford University, 2006-07, notes for EE364b.

[51] "Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource management (Release 14)," 3GPP, Mar. 2017. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/36_series/36.133/36133-e30.zip

[52] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M.-S. Alouini, "Resource allocation in heterogeneous cloud radio access networks: Advances and challenges," *IEEE Wireless Commun. Mag.*, vol. 22, no. 3, pp. 66–73, Jun. 2015.

[53] "Small cell enhancements for E-UTRA and E-UTRAN-physical layer aspects (release 12)," 3GPP, Sep. 2013. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/36_series/36.872/36872-c10.zip

[54] R. Jain, *The Art of Computer Systems: Performance Analysis*. John Wiley & Sons, 1991.

[55] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Math. Program.*, vol. 39, no. 2, pp. 117–129, Jun. 1987.

[56] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[57] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 583–596, Feb. 2014.

Ying Loong Lee (S'13 – M'17) received the B.Eng. (Hons) degree in electronics majoring in telecommunications and the Ph.D. degree in wireless communications from Multimedia University, Cyberjaya, Selangor, Malaysia, in 2012 and 2017, respectively.

He was a short-term student intern in the Department of Electrical and Computer Engineering, National Chiao Tung University in 2015. He is currently working as a Research Officer at Multimedia University. His current research interests fall in the areas of 5G wireless communications which include network slicing, resource management and load balancing for heterogeneous cellular networks and cloud radio access networks.

Dr. Lee received the Yayasan Universiti Multimedia Postgraduate Scholarship in 2012, the Fundamental Research Grant Scheme Scholarship in 2015, the International Teletraffic Congress Student Travel Grant Award in 2016, and the Multimedia University Best Thesis Award for Ph.D. in engineering in 2017.

Jonathan Loo received his M.Sc. degree in Electronics (with Distinction) and the Ph.D. degree in Electronics and Communications from the University of Hertfordshire, Hertfordshire, U.K., in 1998 and 2003, respectively. Between 2003 and 2010, he was a Lecturer in Multimedia Communications with the School of Engineering and Design, Brunel University, Uxbridge, U.K. Between June 2010 and May 2017, he was an Associate Professor in Communication Networks at the School of Science and Technology, Middlesex University, London, U.K. From June 2017, he is a Chair Professor in Computing and Communication Engineering at the School of Computing and Engineering, University of West London, United Kingdom. His research interests include information centric networking, cloud computing, wireless/mobile networks and protocols, network security, wireless communications, IoT/cyber-physical systems, and embedded systems. He has successfully graduated 18 Ph.D. students as their principal supervisor, and has co-authored more than 240 journal and conference papers in the aforementioned specialized areas. Dr. Loo has been an Associate Editor for Wiley *International Journal of Communication Systems* since 2011. He was the Lead Editor of the book entitled "Mobile Ad Hoc Networks: Current Status and Future Trends" (CRC Press, November 2011).

Teong Chee Chuah received the B.Eng. degree (first-class honors) in electrical and electronic engineering and the Ph.D. degree in digital communications from Newcastle University, UK, in 1999 and 2002, respectively. Since 2003, he has been with the Faculty of Engineering, Multimedia University, Malaysia. His current research interests include signal processing and optimization algorithms for wireless and wired xDSL broadband access networks.

Li-Chun Wang (M'96 – SM'06 – F'11) received Ph.D. degree from the Georgia Institute of Technology, Atlanta, in 1996. From 1996 to 2000, he was with AT&T Laboratories, where he was a Senior Technical Staff Member in the Wireless Communications Research Department. Since August 2000, he has joined the Department of Electrical and Computer Engineering of National Chiao Tung University in Taiwan and is jointly appointed by Department of Computer Science and Information Engineering of the same university.

Dr. Wang was elected to the IEEE Fellow in 2011 for his contributions to cellular architectures and radio resource management in wireless networks. He won the Distinguished Research Award of National Science Council, Taiwan (2012). He was the co-recipients of IEEE Communications Society Asia-Pacific Board Best Award (2015), Y. Z. Hsu Scientific Paper Award (2013), and IEEE Jack Neubauer Best Paper Award (1997).

His current research interests are in the areas of software-defined mobile networks, heterogeneous networks, and data-driven intelligent wireless communications. He holds 19 US patents, and have published over 200 journal and conference papers, and co-edited a book, "Key Technologies for 5G Wireless Systems," (Cambridge University Press 2017).