A Dynamic Resource Sharing Mechanism for Cloud Radio Access Networks

Binglai Niu, Yong Zhou, *Member, IEEE*, Hamed Shah-Mansouri, *Member, IEEE*, and Vincent W.S. Wong, *Fellow, IEEE*

***Abstract*—Cloud radio access network (C-RAN) as a promising and cost-efficient cellular architecture has been proposed to meet the increasing demand of wireless data traffic. The main concept of C-RAN is to decouple the baseband unit (BBU) and the remote radio head (RRH), and place the BBUs in a data center for centralized control and processing. In this paper, we study the resource sharing problem in a fronthaul constrained C-RAN, where multiple service providers lease radio resources from a network operator to serve their subscribers. To provide isolation among different service providers, we introduce a threshold-based policy to control the interference among RRHs, and define a new metric to provide minimum resource guarantee for service providers. By leveraging a mobility prediction method, the user locations are predicted for traffic demand estimation and interference control. We propose a multi-timescale resource sharing mechanism, which consists of a global resource allocation process and multiple local resource allocation processes that are performed at different time scales. Simulation results show that the proposed mechanism achieves efficient resource sharing and isolation among service providers.**

***Index Terms*—C-RAN, resource sharing, virtualization, interference control, mobility prediction.**

I. Introduction

Cloud radio access network (C-RAN) has recently been proposed as a cost-efficient solution to meet the increasing mobile data traffic demand [1], [2]. In general, a C-RAN consists of a baseband unit (BBU) pool placed in a cloud-based data center, and a large number of low-cost remote radio heads (RRHs) each deployed in a small cell [3]. The BBUs and RRHs are connected through high-speed optical fronthaul links. By leveraging the cloud computing technique, the BBU pool performs centralized signal processing and provides coordinated radio resource and interference management. The advantages of C-RANs include reducing the capital expenses (CAPEX) and operational expenses (OPEX) for system upgrade and maintenance [4], and improving the spectral efficiency via centralized interference control and coordinated multi-point transmission (CoMP) [5], [6].

Despite these advantages, a practical fronthaul is always capacity and delay constrained, which can significantly reduce the spectrum efficiency gain achieved by C-RAN [7]. The authors in [8] propose a joint precoding and fronthaul compression strategy for downlink transmission, in which the BBUs jointly compress precoded signals for different users to restrict the impact of quantization noise on the aggregate transmission rate. The fronthaul capacity and delay constraints can also be alleviated by flexibly splitting the baseband processing between the BBUs and RRHs (e.g., moving parts of baseband processing functionalities from the BBUs to the RRHs) [9]. Another efficient method is to design advanced resource allocation and optimization mechanisms while taking into account the fronthaul capacity constraint [10], [11]. To maximize the energy efficiency under the fronthaul capacity and queue stability constraints, a dynamic resource optimization problem is formulated in [10]. Based on both the channel and queue states, the authors in [11] devise a fronthaul allocation policy for uplink transmission to minimize the average delay by formulating a stochastic optimization problem.

Supporting radio access network sharing among multiple service providers is an important use case of C-RANs [12], [13]. Multiple service providers lease radio resources from a network operator to serve their subscribed users, which reduces the CAPEX and OPEX of deploying the network infrastructure for each service provider. To enable network sharing, a network operator needs to dynamically allocate radio resources among service providers, which can be achieved by creating multiple virtual radio access networks (vRANs) overlaying the physical infrastructure and ensuring service isolation across the vRANs [14]. Specifically, each vRAN shares a certain amount of radio resources and the operation in one vRAN should not affect other vRANs. To offer flexible customization capability and to efficiently utilize radio resources, several principles for designing virtualization mechanisms have been proposed [15].

Resource sharing in cellular networks has recently been studied. Kokku *et al*. in [16] propose a network virtualization substrate, where a slice scheduler is integrated into the base station's scheduling component for managing the resource slicing and sharing among different service providers. This approach provides flow level isolation and customization by partitioning available channels into non-overlapping slices. Since this approach focuses on resource virtualization in a single base station, it cannot be directly extended to a C-RAN with densely deployed RRHs. To mitigate the inter-tier interference between the base station and the RRHs, the authors in [17] propose a contract-based interference coordination framework, which exploits the time domain by introducing an interference-free transmission interval for the

Manuscript received on May 22, 2015; revised on Jan. 26, 2016 and Jul. 22 2016; accepted on Sept. 16, 2016. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The review of this paper was coordinated by Prof. Sunghyun Choi.

B. Niu is with Arista Networks, Burnaby, BC, V5J 5J8, Canada (e-mail: bniu@ece.ubc.ca). Y. Zhou, H. Shah-Mansouri, and V.W.S. Wong are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, V6T 1Z4, Canada (e-mail:{zhou, hshahmansour, vincentw}@ece.ubc.ca).

RRHs. To reduce the coordination overhead, the size of a C-RAN cluster should be limited, which introduces the intra-tier interference among the RRHs [18]. Such interference should be considered when designing resource sharing mechanisms. Another category of network sharing mechanisms is developed based on the dynamic allocation of spectrum resources [19]–[21]. In these approaches, the physical infrastructure is shared by all service providers, and the service isolation is achieved by allocating non-overlapping spectrum to each service provider. Fu *et al*. in [19] map the wireless spectrum resources into a rate region, and propose a sequential auction game framework to allocate the resources to several competing service providers. An opportunistic resource sharing scheme is proposed in [20], which explores the varying traffic patterns and allows different traffic flows to access the same channel opportunistically. In [21], Hou *et al*. formulate the spectrum sharing problem for multi-hop software defined radio networks as a mixed integer non-linear programming (MINLP) problem, and design an algorithm to find the near optimal solution. Although these approaches guarantee service isolation, they exclude the possibility of spectrum reuse among service providers. Resource sharing in cloud-based networks has also been studied [22], [23]. A number of sharing policies, such as network proportionality, have been proposed and evaluated in [22]. An efficient bandwidth reservation scheme with varying traffic demands is studied in [23] for resource sharing in data centers. However, these works do not consider the sharing of radio resources, which cannot be directly applied to C-RANs.

Different from existing works, in this paper, we design an efficient resource sharing mechanism to support multiple service providers in a C-RAN with capacity constrained fronthaul links. Designing such a mechanism is challenging due to the following reasons. First, to improve the spectral efficiency of a C-RAN with densely deployed RRHs, interference coordination can limit the co-channel interference by imposing restrictions to the resource sharing mechanism. Hence, interference coordination plays an important role in determining the performance of resource sharing. On the other hand, scheduling decisions for users subscribed to different service providers are coupled since concurrent transmissions can cause co-channel interference. Although interference coordination is an effective method to limit the co-channel interference, it can affect the scheduling decisions of one service provider, which in turn affect the scheduling decisions of other service providers. Such a coupling of scheduling decisions is not desirable for providing service isolation among different service providers. As a result, an efficient resource sharing mechanism should take into account both aspects of interference coordination (i.e., efficient utilization of radio resources and service isolation among service providers). Second, in C-RANs, user mobility triggers frequent handoff across small cells, which further complicates the coordination of the intra-tier interference among RRHs and the isolation among service providers. Moreover, due to the variation of users' traffic demand and locations, the amount of resources required at a small cell by each service provider changes, which requires dynamic update of vRANs. Hence, we are motivated to develop an efficient resource sharing mechanism to dynamically allocate radio resources while providing service isolation among service providers and taking into account fronthaul capacity constraint, intra-tier interference, user mobility, and traffic variation. The major contributions are summarized as follows:

- We propose a user-centric resource sharing scheme for a C-RAN with capacity constrained fronthaul links, in which the network operator jointly determines the resource allocation as well as user admission and association. To guarantee service isolation, we introduce an interference threshold to limit the maximum interference at each user. We define a novel metric to determine the minimum aggregate data rate to be allocated to each service provider based on users' QoS requirements and their maximum achievable data rates. The proposed scheme also employs a mobility prediction approach to estimate the locations of users in a short period and uses this information for traffic demand estimation and interference control.
- We design an efficient resource allocation algorithm to assist the resource sharing process. We formulate the resource allocation problem as an MINLP problem, and transform it into a mixed-integer linear programming (MILP) problem using a linearization technique. We propose an increment-based greedy allocation algorithm to obtain a suboptimal solution, which is more time-efficient than the standard techniques.
- To address the issue of traffic variation and user mobility, we propose a multi-timescale resource allocation mechanism. This mechanism consists of global resource allocation and local resource updates to deal with the variation of the network status, which is more efficient than performing network-wide optimization only.
- We discuss possible extensions of the proposed mechanism for uplink transmission and revenue maximization. We show that the optimization problem in the uplink can be transformed into an MILP problem, and can be solved using similar approaches as the downlink scenario. We also show that by adjusting the weighting factors in the objective function, we can achieve revenue maximization under both fixed-rate pricing and tiered pricing schemes.
- Through extensive simulations, we show that our proposed mechanism achieves efficient resource utilization and the isolation among service providers under various network situations with different traffic loads. It achieves higher throughput than an existing proportional spectrum sharing mechanism. When users are mobile, we also show that the proposed mechanism is more robust than the mechanisms without mobility prediction.

The rest of this paper is organized as follows. In Section II, we describe the system model and user-centric network sharing scheme, and formulate the resource allocation problem. In Section III, we propose an efficient algorithm to solve the resource allocation problem. Section IV describes the proposed multi-timescale resource sharing mechanism. Further extensions are discussed in Section V. Performance of the proposed mechanism is evaluated in Section VI. Conclusions are drawn in Section VII. A list of key notations is shown in Table I.

TABLE I
LIST OF KEY NOTATIONS

Symbol	Definition
$a_{u,j,\tau}$	Indicator that user u is served by RRH j at the τth predicted location
$\mathcal{A}_{j,k,\tau}$	Possible user association profile at RRH j at the τth prediction in channel k
B	Bandwidth of each orthogonal channel
B_j^{th}	Capacity of a fronthaul between RRH j and data center
$C_{j,k}$	Indicator that channel k is allocated to RRH j
$g_{u,j,\tau}$	Channel gain between user u and RRH j at the τth predicted location
$\mathcal{M}_{u,\tau}$	Set of RRHs to be associated for user u at the τth predicted location
n_p	Pre-determined number of locations used for estimation
$\mathcal{N}$	Set of orthogonal channels
P_{RRH}	Transmission power of the RRH
$P_u^{\max}$	Maximum transmission power of user u
R_{ref}	Reference data rate for a user
$R_s^{\min}$	Minimum aggregate data rate guaranteed for service provider s
R_s^{ref}	Reference minimum aggregate data rate for service provider s
R_u	Data rate for user u
$\widetilde{R}_{u,j,\tau}$	Unit data rate from RRH j to user u at the τth predicted location
$\mathcal{S}$	Set of service providers
T_G	Number of time slots of a global resource allocation process
T_L	Number of time slots of a local resource allocation process
$\widetilde{\mathcal{U}}_{j,\tau}$	Set of users associated with RRH j at the τth predicted location
$\mathcal{U}_s$	Set of subscribed users for service provider s
$w_{u,j,\tau}$	Proportion of time that user u can access a channel at RRH j at the τth predicted location
z_u	Indicator that user u is admitted by the corresponding service provider
β_u	Weighting factor for user u
Δt	Number of time slots within one prediction location
ϵ	Interference threshold

II. SYSTEM MODEL

We consider a C-RAN as shown in Fig. 1, which consists of a radio access network (RAN) and a cloud-based data center. The RAN consists of a set of small cell RRHs (i.e., microcell or picocell RRHs), which is denoted as $\mathcal{M}$. All RRHs are connected to the data center via high-speed optical fiber. The data center performs centralized resource management and control operation for the RAN. The available spectrum for this system is divided into N orthogonal channels with equal bandwidth B. Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of channels. We consider path loss and shadowing effect of the wireless channel, and the average channel gain between a user and an RRH is distance-dependent. The system is time-slotted, where only one user can access a particular channel from an RRH during one time slot. However, different users may share the same channel via time division multiple access. The same channel can be reused by multiple RRHs to improve the system throughput under certain interference constraints specified by the network operator.

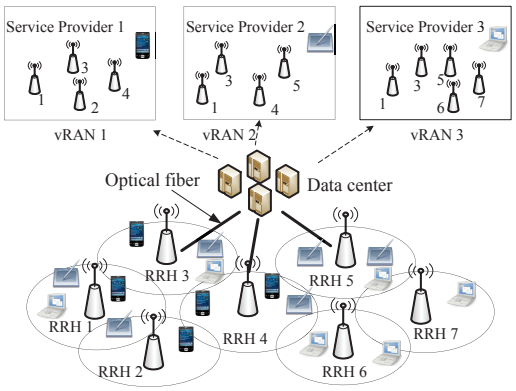


Fig. 1. Illustration of a cloud radio access network. Multiple service providers lease the infrastructure and radio resource from a network operator. The network operator creates a vRAN for each service provider by assigning a number of RRHs and the corresponding channels. Each RRH can serve users belonging to different service providers.

We consider a number of service providers share the C-RAN. The network operator owns the infrastructure and the spectrum, and can lease these resources to a set of service providers, denoted as $\mathcal{S}$. Each service provider $s \in \mathcal{S}$ provides data services with certain quality-of-service (QoS) requirements, such as video streaming and sports TV broadcasting, to a set of subscribed users, which is denoted as $\mathcal{U}_s$. To enable resource sharing among service providers, the network operator employs virtualization techniques to create vRANs and allocate corresponding resources periodically. Specifically, at the beginning of the virtualization process, each service provider sends a reservation request for certain traffic demand (or aggregate data rate required from the users) to the network operator. The network operator creates a vRAN for each service provider, which consists of a number of virtual RRHs (vRRHs) and BBUs in the data center as shown in Fig. 1. Each vRRH can be mapped to a real RRH in the RAN. Multiple vRRHs from different service providers can be mapped to the same RRH and they share the resources available at that RRH. The network operator also determines the amount of channel resources allocated to each vRRH. After that, each service provider performs scheduling and data transmission in its corresponding vRAN. The vRAN created for each service provider remains unchanged until the next resource allocation is performed at the network operator.

The resource allocation process is performed every T time slots, and we refer T time slots as a resource sharing period. We consider the resource allocation at the beginning of a T time slot period $[t, t+T)$, and define $C_{j,k} \in \{0, 1\}$ as the channel allocation variable, where $C_{j,k} = 1$ indicates channel $k \in \mathcal{N}$ is allocated to RRH $j \in \mathcal{M}$, and $C_{j,k} = 0$ otherwise.

A. User-Centric Resource Sharing Scheme

In conventional resource sharing schemes, the service providers need to estimate their users' resource demand before sending the reservation requests. Performing such an estimation usually requires the knowledge of user association decision to estimate the channel gain for each user and the available channel information at each base station to estimate

the corresponding interference. Most of the existing estimation approaches assume simple resource allocation and user association decisions, i.e., the channels available at each base station are either orthogonally allocated or follow a fixed reuse pattern. This may lead to unbalanced resource reservation in the network and competition among the service providers when the users are not uniformly distributed. In this paper, taking advantages of the cloud computing capability in C-RANs, we shift the resource demand estimation task from the service providers to the network operator, and propose a user-centric resource sharing mechanism[1]. The basic idea is to let the service providers send the QoS requirements for their subscribed users to the network operator. Based on the users' information, the network operator performs joint admission control, user association, and resource allocation to create the vRANs with guaranteed QoS for admitted users. Different from the existing resource sharing schemes, we introduce the following metrics and approaches during the network-wide sharing process:

1) Interference threshold: Due to channel reuse among the RRHs, different scheduling and transmission decisions at one service provider may affect the level of interference experienced in other vRANs, which in turn affect other service providers' decisions. To decouple the transmission decisions in service providers, we limit the interference among vRANs within a controllable range. Specifically, we introduce an *interference threshold* ϵ to assist the service isolation process. During resource allocation, we require that the interference experienced at each user in a vRAN should not exceed this threshold. Otherwise, transmission scheduling in one vRAN may affect the transmission scheduling and decisions in other vRANs and the network provider cannot benefit from the advantages of virtualization. By introducing the interference threshold, a service provider can make scheduling decisions based on the threshold instead of the actual interference originated from other vRANs. The value of the threshold can be selected as the one that maximizes the average system throughput, which can be obtained via computer simulations.

2) Rate estimation with mobility prediction: The resource demand estimation is performed by the network operator at the beginning of a resource sharing period, i.e., at time slot t in period $[t, t+T)$. When a user is mobile, the results obtained based on the information (e.g., channel gain) at time slot t may not be sufficient to guarantee its QoS during the entire period $[t, t+T)$. To address this issue, we exploit the users' mobility information when estimating their achievable data rates. Specifically, at time slot t, we consider the network operator can predict the locations of the user at future time slots $t+\Delta t, t+2\Delta t, \ldots, t+T-\Delta t$ according to a certain mobility prediction mechanism [26], where $\Delta t = T/n_p$ is the number of time slots in each prediction period and is assumed to be integer. Furthermore, n_p is a pre-determined positive integer that represents the number of locations we used for estimation. We denote the predicted location at time slot $t+\tau\Delta t$ as the τth predicated location, where τ takes values from set $\mathcal{T} = \{0, 1, \ldots, n_p - 1\}$.

To determine the resource allocation, we define $\mathcal{M}_{u,\tau}, \forall\, \tau \in \mathcal{T}$ as the possible set of RRHs to be associated for user $u \in \mathcal{U}_s$ at the τth predicted location. We denote $a_{u,j,\tau} \in \{0,1\}$ as the corresponding association variable where $a_{u,j,\tau} = 1$ indicates user u is served by RRH $j \in \mathcal{M}_{u,\tau}$ at the τth predicted location, and $a_{u,j,\tau} = 0$ otherwise. We further denote discrete variable $w_{u,j,\tau} \in \{0, \frac{1}{T}, \ldots, \frac{\Delta t}{T}\}, \forall\ j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T}, u \in \mathcal{U}_s, s \in \mathcal{S}$ as the resource sharing variable, which indicates the proportion of time that user u can access a channel at RRH j during τth period. Note that $w_{u,j,\tau} = \frac{1}{T}$ means user u is allocated a channel at RRH j in only one time slot, whereas the user is allocated all Δt time slots during τth period when $w_{u,j,\tau} = \frac{\Delta t}{T} = \frac{1}{n_p}$. The variables $a_{u,j,\tau}$ and $w_{u,j,\tau}$ remain unchanged within Δt time slots. Since the minimum resource unit is one time slot and the maximum number of time slots available for each user in $[t+\tau\Delta t, t+(\tau+1)\Delta t)$ is $\Delta t = T/n_p$, we have $1/T \le w_{u,j,\tau} \le \Delta t/T = 1/n_p$ for user u if it is served by RRH $j \in \mathcal{M}_{u,\tau}$. Thus,

$$\frac{a_{u,j,\tau}}{T} \le w_{u,j,\tau} \le \frac{a_{u,j,\tau}}{n_p}, \forall\, j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T}, u \in \mathcal{U}_s, s \in \mathcal{S}. \quad (1)$$

Equivalently, if user u is not associated with any RRH during τth period (e.g., $a_{u,j,\tau} = 0$), resource sharing variable $w_{u,j,\tau}$ has to be zero. To ensure that each user can only access one channel at a time, we require

$$\sum_{j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T}} w_{u,j,\tau} \le 1, \quad \forall\ u \in \mathcal{U}_s, s \in \mathcal{S}. \quad (2)$$

Note that the maximum number of time slots in a resource sharing period is T. In addition, the resource sharing scheme should ensure that the total resources allocated to users from each RRH does not exceed its available resources. During each time slot, a channel at each RRH can be allocated to at most one user[2]. Thus, during any period $[t+\tau\Delta t, t+(\tau+1)\Delta t), \forall \tau \in \mathcal{T}$,

$$\sum_{u \in \widetilde{\mathcal{U}}_{j,\tau}} w_{u,j,\tau} \le \frac{1}{n_p} \sum_{k \in \mathcal{N}} C_{j,k}, \quad \forall\ j \in \mathcal{M}, \tau \in \mathcal{T}, \quad (3)$$

where $\widetilde{\mathcal{U}}_{j,\tau} \triangleq \{u \mid j \in \mathcal{M}_{u,\tau}, u \in \mathcal{U}_s, s \in \mathcal{S}\}$ is the set of users associated with RRH j at the τth predicted location. Note that $C_{j,k} = 1$ indicates that channel k is allocated to RRH j. Therefore, the right hand side of constraint (3) is the total resources available at RRH j during τth period.

We now determine the estimated data rate for user $u \in \mathcal{U}_s$ during one resource sharing period as follows.

$$R_u = \sum_{j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T}} w_{u,j,\tau} B \log_2 \left(1 + \frac{P_{\text{RRH}}\, g_{u,j,\tau}}{\sigma^2 + \epsilon}\right),$$

where $g_{u,j,\tau}$ is the channel gain between user u and RRH j at the τth predicted location, P_{RRH} is the transmission power of

[1]In this paper, we focus on developing an efficient resource sharing mechanism for an already deployed C-RAN, where the network operator knows the locations of all RRHs. Such a network setting is different from that of stochastic geometry based analytical frameworks [24], [25], in which the network performance is analyzed to provide useful insights on network deployments by modeling the spatial locations of the RRHs as a Poisson point process.

[2]By utilizing techniques such as beamforming, users from different service providers may be able to concurrently access the same channel of an RRH. However, for the sake of tractability of the analysis, we assume that at most one user can access a channel of an RRH during each time slot.

the RRH, and σ^2 is the noise power. Let $z_u \in \{0,1\}$ denote the admission control variable for user u, where $z_u = 1$ indicates that user u is admitted by the corresponding service provider. Each user u is admitted if there exists at least one period $\tau \in \mathcal{T}$ in an RRH such that $w_{u,j,\tau} > 0$. Equivalently,

$$z_u \leq \sum_{j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T}} a_{u,j,\tau}, \quad \forall\ u \in \mathcal{U}_s, s \in \mathcal{S}. \tag{4}$$

If user u is not allocated a channel in any time slot, the right hand side of constraint (4) will be zero. The QoS requirement of each admitted user u is satisfied when its achievable data rate is within a certain range denoted as $[R_u^{\min}, R_u^{\max}]$. The following constraint ensures that the QoS requirements of all users are achieved.

$$z_u R_u^{\min} \leq R_u \leq z_u R_u^{\max}, \quad \forall\ u \in \mathcal{U}_s, s \in \mathcal{S}. \tag{5}$$

The estimated data rate R_u is obtained based on the interference threshold policy to take the advantages of virtualizaton. Therefore, for any user u associated with RRH j, the aggregate interference induced by other RRHs at any allocated channel should not exceed the threshold value ϵ. Mathematically,

$$a_{u,j,\tau} C_{j,k} \textstyle\sum_{l \in \mathcal{M} \setminus \{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau} \leq \epsilon, \\ \forall\ k \in \mathcal{N}, j \in \mathcal{M}, \tau \in \mathcal{T}, u \in \mathcal{U}_s, s \in \mathcal{S}. \tag{6}$$

3) Dynamic rate guarantee: The network operator should reserve some resources for each service provider to guarantee the minimum aggregate data rate of users. However, since users are mobile, the amount of radio resources consumed for guaranteeing a fixed aggregate data rate for a service provider varies, which may result in unbalanced resource allocation and low system throughput. For example, a service provider may consume a large amount of wireless resources when the subscribed users experience poor channel conditions. Thus, the minimum aggregate data rate guaranteed for each service provider should be dynamically adjusted in order to improve the system throughput. In this paper, we propose the following metric to determine the minimum aggregate data rate guaranteed for each service provider. We define R_s^{ref} as the reference minimum aggregate data rate for service provider $s \in \mathcal{S}$, and R_{ref} as the data rate that a user can achieve at a reference distance to an RRH with a given bandwidth, which is used as a reference to facilitate problem formulation. R_s^{ref} is the upper bound of the minimum rate guaranteed by the network operator, which is specified in the service agreement. R_{ref} is calculated as the data rate of a user at a reference distance (i.e., 20 m) to an RRH. We denote the maximum achievable data rate of user $u \in \mathcal{U}_s$ as

$$R_u^* = \frac{1}{n_p} \sum_{\tau \in \mathcal{T}} B \log_2(1 + (P_{\text{RRH}} g_{u,j^*,\tau})/(\sigma^2 + \epsilon)),$$

where $g_{u,j*,\tau}$ is the channel gain between user u and its closest RRH j^* at the τth predicted location. Moreover, $\frac{1}{n_p}$ is the maximum value of $w_{u,j,\tau}$ as all the time slots in τth period are allocated to user u. Then, the minimum aggregate data rate guaranteed for service provider s, denoted as $R_s^{\min}$, is determined according to the following rules: (i) $R_s^{\min}$ should be no larger than the reference value R_s^{ref}. (ii) $R_s^{\min}$ should be no larger than the maximum traffic demand from the subscribed users, which is $\sum_{u \in \mathcal{U}_s} R_u^{\max}$. (iii) When the average value of users' maximum achievable data rates, $(1/|\mathcal{U}_s|) \sum_{u \in \mathcal{U}_s} R_u^*$, is smaller than the reference data rate R_{ref}, it implies that on average the users are relatively far from their closest RRH and guaranteeing R_s^{ref} consumes more resources than expected. In this case, the network operator only guarantees a lower data rate (down scale R_s^{ref} by a factor of $(1/|\mathcal{U}_s| \sum_{u \in \mathcal{U}_s} R_u^*)/R_{\text{ref}}$) for the service provider to save some resources. In summary, we have

$$R_s^{\min} = \min \left\{ R_s^{\text{ref}},\ \sum_{u \in \mathcal{U}_s} R_u^{\max},\ \frac{\sum_{u \in \mathcal{U}_s} R_u^*}{|\mathcal{U}_s| R_{\text{ref}}} R_s^{\text{ref}} \right\}.$$

The minimum resource guarantee for each service provider is specified as follows.

$$\sum_{u \in \mathcal{U}_s} R_u \geq R_s^{\min}, \quad \forall\ s \in \mathcal{S}. \tag{7}$$

4) Fronthaul constraints: We consider a fronthaul-constrained C-RAN. Fronthaul links generally suffer from capacity constraint and latency[3] for collecting the users' information (e.g., user locations and channel gains). The mobility prediction scheme employed in this paper can mitigate the effect of delay in fronthaul as we update the users' information less frequently. During each resource sharing period, the user locations will be predicted, from which the channel gains can be determined. However, the limited capacity of fronthaul links affects the resource allocation. We denote the capacity of an optical fronthaul link which connects RRH $j \in \mathcal{M}$ to the data center as B_j^{fh}. The aggregate data rate transmitted over the fronthaul link of RRHs should satisfy the following constraint.

$$\sum_{u \in \widetilde{\mathcal{U}}_{j,\tau}} n_p w_{u,j,\tau} B \log_2 \left(1 + \frac{g_{u,j,\tau} P_{\text{RRH}}}{\sigma^2 + \epsilon}\right) \leq B_j^{\text{fh}},\ j \in \mathcal{M}, \tau \in \mathcal{T}. \tag{8}$$

Note that $n_p w_{u,j,\tau}$ in the left hand side of constraint (8) is the number of time slots allocated to user u during τth period.

5) Optimization objective: The network operator maximizes the system throughput under different traffic load situations, which can be characterized using the following objective function

$$f = \sum_{u \in \mathcal{U}_s, s \in \mathcal{S}} \beta_u R_u, \tag{9}$$

where β_u is a time-dependent and user-dependent weighting factor. The objective function characterizes the weighted sum rate of the system, which is equivalent to the system throughput when $\beta_u = 1, \forall\ u \in \mathcal{U}_s, s \in \mathcal{S}$. The weighting factor β_u is used to characterize the importance of the users and can be interpreted as the importance of admission control for different users.

Based on the previous discussion, in the proposed scheme, the network operator determines the number of channels

[3]Two types of delay in the fronthaul can affect the resource allocation in C-RANs. First, delay in the fronthaul uplink for transmission of users' control information may affect the resource allocation problem and degrade the system throughput. Second, delay may occur in the fronthaul downlink for transmission of baseband signals. In this paper, we assume that the latter delay is negligible. The same assumption is used in [6] and [17].

allocated to each RRH, the amount of resources shared by each service provider, as well as the admission control and resource sharing variables.

B. Resource Allocation Problem at the Operator

To formulate the resource allocation problem, we define $\mathbf{C} = (C_{j,k}, j \in \mathcal{M}, k \in \mathcal{N})$ and $\mathbf{z}_s = (z_u, u \in \mathcal{U}_s)$ as channel allocation and admission control variables. We further denote the user association variables and resource sharing variables for service provider $s \in \mathcal{S}$ as $\boldsymbol{a}_s = (a_{u,j,\tau}, u \in \mathcal{U}_s, j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T})$ and $\mathbf{w}_s = (w_{u,j,\tau}, u \in \mathcal{U}_s, j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T})$, respectively. Then, the resource allocation problem is to determine the decision variables $\mathbf{z}_s$, $\boldsymbol{a}_s$, $\mathbf{w}_s$, $\forall\, s \in \mathcal{S}$, and $\mathbf{C}$, which can be formulated as the following optimization problem.

$$\underset{\mathbf{z}_s,\boldsymbol{a}_s,\mathbf{w}_s,s\in\mathcal{S},\mathbf{C}}{\text{maximize}} \quad f = \sum_{u\in\mathcal{U}_s,s\in\mathcal{S}} \beta_u R_u \tag{10a}$$

$$\text{subject to} \quad \text{constraints (1)–(8)}, \tag{10b}$$

$$C_{j,k} \in \{0,1\}, \qquad j \in \mathcal{M}, k \in \mathcal{N}, \tag{10c}$$

$$z_u \in \{0,1\}, \qquad u \in \mathcal{U}_s, s \in \mathcal{S}, \tag{10d}$$

$$a_{u,j,\tau} \in \{0,1\}, u \in \mathcal{U}_s, s \in \mathcal{S}, j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T}, \tag{10e}$$

$$w_{u,j,\tau} \in \left\{0, \frac{1}{T}, \ldots, \frac{\Delta t}{T}\right\}, \quad u \in \mathcal{U}_s, s \in \mathcal{S}, j \in \mathcal{M}_{u,\tau}, \tau \in \mathcal{T}. \tag{10f}$$

Constraint (6) is non-convex and problem (10) is an MINLP problem with non-convex constraints, which is difficult to solve in practice. In the following sections, we design an efficient algorithm to find a suboptimal solution.

III. Efficient Resource Allocation Algorithm

In this section, we first transform problem (10) into an MILP problem, and then design an efficient algorithm to find its solution. Specifically, we convert the non-convex constraint (6) into the following linear constraint

$$\sum_{l\in\mathcal{M}\setminus\{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau} \le \frac{(a_{u,j,\tau} + C_{j,k})\epsilon}{2} + (2 - a_{u,j,\tau} - C_{j,k})D, \quad \forall\, k \in \mathcal{N}, j \in \mathcal{M}, \tau \in \mathcal{T}, u \in \mathcal{U}_s, s \in \mathcal{S}, \tag{11}$$

where D is a large constant. Regarding this linearization, we have the following theorem.

Theorem 1: Constraint (11) is equivalent to constraint (6) if D and ϵ satisfy

$$D \ge \max\left\{\Gamma - \frac{\epsilon}{2}, \frac{\Gamma}{2}\right\}, \tag{12}$$

where

$$\Gamma = (|\mathcal{M}| - 1) P_{\text{RRH}} \max_{u\in\mathcal{U}_s, s\in\mathcal{S}, j\in\mathcal{M}, \tau\in\mathcal{T}} \{g_{u,j,\tau}\}. \tag{13}$$

Proof: We show constraints (6) and (11) are equivalent under condition (12) when $a_{u,j,\tau}$ and $C_{j,k}$ take any feasible values from $\{0,1\}$. First, when $a_{u,j,\tau}C_{j,k} = 1$ (i.e., $a_{u,j,\tau} = 1$ and $C_{j,k} = 1$), constraints (6) and (11) are the same. Second, when $a_{u,j,\tau}C_{j,k} = 0$, (6) is always satisfied. Thus, we only need to show that (11) is always satisfied when $a_{u,j,\tau}C_{j,k} = 0$. We consider the following cases for $a_{u,j,\tau}C_{j,k} = 0$: When $a_{u,j,\tau} = 0$ and $C_{j,k} = 0$, constraint (11) becomes $\sum_{l\in\mathcal{M}\setminus\{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau} \le 2D$. When $a_{u,j,\tau} = 1$ and $C_{j,k} = 0$ or $a_{u,j,\tau} = 0$ and $C_{j,k} = 1$, constraint (11) becomes $\sum_{l\in\mathcal{M}\setminus\{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau} \le \epsilon/2 + D$. Based on (13), we have $\Gamma \ge \sum_{l\in\mathcal{M}\setminus\{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau}$. Thus, according to condition (12), it can be verified that (11) is always satisfied in all three cases. This completes the proof. ■

With the linear constraint (11), problem (10) becomes

$$\underset{\mathbf{z}_s,\mathbf{w}_s,\boldsymbol{a}_s,s\in\mathcal{S},\mathbf{C}}{\text{maximize}} \quad f \tag{14a}$$

$$\text{subject to} \quad \text{constraints (1)–(5), (7)–(8), (11), (10c)–(10f)}. \tag{14b}$$

Problem (14) is an MILP, which can be solved by applying standard techniques such as the branch and bound method. However, the computational complexity of these techniques increases significantly as the size of the problem increases. Therefore, in a system with many users, finding the optimal solution to problem (14) by applying the standard techniques may consume a large amount of time, which may not be practical for real-time processing. To address this issue, in the following, we propose a fast algorithm to find an efficient suboptimal solution.

To jointly determine the channel allocation, resource sharing, and user admission and association decisions, we propose an increment-based greedy allocation (IBGA) algorithm. The basic idea of the proposed algorithm is to allocate the available channels to the RRHs one by one. For each channel, we allocate it to the RRHs iteratively, where in each iteration we select the RRH that has the largest increment of the objective value while satisfying the interference constraint (11). The allocation of a channel terminates when no more RRH can use this channel under the interference constraint. Once the channel allocation is fixed, the user admission and association are also determined accordingly. To characterize the increment of the objective value, we first relax the binary variable z_u to be a continuous variable $\widetilde{z}_u \in [0,1]$, where $\widetilde{z}_u = \min\{R_u/R_u^{\min}, 1\}$. We denote ΔR_u as the increment of data rate when user u is allocated additional resources, and further denote $\Delta R_u^{\min} = R_u^{\min} - R_u$. Then, we have

$$\Delta\widetilde{z}_u = \frac{\min\{\Delta R_u, \Delta R_u^{\min}\}}{R_u^{\min}}. \tag{15}$$

We further define $\widetilde{R}_{u,j,\tau} \triangleq B\log_2(1 + (P_{\text{RRH}} g_{u,j,\tau})/(\sigma^2 + \epsilon))$ as the unit data rate from RRH j to user u at the τth predicted location, and denote $\Delta w_{u,j,\tau}$ as the corresponding additional resources allocated from RRH j to user u at the τth predicted location. Then, for user u associated with RRH j, the increment of data rate at the τth predicted location is

$$\Delta R_{u,j,\tau}(\Delta w_{u,j,\tau}) = \Delta w_{u,j,\tau} B\log_2\left(1 + \frac{P_{\text{RRH}} g_{u,j,\tau}}{\sigma^2 + \epsilon}\right) = \Delta w_{u,j,\tau}\widetilde{R}_{u,j,\tau}. \tag{16}$$

We define $\Delta f_{j,k} \triangleq \sum_{\tau\in\mathcal{T}} \Delta f_{j,k,\tau}$ as the increment of the objective value when channel k is allocated to RRH j, where $\Delta f_{j,k,\tau}$ denotes the maximum increment of objective value when RRH j allocates Δt time slots for its associated users at

the τth predicted locations. The objective value is the weighted sum of data rate increment for the users associated with RRH j at the τth prediction in channel k. Thus, $\Delta f_{j,k,\tau}$ is the optimal value of the following problem

$$\underset{\Delta w_{u,j,\tau}, u\in\mathcal{A}_{j,k,\tau}}{\text{maximize}} \quad \sum_{u\in\mathcal{A}_{j,k,\tau}} \beta_u \Delta R_{u,j,\tau}(\Delta w_{u,j,\tau}) \tag{17a}$$

$$\text{subject to} \quad \sum_{u\in\mathcal{A}_{j,k,\tau}} (w_{u,j,\tau} + \Delta w_{u,j,\tau}) \leq \frac{\Delta t}{T}, \tag{17b}$$

$$\sum_{u\in\mathcal{A}_{j,\tau}\cup\mathcal{A}_{j,k,\tau}} (w_{u,j,\tau}+\Delta w_{u,j,\tau})\tilde{R}_{u,j,\tau} \leq B_j^{\text{fh}}, \tag{17c}$$

$$\Delta w_{u,j,\tau} \in \left\{0, \frac{1}{T}, \ldots, \frac{\Delta t}{T}\right\}, \tag{17d}$$

where $\mathcal{A}_{j,k,\tau}$ is the possible user association profile at RRH j at the τth prediction in channel k. Moreover, $\mathcal{A}_{j,\tau}$ is the set of users who have already allocated with other channels at RRH j. We determine $\mathcal{A}_{j,k,\tau}$ by checking constraints (6) and (7). Specifically, we denote $I_{\text{RRH}}(j,k)$ as an indicator function where $I_{\text{RRH}}(j,k)=0$ indicates that constraint (11) cannot be satisfied if channel k is allocated to RRH j, and $I_{\text{RRH}}(j,k)=1$ otherwise. We denote $I_{\text{user}}(u,j,\tau)$ as an indicator function to show whether constraint (6) is satisfied at user u served by RRH j at the τth predicted location. We further define I_{sp} as an indicator function to show whether constraint (7) is satisfied, and denote $\mathcal{S}'$ as the set of service providers which do not satisfy constraint (7). Then, we have

$$\mathcal{A}_{j,k,\tau} = \begin{cases} \{u \mid u\in\widetilde{U}_{j,\tau}, I_{\text{user}}(u,j,\tau)=1, s(u)\in\mathcal{S}'\}, \\ \qquad \text{if } I_{\text{RRH}}(j,k)=1 \text{ and } I_{\text{sp}}=0, \\ \{u \mid u\in\widetilde{U}_{j,\tau}, I_{\text{user}}(u,j,\tau)=1\}, \\ \qquad \text{if } I_{\text{RRH}}(j,k)=1 \text{ and } I_{\text{sp}}=1, \\ \emptyset, \qquad \text{otherwise}, \end{cases} \tag{18}$$

where $s(u)$ is the service provider that user u subscribed to. The set in (18) implies that when constraint (7) is not satisfied, we only consider allocating resources to users subscribed to the service providers which have not achieved the minimum guaranteed resources. Otherwise, all users are considered eligible for additional resources. It can be verified that $\Delta\widetilde{z}_u(\Delta w_{u,j,\tau})$ is a piece-wise concave function with respect to $\Delta w_{u,j,\tau}$. Thus, problem (17) can be solved using standard techniques such as branch and bound. Problem (17) is a mixed integer programming problem. The size of this problem (i.e., the number of variables), which is equal to $|\mathcal{A}_{j,k,\tau}|$, is at most the number of users located in the coverage area of RRH j. However, this is conservative as each RRH can use multiple channels. Although the worst case complexity of solving problem (17) is exponential in $|\mathcal{A}_{j,k,\tau}|$, we will show in Section VI that the running time of our proposed resource allocation algorithm varies linearly with the number of users. Note that we have relaxed the admission control variable z_u when determining the increment of objective value. After the resource allocation process, we convert $\widetilde{z}_u$ back to binary variable z_u as

$$z_u = \begin{cases} 1, & \text{if } \widetilde{z}_u = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Note that if $\widetilde{z}_u = \min\{R_u/R_u^{\min}, 1\} < 1$, to guarantee constraint (5), we need to set $z_u = 0$. The post-processing of z_u does not affect the amount of resources each service provider obtained from the network operator. With the above definitions, the proposed IBGA algorithm is shown in Algorithm 1.

In Algorithm 1, we allocate available channels to the RRHs. For each channel $k\in\mathcal{N}$, we first determine the possible set of users to be associated with each RRH in Step 3. We then allocate channel k iteratively to the RRHs. In each iteration, we first calculate the increment of objective value, $\Delta f_{j,k}$, according to (17). Then, we allocate channel k to the RRH with the largest value of increment in Step 7. Next, we update the user association variables and resource sharing variables according to the solution of problem (17) in Step 9. Then, we update the possible user association profile $\mathcal{A}_{j,k,\tau}$ from Steps 11 to 23. We remove the service provider s from set $\mathcal{S}'$ which has achieved the minimum rate guarantee in Step 14, and update $\mathcal{A}_{j,k,\tau}$ by removing corresponding users in set $\mathcal{U}_s$ in Step 15. Once all the service providers achieved the minimum rate guarantee, we set $I_{\text{sp}}=1$ in Step 17. By using Steps 11 to 18, we aim to allocate resources to service providers which have not achieved the minimum rate guarantee with a higher priority. Finally, we check whether this channel can be allocated to other RRHs under the interference constraints from Steps 19 to 23. If it still can be allocated to RRH j, we update $\mathcal{A}_{j,k,\tau}, \forall\ \tau\in\mathcal{T}$ for RRH j considering the interference constraints at the users in Step 18. Steps 7 to 23 will be repeated until channel k cannot be further allocated to any RRH. Using this algorithm, we obtain the desired channel allocation decision, user association decision, resource sharing decision, and determine the admission control decision.

It can be seen that by using Algorithm 1, we can obtain a solution to problem (14) within a small number of iterations, i.e., $N|\mathcal{M}|$ iterations in total. Therefore, it is more efficient compared to the standard techniques for large-scale networks. However, since we use a greedy algorithm when allocating each channel, and apply relaxation to the admission control variables, the solution obtained using Algorithm 1 is suboptimal. By using this algorithm, we can obtain the resources allocated to each service provider $s\in\mathcal{S}$ at each RRH $j\in\mathcal{M}$ as $W_{j,s} = \sum_{u\in\mathcal{U}_s,\tau\in\mathcal{T}} w_{u,j,\tau}$. Then, the network operator can create the vRAN for each service provider accordingly. We will evaluate the performance of Algorithm 1 in Section VI.

IV. Multi-Timescale Resource Sharing Mechanism

In the previous section, we have designed an efficient resource sharing algorithm to assist the virtualization process at the network operator, where the decisions remain unchanged for T time slots. A typical challenge in designing practical virtualization mechanism is to adapt to changes of the network status, such as the traffic variation and user mobility. Intuitively, when the users' locations and their traffic demand do not vary, the network operator does not need to update the vRAN for each service provider. In this scenario, we can select a large value of T to reduce computation and communication cost. On the contrary, when the network status

Algorithm 1: Increment-based greedy allocation (IGBA) algorithm.

Initialize variables $\mathbf{z}_s$, $\boldsymbol{a}_s$, $\mathbf{w}_s$, $s \in \mathcal{S}$, $\mathbf{C}$, f, I_{sp} to be all zeros, $\mathcal{S}' := \mathcal{S}$, and $\mathcal{A}_{j,\tau} = \emptyset$, $j \in \mathcal{M}, \tau \in \mathcal{T}$
for each channel $k \in \mathcal{N}$ **do**
  Determine $\mathcal{A}_{j,k,\tau}, j \in \mathcal{M}, \tau \in \mathcal{T}$ according to (18).
  Initialize set $\mathcal{M}_k := \mathcal{M}$.
  while $\mathcal{M}_k \neq \emptyset$ **do**
    Solve problem (17) for $j \in \mathcal{M}_k$ and calculate the value of $\Delta f_{j,k}, j \in \mathcal{M}_k$.
    Find the RRH with the largest increment of objective value: $q = \arg\max_{j \in \mathcal{M}_k} \Delta f_{j,k}$.
    Set $\mathcal{M}_k := \mathcal{M}_k \setminus \{q\}$ and $f := f + \Delta f_{q,k}$.
    Set $C_{q,k} := 1$ and update $a_{u,q,\tau}, w_{u,q,\tau}, u \in \mathcal{A}_{q,k,\tau}, \tau \in \mathcal{T}$ according to the solution in Step 6.
    Set $\mathcal{A}_{j,\tau} := \mathcal{A}_{j,\tau} \cup \{u \mid w_{u,j,\tau} > 0\}$, $j \in \mathcal{M}, \tau \in \mathcal{T}$.
    if $I_{\text{sp}} = 0$ **then**
      for $s \in \mathcal{S}'$ **do**
        if $\sum_{u \in \mathcal{U}_s, s \in \mathcal{S}'} R_u \geq R_s^{\min}$ **then**
          Set $\mathcal{S}' := \mathcal{S}' \setminus \{s\}$.
          Set $\mathcal{A}_{j,k,\tau} := \mathcal{A}_{j,k,\tau} \setminus \{u\}, u \in \mathcal{U}_s, j \in \mathcal{M}_k$.
      if $\mathcal{S}' = \emptyset$ **then**
        Set $I_{\text{sp}} := 1$.
        Update $\mathcal{A}_{j,k,\tau}$ according to (18).
    for RRH $j \in \mathcal{M}_k$ **do**
      if $I_{\text{RRH}}(j,k) = 0$ or $\mathcal{A}_{j,k,\tau} = \emptyset, \tau \in \mathcal{T}$ **then**
        Set $\mathcal{M}_k := \mathcal{M}_k \setminus \{j\}$.
      else
        Update $\mathcal{A}_{j,k,\tau}$ according to (18).
Determine the admission control variable $z_u, u \in \mathcal{U}_s, s \in \mathcal{S}$ according to (19).

changes frequently, the amount of resources required at each service provider may vary, which requires update of the vRAN frequently in a small time scale, i.e., we need to choose a small value of T, which results in high computation cost.

In this paper, we propose a multi-timescale resource sharing mechanism as shown in Fig. 2 to address this issue. This mechanism consists of a global resource allocation process which is performed every T_G time slots, and a number of local resource allocation processes performed every $T_L = T$ time slots (where $T_G = n_L T_L$ and n_L is a positive integer) between two consecutive global resource allocation. The allocated resource for each service provider remains unchanged during T_L time slots. In the global resource allocation process, all users in the system are involved in the optimization, and the available resources include all the channels in set $\mathcal{N}$, as shown in problem (10). However, in the local resource allocation process, only users whose locations and traffic demand have changed are considered.

The available resources for local allocation process only include the remaining resources (e.g., channels which have not been utilized) in the system. Without loss of generality, we consider global resource allocation is performed at time slot t_0, and the local resource allocation is performed at time slot $t_1 = t_0 + T$. All through this section, we use superscript to denote the time period when the decision is made. The remaining resources in the system can be classified into three types. The first type is the set of channels that can further be assigned to RRHs without violating the interference constraints of the existing users, which is denoted as $\mathcal{N}^{t_1}$. The second type of remaining resources is the amount of channel resources that are assigned to some RRHs but not fully utilized by the service providers. We denote this type of remaining resources at RRH $j \in \mathcal{M}$ for the τth prediction period as $r_{j,\tau}^{t_1}$. The last type of remaining resources corresponds to the amount of resources released from the service providers to the network operator. This may happen when some of the users subscribed to a service provider have finished their transmission or lower their traffic demand, and the service provider has more than enough resources to satisfy the maximum traffic demand of all its subscribed users. In this case, the service provider can decide to release the additional amount of resources back to the network operator in order to reduce expenses. We denote the amount of resources released from service provider s for RRH j for the τth prediction period as $\widetilde{W}_{s,j,\tau}^{t_1}$. The optimization variables at time slot t_1 include $C_{j,k}^{t_1}$, $\forall\ k \in \mathcal{N}^{t_1}, j \in \mathcal{M}$, $z_u^{t_1}$, $w_{u,j,\tau}^{t_1}$, $a_{u,j,\tau}^{t_1}$, $\forall\ j \in \mathcal{M}_u^{t_1}, \tau \in \mathcal{T}, u \in \mathcal{U}_s^{t_1}$, $s \in \mathcal{S}$, which are defined similarly to those in the global optimization. The optimization problem for local resource allocation is similar to the global optimization problem except that the superscript t is replaced by t_1 and constraint (3) is changed to

$$\sum_{u \in \mathcal{U}_j^t} w_{u,j,\tau} \leq \frac{1}{n_p} \sum_{k \in \mathcal{N}^{t_1}} C_{j,k}^{t_1} + r_{j,\tau}^{t_1} + \sum_{s \in \mathcal{S}} \widetilde{W}_{s,j,\tau}^{t_1}, \quad \forall\ \tau \in \mathcal{T}, j \in \mathcal{M}. \tag{20}$$

Therefore, the local optimization problem at time slot t_1 can be formulated as

$$\underset{\mathbf{z}_s^{t_1}, \mathbf{w}_s^{t_1}, \boldsymbol{a}_s^{t_1}, s \in \mathcal{S},\ \mathbf{C}^{t_1}}{\text{maximize}} \quad f^{t_1} \tag{21a}$$

$$\text{subject to} \quad \text{constraints (1)–(2), (4)–(5), (7)–(8), (10c)–(10f), (11), with superscript } t_1, \text{ and (20).} \tag{21b}$$

Problem (21) can be solved by applying Algorithm 1 with proper adjustment. Specifically, when initializing values of the decision variables and calculating the utility increment of the RRHs, only the amount of remaining resources and users in $\mathcal{U}_s^{t_1}$ are considered. After obtaining the solution to problem (21), the network operator updates the resources allocated to

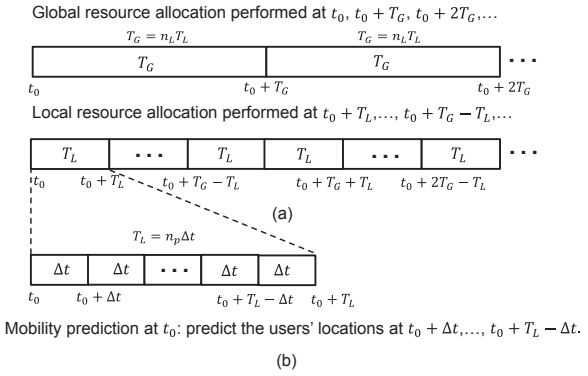


Fig. 2. (a) Multi-timescale resource sharing framework: Global resource allocation performed every T_G time slots, and local resource allocation performed every T_L time slots. (b) Mobility prediction: predicting the locations of mobile users every Δt time slots during the next T_L time slots.

Algorithm 2: Multi-timescale dynamic resource allocation algorithm during $[t_0, t_0 + T_G)$.

1 **for** $t := t_0$ to T_G **do**
2 **if** $t = t_0$ **then**
3 Collect information from service providers, including users' locations and their QoS requirements.
4 Predict the locations of mobile users for time slot $t + \tau\Delta t$, $\tau = 1, 2, \ldots, n_p - 1$
5 Find resource allocation decisions by solving problem (14) using Algorithm 1.
6 Calculate the resources allocated to each service provider $s \in \mathcal{S}$, $W^t_{s,j} = \sum_{u \in \mathcal{U}^t_s, \tau \in \mathcal{T}} w^t_{u,j,\tau}, j \in \mathcal{M}$.
7 Create a vRAN for each service provider $s \in \mathcal{S}$.
8 **if** $t = t_0 + \tau\Delta t$, $\tau = 1, 2, \ldots$ **then**
9 Update the location of all users in the system.
10 **if** $t = t_0 + mT$, $m = 1, 2, \ldots, n_L - 1$ **then**
11 Find the remaining channels that can further be allocated, $\mathcal{N}^t$, using exhaustive search.
12 Set $r^t_{j,\tau} = 1/n_p \sum_{k \in \mathcal{N}^{t-T}} C^{t-T}_{j,k} - \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}^{t-T}_s} w^{t-T}_{u,j,\tau}, j \in \mathcal{M}$.
13 Find the set of users for resource allocation, $\mathcal{U}^t_s, \forall\ s \in \mathcal{S}$.
14 Predict the locations of mobile users for time slot $t + \tau\Delta t$, $\tau = 1, 2, \ldots, n_p - 1$.
15 Find resource allocation decisions by solving problem (21).
16 Update the vRAN resources for each service provider $s \in \mathcal{S}$ according to (22).

each service provider $s \in \mathcal{S}$ at RRH $j \in \mathcal{M}$ at time slot t_1 as

$$W^{t_1}_{s,j} = W^{t_0}_{s,j} - \sum_{\tau \in \mathcal{T}} \widetilde{W}^{t_1}_{s,j,\tau} + \sum_{\tau \in \mathcal{T}, u \in \mathcal{U}^{t_1}_s} w^{t_1}_{u,j,\tau}. \tag{22}$$

Based on the previous discussion, the procedures of the proposed dynamic resource sharing mechanism during time slot $[t_0, t_0 + T_G)$ are shown in Algorithm 2. In Algorithm 2, the parameters $n_L = T_G/T$ and $n_p = T/\Delta t$ are predetermined integers. Steps 3 to 7 in Algorithm 2 correspond to the global resource allocation process, where the network operator creates vRAN for each service provider based on their reservation requests. Step 9 is to update the locations of users in the system, which can be achieved via location monitoring techniques such as global provisioning system (GPS). In Steps 11 to 16, the network operator performs local resource update for service providers by solving problem (21) every T time slots. In Steps 4 and 14, the network operator needs to predict the next $n_p - 1$ locations for each mobile user. Such prediction can be achieved by applying any existing mobility prediction algorithm such as the order-2 Markov predictor [26]. Note that when there is no remaining resources in the system or requests from the service providers, the network operator does not perform local optimization.

V. Further Extension

In this section, we discuss possible extensions of the proposed multi-timescale resource sharing mechanism to address some related issues of resource sharing.

A. Dynamic Resource Sharing for Uplink

Although the dynamic resource sharing mechanism proposed in Section IV is based on downlink communication, it can also be applied to the uplink with proper adjustment. In the uplink, users transmit data to the RRHs, and each RRH may experience a different level of interference. Similar to the downlink scenario, to provide service isolation among different service providers, we restrict that the aggregate interference experienced at each RRH $j \in \mathcal{M}$ is no larger than a threshold ϵ'. We also define the maximum transmission power allowed at user $u \in \mathcal{U}_s$ as $P^{\max}_u$, which depends on the device and applications that the user is using. For simplicity, we reuse the symbols for other variables defined in the downlink scenario. Then, the interference constraints at the RRHs are represented as

$$C_{j,k} \sum_{l \in \mathcal{M} \setminus \{j\}} C_{l,k} \max_{u \in \widetilde{\mathcal{U}}_l, \tau \in \mathcal{T}} \{a_{u,l,\tau} P^{\max}_u g_{u,j,\tau}\} \le \epsilon', \quad \forall\ k \in \mathcal{N}, j \in \mathcal{M}. \tag{23}$$

It can be verified that (23) is a non-convex constraint. To linearize (23), we introduce auxiliary variables $x_{l,j}$ and $y_{l,j,k}$ for all $j, l \in \mathcal{M}$, $k \in \mathcal{N}$, where

$$x_{l,j} = \max_{u \in \widetilde{\mathcal{U}}_l, \tau \in \mathcal{T}} \{a_{u,l,\tau} P^{\max}_u g_{u,j,\tau}\}, \tag{24}$$

$$y_{l,j,k} = C_{l,k} x_{l,j}. \tag{25}$$

With the auxiliary variables, for any $k \in \mathcal{N}$ and $j \in \mathcal{M}$, constraint (23) can be transformed into the following constraints

$$\sum_{l \in \mathcal{M} \setminus \{j\}} y_{l,j,k} \le \epsilon' + (1 - C_{j,k}) D', \tag{26a}$$

$$x_{l,j} \ge a_{u,l,\tau} P^{\max}_u g_{u,j,\tau},\ u \in \widetilde{\mathcal{U}}_l, \tau \in \mathcal{T}, l \in \mathcal{M} \setminus \{j\}, \tag{26b}$$

$$y_{l,j,k} \ge x_{l,j} - (1 - C_{l,k}) \max_{u \in \widetilde{\mathcal{U}}_l, \tau \in \mathcal{T}} \{P^{\max}_u g_{u,j,\tau}\},\ l \in \mathcal{M} \setminus \{j\}, \tag{26c}$$

$$0 \le y_{l,j,k} \le x_{l,j}, \qquad l \in \mathcal{M} \setminus \{j\}, \tag{26d}$$

$$y_{l,j,k} \le C_{l,k} \max_{u \in \widetilde{\mathcal{U}}_l, \tau \in \mathcal{T}} \{P^{\max}_u g_{u,j,\tau}\}, \qquad l \in \mathcal{M} \setminus \{j\}, \tag{26e}$$

where D' is a large positive constant. We have the following theorem.

Theorem 2: For any $j \in \mathcal{M}$ and $k \in \mathcal{N}$, constraint (23) is equivalent to constraint (26) if D' and ϵ satisfy

$$D' \ge (|\mathcal{M}| - 1) \max_{u \in \mathcal{U}_s, s \in \mathcal{S}, \tau \in \mathcal{T}} \{P^{\max}_u g_{u,j,\tau}\} - \epsilon'. \tag{27}$$

Proof: First, we show that constraint (23) is equivalent to constraints (24)–(26a). It can be verified that when $C_{j,k} = 1$, constraint (23) is equivalent to constraint (26a) by substituting (24) and (25). When $C_{j,k} = 0$, constraint (23) is always satisfied. In this scenario, constraint (26a) becomes

$$\sum_{l \in \mathcal{M} \setminus \{j\}} y_{l,j,k} \le \epsilon' + D', \tag{28}$$

which is also always satisfied when $D' \ge (|\mathcal{M}| - 1) \max_{u \in \mathcal{U}_s, s \in \mathcal{S}, \tau \in \mathcal{T}} \{P^{\max}_u g_{u,j,\tau}\} - \epsilon'$. Thus, constraint (23) is equivalent to constraints (24)–(26a). Next, it can be seen that (24) is equivalent to (26b). Finally, we show that (25) is equivalent to (26c)–(26e). When $C_{l,k} = 0$, we have $y_{l,j,k} = 0$

from (25). From (26d) and (26e), we also have $y_{l,j,k} = 0$, which implies that (25) is equivalent to (26d) and (26e) in this scenario. When $C_{l.k} = 1$, we have $y_{l,j,k} = x_{l,j}$ from (25). Meanwhile, from (26c) and (26d), we also have $y_{l,j,k} = x_{l,j}$. Therefore, (25) is equivalent to (26c)–(26e) under any value of $C_{l,k}$. In summary, constraint (23) is equivalent to constraints (26a)–(26e). This completes the proof. ■

According to Theorem 2, we have transformed the non-convex constraint (23) into a set of linear constraints (26). Note that the objective function and all other constraints in the uplink resource allocation problem is the same as those in the downlink scenario. Therefore, the global resource allocation problem for the uplink can be formulated as

$$\underset{\mathbf{z}_s, \mathbf{w}_s, \boldsymbol{\alpha}_s, s\in\mathcal{S}, \mathbf{C}}{\text{maximize}} \quad f \tag{29a}$$

$$\text{subject to} \quad \text{constraints (1)–(5), (7)–(8), (10c)–(10f)} \quad \text{with } \epsilon = \epsilon', P_{\text{RRH}} = P_u^{\max}, \text{ and (26).} \tag{29b}$$

Similar to the downlink scenario, we define $\widetilde{I}_{\text{RRH}}(j,k)$ as an indicator function to show whether channel k can be allocated to RRH j without violating constraints in (26). We further define $\widetilde{I}_{\text{user}}(u,j,\tau)$ as the indicator function whether user u can be associated with RRH j in the τ prediction period without violating the interference constraints. Then, the global resource allocation problem can be solved using Algorithm 1 by replacing ϵ, P_{RRH}, $I_{\text{RRH}}(j,k)$, and $I_{\text{user}}(u,j,\tau)$ with ϵ', $P_u^{\max}$, $\widetilde{I}_{\text{RRH}}(j,k)$, and $\widetilde{I}_{\text{user}}(u,j,\tau)$, respectively. Similarly, we can obtain the local resource allocation decisions for the uplink using Algorithm 1 with the aforementioned changes. Therefore, the proposed mechanism in Algorithm 2 can also be applied for uplink resource allocation.

B. Revenue Maximization for On-Demand Service

In the proposed resource sharing mechanism, we optimize an objective function that characterizes the weighted sum rate of the system. In this subsection, we show that the objective function (9) can be extended to solve revenue maximization problem for on-demand resource sharing, where the service providers pay for the amount of resources they reserved for a certain period. We consider two different pricing schemes, fixed-rate pricing scheme and tiered pricing scheme, respectively. For fixed-rate pricing scheme, we assume the price for reserving data rate R_0 is ρ_0. Then, the revenue maximization objective is

$$f_0 = \rho_0 \sum_{u\in\mathcal{U}_s, s\in\mathcal{S}} \frac{R_u}{R_0}. \tag{30}$$

It can be seen that (30) can be obtained by setting $\beta_u = \rho_0/R_0, \forall\ u \in \mathcal{U}_s, s \in \mathcal{S}$ from (9). With the fixed-rate pricing scheme, the network operator only concerns the throughput of the system, which may result in unbalanced resource allocation among the service providers and users, e.g. users with better channel condition obtain more resources, while users with poor channel condition may not be admitted for service.

We address this issue by adopting a tiered pricing scheme. In this scheme, the price ρ_u for reserving data rate R_u for each user is a piece-wise function

$$\rho_u = \begin{cases} 0, & \text{if } R_u < R_u^{\min}, \\ \frac{\rho_1 R_u^{\min}}{R_0} + \frac{\rho_2 (R_u - R_u^{\min})}{R_0}, & \text{if } R_u^{\min} \le R_u \le R_u^{\max}, \\ \frac{\rho_1 R_u^{\min}}{R_0} + \frac{\rho_2 (R_u^{\max} - R_u^{\min})}{R_0}, & \text{otherwise}, \end{cases} \tag{31}$$

where ρ_1, ρ_2 are constant unit prices that satisfy $\rho_1 > \rho_2$. (31) implies that payment is only made when the reserved data rate satisfies the user's QoS requirement. When the data rate is greater than the upper bound $R_u^{\max}$, no more payment is made for additional data rate reserved for this user. With this pricing scheme, the revenue maximization objective becomes

$$f_1 = \sum_{u\in\mathcal{U}_s, s\in\mathcal{S}} \left(\frac{\rho_1 z_u R_u^{\min}}{R_0} + \frac{\rho_2 z_u (R_u - R_u^{\min})}{R_0} \right). \tag{32}$$

Note that we have constraint (5) to restrict the data rate for an admitted user to be within the $[R_u^{\min}, R_u^{\max}]$. With constraint (5), (32) is equivalent to

$$f_1 = \sum_{u\in\mathcal{U}_s, s\in\mathcal{S}} \left(\frac{(\rho_1 - \rho_2) z_u R_u^{\min}}{R_0} + \frac{\rho_2 R_u}{R_0} \right), \tag{33}$$

where we omit the z_u in the second term within the bracket due to (5). Notice that if $z_u = 0$, constraint (5) forces R_u to be zero as well. Therefore, by using (33) as the objective function, we can formulate the revenue maximization problem under tiered pricing scheme.

VI. Performance Evaluation

We evaluate the performance of the proposed dynamic resource sharing mechanism via simulations. We consider three service providers sharing a C-RAN to serve their subscribed users in a residential area. The system consists of 16 cells, where 16 RRHs are placed in a 4×4 grid. The distance between two adjacent RRHs is 30 m. There are 25 channels available, each with a bandwidth of 180 kHz. The wireless channel model follows [27]. The path loss exponent is 4. The RRH's transmission power and noise power are 250 mW and -90 dBm, respectively. Each time slot is 100 ms. Unless specified, we set $\beta_u = 1$, $R_s^{\text{ref}} = 25$ Mbps, $u \in \mathcal{U}_s, s \in \mathcal{S}$, $R_{\text{ref}} = 800$ kbps, $\epsilon = -75$ dBm, $T = 100$, $\Delta t = 20$, $B_j^{\text{th}} = 100$ Mbps, and $D = 1$. We further set $n_p = 5$ as inspired by [2] and $n_L = 10$. The value of n_L is determined by the network operator, where smaller values of n_L can track the changes in the network more quickly with the cost of a higher computational complexity. The results in this section are obtained by averaging the outcome of 50 simulation runs with different user topologies.

We evaluate the efficiency of the proposed IBGA algorithm and compare its performance with that of a standard branch and bound algorithm solved by the MATLAB MILP solver. As the complexity of the branch and bound algorithm increases significantly with respect to the network size, in this evaluation, we consider part of the simulation model, i.e., a 3×3 grid with 9 RRHs and 9 channels. We vary the number of users subscribed to each service provider from 1 to 10. All users are stationary and randomly distributed within the 100×100 m^2 area that covers the 3×3 grid. The QoS requirement of

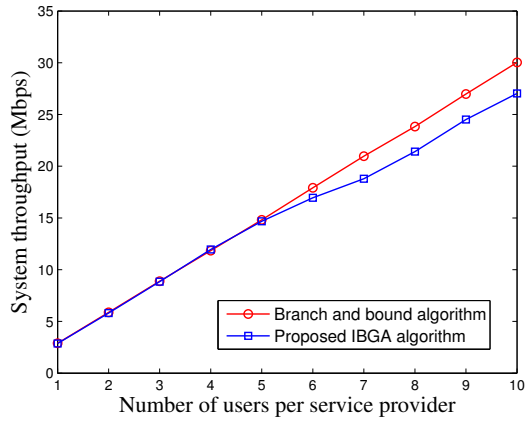


Fig. 3. System throughput versus number of users per service provider.

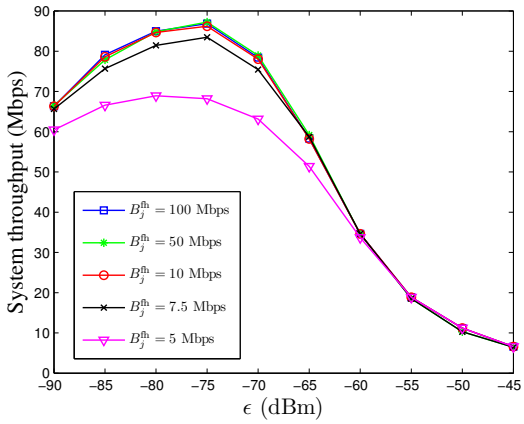


Fig. 5. System throughput versus interference threshold ϵ for different fronthaul link capacity.

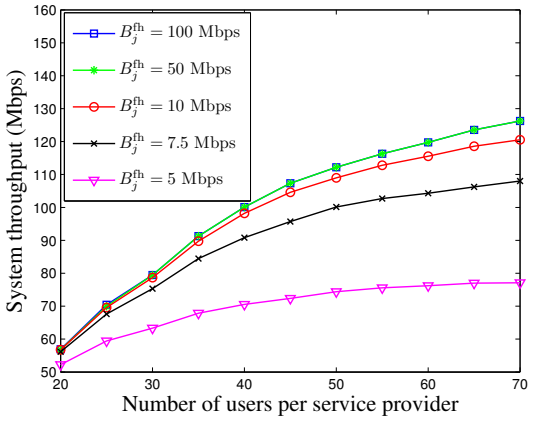


Fig. 4. System throughput with different fronthaul capacity versus number of users per service provider.

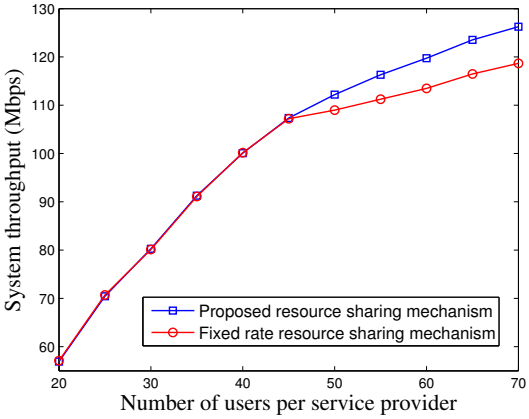


Fig. 6. System throughput of proposed and fixed rate resource sharing mechanisms.

each user is set to be $\{800, 1000\}$ kbps. The average system throughput and the running time for both algorithms are shown in Fig. 3 and Table II, respectively. As shown in Fig. 3, the system throughput of both algorithms are almost the same when the number of users per service provider is less than 5. When the number of users is greater than 5, the proposed IBGA algorithm achieves no less than 90% of the performance achieved by the branch and bound algorithm. However, as shown in Table II, the running time of the proposed IBGA algorithm is significantly lower than that of the MILP solver when the number of users is greater than 7.

We now evaluate the performance of the proposed algorithm for different values of fronthaul capacity B_j^{fh}. The system throughput of the proposed algorithm is shown in Fig. 4, where we vary the number of users per service provider from 20 to 70 and set $R_s^{\text{ref}} = 40$ Mbps. It is shown that the system throughput increases with the number of users as the traffic load increases. However, when the fronthaul capacity is small, the number of users associated with each RRH is limited, which restricts the system throughput. It is observed that the system throughput increases with the fronthaul capacity. When the fronthaul capacity is large enough to satisfy the QoS requirements of all users, the system throughput does not increase further with the fronthaul capacity.

We further evaluate the performance of the proposed mechanism with different values of interference threshold ϵ and fronthaul capacity B_j^{fh}. We fix the number of users for each service provider to be 30. Fig. 5 shows that the system throughput first increases with ϵ and then decreases when ϵ is larger than -75 dBm. The reason is as follows. When ϵ is small, few channels can be reused among different RRHs since the interference constraint cannot be satisfied. Some users cannot be admitted due to the limited amount of resources available. As ϵ increases, channel reuse among RRHs becomes possible and more users can be admitted, which increases the system throughput. However, the transmission rate R_u from a service provider to a user becomes smaller as ϵ increases. Thus, the system throughput decreases when ϵ exceeds a certain value. Similarly, the system throughput increases with the fronthaul capacity when ϵ is smaller than -60 dBm. In practice, ϵ can be selected based on the simulation results under different network settings.

We compare the proposed mechanism (with IGBA algorithm) with a mechanism where each service provider is guaranteed a fixed aggregate data rate of 40 Mbps. We set $R_s^{\text{ref}} = 40$ Mbps. We restrict users subscribed to service provider 3 be located in the boundary area (within 30 m to the area boundary). Fig. 6 shows the system throughput increases as the number of users per service provider increases. It can be seen that both mechanisms achieve similar performance when the number of users per service provider is under 45. This is because in these scenarios, the resources in the system

TABLE II
AVERAGE RUNNING TIME FOR DIFFERENT ALGORITHMS.

Number of users per service provider	1	2	3	4	5	6	7	8	9	10
Running time for IBGA (s)	0.032	0.057	0.108	0.136	0.16	0.226	0.255	0.358	0.404	0.446
Running time for branch and bound (s)	0.073	0.116	0.178	0.221	0.409	0.708	2.753	42.655	865.83	6758.2

are sufficient to satisfy the QoS requirements of almost all users, and the total minimum rate demand for each service provider is no larger than the minimum rate guarantee (40 Mbps). Thus, using fixed or dynamic rate guarantee does not affect the system throughput. However, as the number of users increases further, the proposed mechanism with dynamic rate guarantee achieves higher throughput. The reason is that as the number of users increases, the resources become stringent and may not be sufficient to satisfy users' QoS requirement, especially for those subscribed to service provider 3. The proposed mechanism reduces the minimum resource guarantee for service provider 3 accordingly to save some resources for other users who are close to the RRHs. On the contrary, using fixed guarantee requires much more resources for service provider 3, which results in inefficient utilization and reduces the system throughput.

We compare the proposed mechanism (with IGBA algorithm) with the proportional spectrum sharing mechanism. In the proportional spectrum sharing mechanism, each RRH is allocated one channel, and this channel is shared by three service providers proportionally according to their traffic demand. The traffic demand is estimated by assuming users are associated with their closest RRHs. The allocated resources remain unchanged until the next allocation process is performed. The simulation setting is the same as that in Fig. 6. Fig. 7 shows the system throughput with respect to different number of users per service provider. It can be seen that the proposed resource sharing mechanism achieves higher system throughput than the proportional resource sharing mechanism. This is because the proportional resource sharing mechanism does not explore dynamic channel reuse and user association, which is less efficient than the proposed mechanism.

We let each service provider serve 30 stationary users. Each user has a QoS requirement of $\{800, 1000\}$ kbps at the beginning. Then, every 100 time slots from time slot 100 to time slot 500, we increase the QoS requirement of service provider 1's users by 45 kbps one at a time. Next, from time slot 600 to 1000, we decrease the QoS requirement of service provider 2's users by 45 kbps one at a time every 100 time slots. We compare the proposed multi-timescale mechanism with a single-timescale mechanism. The single-timescale mechanism performs global optimization every 400 time slots, while the multi-timescale mechanism also performs local optimization every 100 time slots. Fig. 8 shows the achievable throughput over each time period with respect to the traffic variation, which is calculated according to the allocated resources. It can be seen that as the resource demand from the users varies, the system throughput of the proposed mechanism changes accordingly every 100 time slots. This is because the proposed

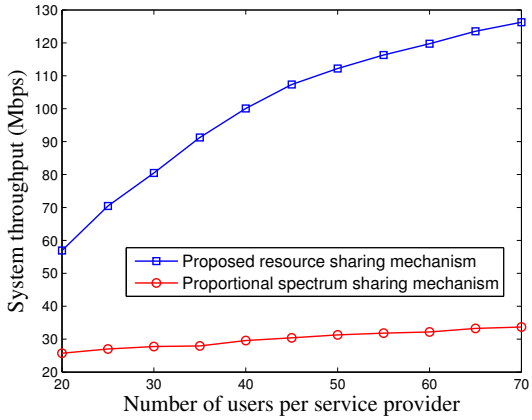


Fig. 7. System throughput of proposed resource sharing and proportional spectrum sharing mechanisms.

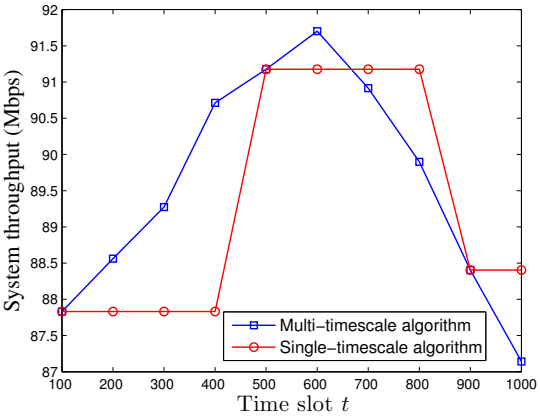


Fig. 8. System throughput of multi-timescale and single-timescale algorithms versus time slot.

mechanism has a local resource allocation procedure which updates the resources allocated to each service provider every $T = 100$ time slots. However, the system throughput of the single-timescale mechanism is updated every 400 time slots due to the global resource allocation, and users' varying resource demand between two global resource allocation processes may not be satisfied. As shown in Fig. 8, the system throughput of the proposed algorithm is lower than that of the single-timescale algorithm between time slots 700 and 1000. This is because the proposed algorithm responds faster to traffic variation and achieves higher system throughput in the first 600 time slots, which reduces the remaining backlog traffic required to be transmitted between time slots 700 and 1000. This demonstrates that the proposed multi-timescale mechanism with local resource update can adapt to frequent traffic variation and can provide on-demand services.

We consider each service provider serves 15 stationary users and 15 mobile users each with QoS requirement $\{600, 800\}$

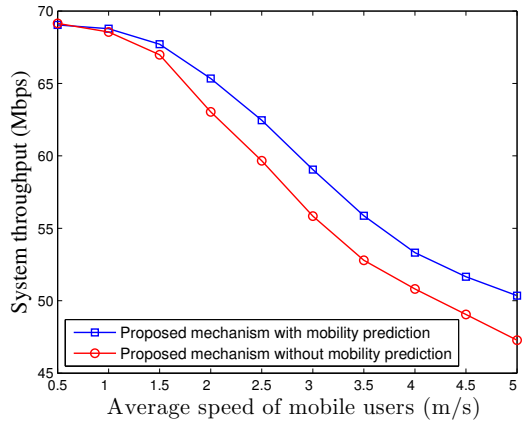


Fig. 9. System throughput versus average speed of mobile users.

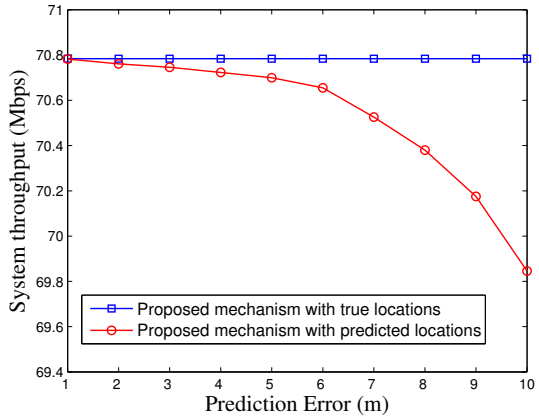


Fig. 10. System throughput versus the prediction error.

kbps. We adopt the 2D Gauss-Markov movement model [28], where the velocity of a mobile user is correlated in time. User u's velocity in each dimension v_u^t at time t is given by $v_u^t = \gamma_v v_u^{t-1} + (1-\gamma_v)\mu_v + \sqrt{1-\gamma_v^2}x^{t-1}$, where $\gamma_v \in [0,1]$ is the velocity memory factor, μ_v is the asymptotic mean of v_u^t and x is an independent and stationary Gaussian random variable with zero mean and standard deviation σ_v. We set $\gamma_v = 0.9$, $\mu_v = 1$ m/s, and $\sigma_v = 1$. The initial speed of each user is selected as 0.25 m/s at a random direction and is updated every four seconds. For location prediction, we implement an order-2 Markov predictor [26], which predicts the next movement in one direction based on the most recent two movements in the same direction. Each movement is measured by a speed change in horizontal direction and vertical direction, respectively, with step size selected from $[-2.5, -2.25, \ldots, 2.25, 2.5]$ m/s. The parameter of the Markov predictor is obtained via simulation for 10^5 time slots. We predict four future positions for each user during the optimization process.

Fig. 9 shows the system throughput with respect to different average speeds of mobile users. The system throughput decreases for both mechanisms with or without mobility prediction. This is because as the users move faster, the amount of resources required at each service provider from different RRHs varies more quickly. However, the amount of resources allocated to each service provider remain unchanged for a certain period, which may not be sufficient to satisfy the QoS requirements of all mobile users. Thus, the system throughput decreases since some mobile users are not admitted for services. Nevertheless, with the mobility prediction, the proposed mechanism achieves around 5% throughput improvement as the users moves faster than 3.5 m/s.

We also evaluate the effect of prediction error on the proposed mechanism, where we manually add prediction error to the actual locations of users during each prediction period. Fig. 10 shows the system throughput with respect to different prediction errors with users' average speed of 2 m/s. It can be seen that as the error increases, the system throughput decreases. However, the decrease is within 1% as long as the prediction error is within 10 m.

VII. Conclusion

In this paper, we proposed a multi-timescale dynamic resource sharing mechanism. The network operator performs a global resource allocation at a relatively large time scale, and performs local resource allocation based on the changes of network status such as traffic variation and user mobility. We have introduced a threshold-based policy to limit the aggregate interference observed at each user and provide isolation among service providers. We have also employed a mobility prediction approach to facilitate the estimation of traffic demand. We have formulated a resource allocation problem that jointly optimizes the channel allocation, user association, and admission control decisions and developed an efficient algorithm to solve it. We have discussed possible extensions of the proposed mechanism for uplink transmission and revenue maximization. Through simulations, we have shown that the proposed mechanism achieves service isolation and efficient resource sharing among service providers. It can adapt to traffic variation, and achieves robust performance under user mobility. In future, we aim to utilize transmission power control to further alleviate the intra-tier interference.

References

[1] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G. K. Chang, "The case for re-configurable backhaul in cloud-RAN based small cell networks," in *Proc. of IEEE INFOCOM*, Turin, Italy, Apr. 2013.

[2] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, "FluidNet: A flexible cloud-based radio access network for small cells," in *Proc. of ACM MobiCom*, Miami, FL, Sept. 2013.

[3] China Mobile, "C-RAN: The road towards green radio access networks," *white paper*, Version 3.0, Dec. 2013.

[4] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

[5] 3GPP TR 36.814, V2.0.0, "Further advancements for E-UTRA physical layer aspects," 2010. [Online]. Available: www.3gpp.org.

[6] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[7] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun. Mag.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[8] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[9] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN," *IEEE Trans. Signal Process.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[10] H. Xiang, Y. Yu, Z. Zhao, Y. Li, and M. Peng, "Tradeoff between energy efficiency and queues delay in heterogeneous cloud radio access networks," in *Proc. of IEEE ICC*, London, UK, Jun. 2015.

[11] W. Wang, V. K. Lau, and M. Peng, "Delay-aware uplink fronthaul allocation in cloud radio access networks," *arXiv preprint arXiv:1502.07966*, 2015.

[12] A. Khan, W. Kellerer, K. Kozu, and M. Yabusaki, "Network sharing in the next mobile network: TCO reduction, management flexibility, and operational independence," *IEEE Wireless Commun. Mag.*, vol. 49, no. 10, pp. 134–142, Oct. 2011.

[13] P. Rost, C. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wbben, "Cloud technologies for flexible 5G radio access networks," *IEEE Wireless Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[14] A. Tzanakaki, M. P. Anastasopoulos, G. Zervas, B. R. Rofoee, R. Nejabati, and D. Simeonidou, "Virtualization of heterogeneous wireless-optical network and IT infrastructures in support of cloud and mobile cloud services," *IEEE Wireless Commun. Mag.*, vol. 51, no. 8, pp. 155–161, Aug. 2013.

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

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

[17] M. Peng, X. Xie, Q. Hu, J. Zhang, and H. V. Poor, "Contract-based interference coordination in heterogeneous cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1140–1153, Jun. 2015.

[18] M. Hong, R. Sun, H. Baligh, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, Feb. 2013.

[19] F. Fu and U. C. Kozat, "Wireless network virtualization as a sequential auction game," in *Proc. of IEEE INFOCOM*, San Diego, CA, Mar. 2010.

[20] S. Zhang, Z. Qian, J. Wu, and S. Lu, "An opportunistic resource sharing and topology-aware mapping framework for virtual networks," in *Proc. of IEEE INFOCOM*, Orlando, FL, Mar. 2012.

[21] Y. T. Hou, Y. Shi, and H. D. Sherali, "Optimal spectrum sharing for multi-hop software defined radio networks," in *Proc. of IEEE INFOCOM*, Anchorage, AK, May 2007.

[22] L. Popa, G. Kumar, M. Chowdhury, A. Krishnamurthy, S. Ratnasamy, and I. Stoica, "FairCloud: Sharing the network in cloud computing," in *Proc. of ACM SIGCOMM*, Helsinki, Finland, Aug. 2012.

[23] J. Guo, F. Liu, X. Huang, J. C. Lui, M. Hu, Q. Gao, and H. Jin, "On efficient bandwidth allocation for traffic variability in datacenters," in *Proc. of IEEE INFOCOM*, Toronto, Canada, Apr. 2014.

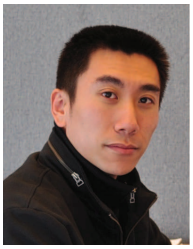
[24] Z. Ding and H. V. Poor, "The use of spatially random base stations in cloud radio access networks," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1138–1141, Nov. 2013.

[25] M. Peng, S. Yan, and H. V. Poor, "Ergodic capacity analysis of remote radio head associations in cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 365–368, Aug. 2014.

[26] C. Cheng, R. Jain, and E. V. D. Berg, *Location Prediction Algorithms for Mobile Wireless Systems*. CRC Press, 2003.

[27] ITU-R Std. Recommendation ITU-R M.1225, "Guidelines for evaluation of radio transmission technologies for IMT-2000," Feb. 1997.

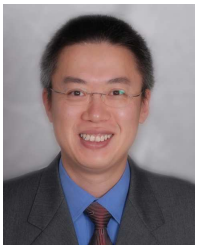
[28] B. Liang and Z. J. Haas, "Predictive distance-based mobility management for multidimensional PCS networks," *IEEE/ACM Trans. on Networking*, vol. 11, no. 5, pp. 718–732, Oct. 2003.



Binglai Niu (S'11) received the B.S. degree in electronics engineering from Fudan University, Shanghai, China, in 2008, the M.Sc. degree in electrical engineering from the University of Alberta, Edmonton, Alberta, Canada, in 2010, and the Ph.D. degree at the University of British Columbia, Vancouver, British Columbia, Canada in 2015. He is currently working at Arista Networks in Vancouver, BC, Canada. His research interests include optimization, game theory, resource allocation and interference management in wireless networks, and incentive mechanism design for cooperative communications.

Yong Zhou (S'13, M'16) received the Ph.D. degree at the Department of Electrical and Computer Engineering, the University of Waterloo, Canada in 2015, and the M.Eng. and B.Sc. degrees from Shandong University, China in 2011 and 2008, respectively. Since 2015, he has been a post-doctoral fellow with the Department of Electrical and Computer Engineering, the University of British Columbia, Canada. His research interests include performance analysis for cooperative communications and resource allocation in 5G networks.

Hamed Shah-Mansouri (S'06-M'14) received the B.Sc., M.Sc., and Ph.D. degrees from Sharif University of Technology, Tehran, Iran, in 2005, 2007, and 2012, respectively all in electrical engineering. He ranked first among the graduate students. From 2012 to 2013, he was with Parman Co., Tehran, Iran. Currently, Dr. Shah-Mansouri is a post-doctoral research and teaching fellow at the University of British Columbia, Vancouver, Canada. His research interests are in the area of stochastic analysis, optimization and game theory and their applications in economics of cellular networks and mobile cloud computing systems. He has served as the publication co-chair for the IEEE Canadian Conference on Electrical and Computer Engineering 2016 and as the technical program committee (TPC) member for several conferences including the IEEE Global Communications Conference (GLOBECOM) 2015 and the IEEE Vehicular Technology Conference (VTC2016-Fall).



Vincent W.S. Wong (S'94, M'00, SM'07, F'16) received the B.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2000. From 2000 to 2001, he worked as a systems engineer at PMC-Sierra Inc. (now Microsemi). He joined the Department of Electrical and Computer Engineering at UBC in 2002 and is currently a Professor. His research areas include protocol design, optimization, and resource management of communication networks, with applications to wireless networks, smart grid, mobile cloud computing, and Internet of Things. Dr. Wong is an Editor of the *IEEE Transactions on Communications*. He is a Guest Editor of *IEEE Journal on Selected Areas in Communications*, special issue on "Emerging Technologies" in 2016. He has served on the editorial boards of *IEEE Transactions on Vehicular Technology* and *Journal of Communications and Networks*. He has served as a Technical Program Co-chair of *IEEE SmartGridComm'14*, as well as a Symposium Co-chair of *IEEE SmartGridComm'13* and *IEEE Globecom'13*. He is the Chair of the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications and the IEEE Vancouver Joint Communications Chapter. He received the 2014 UBC Killam Faculty Research Fellowship.