IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 May 2023, accepted 16 June 2023, date of publication 23 June 2023, date of current version 10 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3288996

RESEARCH ARTICLE

Offline and Real-Time Deadline-Aware Scheduling and Resource Allocation Algorithms Favoring Big Data Transmission Over Cognitive CRANs

MOHAMMAD BIGDELI[1], BAHMAN ABOLHASSANI[1], SHAHROKH FARAHMAND[1], AND CHINTHA TELLAMBURA[2], (Fellow, IEEE)
[1]School of Electrical Engineering, Iran University of Science and Technology (IUST), Tehran 16846-13114, Iran
[2]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada

Corresponding author: Shahrokh Farahmand (shahrokhf@iust.ac.ir)

ABSTRACT Big data is generated from various sources, such as the Internet of things, social media, databases, wearables, smart cars, and so on, and is characterized by five V's: volume, value, variety, velocity, and veracity. Transmitting big data to secondary users (SUs) over a cognitive cloud radio access network (CRAN) offers multiple benefits and critical challenges. To address these limitations, we have designed two deadline-aware, non-preemptive algorithms that maximize the sum of weighted data transferred by the network over admission, time scheduling, spectrum, and remote radio head (RRH) allocation decisions. Each data request can have a different size, target bit error rate (BER), minimum signal-to-noise ratio (SNR) requirement, and deadline, incorporating the simultaneous provision of various types of big data and ordinary data jointly. Furthermore, our formulation considers all five V's of big data. The first algorithm we propose is an offline batch scheduling (OFB) algorithm, which assumes that all data requests are available at the time of optimization. While this sub-optimal algorithm has a lower complexity and can be implemented in larger networks than the global optimum algorithm, it is not practical for real-time applications since it requires collecting all data requests beforehand for joint scheduling. Thus, our second one is a sub-optimal online real-time scheduling (ONR) algorithm that performs admission and resource allocation on-the-fly using predictions of upcoming data requests and future availability of spectrum channels. After deriving these two algorithms, we conduct a thorough performance analysis and derive bounds on their objective values compared to the global optimum. We then demonstrate their effectiveness in achieving higher weighted sums of transferred data and prioritizing SUs with big data requests over existing alternatives through extensive numerical comparisons.

INDEX TERMS Scheduling, resource allocation, user selection, cloud radio access network (CRAN), big data, total transferred data.

I. INTRODUCTION

We are in the big data era. In the past decade, immense amount of new data generated by the proliferation of smart mobile phones, the internet of things, wireless smart meters and cloud computing has led to wireless big data [1], [2]. Data generation rates are neither decreasing nor stable [3], and on the contrary, it is expected that wireless networks face significant growth in wireless big data due to future emerging services such as the internet of everything (IoE) and holographic telepresence. According to the International Telecommunication Union Radio (ITU-R) [4], total mobile data traffic is expected to experience 77-fold growth in ten years such that it increases from 57 exabytes (10^{18} bytes)

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

per month in 2020 to 4394 exabytes per month in 2030. The accuracy of this estimation is confirmed in [5] where it reports that the total mobile traffic has reached 59 exabytes per month at the end of 2020. In fact, Quarter 3 of 2021 has witnessed a data generation rate of 80 exabytes per month. It is predicted that each subscriber will demand and/or generate almost 257.1 gigabytes of data traffic per month by 2030 [4].

Big data refers to large, complex datasets that are difficult to process and analyze using traditional methods. Some of the main categories of big data include:

(i) High-resolution audio and video streaming (ii) Data generated by social networking websites such as Instagram, Facebook, Twitter, and Flickr, (iii) Mobile TV, (iv) Real-time gaming and control, (v) High-speed downloading, (vi) Online remote monitoring.

These categories are expected to continue growing in the coming years, and they are characterized by five main features, often referred to as the ''5 V's of big data'' [6], [7]:

1) *Volume:* Big data sets are typically massive, ranging from hundreds of gigabytes to petabytes in size.
2) *Velocity:* Big data must be transmitted quickly to meet the time-sensitive needs of various applications.
3) *Variety:* Big data comes in many different forms, from structured data in databases to unstructured data in social media feeds.
4) *Value:* Big data has significant priority and usefulness with the potential to create value for businesses and organizations, but it must be properly analyzed and interpreted.
5) *Veracity:* Big data quality can be compromised by errors, inconsistencies, or biases, so it's important to ensure data accuracy and reliability.

A fundamental challenge is to support these big data characteristics in future wireless networks. In fact, they impose tremendous technical burdens on designing efficient networks [8]. Traditional networks are inadequate for dealing with big data. The traditional cellular network, also known as Radio Access Networks (RAN), consists of numerous standalone base stations (BSs). Each BS covers a limited geographical area, and multiple BSs work together to provide seamless network coverage. Each BS is responsible for processing and transmitting its own signal to and from the mobile device, and forwarding data to and from the mobile device to the core network through the backhaul. However, the current RAN architecture has some drawbacks. Each BS has its own cooling system, backhaul transportation, backup battery, and monitoring system, which can be costly to build and maintain. Moreover, due to limited spectral resources, network operators ''reuse'' the frequency among different base stations, which can lead to interference between neighboring cells and affect network performance.

To address the challenges posed by big data, new technologies such as cloud radio access network (CRAN) [9] offer a flexible and promising infrastructure. The CRAN comprises three main components: a centralized pool of baseband processing units (BBUs), distributed remote radio heads (RRHs), and high-bandwidth, low-latency wired or wireless fronthaul links that connect the BBU pool and RRHs. In contrast to traditional base stations, the BBU is separated from its corresponding RRH, providing an efficient structure for cloud-based resource sharing.

The CRAN architecture has several distinct characteristics that set it apart from other cellular architectures. First, it promotes large-scale centralized deployment by enabling many RRHs to connect to a centralized BBU pool. Second, it supports collaborative radio technologies, allowing any BBU to communicate with any other BBU within the BBU pool with high bandwidth and low latency. Finally, it provides real-time virtualization capability, which ensures that resources in the pool can be dynamically allocated to base station software stacks, such as 4G/3G/2G function modules from different vendors, based on network load.

This paper tackles the significant challenges of downlink big data transmission for unlicensed or secondary users (SUs) in cognitive CRANs. The aim is to maximize the total sum weighted transferred data while taking into account the five V characteristics of big data. To achieve this, the paper simultaneously optimizes SU selection, the association of remote radio heads (RRHs) with selected SUs, allocation of temporarily available spectrum, deadline-aware non-preemptive time scheduling, and adaptive modulation to account for time-varying channels between each SU and the connected RRHs. This is a complex and challenging problem, involving a high-dimensional mixed continuous and integer program of highly non-convex nature.

Before summarizing the contributions of this paper, we review current prior art on this topic.

A. RELATED WORKS

Given our design focus, we classify the relevant literature into four categories: Big data transmission, RRH and spectrum allocation, user selection, and time scheduling.

1) BIG DATA TRANSMISSION

Reference [10] focuses on utilizing big data for machine learning applications that require large amounts of data. To reduce the transmission of wasteful data that does not significantly impact the learning algorithm's performance, they combine edge and cloud computing. This approach involves caching selected data content on various RRHs and BBU pools, which is determined based on predictions of the demanded data's content.

In [11], the big data transmission problem in a wireless network is addressed, taking into account link capacity constraints, current loads of links, requested data sizes, and network delay limits. The goal is to optimize service/waiting time and throughput of the network. To achieve this, a new centralized algorithm is designed to carry out routing and scheduling simultaneously.

In multimedia big data wireless services, meeting deterministic constraints on service delay is challenging, especially when bandwidth and transmit power are constrained.

To tackle this issue, reference [12] substitutes the deterministic delay constraint with a statistical one for software-defined radios over 5G networks. They solve the optimization problem over routing, cache placement, and power allocation decisions and demonstrate that three techniques should be jointly utilized. Specifically, (i) network function virtualization is exploited to find optimal data transmission paths, (ii) information-centric network concept derives optimal caching locations for big data, and (iii) software-defined networks (SDN) help allocate resources dynamically.

Overall, these references propose innovative approaches to address the challenges of big data transmission in wireless networks. By utilizing edge and cloud computing, designing centralized algorithms for routing and scheduling, and leveraging techniques such as network function virtualization, information-centric networking, and SDN, these approaches aim to optimize performance and throughput while reducing wasteful data transmission and meeting service delay constraints.

A variety of techniques have been proposed to address the challenges associated with transmitting big data wirelessly. For instance, Terahertz transmission has been suggested in [13] as a way to communicate big data between autonomous vehicles, thereby increasing network capacity due to its tremendous bandwidth. In [14], the authors study multiple parts of a wireless network infrastructure to efficiently transmit geographically distributed big data to data centers, including servers inside a data center, different data centers, backbone, and access networks.

Big data transmission has also been investigated under different wireless network architectures such as CRANs, SDNs, 5G, wireless sensor networks, D2D communication, and 6G integrated space-air-ground networks, as discussed in [2], [15], [16], [17], [18], [19], and [20], respectively. Reference [21] introduces a cooperative cache-based strategy on ground stations to reduce the load on satellite links and their latency. To ensure the confidentiality of big data transmission while sharing tasks between graphic processing units across various ground stations, compression techniques were adopted.

Moreover, transfer control protocol (TCP) with simultaneous data transmission in multiple paths is introduced as a promising transport layer protocol for big data applications in [22]. This approach offers improved reliability and throughput over traditional TCP, making it a suitable candidate for large-scale big-data transmission.

Reference [23] treats video traffic as the dominating real-time big data application, and designs a new scheduling policy for packet transmission such that more users are simultaneously served without degrading current users' experiences. This algorithm offers a guaranteed improvement in the total number of served users. This achievement is a result of the proper assignment of big data requests and the corresponding bandwidth on each server on a small time scale. The problem of deadline-aware bandwidth allocation is investigated in a wired setup by [24], where both an offline batch scheduling algorithm and online dynamic scheduling were derived to ensure acceptance of a maximum number of big data requests. Upon solving the posed optimization problem, admission and scheduling decisions, data rates, and path selection for every admitted request are determined. The allocated bandwidth may be varied in an adaptive fashion at any time during big data transmission. Contrary to [24], we consider a scenario where both big data and non-big data requests arrive simultaneously and we aim to assign a larger priority to big data requests. Furthermore, our model is wireless instead of wired. Finally, we strive to maximize the weighted sum of transferred data instead of the number of served users.

2) RRH AND SPECTRUM ALLOCATION

Reference [25] jointly assigns RRHs and allocates virtual machines (VMs) to minimize the total delay including task execution time on BBU pool and transmission delay to the corresponding RRH cluster over programmable hierarchical CRANs. Energy consumption for CRANs is minimized in [26], [27], [28], and [29] where RRH selection is considered. The BBU pool performs joint RRH selection, RRH-user association, transmit beamforming, and VM allocation in [26] over CRANs with limited fronthaul capacity. A new model of energy usage for the BBU pool is derived by using collected empirical data from a programmable CRAN testbed in [27]. Upon model fixation, power-bandwidth assignment and active VMs selection are carried out. The goal of the power-bandwidth assignment is to meet the quality of service (QoS) for users, while VM assignment is performed to minimize energy usage. Heuristic green energy-aware RRHs selection algorithm is derived for coordinated multi-point (CoMP) communication over CRANs in [28]. In [30], a RRH clustering algorithm is proposed to jointly perform load balancing and maximize coverage range in the CRAN. RRH clusters are formed by mapping as large a number of RRHs as possible with different traffic to each BBU while minimizing the number of active BBUs. Furthermore, the optimal spectrum assignment problem is solved in each cluster by a genetic algorithm to maximize communicated traffic load under overall energy consumption. The RRH-BBU mapping is also studied in [31] and [32]. A traffic anticipation model is leveraged to assign every BBU with certain RRHs in [31]. In addition to this offline approach, a real-time BBU-RRH mapping is also derived to provide load balancing while maintaining QoS upon the arrival of every data request. Joint user association and RRH-BBU mapping subject to QoS constraints are carried out in [32]. Orthogonal frequency-division multiple access (OFDMA) based CRAN is used for downlink data transmission in [33], where the weighted sum rate is maximized in two successive steps. First, RRHs, spectrum, and users are allocated given a fixed transmission power. Then, transmit power is optimized for the given spectrum, RRHs, and users. Reference [34] enhances sum capacity by jointly assigning time-frequency resources and RRHs where RRH cooperation, i.e., CoMP, is assumed over CRAN. Spectrum trading between network and service providers in a virtual CRAN is investigated by [35]. Virtual CRAN is comprised of a set of separate RRH-BBUs but one assumes that the BBUs are integrated into one BBU pool. Full duplex CRANs are looked at by [36] and [37], where RRH selection is carried out.

3) USER SELECTION

The concept of user selection in wireless communications involves selecting the users with the best channel quality at any given time to allocate system resources to those who can best exploit them. This approach leads to improved system capacity and performance. While this concept has been around for a long time, recently machine learning has been deployed to reach it. For example, in [38], power allocation using deep unsupervised learning is performed first, followed by user selection.

In addition to this, several studies have been conducted on user selection in CRANs. One such study [39] focuses on maximizing the weighted sum rate in CRAN by jointly selecting users and their corresponding beamforming vectors. To achieve this, the study finds the maximal independent sets in the user selection graph while optimizing the beamforming vectors for every possible user to multi-antenna RRH assignments. Similarly, user selection has also been performed in conjunction with RRH and spectrum allocation [33]. Additionally, another study [40] performs user selection to minimize network power consumption in full duplex CRANs while meeting QoS requirements.

4) TIME SCHEDULING

Reference [41] performs time and power allocation when users' requests arrive in real-time and must be served within a specific deadline and signal-to-interference plus noise ratio (SINR). This work strives to maximize power efficiency while minimizing per-processor power consumption. Thus, it formulates a maximization problem with a weighted sum of power efficiency and processors power consumption. Optimization parameters are power allocation and processor scheduling in CRANs. A maximum transmission time minimization with constraints on spectrum and power resources and tolerable delay is studied in [42] where VMs are optimally allocated. A real-time BBU and RRH assignment is considered in [43]. The backhaul design problem of the CRAN is formulated in [44] to maximize a weighted sum of energy and spectral efficiency by a joint allocation of power and time slots for RRHs.

B. OUR CONTRIBUTION

Our proposed approach addresses several challenges in wireless big data transmission that have not been jointly investigated in existing literature. Specifically, we optimize jointly over SU selection, RRHs association to selected SUs, allocation of temporarily available spectrum, deadline-aware non-preemptive time scheduling, and adaptive modulation to maximize the weighted sum of transferred data. In addition, we take into account the 5 V features of wireless big data in our optimization problem. To address the *Volume* feature, we include data size in the objective function. To address the *Value* feature, we assign different priorities to each data request. To address the *Velocity*, *Variety*, and *Value* features, we consider different hard deadlines for the completion of data delivery to various users. Finally, to address the *Veracity* and *Variety* features, we use minimum signal-to-noise ratio (SNR) and a target bit error rate. By considering these factors in our optimization problem, we can better allocate resources and improve the efficiency of wireless big data transmission.

Wireless big data transmission requires a significant amount of bandwidth, which makes unlicensed spectrum allocation particularly challenging. The spectrum crunch is caused by both primary user activity and spectrum scarcity. To address this issue, we propose an offline, non-preemptive scheduling algorithm that assumes all requests are collected by the BBU pool and then jointly scheduled. However, real-time user admission and online resource allocation are also needed. Therefore, we leverage predictions of possible upcoming data requests and spectrum availability to make decisions on the fly. For example, by analyzing data request history, we can reserve resources for SUs that have a higher impact on the objective function. In addition, we constrain the maximum number of SUs that an RRH can serve based on the energy supply of each RRH. Finally, we use adaptive modulation to adjust the transmission rate due to the variable nature of RRH-SU links. To summarize, we face several challenges in wireless big data transmission, and we propose remedies for each of these challenges, which are summarized in Table 1.

Before delving into our proposed algorithms, it is important to highlight the main differences between our work and the algorithm proposed in [45]. While our system model and optimization problem remain the same, our contribution lies in the development of new algorithms. Reference [45] presents a dynamic programming algorithm that achieves global optimality but suffers from high complexity, making it suitable only for small networks. Additionally, it is an offline algorithm that requires all upcoming data requests to be collected before scheduling, causing unacceptable delays for real-time applications.

In contrast, we present two new algorithms: an offline sub-optimal algorithm with lower complexity that can handle larger networks than [45], and an online algorithm that can schedule new requests as they arrive without any delay. Our main novelty lies in these new algorithms, rather than the system model. However, there are two minor differences in our system model compared to [45]. Firstly, we consider a frequency-selective fading model, whereas [45] assumes frequency-flat fading across all subcarriers. Secondly, our proposed algorithms are non-pre-emptive, meaning they can stop serving a low-priority user midway to serve a higher-priority user, while [45] uses a pre-emptive algorithm that

TABLE 1. Challenges and our proposed solutions.

Challenges		Our proposed solutions
5*V*s	*Volume*	Place *Volume* of data in objective function Overall weighted transfer data is objective function Customize proposed algorithms to favor big data requests
	Velocity	Place hard deadline for data delivery in constraints Deadline-aware scheduling
	Value	Use weight to determine *value* of data
	Veracity *Variety*	Consider minimum required SNR Consider target bit error rate Consider data type dependent resource allocation
Spectrum sharing for SUs		Non-preemptive scheduling (because frequency spectrum is temporarily available)
Online design		History-aware scheduling (history of spectrum availability and SUs activity)
Different traffic prediction algorithms		Universal model that is compatible with different traffic prediction algorithm
SU selection RRH association Spectrum allocation		Select disjunctive set of SUs, i.e., set of SUs which their corresponding allocated resources are compatible with each other and with available resources
Variable nature of RRH-SU links		Use adaptive modulation to adjust transmission rate
Computational complexity		Two low-complexity algorithms: OFB and ONR

serves each admitted user completely before serving the next.

Regarding the objective function, [45] considers the weighted sum of data transferred divided by the largest service time of admitted users. In this work, we focus solely on the weighted sum of data transferred, as omitting the largest service time from the objective function leads to more favorable, low-complexity, and real-time algorithms. Nevertheless, our objective function still emphasizes serving big data requests in two ways. Firstly, big data requests lead to a large increase in transferred data, which explicitly appears in the objective. Secondly, we can assign higher priority weights to big data requests to ensure they are served first.

Given the points mentioned above, the contributions of our work can be summarized as follows:

- We propose an objective function that prioritizes big data requests while still accommodating ordinary data requests. Our objective function maximizes the weighted sum of transferred data over decisions involving SU selection, SU-RRH associations, channel allocations, and deadline-aware time scheduling, subject to minimum SNR and maximum target bit error rate (BER) constraints. We have incorporated all five characteristics of big data in our optimization problem as follows:
    1) Volume: The objective function directly encourages larger volumes of data.
    2) Velocity: We have incorporated velocity by considering the deadline parameter of each data request.
    3) Variety: We have modeled variety by allowing for different types of data demands with varying BER, SNR requirements, and deadlines.
    4) Value: The priority factor assigned to each user in the objective function captures the value aspect of big data.
    5) Veracity: The priority factor, target BER, and deadline for each data request are used to incorporate veracity.

  To the best of our knowledge, the proposed objective function has not been previously reported in big data literature. Furthermore, the set of parameters we optimize over is novel and not covered in the literature except for [45], whose major differences with current work were elaborated before.
- We present two algorithms to solve the optimization problem in different scenarios. Firstly, assuming all data requests arrive before running the scheduling algorithm, we propose an offline batch method to sub-optimally solve the problem. Secondly, we consider a scenario where real-time decisions are made on admission and resource allocation upon the arrival of every data request. In this case, we propose an online algorithm that takes advantage of probabilistic predictions of upcoming data requests and the availability of channels. Our online algorithm is designed to adapt to any prediction method, regardless of its quality.
- We rigorously analyze the performance of both the batch and real-time algorithms, and derive a bound on their performance compared to the globally optimal solution. We also evaluate the complexities of both algorithms. To further validate our proposed algorithms, we conduct extensive simulations and compare their performance to existing alternatives using various metrics. Our simulation results demonstrate the superior performance of our proposed algorithms over existing alternatives.

C. PAPER ORGANIZATION

The rest of this paper is organized as follows. Section II introduces the system model. Section III poses the optimization problem. Section IV presents the proposed offline

TABLE 2. Definitions of all acronyms used in the paper.

Acronyms	Terms
AMC	Adaptive Modulation and Coding
BBU	Baseband Processing Unit
BER	Bit Error Rate
BS	Base Station
CoMP	Coordinated Multi-Point
CRAN	Cloud Radio Access Network
EDF	Earliest Deadline First
EEF	Earliest Ending time First
IoE	Internet of Everything
ITU-R	International Telecommunication Union Radio
OFB	Offline Batch Scheduling
OFDMA	Orthogonal Frequency-Division Multiple Access
ONR	Online Real-Time
PU	Primary User
QoS	Quality of Service
RAN	Radio Access Network
RRH	Remote Radio Head
SDN	Software-Defined Networks
SINR	Signal-to-Interference plus Noise Ratio
SNR	Signal-to-Noise Ratio
SU	Secondary User
TCP	Transfer Control Protocol
VM	Virtual Machine

batch (OFB) scheduling algorithm, while online real time (ONR) scheduling algorithm is derived in Section V. Section VI carries out the rigorous analysis of our two proposed algorithms in terms of both performance and complexity. Simulation results are illustrated in Section VII and conclusions are drawn in Section VIII.

All the acronyms used in the paper are enlisted in Table 2. Table 3 presents the notations of the following sections.

II. SYSTEM MODEL

Our network is composed of a macro cell, which is overlaid with the cognitive CRAN architecture based on set $\mathcal{R}$ of small cell RRHs. Macro and small cells are deployed to serve licensed primary users (PUs) and unlicensed SUs belonging to set $\mathcal{U}$, respectively. This model uses mutually synchronized time slot structure for PUs and SUs, in which time slot t spans the time interval $[(t-1)\Delta t, t\Delta t)$. The value of Δt generally depends on subcarrier spacing, for example in IEEE 802.11 family, $\Delta t = 9$ μseconds [46], and recommended Δt for 5G is reported in [47]. Time is divided into periods, where each period is comprised of many time slots. It is assumed that data requests at each period are scheduled independently, set aside our online algorithm, and thus the proposed optimization is carried out independently for each period. To serve selected SUs, RRHs are distributed in the service area and connected to the BBU pool via high speed and low latency ideal fronthaul links [48]. The symbol $r^{n,t}$ denotes the maximum number of SUs that RRH $r \in \mathcal{R}$ can serve in time slot t of period n. The BBU pool has perfect knowledge of path loss and shadowing between every SU and all RRHs. However, it has access to statistics of small scale fading only. It should be mentioned that once scheduling is completed, every RRH estimates the channel to its assigned SUs with almost perfect accuracy in order to carry out the needed precoding. However, only statistical knowledge is utilized for our scheduling algorithms.

The available spectrum in the network is divided into S equal channels, each with bandwidth Δf. The channel $s \in \{1, \dots, S\}$ is denoted by Δf_s. The unused channels are available for SUs and are arranged in the spectrum pool [49], [50], where set $\mathcal{S}^{n,t}$ denotes these available channels in time slot t of period n. To ensure tractability of the problem formulation, we utilize an orthogonal multiple access scheme, where only one SU or PU can use a particular frequency band s belonging to $\{1, 2, \dots, S\}$ at each time slot. Every SU u may request a different types of data with different QoS requirements. We model the QoS requirements by target bit error rate, $\text{BER}_u^{\text{tar}}$, minimum satisfactory SNR, γ_u, priority of SU, α_u, and T_u^n as the deadline to receive the whole requested data in period n. We suppose SU u requests data with length $L_u \times L$, where $L_u \in \mathbb{N}$ presents number of data frames, and L is the standard frame size. Frame size is about 1500 bytes for Ethernet II and IEEE 802.3, or 2304 bytes for WLAN, and may be higher for extended versions [46]. This user can start to receive data from time slot t_u^n of period n. Subsequently, we describe every user's QoS demand with a 6-tuple: $\left(\text{BER}_u^{\text{tar}}, L_u \times L, t_u^n, T_u^n, \gamma_u, \alpha_u\right)$.

Let $\mathcal{R}_u^{n,t}$ and $\mathcal{S}_u^{n,t}$ denote allocated RRHs and channels to SU u in time slot t of period n. Also, define $\mathcal{R}_u^n := \underset{t}{\cup} \mathcal{R}_u^{n,t}$ and $\mathcal{S}_u^n = \underset{t}{\cup} \mathcal{S}_u^{n,t}$ as allocated resources to SU u in period n. To ensure fairness in resource allocation, maximum number of channels allocated to each SU is limited to $s^{\max}$. Moreover, the set of time slots that u receives service in period n is denoted by $\mathcal{T}_u^n := \left\{t \mid \mathcal{R}_u^{n,t} \neq \emptyset \wedge \mathcal{S}_u^{n,t} \neq \emptyset\right\}$. $\mathcal{T}_u^n$ may be comprised of several separate time slots due to unavailability of spectrum for unlicensed users in certain time slots.

For simplicity, we assume all RRHs and SUs have a single antenna. Let us denote small-scale fading between RRH r and SU u in frequency band s by $h_{r,u}^s \in \mathbb{C}$. Furthermore, we represent the combined effects of transmit and receive antenna gains as well as path loss and shadowing by $d_{r,u}^s \in \mathbb{R}^+$. The instantaneous SNR in the receiver of SU $u \in \mathcal{U}$ when associated with RRH r and channel s, $\gamma_{r,u}^s$, is given as

$$\gamma_{r,u}^s = \frac{d_{r,u}^s \mid h_{r,u}^s \mid^2 P_{r,u}}{\Gamma \sigma^2 \Delta f_s}. \tag{1}$$

In (1), $P_{r,u}$ is the transmit power of RRH r to SU u, σ^2 denotes the background noise power spectral density, and Γ is the SNR gap which represents the mismatch between theoretical and practical SNR values for achieving a given information rate [51], [52]. Assuming adaptive modulation and coding (AMC) is utilized for each SU and maximum ratio beamforming is performed by the associated RRHs to each SU, the approximated spectral efficiency for user u at frequency s at period n and time slot t, defined simply by

TABLE 3. List of symbols used in the paper.

Symbol	Description
$\mathcal{R}$	Set of small cell RRHs
$\mathcal{U}$	Set of SUs
t	Time slot t spans the time interval $[(t-1)\Delta t, t\Delta t)$
Δt	Time range of each time slot
n	Index of CRAN operating period
r	RRH $r \in \mathcal{R}$
$r^{n,t}$	Maximum number of SUs that r can serve in time slot t of period n
S	Number of channels
Δf	Bandwidth of each channel
s	$s \in \{1, \ldots, S\}$
Δf_s	Cannel s
$\mathcal{S}^{n,t}$	Available channels in time slot t of period n
u	$u \in \mathcal{U}$
$\mathrm{BER}_u^{\mathrm{tar}}$	Required target bit error rate for request of SU u
γ_u	Minimum satisfactory SNR that is required for SU u
α_u	Priority of SU u
T_u^n	Deadline to receive the whole requested data of SU u in period n
$L_u \times L$	Requested data length of SU u, where $L_u \in \mathbb{N}$
L	Standard frame size
t_u^n	SU u can start to receive data from time slot t_u^n of period n
$\mathcal{R}_u^{n,t}$	Allocated RRHs to SU u in time slot t of period n
$\mathcal{S}_u^{n,t}$	Allocated channels to SU u in time slot t of period n
$\mathcal{R}_u^n$	Allocated RRHs to SU u in period n
$\mathcal{S}_u^n$	Allocated channels to SU u in period n
$s^{\max}$	Maximum number of channels that can be allocated to each SU
$\mathcal{T}_u^n$	Set of time slots that u receives service in period n
$h_{r,u}^s$	Small-scale fading factor between RRH r and SU u in frequency band s
$d_{r,u}^s$	Combined effects of transmit and receive antenna gains, path loss, and shadowing between RRH r and SU u in frequency band s
$\gamma_{r,u}^s$	Instantaneous SNR in the receiver of SU $u \in \mathcal{U}$ when associated with RRH r and channel s
$P_{r,u}$	Transmit power of RRH r to SU u
σ^2	Background noise power spectral density
Γ	Mismatch between theoretical and practical SNR values for a given information rate
$k_u^{n,t,s}$	Approximated spectral efficiency for user u at frequency s at period n and time slot t
$\mathbf{h}$	Set of all small-scale fading coefficients $h_{r,u}^s$
$I_X(x)$	Indicator function $I_X(x)$ which returns 1 if $x \in X$ is true, and 0 otherwise
$L_u^{\mathcal{R}_u^{n,t}, \mathcal{S}_u^{n,t}}$	Number of bits communicated to user u at time slot t of period n
U	Disjunctive set of SUs
$\subseteq_D$	Disjunctive subset
U_n^*	Optimal selected SUs in period n
u^*	SU u which is selected by optimal resource scheduling
$\preceq$	Sorting order
U_n^{OFB}	Set of admitted SUs by OFB in period n
t_e	Track of the time slot currently being considered when Algorithm 1 is running
s_u	Subset of of $\{1, \ldots, S\}$ with size $s^{\max}$
t_1' and t_1	Auxiliary variables to store candidate starting time slot for considered SU in Algorithm 1
ζ	$0 \leq \zeta < 1$, replacement factor (used by OFB)
U'	Subset of admitted SUs that may be removed from U_n^{OFB} to provide enough resources to serve another SU
$\mathcal{K}$	Auxiliary set for storing time slots in Algorithm 1
k	Time slot k, which $k \in \mathcal{K}$
U_k'	Candidate SUs for replacement until time slot k
w_k	Contribution of U_k' to objective
$w_{\min}$	$w_{\min} = \min_{k \in \mathcal{K}} w_k$
$U_{\min}'$	Set of SUs corresponding to $w_{\min}$
$\mathcal{U}_n^{\mathrm{temp}}$	Auxiliary set that stores all those SUs that were admitted at least once when OFB is running
$P_n(u)$	Probability for arrival of a request by SU u in period n
$P_n(s)$	Probability for availability of channel s in period n
α	$\alpha \geq 1$, Confidence interval factor for ratio of $P_n(u)$ and $P_{n+1}(u)$

TABLE 3. ***(Continued.)*** **List of symbols used in the paper.**

Symbol	Description
β	$\beta \geq 1$, Confidence interval factor for ratio of $P_n(s)$ and $P_{n+1}(s)$
$\mathcal{U}^{\text{op}}_{n+1}$	Candidate set of SUs for acceptance in period $n+1$ when ONR is running
p	Success probability in Bernoulli experiment
$\mathcal{U}^{\text{sp}}_{n+1}$	Sparse set of SUs in period $n+1$ when ONR is running
U^{ONR}_{n+1}	Set of admitted SUs by ONR algorithm in period $n+1$
ρ	Sub-intervals that requested data length belongs to that
$d_{r,u}$	Distance between RRH r and SU u
a_0	Correction factor that accounts for different RRH and SU antenna heights
η	Availability of channels
$\mathcal{S}_a$	Subset of $\mathcal{S}$ of size $\lvert\mathcal{S}_a\rvert = s^{\max}$
$\mathcal{A}_u$	Event that u is disjunctive with $U_n^{OFB} \setminus \{u\}$

$k_u^{n,t,s}$, is given by

$$k_u^{n,t,s} = \mathbb{E}_{\mathbf{h}}\left[\log_2\left(1 + \frac{1.5}{\ln \frac{0.2}{\mathrm{BER}_u^{\mathrm{tar}}}} \times \sum_{r \in \mathcal{R}_u^{n,t}} \gamma_{r,u}^{s} I_{\mathbb{R}^+}\left(\gamma_{r,u}^{s} - \gamma_u\right)\right)\right], \tag{2}$$

where $\mathbf{h}$ denotes the set of all small-scale fading coefficients $h_{r,u}^{s}$. It should be mentioned that we also use an indicator function $I_X(x)$ which returns 1 if $x \in X$ is true, and 0 otherwise. Subsequently, the number of bits communicated to user u at time slot t of period n is given by

$$L_u^{\mathcal{R}_u^{n,t}, \mathcal{S}_u^{n,t}} = \sum_{s \in \mathcal{S}_u^{n,t}} \Delta t \Delta f_s k_u^{n,t,s} = \sum_{s \in \mathcal{S}_u^{n,t}} \Delta t \Delta f_s \times \mathbb{E}\left[\log_2\left(1 + \frac{1.5 \sum_{r \in \mathcal{R}_u^{n,t}} \gamma_{r,u}^{s} I_{\mathbb{R}^+}\left(\gamma_{r,u}^{s} - \gamma_u\right)}{\ln \frac{0.2}{\mathrm{BER}_u^{\mathrm{tar}}}}\right)\right]. \tag{3}$$

III. PROBLEM FORMULATION

Our optimization problem simultaneously performs SU admission as well as assignment of RRHs, channels, and time slots to admitted SUs so that their QoS demands are satisfied. To rigorously define our optimization problem, we need to provide the concept of a disjunctive set of SUs. In period n, set $U \subseteq_D \mathcal{U}$ is a disjunctive set of SUs if they can be served simultaneously in that single period with the available resources. Thus, any disjunctive set of SUs should satisfy the following constraints for a given set of $\mathcal{R}_u^{n,t}$, $\mathcal{S}_u^{n,t}$, and $\mathcal{T}_u^{n}$:

$$\mathcal{T}_u^{n} \subseteq [t_u^n, T_u^n], \tag{4a}$$

$$\sum_{t \in \mathcal{T}_u^n} L_u^{\mathcal{R}_u^{n,t}, \mathcal{S}_u^{n,t}} \geq L_u L, \tag{4b}$$

$$\mathcal{S}_u^{n,t} \bigcap \mathcal{S}_v^{n,t} = \emptyset, \quad \forall u \neq v \in U, \quad \forall t, \tag{4c}$$

$$\bigcup_{u \in U} \mathcal{S}_u^{n,t} \subseteq \mathcal{S}^{n,t}, \quad \forall t, \tag{4d}$$

$$|\mathcal{S}_u^{n,t}| \leq s^{\max}, \quad \forall t, \tag{4e}$$

$$\sum_{u \in U} I_{\mathcal{R}_u^{n,t}}(r) \leq r^{n,t}, \quad \forall r, t. \tag{4f}$$

Equation (4a) ensures that the service time slots for SU u all fall in the acceptable integer interval given by $[t_u^n, T_u^n]$. Constraint (4b) ensures that the allocated resources to SU u is sufficient for communicating all its requested data bits. Equation (4c) enforces orthogonal frequency allocation among SUs, while (4d) guarantees that only PU's unused spectrum bands are allocated to SUs. Constraint (4e) limits the number of channels allocated to SU u by $s^{\max}$. Finally, (4f) ensures that every RRH does not exceed its service capacity. We utilize the symbol $\subseteq_D$ to denote a disjunctive subset.

Our optimization goal is to find a disjunctive set of SUs and their corresponding resource allocation and schedules such that sum weighted data transfer is maximized in a given period

$$\max_{U \subseteq_D \mathcal{U},\ \{\mathcal{R}_u^n, \mathcal{S}_u^n, \mathcal{T}_u^n\}_{u \in U}} \sum_{u \in U} \alpha_u L_u \,. \tag{5}$$

The QoS for the admitted SUs are guaranteed as we optimize over disjunctive sets only. In period n, the optimal disjunctive set of the selected SUs is denoted by U_n^*, and corresponding optimal allocated resources are shown by $\mathcal{R}_{u^*}^n$, $\mathcal{S}_{u^*}^n$, and $\mathcal{T}_{u^*}^n$ for $u^* \in U_n^*$. When available time/spectrum/RRH resources are sufficient to serve all SUs, the maximum value for the objective function is achieved which equals to $\sum_{u \in \mathcal{U}} \alpha_u L_u$, i.e., $U_n^* = \mathcal{U}$. However, resource scarcity introduces a bottleneck. Thus, only a subset of SUs is usually admitted and served. Given that L_u appears in the objective in (5), big data requests are favored as serving them will lead to larger objective values. Priority coefficients α_us add another degree of flexibility to our optimization. These coefficients allow us to change the priorities of different SUs as necessary. For example, they can be set to favor big data users, or to favor a subset of premium users over others and so on.

The problem investigated in this paper is similar to the one studied in [45], where it was proven to be NP-hard. Although [45] solved the problem to global optimality using dynamic programming, their approach is only suitable for small networks with few resources and SUs. To address this limitation, we propose a low-complexity sub-optimal offline batch (OFB) scheduling algorithm that aims to solve the optimization problem in a greedy manner. However, to prevent a significant loss in performance compared to the global optimum, we also incorporate a substitution mechanism into OFB. This technique allows OFB to replace previously admitted users with low utilities with a user of significantly higher utility [53].

The OFB algorithm assumes that all data requests from the secondary users for a given period (denoted as n) are received during the previous period (denoted as $n-1$). These requests are then processed jointly in a batch mode to determine their admission and scheduling variables. However, this approach can become a bottleneck, particularly when the period length is long. To address this issue, we propose an online real-time (ONR) scheduling algorithm that evaluates and either accepts or rejects new requests as soon as they arrive. Additionally, the required resources are reserved immediately. In this paper, we provide a detailed description of the OFB algorithm in Section IV and introduce the ONR algorithm in Section V. We also conduct a comprehensive performance evaluation of both algorithms in Section VI.

IV. PROPOSED OFFLINE BATCH SCHEDULING (OFB)

Both the OFB and ONR scheduling algorithms are non-preemptive, which means that data transmission to any user can be delayed or interrupted to serve other users. These delays may occur if other users have stricter deadlines, higher priorities for receiving service, or there is a lack of spectrum channels due to PUs' activity.

OFB operates at the Baseband Unit (BBU) pool where all incoming data requests are collected for the next service period. At the end of the current service period, OFB schedules all requests jointly and provides the list of admitted users along with their allocated RRHs, spectrum channels, and time slots for the next service period. By using this approach, OFB can optimize system performance by considering all incoming data requests together and allocating resources accordingly. However, since it is non-preemptive, there is a possibility of some requests being delayed or interrupted, which can result in higher latency for some users.

Scheduling algorithms sort and schedule SUs in some order based on some criterion. The sorting criterion varies greatly for different algorithms. For example, SUs may be sorted based on T_u^n in an ascending order, which gives priority to the SUs with the earliest deadline. This leads to the well-known offline greedy earliest deadline first (EDF) algorithm [54], [55]. In another approach, SUs are sorted based on their achievable data rate per unit resource, which is equivalent to greedily solving a Knapsack problem. Based on our objective function in (5), we use scaled requested data size or $\alpha_u L_u L$ as our sorting criterion. Upon denoting sorting order by $\preceq$, we have $u \preceq u'$ if $\alpha_u L_u \geq \alpha_{u'} L_{u'}$. It means that u has priority over u' and should be scheduled first. We hasten to add that when an algorithm reaches global optimum, as in [45], sorting is unnecessary. For sub-optimal approaches, initialization is critical as it can lead to sub-optimal solutions with significantly different objective values. Thus, sorting ensures that we start the algorithm with a good initialization.

Algorithm 1 The Proposed OFB for Period n.

```
Input: $\forall u \in \mathcal{U} : \big(\mathrm{BER}_u^{\mathrm{tar}}, L_u \times L, t_u^n, T_u^n, \gamma_u, \alpha_u\big)$,
       $\forall t : \mathcal{S}^{n,t}$, $\forall (r,t) : r^{n,t}$, and $\zeta$.
Output: $U_n^{\mathrm{OFB}}$ and corresponding allocated
        resources.
$U_n^{\mathrm{OFB}} \leftarrow \emptyset, \mathcal{U}_n^{\mathrm{temp}} \leftarrow \emptyset$
Sort $\mathcal{U}$ based on $\alpha_u L_u$ in a descending order
$t_e \leftarrow \min_{u \in \mathcal{U}} t_u^n$
while $t_e \leq \max_{u \in \mathcal{U}} T_u^n$ do
    forall $u \in \mathcal{U}$ do
        if $t_u^n \leq t_e \leq T_u^n$ then
            $t_1 \leftarrow -\infty$
            Case 1
                /* Are remaining resources
                sufficient to serve $u$? */
            if $t_1 = -\infty$ then
                Case 2
                    /* Is it possible to
                    replace some of already
                    selected SUs by this
                    present SU $u$ who has
                    faced insufficient
                    resources? */
    $t_e \leftarrow t_e + 1$
return $U_n^{OFB}$, and for $\forall u \in U_n^{OFB} : (\mathcal{R}_u^n, \mathcal{S}_u^n, \mathcal{T}_u^n)$
```

Algorithm 1 summarizes the OFB scheduling algorithm. First, SUs are sorted based on the earliest time slot that they can receive service which is t_u^n. OFB starts from the earliest time slot and iteratively increments time slots. At each time slot, OFB attempts to schedule as many SUs as possible up to the current time slot by considering all unscheduled SUs in the sorted order.

For every unscheduled SU and every time slot, OFB goes through two cases. In case 1, OFB attempts to schedule the SU whose turn has come by utilizing the remaining available resources. If there are enough resources, the SU is scheduled, and OFB moves to the next unscheduled SU. If the remaining resources are not sufficient, case 2 is invoked. In case 2, OFB checks to see if any set of previously admitted users, whose contribution to the objective is considerably lower than the current SU, can be dismissed so that the current SU can be scheduled instead.

Once all SUs have been considered, OFB returns the set of admitted users, denoted by U_n^{OFB}, along with the corresponding resource allocations. OFB is run independently at the beginning of each service period. Overall, the OFB scheduling algorithm aims to optimize the cognitive CRAN's

Case 1 in Algorithm 1

```
8  foreach $s_u \subseteq \{1,\dots,S\}$ with $|s_u| = s^{\max}$ do
9      $t_1' \leftarrow -\infty$
10     $t_1' \leftarrow \max_i \left\{\sum_{t=i}^{t_e} \sum_{s\in s_u} L_u^{s,t} \geq LL_u\right\}$
11     if $t_1' > t_1 \wedge t_1' \geq t_u^n$ then
12         $t_1 \leftarrow t_1'$
13         for $t \leftarrow t_1$ to $t_e$ do
14             $\mathcal{S}_u^{n,t} \leftarrow \mathcal{S}^{n,t} \cap \{\Delta f_s \mid s \in s_u\}$
15             $\mathcal{R}_u^{n,t} \leftarrow$
                 $\left\{r \mid I_{\mathbb{R}^+}\left(\gamma_{r,u}^s - \gamma_u\right) I_{\mathbb{N}}\left(r^{n,t}-1\right) = 1\right\}$
16             $\mathcal{T}_u^n \leftarrow \{t \mid \mathcal{S}_u^{n,t} \neq \emptyset \wedge \mathcal{R}_u^{n,t} \neq \emptyset\}$
17 if $t_1 \neq -\infty$ then
18     $U_n^{\text{OFB}} \leftarrow U_n^{\text{OFB}} \cup \{u\}$
19     $\mathcal{U} \leftarrow \mathcal{U} \setminus \{u\}$
20     $\mathcal{U}_n^{\text{temp}} \leftarrow \mathcal{U}_n^{\text{temp}} \cup \{u\}$
21     $\forall t \notin \mathcal{T}_u^n : \mathcal{S}_u^{n,t} \leftarrow \emptyset, \mathcal{R}_u^{n,t} \leftarrow \emptyset$
22     for $t \leftarrow t_1$ to $t_e$ do
23         $\mathcal{S}^{n,t} \leftarrow \mathcal{S}^{n,t} \setminus \mathcal{S}_u^{n,t}$
24         $\forall t \in \mathcal{R}_u^{n,t}$: $r^{n,t} \leftarrow r^{n,t} - 1$
```

Case 2 in Algorithm 1

```
26 Sort $U_n^{\text{OFB}}$ based on starting time of service in a
     descending order
27 $\mathcal{K} \leftarrow \{+\infty\}$
28 for $u' \in U_n^{OFB}$ do
29     if $\mathcal{T}_{u'}^n \cap [t_u^n, t_e] \neq \emptyset$ then
30         $\mathcal{K} \leftarrow \mathcal{K} \cup \left\{\min_{t\in\mathcal{T}_{u'}^n} t + 1\right\}$
31 $w_{+\infty} \leftarrow 0, U_{+\infty}' = \emptyset, w_{\min} \leftarrow +\infty$
32 $\forall k \in \mathcal{K} \setminus \{+\infty\}$: $w_k \leftarrow +\infty, U_k' \leftarrow \emptyset$
33 for $u' \in U_n^{OFB}$ do
34     $H = \min_{k\in\mathcal{K}}\{k > \max_{t\in\mathcal{T}_{u'}^n} t\}$
35     for $k \in \mathcal{K} \cap \mathcal{T}_{u'}^n$ do
36         $w_k \leftarrow \min\{w_H + \alpha_{u'} L_{u'}, w_k\}$
37         If the first term in the RHS of the above
             relation is selected, $U_k' \leftarrow U_H' \cup \{u'\}$
38 Sort $w_k$s in an ascending order, and apply this order to
     $U'$
39 for $k \in \mathcal{K}$ do
40     Sort $U_k'$ based on starting service times in an
         descending order
41     $\forall t : \mathcal{R}_u^{n,t} \leftarrow \emptyset, \mathcal{S}_u^{n,t} \leftarrow \emptyset, \mathcal{S}_k^{n,t} \leftarrow \mathcal{S}^{n,t}$,
         $r_k^{n,t} \leftarrow r^{n,t}$
42     for $u' \in U_k'$ do
43         $\forall t : \mathcal{S}_k^{n,t} \leftarrow \mathcal{S}_k^{n,t} \cup \mathcal{S}_{u'}^{n,t}$
44         $\forall (r,t) \mid r \in \mathcal{R}_{u'}^{n,t} : r_k^{n,t} \leftarrow r_k^{n,t} + 1$
45     Run Lines 8-16 in Case 1, by considering $r_k^{n,t}$ and
         $\mathcal{S}_k^{n,t}$ as available resources
46     if $t_1 \neq -\infty$ then
47         $w_{\min} = w_k$
48         $U' = U_k'$
49         go to Line 50
50 if $w_{\min} < \zeta \alpha_u L_u L$ then
51     $\mathcal{U} \leftarrow \mathcal{U} \cup U' \setminus \{u\}$
52     $U_n^{\text{OFB}} \leftarrow U_n^{\text{OFB}} \cup \{u\} \setminus U'$
53     $\mathcal{U}_n^{\text{temp}} \leftarrow \mathcal{U}_n^{\text{temp}} \cup \{u\}$
54     $\forall t, \mathcal{S}^{n,t} \leftarrow \mathcal{S}_k^{n,t} \setminus \mathcal{S}_u^{n,t}$
55     $\forall (r,t) | r \in \mathcal{R}_u^{n,t} : r^{n,t} \leftarrow r_k^{n,t} - 1$
```

performance by efficiently allocating resources to all SUs in order to maximize the weighted sum rate of SUs.

We elaborate on OFB algorithm pseudo-code next. OFB first sorts the SUs in a descending order of $\alpha_u L_u$ in line 2. Scheduling is performed iteratively for time slots between $\min_{u\in\mathcal{U}} t_u^n$ and $\max_{u\in\mathcal{U}} T_u^n$. OFB begins with the smallest acceptable time slot $\min_{u\in\mathcal{U}} t_u^n$, checks if it can schedule any new SUs and then increments the time slot until it reaches the largest value $\max_{u\in\mathcal{U}} T_u^n$. The parameter t_e keeps track of the time slot currently being considered. In Line 5, every unscheduled SU is considered in the sorted order. In Line 6, those SUs whose acceptable data communication interval $[t_u^n, T_u^n]$ contains t_e but are not yet scheduled, are considered. For any such SU, Case 1 and Case 2 are performed successively.

In Case 1, each subset s_u of $\{1,\dots,S\}$ with size $s^{\max}$ becomes a spectrum resource candidate for SU u. Considering each s_u sequentially, latest possible starting service time of SU u, referred to as t_1', is evaluated such that t_e will become the service ending time slot. The s_u which yields maximum t_1', i.e., $t_1 = \max_{s_u} t_1' = \max_{s_u} \min_{t\in\mathcal{T}_u^n} t$, will be selected as the allocated channels. It is obvious that the starting service time, i.e., $\min_{t\in\mathcal{T}_u^n} t$, should be greater than or equal to t_u^n. For each $t \in [t_1, t_e]$, we store $\{\Delta f_s \mid s \in s_u\} \cap \mathcal{S}^{n,t}$ in $\mathcal{S}_u^{n,t}$, if there is at least one RRH that meets minimum received SNR requirement for this SU. When $\mathcal{S}_u^{n,t} \neq \emptyset$, each RRH r that meets $\gamma_{r,u}^s > \gamma_u$ and has free capacity to serve SU u, is associated with u, and its identity is stored in $\mathcal{R}_u^{n,t}$. Finally, if $\mathcal{S}_u^{n,t} \neq \emptyset$, t is stored in $\mathcal{T}_u^n$.

If no such combination of s_u and assigned RRHs can be found that can complete serving u before t_e, then Case 2 is executed. We define the replacement factor, $0 \leq \zeta < 1$, where $\zeta = 0$ enforces no substitution, and ζ near 1 increases chance of replacement. Case 2 determines the subset of admitted SUs, with minimum sum of weighted data lengths, denoted by U', which can be removed from U_n^{OFB} in order to provide enough resources to serve SU u. The following optimization problem is solved to determine aforementioned U' if it exists:

$$\min_{U'\subseteq U_n^{\text{OFB}}} \sum_{u'\in U'} \alpha_{u'} L_{u'}, \tag{6a}$$

$$\text{s.t. } \{U_n^{\text{OFB}} \setminus U'\} \cup \{u\} \subseteq_D \mathcal{U}, \tag{6b}$$

$$\mathcal{T}_u^n \subseteq [t_u^n, t_e], \tag{6c}$$

$$\sum_{u' \in U'} \alpha_{u'} L_{u'} < \zeta \alpha_u L_u. \tag{6d}$$

Search space for finding U' can be further limited by implicit constraints. Constraint (6d) means that every U' such that $\exists u' \in U' : \alpha_{u'} L_{u'} > \zeta \alpha_u L_u$ can not be substituted. Moreover, we should exclude U' that $\exists u' \in U' : \mathcal{T}_{u'}^n \cap [t_u^n, t_e] = \emptyset$. Therefore, U' should also satisfy the following:

$$\begin{aligned} \alpha_{u'} L_{u'} &< \zeta \alpha_u L_u, \quad \forall u' \in U', \\ \mathcal{T}_{u'}^n \cap [t_u^n, t_e] &\neq \emptyset, \quad \forall u' \in U'. \end{aligned} \tag{7}$$

The substitution of U' found in Case 2 with the current SU u is performed only if $\sum_{u' \in U'} \alpha_{u'} L_{u'}$ is smaller than $\zeta \alpha_u L_u$. This means that substitution is carried out if the objective function is increased by at least $(1 - \zeta)\alpha_u L_u$. Next, let us elaborate on how Case 2 works. By executing Lines 26-49, OFB finds a "sub-optimal" solution for U' in the following manner. First, for each SU u' such that $\mathcal{T}_{u'}^n \cap [t_u^n, t_e] \neq 0$, time slot $\min_{t \in \mathcal{T}_{u'}^n} t + 1$ is stored in auxiliary set $\mathcal{K}$ for future processing, in Line 30. The stored time slots in $\mathcal{K}$ are sorted in a decreasing order. The dynamic program is performed in Lines 31-38. For each $k \in \mathcal{K}$, we will form candidate SUs for replacement or U_k' to free the time slots $[k, t_e]$. Candidate U_k's are initialized by empty set, and their corresponding contribution to objective in (5), denoted by w_k, is initialized to $+\infty$. Utilizing the dynamic relation in Line 36, U_k' and w_k for all $k \in \mathcal{K}$ are iteratively updated. Once these iterations are completed, U_k's are sorted in ascending order of their weights w_k. Beginning with the smallest weight $w_{\min}$, one checks if dismissing the corresponding set $U_{\min}'$ can free up enough resources to serve u. This is checked in Lines 45-49. If the answer is positive and if ζ times $\alpha_u L_u L$ has a greater value than $w_{\min}$, then the substitution is carried out in Lines 51 to 55. Otherwise, next smallest weight w_k is checked in Lines 38 and 39. The set $\mathcal{U}_n^{\text{temp}}$ is an auxiliary set that stores all those SUs that were admitted at least once when OFB was running. If a SU is deleted from U_n^{OFB} in Case 2, it is not deleted from $\mathcal{U}_n^{\text{temp}}$. This set will be used for performance evaluation in Section VI.

Our proposed OFB is a generalization of the greedy method in [53] for weighted interval selection problem. OFB is sub-optimal from several aspects: (i) It greedily schedules SUs who can be served at the earliest deadline. (ii) Rejections are greedy and permanent at every given time slot t_e. Once rejected for a given t_e, the SU should wait for t_e to be incremented before it gets a second chance of being scheduled. (iii) The dynamic program in Case 2 provides only a sub-optimal solution of (6). Still, it performs satisfactorily in our numerical results compared to existing schemes.

A. A SIMPLE EXAMPLE FOR CASE 2 OF ALGORITHM 1

We consider a simple CRAN where $S = 3$, $s^{\max} = 1$, and $\forall u \in \mathcal{U} : \alpha_u = 1$. To maintain the simplicity of exposition, we assume all three spectrum channels are available in all time slots. Furthermore, we assume RRHs have enough capacity to serve all demanding SUs in every time slot as long as the minimum SNR requirement is satisfied. Since we focus on Case 2 in this example, we assume that the answer to Case 1 was negative meaning that there are not enough channels to serve SU u alongside the already scheduled users. Therefore, Case 2 aims to find a subset of low-utility users which can be dropped in favor of the to be scheduled user u thus increasing the objective value.

We assume $U_n^{\text{OFB}} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, and their corresponding number of requested frames are $L_{u'} = \{1, 2, 2, 1, 2, 3, 5, 3, 3, 2, 4, 3\}$. Fig. 1 shows the allocated time slots and channels for these selected SUs. In addition, the time duration that new user u may be scheduled is plotted as a horizontal dotted line on top of the figure. First, SUs in U_n^{OFB} are sorted in descending order of their starting service times, i.e., $\min_{t \in \mathcal{T}_{u'}^n,} t$ for all $u' \in U^{\text{OFB}}$. According to Fig. 1, SUs are sorted as $5, 1, 9, 6, 2, 10, 11, 3, 7, 4, 8, 12$. Then, we determine $\mathcal{K}$ in Lines 27-30 of Algorithm 1, and show this set of time slots in Fig. 1 by vertical dashed lines, and label them by $k_1, k_2, \cdots, k_{12}$. Initialization in Lines 31 and 32 of Case 2 is carried out as $w_{+\infty} \leftarrow 0$, $w_{k_1} \leftarrow +\infty, \cdots, w_{k_{12}} \leftarrow +\infty$, $U_{+\infty}' \leftarrow \emptyset$, $U_{k_1}' \leftarrow \emptyset, \cdots, U_{k_{12}}' \leftarrow \emptyset$. In the following paragraphs, we execute Lines 33-37 of Case 2 for each $u' \in U_n^{\text{OFB}}$, and calculate w_k and U_k' for each $k \in \mathcal{K} \cap \mathcal{T}_{u'}^n$.

For $u' = 5$, we determine that $H = +\infty$. Hence, we have $w_{k_1} = \min\{w_H + L_5, w_{k_1}\} = \min\{0 + 2, +\infty\} = 2$, and $U_{k_1}' = \{1\}$.

For $u' = 1$, we determine that $H = +\infty$. Hence, we have $w_{k_1} = \min\{w_{+\infty} + L_1, w_{k_1}\} = \min\{0 + 1, 2\} = 1$, $U_{k_1}' = \{1\}$, $w_{k_2} = \min\{w_{+\infty} + L_1, w_{k_2}\} = \min\{0 + 1, +\infty\} = 1$, and $U_{k_2}' = \{1\}$.

For $u' = 9$, we determine $H = +\infty$. Hence, we have $w_{k_1} = \min\{w_{+\infty} + L_9, w_{k_1}\} = \min\{0 + 3, 1\} = 1$, $U_{k_1}' = \{1\}$; $w_{k_2} = \min\{w_{+\infty} + L_9, w_{k_2}\} = \min\{0 + 3, 1\} = 1$, $U_{k_2}' = \{1\}$, $w_{k_3} = \min\{w_{+\infty} + L_9, w_{k_3}\} = \min\{0 + 3, +\infty\} = 1$, and $U_{k_3}' = \{9\}$.

For $u' = 6$, we determine $H = k_1$. Hence, we have $w_{k_2} = \min\{w_{k_1} + L_6, w_{k_2}\} = \min\{1 + 3, 1\} = 1$, $U_{k_2}' = \{1\}$, $w_{k_3} = \min\{w_{k_1} + L_6, w_{k_3}\} = \min\{1 + 3, 3\} = 3$, $U_{k_3}' = \{9\}$, $w_{k_4} = \min\{w_{k_1} + L_6, w_{k_4}\} = \min\{1 + 3, +\infty\} = 4$, and $U_{k_4}' = \{1, 6\}$.

For $u' = 2$, we determine $H = k_2$. Hence, we have $w_{k_3} = \min\{w_{k_2} + L_2, w_{k_3}\} = \min\{1 + 2, 3\} = 3$, $U_{k_3}' = \{9\}$, $w_{k_4} = \min\{w_{k_2} + L_2, w_{k_4}\} = \min\{1 + 2, 4\} = 3$, $U_{k_4}' = \{1, 2\}$, $w_{k_5} = \min\{w_{k_2} + L_2, w_{k_5}\} = \min\{1 + 2, +\infty\} = 3$, and $U_{k_5}' = \{1, 2\}$.

For $u' = 10$, we determine $H = k_5$. Hence, we have $w_{k_6} = \min\{w_{k_5} + L_{10}, w_{k_6}\} = \min\{3 + 2, +\infty\} = 5$, and $U_{k_6}' = \{1, 2, 10\}$.

For $u' = 11$, we determine $H = k_6$. Hence, we have $w_{k_7} = \min\{5 + 4, +\infty\} = 9$, and $U_{k_7}' = \{1, 2, 10, 11\}$.

For $u' = 3$, we determine $H = k_6$. Hence, we have $w_{k_7} = \min\{5 + 2, 9\} = 7$, $U_{k_7}' = \{1, 2, 10, 3\}$, $w_{k_8} = \min\{5 + 2, +\infty\} = 7$, and $U_{k_8}' = \{1, 2, 10, 3\}$.

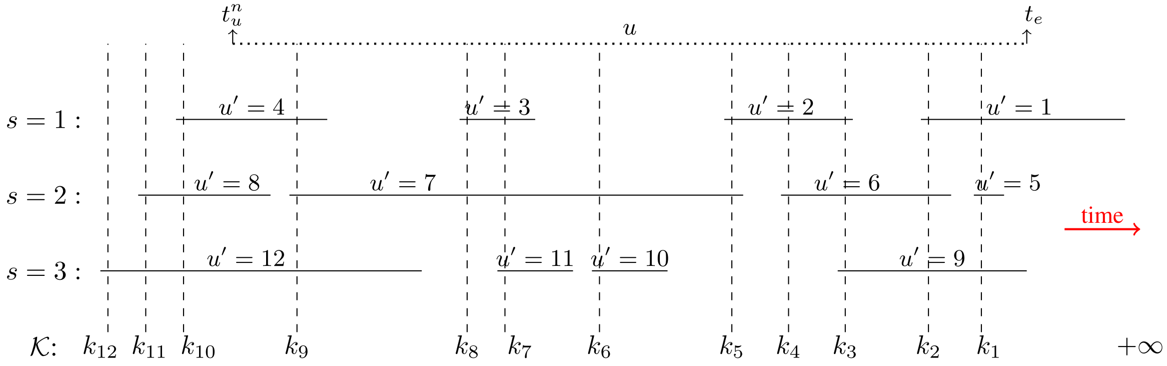


FIGURE 1. Simple example for Case 2 of OFB.

For $u'=7$, we determine $H=k_4$. Hence, we have $w_{k_5}=\min\{3+5,3\}=3$, $U'_{k_5}=\{1,2\}$, $w_{k_6}=\min\{3+5,5\}=5$, $U'_{k_6}=\{1,2,10\}$, $w_{k_7}=\min\{3+5,7\}=7$, $U'_{k_7}=\{1,2,10,3\}$, $w_{k_8}=\min\{3+5,7\}=7$, $U'_{k_8}=\{1,2,10,3\}$, $w_{k_9}=\min\{3+5,+\infty\}=8$, and $U'_{k_9}=\{1,2,7\}$.

For $u'=4$, we determine $H=k_9$. Hence, we have $w_{k_{10}}=\min\{8+1,+\infty\}=9$, and $U'_{k_{10}}=\{1,2,7,4\}$.

For $u'=8$, we determine $H=k_{10}$. Hence, we have $w_{k_{11}}=\min\{9+3,+\infty\}=12$, and $U'_{k_{11}}=\{1,2,7,4,8\}$.

Finally, for $u'=12$, we determine $H=k_8$. Hence, we have $w_{k_9}=\min\{7+3,8\}=8$, $U'_{k_9}=\{1,2,10,3\}$, $w_{k_{10}}=\min\{7+3,9\}=9$, $U'_{k_{10}}=\{1,2,7,4\}$; $w_{k_{11}}=\min\{7+3,12\}=10$, $U'_{k_{11}}=\{1,2,10,3,12\}$; $w_{k_{12}}=\min\{7+3,+\infty\}=10$, and $U'_{k_{12}}=\{1,2,10,3,12\}$.

Subsequently, $\{w_{k_1},\cdots,w_{k_{12}}\}=\{1,1,3,3,3,5,7,7,8,9,10,10\}$, and $\{U'_{k_1},\cdots,U'_{k_{12}}\}=\{\{1\},\{1\},\{9\},\{1,2\},\{1,2\},\{1,2,10\},\{1,2,10,3\},\{1,2,10,3\},\{1,2,10,3\},\{1,2,7,4\},\{1,2,10,3,12\},\{1,2,10,3,12\}\}$.

Finally, Lines 39-49 are executed. By starting from w_{k_1} as the minimum sum weight of dropped users, we check to see if the released resources of users in U'_{k_1} is enough to serve u. If the answer is positive, we store w_{k_1} in $w_{\min}$ and U'_{k_1} in U' as candidate SUs for substitution. If the answer is negative, w_{k_2} and U'_{k_2} are considered next. If the answer is still negative, we repeat this question for w_{k_3} and U'_{k_3} and so on. As soon as the answer for this question becomes positive, we set $w_{\min}$ and U', and go to Line 50 of Algorithm 1. In this Line, we check if $w_{\min}$ is lower than $\zeta\alpha_u L_u L$. If answer is positive, then u is added to U_n^{OFB}, and U' is removed from U_n^{OFB}. This means that the users in U' are not admitted while user u will be accepted in their place.

V. PROPOSED ONLINE REAL-TIME SCHEDULING (ONR)

The OFB scheduling algorithm assumes that all data requests for time period $n+1$ arrive at period n. Hence, in the worst case, a user should wait for one period before it is either scheduled or rejected. If the time periods are large, this waiting time is unacceptable for most real-time applications. Thus, we consider the same optimization problem as in (5) but assume that any arriving request in period $n+1$ should be either scheduled in period $n+1$ or immediately rejected. Unlike the OFB, the ONR assumes that the BBU pool has no prior knowledge of which SUs will request to be served in period $n+1$. Furthermore, the BBU pool has no knowledge about availability of channels in period $n+1$. As soon as the ONR receives a data request with the 6-tuple description $\left(\text{BER}_u^{\text{tar}}, L_u\times L, t_u^{n+1}, T_u^{n+1}, \gamma_u, \alpha_u\right)$ in period $n+1$, it executes a real-time admission control to check if sufficient resources are available to admit this request. If the request is accepted, the ONR allocates the corresponding resources immediately. Lack of prior knowledge on the number and specification of upcoming data requests degrades performance of the ONR compared to that of OFB. To reduce the degradation, statistics of SUs' activities and channels availability, will be exploited in the ONR as we will describe next.

A. SU's ACTIVITIES AND SPECTRUM AVAILABILITY PREDICTION MODEL

To enhance ONR performance, the algorithm employs statistical information of both SUs' request arrivals and channels availability in previous periods. Let $P_n(u)$ and $P_n(s)$ denote the probabilities for arrival of a request by SU u and availability of channel s in period n, respectively. We assume that these probabilities are independent across SUs and channels. In our model, confidence intervals are considered for the ratio of each of these probabilities over two consecutive periods. We assume

$$\frac{1}{\sqrt{\alpha}}\leq\frac{P_{n+1}(u)}{P_n(u)}\leq\sqrt{\alpha}, \tag{8}$$

where $\alpha\geq 1$ is the confidence interval factor. When α is close to one, we have achieved a very good prediction of SU's request arrival probability for the next period. When α gets large, our prediction has a very low accuracy and request arrival probability ranges from near zero to close to one. Nevertheless, our proposed ONR can work with any general prediction algorithm as long as a bound like (8) can be obtained with a specific known α. Similarly, for availability of channel s, we have

$$\frac{1}{\sqrt{\beta}}\leq\frac{P_{n+1}(s)}{P_n(s)}\leq\sqrt{\beta}, \tag{9}$$

where $\beta \geq 1$ is the confidence interval factor for channel availability. Equation (9) accepts the same properties as (8). A similar confidence interval model has been used in [23]. However, the bounds are with respect to the expected values instead of probabilities. In the literature, different models have been investigated for wireless traffic prediction [56]. Yet, our proposed algorithm can work with any general traffic anticipation approach. The inaccuracy of the predictions can be well modeled by α and β. Here, we assume time invariant (or fixed) uncertainty factors for all periods.

ONR is presented as Algorithm 2 and it works as follows. As soon as a SU u''s data request arrives in $t_{u'}^{n+1}$, ONR first runs the OFB on the set of selected SUs in the previous, i.e. nth, period on $U_n^{\text{OFB}} \cup \{u'\}$, by assuming the same availability of channels and RRHs as in the n'th period. In fact, ONR first checks to see that if this request had arrived in the previous period, it would be admitted or not. If the answer is positive, we assign u' to $\mathcal{U}_{n+1}^{\text{op}}$ as a candidate for acceptance in period $n+1$. This decision is made based on the available resources in the nth period, so it will incur a performance degradation in period $n+1$. It is possible that given the already admitted requests in period $n+1$, $\mathcal{U}_{n+1}^{\text{op}}$ is not a disjunctive subset of $\mathcal{U}$. Thus, we perform two more purging steps. If the new data request passes these two steps successfully, it will be admitted and scheduled. First, we run a Bernoulli experiment with success probability p that we will optimally tune later. If the Bernoulli experiment is a success we will keep the new request as a possible scheduling candidate, otherwise the request is rejected. Finally, we run the OFB on the set of already accepted requests U_{n+1}^{ONR} plus the u' given by $U_{n+1}^{\text{ONR}} \cup \{u'\}$ with resources in period n. If the new request is selected by OFB, we will admit the new request. Then, we drop all SUs that belonged to previous U_{n+1}^{ONR} but are no longer in the new U_{n+1}^{ONR}. Resources for this new U_{n+1}^{ONR} are allocated by OFB. The complexity and performance of both OFB and ONR are rigorously derived in the next section.

Algorithm 2 Proposed ONR for Period $n+1$

```
   Input: $U_n^{\text{OFB}}$, and correspondig resource allocation,
          $\alpha, \beta, s^{\max}$
   Output: $U_{n+1}^{\text{ONR}}$, and correspondig resource allocation
1  $U'_{n+1} \leftarrow \emptyset, \mathcal{U}_{n+1}^{\text{op}} \leftarrow \emptyset, \mathcal{U}_{n+1}^{\text{sp}} \leftarrow \emptyset, U_{n+1}^{\text{ONR}} \leftarrow \emptyset$
2  for arriving SU $u'$ do
3      Run Algorithm 1 with inputs: $U_n^{\text{OFB}} \cup \{u'\}$,
        $\forall u \in U_n^{\text{OFB}}$: $t_u^n, T_u^n, L_u, \gamma_u, \forall t : \mathcal{S}^{n,t}, r^{n,t}$ (for $u'$
        use $t_{u'}^{n+1}$ and $T_{u'}^{n+1}$) and store selected SUs in
        $U'_{n+1}$
4      if $u' \in U'_{n+1}$ then
5          $\mathcal{U}_{n+1}^{\text{op}} \leftarrow \mathcal{U}_{n+1}^{\text{op}} \cup \{u'\}$
6          The Bernoulli experiment with success
            probability of $p$ is done
7          if above experiment is successful then
8              $\mathcal{U}_{n+1}^{\text{sp}} \leftarrow \mathcal{U}_{n+1}^{\text{sp}} \cup \{u'\}$
9              Run Algorithm 1 with inputs:
                $\forall u \in U_{n+1}^{\text{ONR}} \cup \{u'\}$: $t_u^{n+1}$ and $T_u^{n+1}, L_u, \gamma_u$,
                $\forall t$: $\mathcal{S}^{n,t}$ and store selected SUs in $U'_{n+1}$
10             if $u' \in U'_{n+1}$ then
11                 $U_{n+1}^{\text{ONR}} \leftarrow U'_{n+1}$
12                 $\mathcal{R}_{u'}^{n+1,t}, \mathcal{S}_{u'}^{n+1,t}$, and $\mathcal{T}_{u'}^{n+1}$ are derived
                    by executing Line 9.
13 return $U_{n+1}^{ONR}$ and for
    $\forall u \in U_{n+1}^{ONR} : (\mathcal{R}_u^{n+1}, \mathcal{S}_u^{n+1}, \mathcal{T}_u^{n+1})$
```

VI. PERFORMANCE AND COMPLEXITY ANALYSIS FOR OFB AND ONR

Here, we rigorously evaluate OFB and ONR performance, where we derive bounds on how far the objective function of these algorithms are from the global optimum given by $\sum_{u \in U_n^*} \alpha_u L_u$. Here, U_n^* denotes the set of admitted users at the global optimum of the n-th period. The following theorem summarizes our results on OFB performance.

Theorem 1: The proposed OFB algorithm is guaranteed to achieve an objective value bounded below by 0.17 times the global optimum of (5), that is

$$\sum_{u \in U_n^{\text{OFB}}} \alpha_u L_u \geq 0.17 \sum_{u \in U_n^*} \alpha_u L_u. \tag{10}$$

where the optimum value for ζ is given by $-1+\sqrt{2} \approx 0.414$.

Proof: Please see Appendix A. ■

Performance analysis for ONR is summarized in the following theorem.

Theorem 2: The proposed ONR algorithm is guaranteed to achieve an expected objective value lower bounded by

$$\mathbb{E}\left(\sum_{u \in U_{n+1}^{\text{ONR}}} \alpha_u L_u\right) \geq \frac{71 - 17\sqrt{17}}{4\left(4\alpha\beta s^{\max}\right)^{\frac{3}{2}}} \mathbb{E}\left(\sum_{u \in U_{n+1}^*} \alpha_u L_u\right). \tag{11}$$

The optimal values for p and ζ are given by $\frac{7-\sqrt{17}}{8\sqrt{\alpha\beta s^{\max}}} \approx \frac{0.36}{\sqrt{\alpha\beta s^{\max}}}$, and $\frac{\sqrt{17}-3}{4} \approx 0.28$, respectively.

Proof: Please see Appendix B. ■

It needs to be mentioned that the bound in Theorem 2 is derived assuming $p < 1$. Thus, the Bernoulli experiment has a nonzero probability of rejecting a particular user. If $p = 1$, the bound in Theorem 2 becomes trivial as it will amount to left hand side of (11) to be greater than some negative value which is obvious; Please check Appendix B to verify this. If a stronger bound is derived then we can also allow for $p = 1$. To summarize, the Bernoulli experiment is not a fundamental block of our proposed algorithm. It only allows us to derive a non-trivial bound on ONR performance.

A. COMPLEXITY ANALYSIS

OFB's complexity is given by $\mathcal{O}\left(|\,\mathcal{U}\,| \log_2(|\,\mathcal{U}\,|) + \max T_u^n \times |\,\mathcal{U}\,| \times (A_1 + A_2)\right)$, where A_1 and A_2 are computational complexity of Case 1 and Case 2, respectively. A_1 is given by $A_1 = \binom{S}{s^{\max}} \times \max\left(T_u^n - t_u^n\right) \times \mid \mathcal{R} \mid$. In Case 2, complexity of **for** in Line 28, and also Line 32 is on the order of $\mid \mathcal{U} \mid$. Moreover, complexity of the sort instruction in Lines 38 and 40 is in order of $\mid \mathcal{U} \mid \log_2 (\mid \mathcal{U} \mid)$. Furthermore, complexities of Lines 33-37, and Lines 39-49 are $\mathcal{O}\left(\mid \mathcal{U} \mid^2\right)$, and $\mathcal{O}\left(\mid \mathcal{U} \mid^2 \times \max T_u^n + \mid \mathcal{U} \mid^2 \log_2 (\mid \mathcal{U} \mid) + \mid \mathcal{U} \mid \times A_1\right)$, respectively. To sum up, overall complexity of OFB is on the order of $\mathcal{O}\left(\max T_u^n \times \mid \mathcal{U} \mid^2 \times \max\{\mid \mathcal{U} \mid \max T_u^n, \mid \mathcal{U} \mid \log_2 \mid \mathcal{U} \mid, A_1\}\right)$.

ONR runs OFB in Lines 3 and 9 for $\mid \mathcal{U} \mid$ times. Thereby, ONR's complexity is on the order of $\mathcal{O}\left(\max T_u^n \times \mid \mathcal{U} \mid^3 \times \max\{\mid \mathcal{U} \mid \max T_u^n, \mid \mathcal{U} \mid \log_2 (\mid \mathcal{U} \mid), A_1\}\right)$.

To find the global optimum of the optimization problem in (5), one should resort to exhaustive search. The number of possible resource allocations for SU u is given by $N_u = \left(\sum_{s=1}^{s^{\max}} \binom{S}{s}\right) \times \left(\sum_{r=1}^{|\mathcal{R}|} \binom{|\mathcal{R}|}{r}\right) \times \left(T_u^n - t_u^n\right)$. So, the computational complexity of an exhaustive search is $\Pi_{u \in \mathcal{U}} N_u$.

VII. NUMERICAL RESULTS

Our proposed OFB and ONR algorithms outperform existing alternatives in the literature, and we present a comprehensive numerical analysis to support this claim. We perform our analysis using two different setups. Firstly, we generate a sample CRAN to demonstrate the significant differences between OFB and the currently available alternatives. Secondly, we conduct Monte Carlo simulations to evaluate the average performance of OFB and ONR. These simulations enable us to assess their performance under various scenarios and network conditions.

We investigate the ratio of transferred data over total data requests, the percentage of scheduled SUs, and the percentage of allocated channels and assigned RRHs as performance metrics in both setups. Moreover, our main focus is on the impact of the proposed algorithms on big data users. To address this, we select a suitable value for L_u in the range of $\left[2^5, 2^{20}\right]$ for each SU u. We partition this range into 5 equal sub-intervals, with each sub-interval represented by a distinct value of ρ. Specifically, we obtain ρ by dividing the rightmost point in each sub-interval by 2^{20}, and the resulting values of ρ are 0.2, 0.4, 0.6, 0.8, and 1, respectively. Notably, larger values of ρ indicate higher data demands for the SUs in that sub-interval, and the sub-interval with $\rho = 1$ corresponds to the big data SUs.

Simulation setups are determined next. We consider the service area of the CRAN to be within a 2000×2000 m^2 area with multiple RRHs serving the SUs and a single RRH serving the PUs. The RRHs and SUs are uniformly and independently distributed within this square area. These SUs are assumed to be either static or have low mobility. Simulation parameters are summarized in Table 4. The capacity of backhaul and fronthaul links are assumed to be sufficiently large to support all data flow in the CRAN with negligible delay. Upon assuming an urban environment, the RRH-SU channel coefficients follow the path loss model $PL[\text{dB}] = 30.58 + 36.7 \log_{10} d_{r,u} - a_0$ where $d_{r,u} > 1.135$ m is the distance between RRH r and SU u. Log-normal shadowing with 8 dB variance is considered [58]. Parameter a_0 is a correction factor that accounts for different RRH and SU antenna heights. The total bandwidth is 20 MHz, which is divided into $S = 100$ channels having equal bandwidth of 200 KHz each. The utilization rate of each channel by PUs varies from 40 to 60 percent [59]. The duration of channel occupation by PU is modeled by an exponential random variable with mean dwell time $10^3 \times \Delta t$. Finally, we consider $\text{BER}_u^{\text{tar}}$ to be from the set $\left\{10^{-3}, 10^{-5}, 10^{-6}\right\}$, and γ_u from the set {0, 3, 5} [dB] for requests with audio, video, and text, respectively [60].

TABLE 4. Simulation parameters and their values.

Parameter	Value
Δt	10 μs [47]
Δf	200 KHz [57]
L_u	$\left[2^5, 2^{20}\right]$ frames
$P_{r,u}$	33 dBm
σ^2	-168 dBm/Hz
CSI	Full
Traffic Model	Full Buffer Node
SNR Gap Γ	0 dB

The global optimum of (5) can be determined through efficient exhaustive search methods like branch and bound. However, these approaches become impractical for medium to large network sizes. Therefore, we compare our proposed methods against existing suboptimal alternatives. Specifically, we evaluate two scheduling algorithms, earliest deadline first (EDF) [55], [61], [62], [63], [64] and earliest ending time first (EEF). EDF has been proven to achieve a total number of admitted SUs at least half of the global optimum [65]. Thus, we compare against three different algorithms: EDF, EDF_ζ, and EEF. EDF_ζ is an algorithm that allows for some users to be dropped in favor of users who increase the objective by at least a $1 - \zeta$ value.

Furthermore, we compare our proposed ONR against ONR/EDF_ζ, ONR/EDF, and ONR/EEF. It should be noted that ONR utilizes successive applications of OFB to determine the admitted users and their resource allocations. Hence, ONR/EDF, for instance, represents the online algorithm that utilizes successive EDF runs instead of OFB runs as the primary building block.

A. ONE CRAN REALIZATION

An instance realization of the coverage area is shown in Fig. 2a for a CRAN with $\mid \mathcal{R} \mid = 20$ RRHs of small cells and $\mid \mathcal{U} \mid = 15$ SUs. Here, we assume there are only $S = 5$ channels with $\Delta f = 1$ MHz. Parameters t_u^n and T_u^n for these SUs are shown in Fig. 2b when they make requests with lengths that are shown in Fig. 2c. Every RRH has enough capacity to simultaneously support all SUs, i.e., $\forall r, t : r^{n,t} \geq 15$.

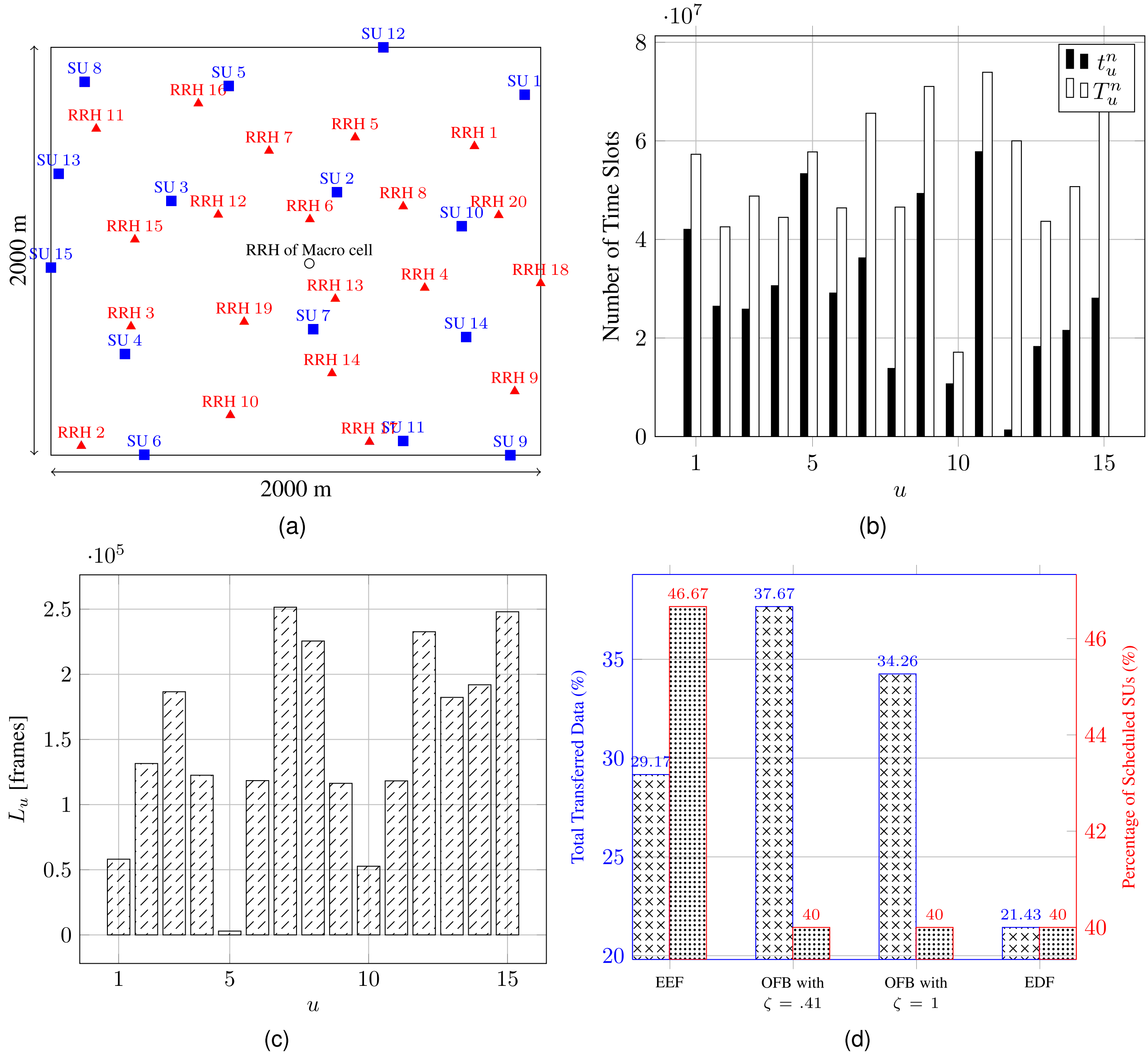
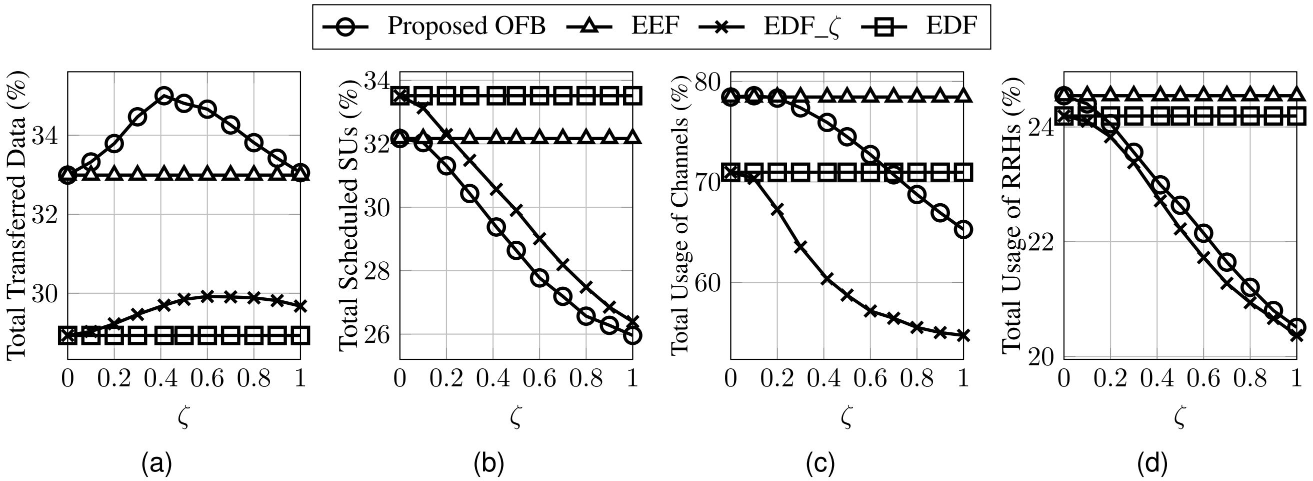


FIGURE 2. (a) Instance of the considered CRAN with 20 small cell RRHs which are labeled by RRH r for $r \in \{1, 2, \cdots, 20\}$, and 15 SUs which are labeled by SU u when $u \in \{1, 2, \cdots, 15\}$, and a single macro cell RRH that is fixed in center of the service area, (b) Values of t_u^n and T_u^n with respect to u, (c) Number of requested frames, L_u, with respect to u, and (d) Percentage of total transferred data as crosshatch bars corresponding to the left vertical axis, and percentage of scheduled SUs as dotted bars corresponding to the right vertical axis, with respect to different resource scheduling algorithms.

Furthermore, when SU u is selected, it is assigned to all RRHs with indicator function $I_{\mathbb{R}^+}(\gamma_{r,u}^s - \gamma_u) = 1$. For this CRAN, the percentage of total transferred data and the percentage of scheduled SUs are shown for different resource scheduling algorithms in Fig. 2d. As illustrated in this figure, the proposed OFBs with $\zeta = 0.41$ and $\zeta = 1$ achieve the highest transferred data percentages, respectively. However, these two algorithms serve a smaller percentages of SUs with respect to the EEF. The selected SUs by the algorithms EEF, OFB with $\zeta = 0.41$, OFB with $\zeta = 1$, and EDF are, respectively $\{10, 13, 4, 1, 5, 9, 11\}$, $\{10, 13, 4, 7, 9, 11\}$, $\{10, 8, 7, 5, 11, 9\}$, $\{10, 2, 1, 5, 9, 11\}$. These results show that algorithms with $\zeta \neq 0$ serve those SUs requesting larger volumes of data with a higher priority, while algorithms with $\zeta = 0$, namely EEF, serve a larger number of SUs.

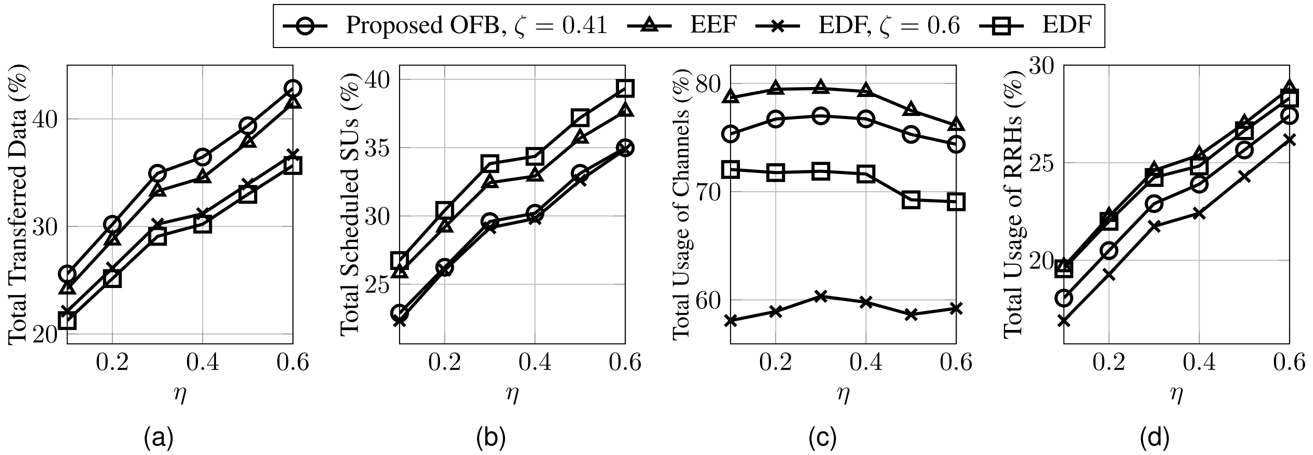
B. MONTE CARLO SIMULATIONS

Next, we evaluate average performance of OFB and ONR over 10^4 random CRAN realizations. These results are averaged over different values of U, R, S, $s^{\max}$, $r^{n,t}$ and availability distribution of channels. As mentioned earlier, the performance of OFB depends on ζ. Fig. 3 illustrates the percentages of total transferred data, scheduled SUs, and usage of channels and RRHs for all offline algorithms with respect to ζ. It should be mentioned that OFB and EDF_ζ with $\zeta = 0$ are equivalent to the EEF and



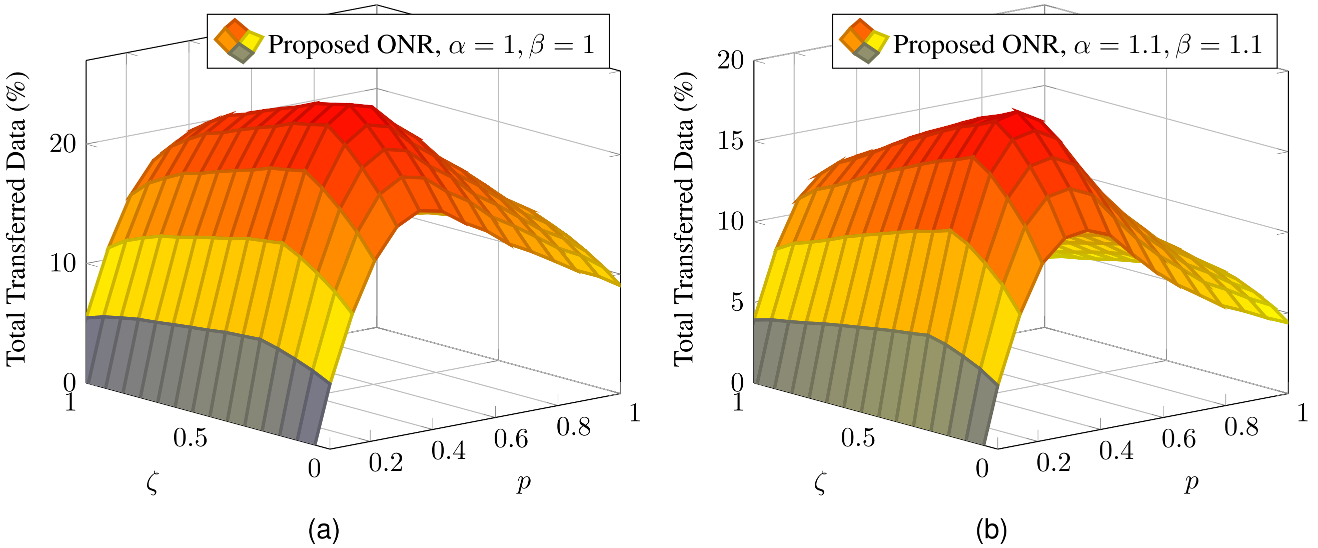
(a) (b) (c) (d)

FIGURE 3. Percentage of total (a) transferred data, (b) scheduled SUs, (c) usage of channels, and (d) usage of RRHs, versus ζ for the offline batch algorithms.



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

FIGURE 4. Percentage of total (a) transferred data, (b) scheduled SUs, (c) usage of channels, and (d) usage of RRHs, for the offline batch algorithms (in optimum ζ) with respect to η.



(a) (b)

FIGURE 5. Total transferred data by the proposed ONR for (a) $(\alpha, \beta) = (1, 1)$, and (b) $(\alpha, \beta) = (1.1, 1.1)$, versus ζ and p.

EDF, respectively. Fig. 3 plots our four performance criteria for various algorithms versus ζ. It is demonstrated that OFB performs better in the percentage of total transferred data over the whole range of ζ and its maximum occurs

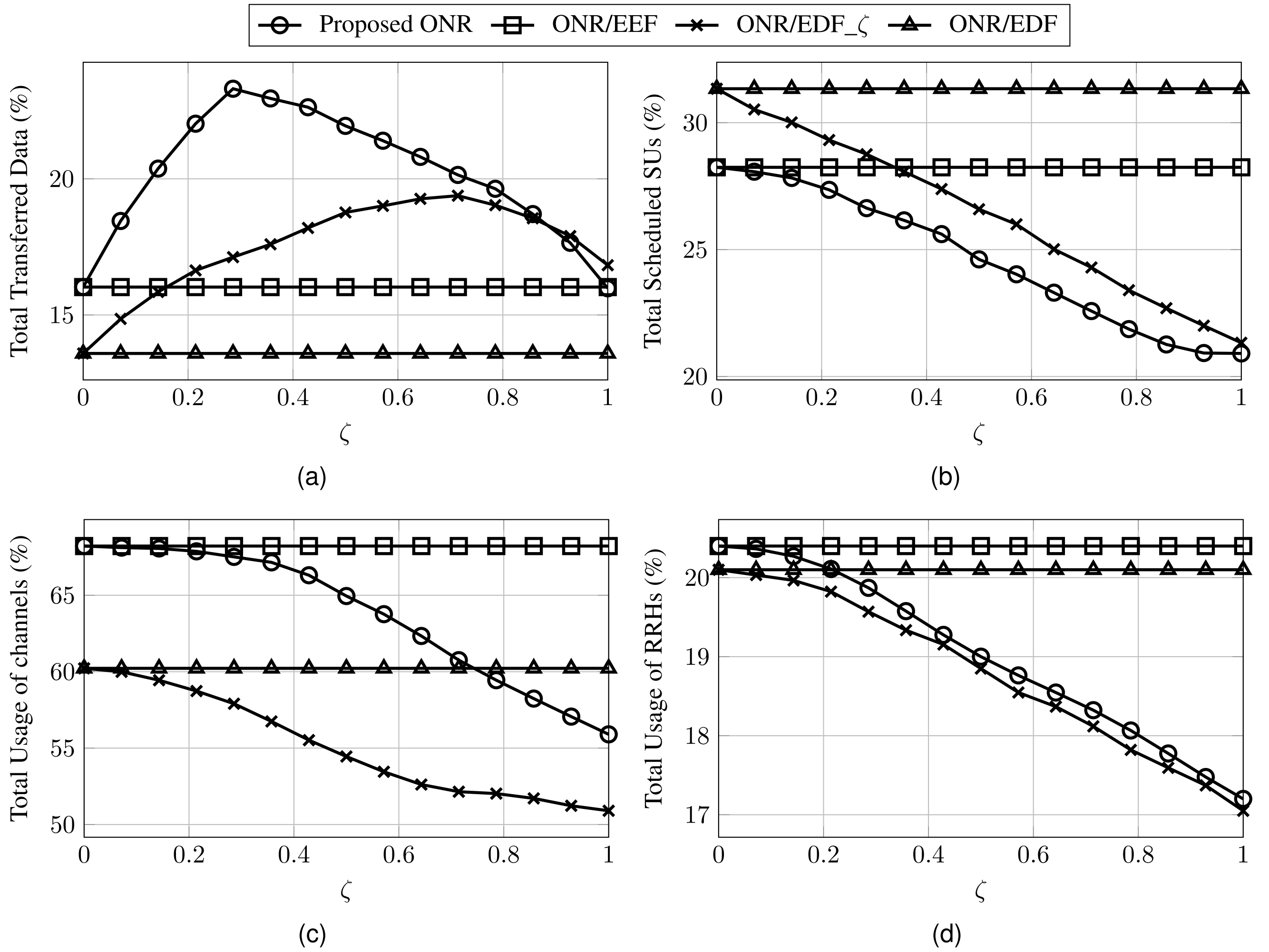


FIGURE 6. Percentage of total (a) transferred data, (b) scheduled SUs, (c) usage of channels, and (d) usage of RRHs, for the online real-time algorithms when $(\alpha, \beta) = (1, 1)$ with respect to ζ for optimum value of $p = 0.36$.

at $\zeta \approx 0.41$ that is also expected from Theorem 1. This improvement in OFB's performance is also a direct consequence of the fact that OFB makes a more efficient utilization of spectrum as corroborated in Fig. 3c. The best percentage of the transferred data are 35.00%, 32.99%, 29.92%, and 28.93% which are achieved by OFB with $\zeta = 0.41$, EEF, EDF_$\zeta = 0.6$, and EDF, respectively. Fig. 3b shows that EDF_ζ achieves approximately 2% more total scheduled SUs compared to OFB. However, OFB performs better in terms of transferred data percentage as it achieves 29.38% versus EDF_ζ's 29.00%. Upon increasing ζ, both algorithms become inclined to schedule SUs with higher volumes of data requests. As a result, the percentage of the total scheduled SUs decreases. We have observed that the EEF and EDF algorithms utilize more channels and RRHs compared to our proposed OFB and ONR algorithms. However, they lack the flexibility to properly select SUs with higher data requests. These algorithms are designed to maximize the number of scheduled SUs, often at the cost of lower data transfer percentages. This inflexibility results in suboptimal solutions for big data transmission, as they do not take into account the data prioritization and QoS requirements of the selected SUs.

Given that PUs' activity pattern vary in time, they cause the available spectrum for SUs to vary in time as well. We use η to express the availability of channels. The percentages of total transferred data, of scheduled SUs, of usage of channels and RRHs are depicted in Fig. 4 versus η for OFB with $\zeta = 0.41$, EEF, EDF with $\zeta = 0.6$, and EDF. It illustrates that OFB achieves a better percentage of total transferred data while maintaining the percentage of total scheduled SUs near to that of the EDF_$\zeta = 0.6$. By increasing ζ, the percentages of total transferred data, scheduled SUs, and RRHs utilization improve for all algorithms. Yet, OFB maintains the best performance in the percentage of total transferred data for all η values.

The performance of ONR is a function of ζ and p as well as the parameters α and β. Fig. 5 illustrates the percentage of total transferred data with respect to ζ and p for $(\alpha, \beta) = (1, 1)$ and $(1.1, 1.1)$, respectively. It can be observed that maximum performance is achieved when $(\zeta, p) = (0.28, 0.36)$ and $(\zeta, p) = (0.28, 0.33)$, in these two

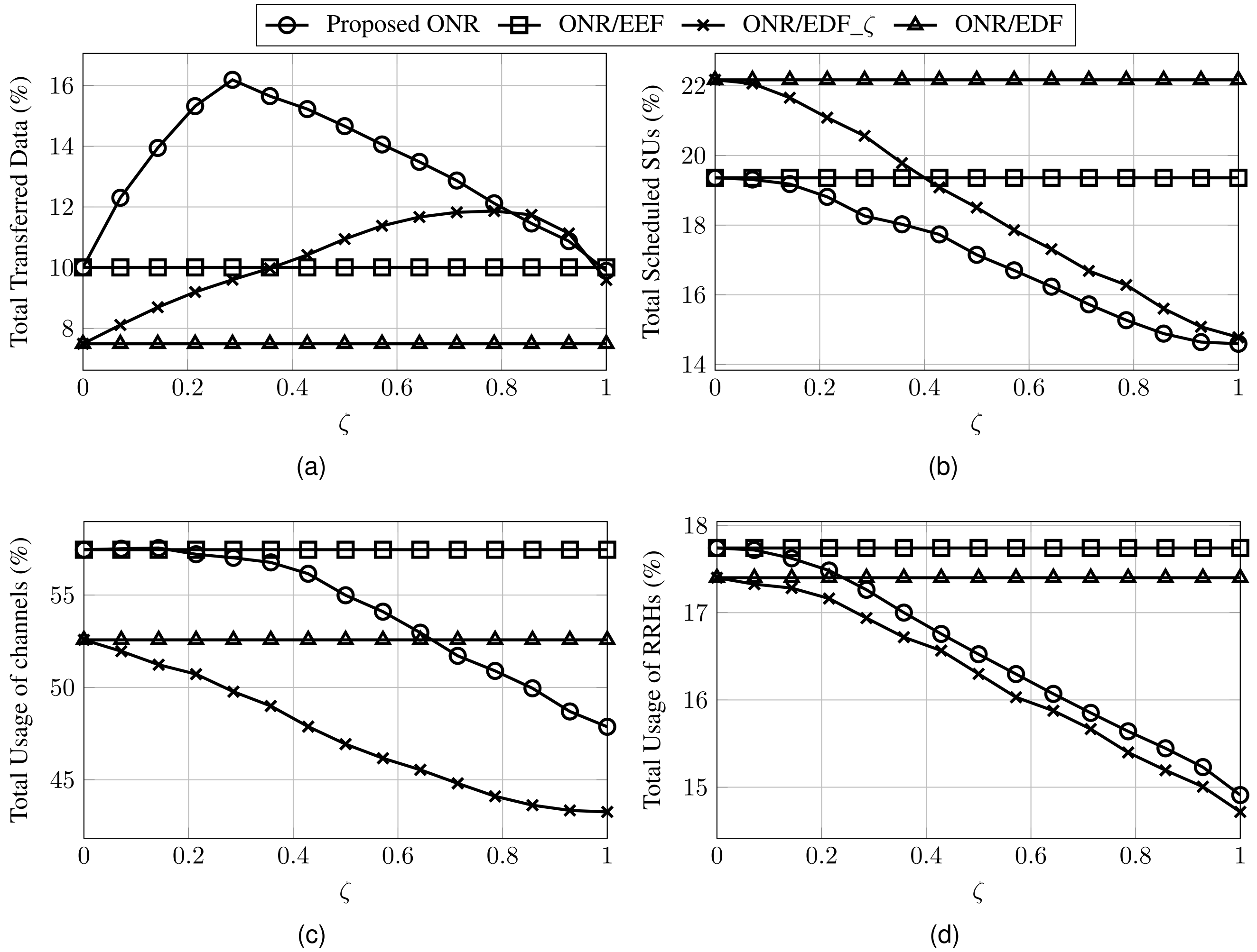
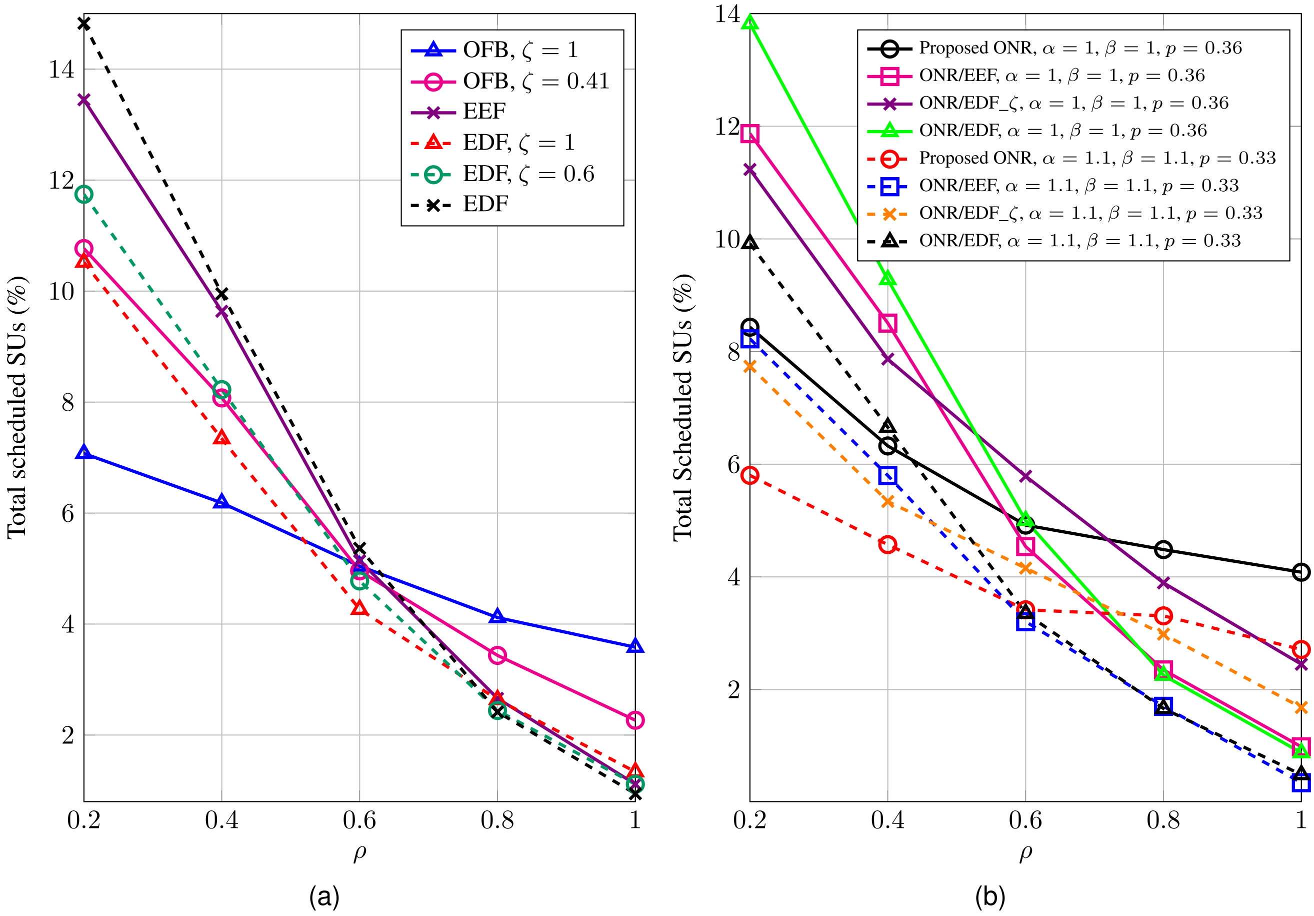


FIGURE 7. Percentage of total (a) transferred data, (b) scheduled SUs, (c) usage of channels, and (d) usage of RRHs, for the online real-time algorithms when $(\alpha, \beta) = (1.1, 1.1)$ with respect to ζ for optimum value of $p = 0.33$.

scenarios. ONR algorithm is simulated with these optimal values of p and plotted in Fig. 6 and Fig. 7 for $(\alpha, \beta) = (1, 1)$ and $(1.1, 1.1)$, respectively. By comparing these two figures, it is deduced that by increasing α and β, all performance criteria degrade. This is a direct consequence of increased uncertainty about requesting SUs and availability of channels in period $n + 1$, which was also predicted by Theorem 2. In Fig. 6, for $(\alpha, \beta) = (1, 1)$, the maximum percentage of total transferred data is given by 23.31%, 16.03%, 19.38%, and 13.59% for ONR, ONR/EEF, ONR/EDF_$\zeta = 0.71$, and ONR/EDF, respectively. In Fig. 7, for $(\alpha, \beta) = (1.1, 1.1)$, the maximum percentage of the total transferred data is given by 16.20%, 10.01%, 11.87%, and 7.49% for ONR, ONR/EEF, ONR/EDF_$\zeta = 0.79$, and ONR/EDF, respectively. The results corroborate a higher percentage of total transferred data for ONR versus all alternatives. This improvement occurs due to a higher utilization of channels, flexibility in SUs' selection due to ζ, and applying our prior knowledge of SUs activity and channels availability probabilities. Similar to offline batch algorithms, ONR/EEF and ONR/EDF have better performances in percentage of total scheduled SUs. According to Fig. 6, the percentage of total scheduled SUs, when $(\alpha, \beta) = (1, 1)$ is 26.63%, 28.24%, 24.29%, and 31.34% for ONR, ONR/EEF, ONR/EDF_$\zeta = 0.71$, and ONR/EDF, respectively. Curiously, ONR performs better than ONR/EDF_ζ. The latter result is also inferred from Fig. 7.

C. BIG DATA REQUESTS

Both OFB and ONR were designed to improve service quality for big data requests. Here, we evaluate both OFB and ONR for big data services. Upon recalling that all requested data sizes are divided into five equal ranges in the interval $[2^5, 2^{20}]$, where each range is recognized by a different ρ, one deduces that the sub-interval with $\rho = 1$ contains 20% of the largest requested data sizes and represents big data users. In Fig. 8, the percentage of the totaled scheduled SUs is plotted versus ρ. The results are plotted for $\zeta = 0, 1$, and ζ's optimal values of Theorems 1 and 2 for OFB and ONR respectively. The results determine that both OFB and ONR schedule more big data requests compared to existing alternatives. By increasing ζ, OFB and ONR exert a higher



(a)

(b)

FIGURE 8. Percentage of total scheduled SUs of the (a) OFB and (b) ONR algorithms, with respect to ρ.

priority for big data requests, so the largest percentage of admitted big data demands occur at $\zeta = 1$. However, $\zeta = 0.41$ also performs satisfactorily on big data. these observations are corroborated numerically in Figs. 8a and 8b for OFB and ONR respectively.

VIII. CONCLUSION

We addressed the problem of selecting SUs, associating them with RRHs, allocating channels, and performing deadline-aware non-preemptive time scheduling over the cognitive CRAN. Our objective is to find an optimal disjunctive set of SUs with corresponding resource allocation to maximize overall weighted data transmission while ensuring QoS parameters for big data transmission. We prioritized SUs based on the requested big data type, which is multiplied by data length in the objective function, to customize this problem for big data transmission. Furthermore, we considered the 5V characteristics of big data in our work.

To solve this problem, we proposed the OFB and ONR algorithms, which support QoS for data requests of selected SUs, including target bit error level, minimum signal-to-noise ratio (SNR), and deadline to receive data. The performance of these algorithms is at most a factor of $3-2\sqrt{2}$ and $\frac{71-17\sqrt{17}}{4(4\alpha\beta^{s^{\max}})^{\frac{3}{2}}}$ away from the globally optimal solutions, respectively.

We evaluated the performance of our proposed algorithms through simulations, which demonstrate that they outperform the EEF and EDF algorithms in total transferred data and big data transmission. Specifically, our proposed algorithms achieve better performance in terms of maximizing overall weighted data transmission, ensuring QoS for data requests, and improving the efficiency of spectrum utilization.

APPENDIX A
PROOF OF THEOREM 1

First, we analyze the relation between U_n^{OFB} and $\mathcal{U}_n^{\text{temp}}$. According to OFB algorithm

$$U_n^{\text{OFB}} \subseteq \mathcal{U}_n^{\text{temp}} \longrightarrow \sum_{u\in U_n^{\text{OFB}}} \alpha_u L_u \leq \sum_{u\in \mathcal{U}_n^{\text{temp}}} \alpha_u L_u. \qquad (12)$$

Each $u \in U_n^{\text{OFB}}$ has been admitted either through Line 18 or 52 of Algorithm 1. Obviously, each SU u that is finally admitted belongs to U_n^{OFB}. These users are all members of $\mathcal{U}_n^{\text{temp}}$ as well. However, $\mathcal{U}_n^{\text{temp}}$ also contains those SUs that were once admitted but were later dropped according to Line 52 of OFB. In this Line, u is accepted and the set U' of previously admitted SUs are rejected if $\zeta\alpha_u L_u > \sum_{u'\in U'} \alpha_{u'} L_{u'}$. Due to this substitution, the objective function increases by at least $(1-\zeta)\,\alpha_u L_u$. We can write this as

$$\begin{aligned}\sum_{u\in\mathcal{U}_n^{\text{temp}}} \alpha_u L_u &\le \sum_{u\in U_n^{\text{OFB}}} \alpha_u L_u + \sum_{u\in\mathcal{U}_n^{\text{temp}}\setminus U_n^{\text{OFB}}} \alpha_u L_u \\ &\le \sum_{u\in U_n^{\text{OFB}}} \alpha_u L_u + \sum_{u\in U_n^{\text{OFB}}} \zeta\,\alpha_u L_u \\ &\le \sum_{u\in U_n^{\text{OFB}}} \alpha_u L_u + \sum_{u\in\mathcal{U}_n^{\text{temp}}} \zeta\,\alpha_u L_u.\end{aligned}$$

Finally, we arrive at

$$\sum_{u\in U_n^{\text{OFB}}} \alpha_u L_u \ge (1-\zeta)\sum_{u\in\mathcal{U}_n^{\text{temp}}} \alpha_u L_u. \tag{13}$$

Next, we derive a bound between $\mathcal{U}_n^{\text{temp}}$ and U_n^*. We can write the following inequality

$$\begin{aligned}\sum_{u\in U_n^*} \alpha_u L_u &= \sum_{u\in U_n^*\cap\mathcal{U}_n^{\text{temp}}} \alpha_u L_u + \sum_{u\in U_n^*\setminus\mathcal{U}_n^{\text{temp}}} \alpha_u L_u \\ &\le \sum_{u\in\mathcal{U}_n^{\text{temp}}} \alpha_u L_u + \sum_{u\in U_n^*\setminus\mathcal{U}_n^{\text{temp}}} \alpha_u L_u. \end{aligned}\tag{14}$$

For every $u \in U_n^* \setminus \mathcal{U}_n^{\text{temp}}$, this SU was not admitted because there was a set of users $U' \in \mathcal{U}_n^{\text{temp}}$ such that $\sum_{u'\in U'} \alpha_{u'} L_{u'} \ge \zeta\alpha_u L_u$. We need to show that for different $u, v \in U_n^* \setminus \mathcal{U}_n^{\text{temp}}$, the corresponding sets $U', V' \subset \mathcal{U}_n^{\text{temp}}$ are disjoint. To show this, we consider two cases. Either u, v schedules in the global optimum share a time slot or do not share any time slots. If they share time slots, then they should be scheduled on different frequency channels. Hence, they will interfere with disjoint U', V'. If they do not share time slots, then they can be scheduled on the same frequency channels. Let us assume there exists a SU $w \in \mathcal{U}_n^{\text{temp}}$ which belongs to both U', V'. Then, either u or v will end before w. According to the while loop in line 4 of Algorithm 1, either u or v should belong to $\mathcal{U}_n^{\text{temp}}$ which is not the case voiding this assumption. As a result, U' and V' are guaranteed to be disjoint. Given the disjoint assumption, we can write $\zeta \sum_{u\in U_n^*/\mathcal{U}_n^{\text{temp}}} \alpha_u L_u \le \sum_{u\in\mathcal{U}_n^{\text{temp}}} \alpha_u L_u$. Combining this with (14), we arrive at

$$\sum_{u\in U_n^*} \alpha_u L_u \le \left(1+\frac{1}{\zeta}\right)\sum_{u\in\mathcal{U}_n^{\text{temp}}} \alpha_u L_u. \tag{15}$$

Finally, we combine (13) and (15) to arrive at

$$\sum_{u\in U_n^{\text{OFB}}} \alpha_u L_u \ge \left(\zeta\frac{1-\zeta}{1+\zeta}\right)\sum_{u\in U_n^*} \alpha_u L_u. \tag{16}$$

By taking the derivative of $\zeta\frac{1-\zeta}{1+\zeta}$ and set it to zero, we obtain two values for ζ as $-1-\sqrt{2}$ and $-1+\sqrt{2}$. The first one is negative and hence not a valid choice. Thus, $\zeta = -1+\sqrt{2}$ leading to $\zeta\frac{1-\zeta}{1+\zeta} = 3-2\sqrt{2} \approx 0.17$.

APPENDIX B
PROOF OF THEOREM 2

To derive the performance bound for ONR, we derive successive bounds on how much objective value we loose in going from U_{n+1}^* to U_{n+1}^{ONR} at every step of Fig. 9. Then, we combine the corresponding losses to derive Theorem 2. This proof idea is borrowed from [66]. However, our ONR is different from their proposed online algorithm and thus demands a separate in-depth analysis. First, we assume that all admitted users can only be scheduled on a set $\mathcal{S}_a \subset \mathcal{S}$ of size $|\mathcal{S}_a| = s^{\max}$. It is notable that data request probability for SUs and availability of channels are independent, so the joint probability of data request by SU u at period n and availability of set $\mathcal{S}_a$ of channels in nth period of the CRAN is given by

$$P_n(u,\mathcal{S}_a) = P_n(u)\,\Pi_{s\in\mathcal{S}_a} P_n(\Delta f_s). \tag{17}$$

Lemma 1: By using (8) and (9) in (17), we have

$$\begin{aligned}\frac{1}{\sqrt{\alpha\beta^{s^{\max}}}} = \frac{1}{\sqrt{\alpha\beta^{\sum_s I_{\mathcal{S}_a}(s)}}} &\le \frac{P_{n+1}(u,\mathcal{S}_a)}{P_n(u,\mathcal{S}_a)} \\ &\le \sqrt{\alpha\beta^{\sum_s I_{\mathcal{S}_a}(s)}} = \sqrt{\alpha\beta^{s^{\max}}}.\end{aligned}\tag{18}$$

First, we characterize the loss in going from U_{n+1}^* to U_n^*.

Lemma 2: The following inequality holds

$$\mathbb{E}\left(\sum_{u\in U_{n+1}^*} \alpha_u L_u \middle| \mathcal{S}_a\right) \le \sqrt{\alpha\beta^{s^{\max}}}\,\mathbb{E}\left(\sum_{u\in U_n^*} \alpha_u L_u \middle| \mathcal{S}_a\right). \tag{19}$$

Proof:

$$\begin{aligned}&\mathbb{E}\left(\sum_{u\in U_{n+1}^*} \alpha_u L_u \middle| \mathcal{S}_a\right) \\ &= \sum_{u\in\mathcal{U}} \mathbb{E}\left(I_{U_{n+1}^*}(u)\alpha_u L_u \middle| \mathcal{S}_a\right) \\ &= \sum_{u\in\mathcal{U}} P\left(I_{U_{n+1}^*}(u)\middle|\mathcal{S}_a\right)\alpha_u L_u \le \sum_{u\in\mathcal{U}} P_{n+1}(u,\mathcal{S}_a)\,\alpha_u L_u \\ &\le \sqrt{\alpha}\sum_{u\in\mathcal{U}} \sqrt{\beta^{\sum_{s\in\mathcal{S}_a} I(s)}}P_n(u,s_u)\,\alpha_u L_u \\ &\le \sqrt{\alpha\beta^{s^{\max}}}\sum_{u\in\mathcal{U}} P_n(u,\mathcal{S}_a)\,\alpha_u L_u \\ &= \sqrt{\alpha\beta^{s^{\max}}}\sum_{u\in\mathcal{U}} P\left(I_{U_n^*}(u)\middle|\mathcal{S}_a\right)\alpha_u L_u \\ &= \sqrt{\alpha\beta^{s^{\max}}}\sum_{u\in\mathcal{U}} \mathbb{E}\left(I_{U_n^*}(u)\alpha_u L_u\middle|\mathcal{S}_a\right) \\ &= \sqrt{\alpha\beta^{s^{\max}}}\,\mathbb{E}\left(\sum_{u\in U_n^*} \alpha_u L_u \middle| \mathcal{S}_a\right),\end{aligned}\tag{20}$$

where in the second inequality, Lemma 1 was applied. ■

Next, we characterize the loss in going from U_n^* to U_n^{OFB}. Taking expected values from both sides of (16) we arrive at

$$\mathbb{E}\left(\sum_{u\in U_n^{\text{OFB}}} \alpha_u L_u \middle| \mathcal{S}_a\right) \ge \zeta\frac{1-\zeta}{1+\zeta}\mathbb{E}\left(\sum_{u\in U_n^*} \alpha_u L_u \middle| \mathcal{S}_a\right). \tag{21}$$

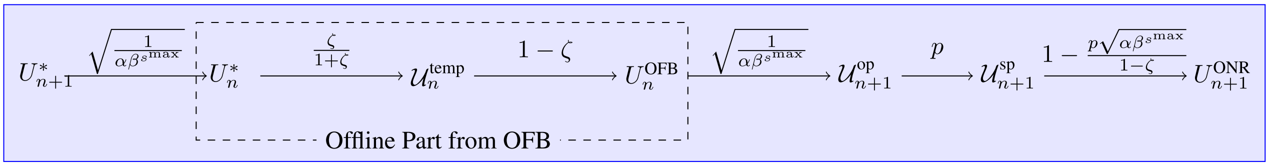

U^*_{n+1}
$\sqrt{\frac{1}{\alpha\beta^{s^{\max}}}}$
U^*_n
$\frac{\zeta}{1+\zeta}$
$\mathcal{U}^{\text{temp}}_n$
$1-\zeta$
U^{OFB}_n
$\sqrt{\frac{1}{\alpha\beta^{s^{\max}}}}$
$\mathcal{U}^{\text{op}}_{n+1}$
p
$\mathcal{U}^{\text{sp}}_{n+1}$
$1-\frac{p\sqrt{\alpha\beta^{s^{\max}}}}{1-\zeta}$
U^{ONR}_{n+1}
Offline Part from OFB

FIGURE 9. Performance degradation of the proposed ONR scheduling algorithm from optimal solution; which shows how much successive bounds on objective value is degraded in going from U^*_{n+1} to U^{ONR}_{n+1} at every step.

The following lemma determines the loss in going from U^{OFB}_n to $\mathcal{U}^{\text{op}}_{n+1}$:

Lemma 3: The following inequality holds

$$\mathbb{E}\left(\sum_{u\in U^{\text{OFB}}_n}\alpha_u L_u\middle|\mathcal{S}_a\right)\leq\sqrt{\alpha\beta^{s^{\max}}}\,\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{op}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right). \tag{22}$$

Proof: We define $\mathcal{A}_u$ as event that u is disjunctive with $U^{\text{OFB}}_n\setminus\{u\}$. Based on Algorithm 2, $u\in U^{\text{OFB}}_n$ if u requests data in period n, $\mathcal{S}_a$ is available in period n, and u is disjunctive with $U^{\text{OFB}}_n\setminus\{u\}$. Subsequently, $P(u\in U^{\text{OFB}}_n)=P_n(u,\mathcal{S}_a,\mathcal{A}_u)$. As well, u is a member of $\mathcal{U}^{\text{op}}_{n+1}$, if u requests data in period $n+1$, $\mathcal{S}_a$ is available in period $n+1$, and u is disjunctive with $U^{\text{OFB}}_n\setminus\{u\}$. As a result, $P(u\in\mathcal{U}^{\text{op}}_{n+1})=P_{n+1}(u,\mathcal{S}_a,\mathcal{A}_u)$. So, we have:

$$\mathbb{E}\left(I_{U^{\text{OFB}}_n}(u)\alpha_u L_u\middle|\mathcal{S}_a\right) \tag{23a}$$
$$\begin{aligned}
&=P\left(I_{U^{\text{OFB}}_n}(u)=1\middle|\mathcal{S}_a\right)\times\alpha_u L_u=P_n\left(u,\mathcal{S}_a,\mathcal{A}_u\right)\alpha_u L_u\\
&=P\left(\mathcal{A}_u\right)P_n\left(u,\mathcal{S}_a\mid\mathcal{A}_u\right)\alpha_u L_u=P\left(\mathcal{A}_u\right)P_n\left(u,\mathcal{S}_a\right)\alpha_u L_u\\
&\leq\sqrt{\alpha\beta^{s^{\max}}}P\left(\mathcal{A}_u\right)P_{n+1}\left(u,\mathcal{S}_a\right)\alpha_u L_u\\
&=\sqrt{\alpha\beta^{s^{\max}}}P\left(\mathcal{A}_u\right)P_{n+1}\left(u,\mathcal{S}_a|\mathcal{A}_u\right)\alpha_u L_u\\
&=\sqrt{\alpha\beta^{s^{\max}}}P_{n+1}\left(u,\mathcal{S}_a,\mathcal{A}_u\right)\alpha_u L_u\\
&=\sqrt{\alpha\beta^{s^{\max}}}P\left(u\in\mathcal{U}^{\text{op}}_{n+1}\middle|\mathcal{S}_a\right)\alpha_u L_u\\
&=\sqrt{\alpha\beta^{s^{\max}}}\mathbb{E}\left(I_{\mathcal{U}^{\text{op}}_{n+1}}(u)\alpha_u L_u\middle|\mathcal{S}_a\right).
\end{aligned} \tag{23b}$$

It should be mentioned that we assume $\mathcal{A}_u$ is independent of u requesting data and availability of channels in n and $n+1$ periods. Summing (23a) and (23b) over all $u\in\mathcal{U}$ will yield the lemma's inequality. ∎

Lemma 4: We have the following equality

$$\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{sp}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right)=p\,\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{op}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right). \tag{24}$$

Proof: We know that if $u\in\mathcal{U}^{\text{op}}_{n+1}$, then $u\in\mathcal{U}^{\text{sp}}_{n+1}$ with probability p. So, we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{sp}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right)&=\sum_{u\in\mathcal{U}}\alpha_u L_u\mathbb{E}\left(I_{\mathcal{U}^{\text{sp}}_{n+1}}(u)\middle|\mathcal{S}_a\right)\\
&=p\sum_{u\in\mathcal{U}}\alpha_u L_u\mathbb{E}\left(I_{\mathcal{U}^{\text{op}}_{n+1}}(u)\middle|\mathcal{S}_a\right)\\
&=p\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{op}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right).
\end{aligned} \tag{25}$$

∎

Lemma 5: We have the following inequality

$$\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{sp}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right)\leq p\sqrt{\alpha\beta^{s^{\max}}}\,\mathbb{E}\left(\sum_{u\in U^{\text{OFB}}_n}\alpha_u L_u\middle|\mathcal{S}_a\right). \tag{26}$$

Proof: Upon applying Lemma 4 to the left hand side (LHS) of (26), it suffices to prove the following

$$\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{op}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right)\leq\sqrt{\alpha\beta^{s^{\max}}}\,\mathbb{E}\left(\sum_{u\in U^{\text{OFB}}_n}\alpha_u L_u\middle|\mathcal{S}_a\right). \tag{27}$$

According to the proof of Lemma 3, we have $P(u\in U^{\text{OFB}}_n)=P_n(u,\mathcal{S}_a,\mathcal{A}_u)$ and $P(u\in\mathcal{U}^{\text{op}}_{n+1})=P_{n+1}(u,\mathcal{S}_a,\mathcal{A}_u)$. So, similar to (23) we have:

$$\mathbb{E}\left(I_{U^{\text{OFB}}_n}(u)\alpha_u L_u\middle|\mathcal{S}_a\right) \tag{28a}$$
$$\begin{aligned}
&=P\left(I_{U^{\text{OFB}}_n}(u)=1\middle|\mathcal{S}_a\right)\times\alpha_u L_u=P_n\left(u,\mathcal{S}_a,\mathcal{A}_u\right)\alpha_u L_u\\
&=P\left(\mathcal{A}_u\right)P_n\left(u,\mathcal{S}_a\mid\mathcal{A}_u\right)\alpha_u L_u=P\left(\mathcal{A}_u\right)P_n\left(u,\mathcal{S}_a\right)\alpha_u L_u\\
&\geq\frac{1}{\sqrt{\alpha\beta^{s^{\max}}}}P\left(\mathcal{A}_u\right)P_{n+1}\left(u,\mathcal{S}_a\right)\alpha_u L_u\\
&=\frac{1}{\sqrt{\alpha\beta^{s^{\max}}}}P\left(\mathcal{A}_u\right)P_{n+1}\left(u,\mathcal{S}_a|\mathcal{A}_u\right)\alpha_u L_u\\
&=\frac{1}{\sqrt{\alpha\beta^{s^{\max}}}}P_{n+1}\left(u,\mathcal{S}_a,\mathcal{A}_u\right)\alpha_u L_u\\
&=\frac{1}{\sqrt{\alpha\beta^{s^{\max}}}}P\left(u\in\mathcal{U}^{\text{op}}_{n+1}\middle|\mathcal{S}_a\right)\alpha_u L_u\\
&=\frac{1}{\sqrt{\alpha\beta^{s^{\max}}}}\mathbb{E}\left(I_{\mathcal{U}^{\text{op}}_{n+1}}(u)\alpha_u L_u\middle|\mathcal{S}_a\right).
\end{aligned} \tag{28b}$$

Summing (28a) and (28b) over all $u\in\mathcal{U}$ will yield (27). ∎

Lemma 6: The following inequality holds

$$\mathbb{E}\left(\sum_{u\in U^{\text{ONR}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right)\geq\left(1-p\frac{\sqrt{\alpha\beta^{s^{\max}}}}{1-\zeta}\right)\times\mathbb{E}\left(\sum_{u\in\mathcal{U}^{\text{sp}}_{n+1}}\alpha_u L_u\middle|\mathcal{S}_a\right). \tag{29}$$

Proof: The set U_{n+1}^{ONR} is obtained when OFB is applied to $\mathcal{U}_{n+1}^{\mathrm{sp}}$ and a disjunctive subset of SUs in $\mathcal{U}_{n+1}^{\mathrm{sp}}$ are selected. Subsequently, we have the following

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{u\in\mathcal{U}_{n+1}^{\mathrm{sp}}}\alpha_u L_u\middle|\mathcal{S}_a\right)\\
&=\mathbb{E}\left(\sum_{u\in U_{n+1}^{\mathrm{ONR}}}\alpha_u L_u\middle|\mathcal{S}_a\right)+\mathbb{E}\left(\sum_{u'\in\mathcal{U}_{n+1}^{\mathrm{sp}}\setminus U_{n+1}^{\mathrm{ONR}}}\alpha_{u'}L_{u'}\middle|\mathcal{S}_a\right)\\
&=\mathbb{E}\left(\sum_{u\in U_{n+1}^{\mathrm{ONR}}}\alpha_u L_u\middle|\mathcal{S}_a\right)\\
&\quad+\mathbb{E}\left(\sum_{u'\in\mathcal{U}_{n+1}^{\mathrm{sp}},\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\middle|\mathcal{S}_a\right) \qquad \text{(30a)}\\
&\le\mathbb{E}\left(\sum_{u\in U_{n+1}^{\mathrm{ONR}}}\alpha_u L_u\middle|\mathcal{S}_a\right)+p\sqrt{\alpha\beta^{s^{\max}}}\\
&\quad\times\mathbb{E}\left(\sum_{u'\in U_{n}^{\mathrm{OFB}},\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\middle|\mathcal{S}_a\right), \qquad \text{(30b)}
\end{aligned}
$$

where $\bar{\mathcal{D}}\left(u', U_{n+1}^{\mathrm{ONR}}\right)$ means that u' is not disjunctive with U_{n+1}^{ONR}. The inequality in (30b) is derived by an application of Lemma 5:

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{u'\in\mathcal{U}_{n+1}^{\mathrm{sp}},\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\middle|\mathcal{S}_a\right)\\
&\quad\le p\sqrt{\alpha\beta^{s^{\max}}}\times\mathbb{E}\left(\sum_{u'\in U_{n}^{\mathrm{OFB}},\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\middle|\mathcal{S}_a\right). \qquad \text{(31)}
\end{aligned}
$$

Now, we simplify the second term in the right hand side (RHS) of (30b). We know that members of $\left\{u' \in U_n^{\mathrm{OFB}}, \bar{\mathcal{D}}\left(u', U_{n+1}^{\mathrm{ONR}}\right)\right\}$ are disjunctive, and are jointly admitted and scheduled in period n given the available resources in period n. Therefore, the reason these SUs do not belong to U_{n+1}^{ONR} is that their weighted data size is smaller than those appearing in U_{n+1}^{ONR}. Next, we assume each SU $v \in U_{n+1}^{\mathrm{ONR}}$ have caused the absence of set $C_v \subseteq U_n^{\mathrm{OFB}}$ in U_{n+1}^{ONR}. In other words, C_v is the part of $\left\{u' \in U_n^{\mathrm{OFB}}, \bar{\mathcal{D}}\left(u', U_{n+1}^{\mathrm{ONR}}\right)\right\}$ that are omitted from U_{n+1}^{ONR} due to not being disjunctive with $v \in U_{n+1}^{\mathrm{ONR}}$. Consequently, we have

$$
\sum_{u'\in C_v}\alpha_{u'}L_{u'} < \zeta\alpha_v L_v.
$$

U_{n+1}^{ONR} is a disjunctive set. So, for two different SUs v and v' in U_{n+1}^{ONR}, we have $C_v \cap C_{v'} = \emptyset$. Therefore, we can write the following

$$
\begin{aligned}
(1-\zeta)&\sum_{u'\in U_n^{\mathrm{OFB}},\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\\
&\le\sum_{u'\in U_n^{\mathrm{OFB}},\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\\
&=\sum_{u'\in \bigcup_{v\in U_{n+1}^{\mathrm{ONR}}} C_v,\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\\
&\le\zeta\sum_{v\in U_{n+1}^{\mathrm{ONR}}}\alpha_v L_v\\
&\le\sum_{v\in U_{n+1}^{\mathrm{ONR}}}\alpha_v L_v\\
&\le\sum_{v\in \mathcal{U}_{n+1}^{\mathrm{sp}}}\alpha_v L_v.
\end{aligned}
$$

Subsequently, we have

$$
\sum_{u'\in U_n^{\mathrm{OFB}},\bar{\mathcal{D}}\left(u',U_{n+1}^{\mathrm{ONR}}\right)}\alpha_{u'}L_{u'}\le\frac{1}{1-\zeta}\sum_{v\in \mathcal{U}_{n+1}^{\mathrm{sp}}}\alpha_v L_v.
$$

Upon substituting this inequality in the second term on the RHS of (30b), proof of Lemma 6 is completed. ■

Next, we combine Lemmas 1-4 and Lemma 6 to arrive at

$$
\begin{aligned}
\mathbb{E}\left(\sum_{u\in U_{n+1}^{\mathrm{ONR}}}\alpha_u L_u\right)\ge&\left(\frac{p\zeta(1-\zeta)}{(1+\zeta)\alpha\beta^{s^{\max}}}\right.\\
&\left.-\frac{p^2\zeta}{(1+\zeta)\sqrt{\alpha\beta^{s^{\max}}}}\right)\mathbb{E}\left(\sum_{u\in U_{n+1}^{*}}\alpha_u L_u\right).
\end{aligned}
\qquad \text{(32)}
$$

We maximize the RHS of bound in (32) with respect to both p and ζ. Taking the derivative of RHS with respect to p and setting it equal to zero will yield $p = \frac{7-\sqrt{17}}{8\sqrt{\alpha\beta^{s^{\max}}}}$. Then, we take the derivative with respect to ζ and set it equal to zero which yields $\zeta = \frac{\sqrt{17}-3}{4}$. Substituting these values for p, ζ into (32) will complete the proof of Theorem 2.

REFERENCES

[1] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Greening big data," *IEEE Syst. J.*, vol. 10, no. 3, pp. 873–887, Sep. 2016.

[2] N. Zhang, P. Yang, J. Ren, D. Chen, L. Yu, and X. Shen, "Synergy of big data and 5G wireless networks: Opportunities, approaches, and challenges," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 12–18, Feb. 2018.

[3] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[4] International Telecommunications Union, "IMT traffic estimates for the years 2020 to 2030," Electron. Publ. Geneva, Switzerland, Tech. Rep., R-REP-M.2370-2015, 2015, pp. 1–51. [Online]. Available: https://www.itu.int/dmspub/itu-r/opb/rep/R-REP-M.2370-2015-PDF-E.pdf

[5] W. Obile, "Ericsson Mobility Report," Ericsson, Stockholm, Sweden, Tech. Rep. EAB-16:001237, Nov. 2016.

[6] V. N. Gudivada, R. Baeza-Yates, and V. V. Raghavan, "Big data: Promises and problems," *Computer*, vol. 48, no. 3, pp. 20–23, Mar. 2015.

[7] K. Wang, Y. Wang, X. Hu, Y. Sun, D. Deng, A. Vinel, and Y. Zhang, ''Wireless big data computing in smart grid,'' *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 58–64, Apr. 2017.

[8] S. Yu, M. Liu, W. Dou, X. Liu, and S. Zhou, ''Networking for big data: A survey,'' *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 531–549, 1st Quart., 2017.

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

[10] W. Chien, C. Lai, and H. Chao, ''Dynamic resource prediction and allocation in C-RAN with edge artificial intelligence,'' *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4306–4314, Jul. 2019.

[11] F. Al-Turjman, L. Mostarda, E. Ever, A. Darwish, and N. Shekh Khalil, ''Network experience scheduling and routing approach for big data transmission in the Internet of Things,'' *IEEE Access*, vol. 7, pp. 14501–14512, 2019.

[12] X. Zhang and Q. Zhu, ''Information-centric virtualization for software-defined statistical QoS provisioning over 5G multimedia big data wireless networks,'' *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1721–1738, Aug. 2019.

[13] C. Zhang, K. Ota, J. Jia, and M. Dong, ''Breaking the blockage for big data transmission: Gigabit road communication in autonomous vehicles,'' *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 152–157, Jun. 2018.

[14] X. Yi, F. Liu, J. Liu, and H. Jin, ''Building a network highway for big data: Architecture and challenges,'' *IEEE Netw.*, vol. 28, no. 4, pp. 5–13, Jul. 2014.

[15] X. Hou, B. Lin, R. He, and X. Wang, ''Infrastructure planning and topology optimization for reliable mobile big data transmission under cloud radio access networks,'' *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, pp. 1–11, Dec. 2016.

[16] M. Bigdeli and B. Abolhassani, ''A new algorithm for maximizing total big data flow in a cloud radio access network,'' in *Proc. 28th Iranian Conf. Electr. Eng. (ICEE)*, Aug. 2020, pp. 1–5.

[17] L. Cui, F. R. Yu, and Q. Yan, ''When big data meets software-defined networking: SDN for big data and big data for SDN,'' *IEEE Netw.*, vol. 30, no. 1, pp. 58–65, Jan. 2016.

[18] H. Peng, Y. Tian, and J. Kurths, ''Semitensor product compressive sensing for big data transmission in wireless sensor networks,'' *Math. Problems Eng.*, vol. 2017, pp. 1–8, 2017.

[19] Y. Yang, J. Xu, Z. Xu, P. Zhou, and T. Qiu, ''Quantile context-aware social IoT service big data recommendation with D2D communication,'' *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5533–5548, Jun. 2020.

[20] C. Jiang and Z. Li, ''Decreasing big data application latency in satellite link by caching and peer selection,'' *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2555–2565, Dec. 2020.

[21] S. Bhattacharjee, L. B. A. Rahim, J. Watada, and A. Roy, ''Unified GPU technique to boost confidentiality, integrity and trim data loss in big data transmission,'' *IEEE Access*, vol. 8, pp. 45477–45495, 2020.

[22] Y. Xing, J. Han, K. Xue, J. Liu, M. Pan, and P. Hong, ''MPTCP meets big data: Customizing transmission strategy for various data flows,'' *IEEE Netw.*, vol. 34, no. 4, pp. 35–41, Jul. 2020.

[23] X. Zheng and Z. Cai, ''Real-time big data delivery in wireless networks: A case study on video delivery,'' *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2048–2057, Aug. 2017.

[24] S. M. Srinivasan, T. Truong-Huu, and M. Gurusamy, ''Deadline-aware scheduling and flexible bandwidth allocation for big-data transfers,'' *IEEE Access*, vol. 6, pp. 74400–74415, 2018.

[25] W. Xia, T. Q. S. Quek, J. Zhang, S. Jin, and H. Zhu, ''Programmable hierarchical C-RAN: From task scheduling to resource allocation,'' *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 2003–2016, Mar. 2019.

[26] P. Luong, F. Gagnon, C. Despins, and L. Tran, ''Joint virtual computing and radio resource allocation in limited fronthaul green C-RANs,'' *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2602–2617, Apr. 2018.

[27] A. Younis, T. X. Tran, and D. Pompili, ''Bandwidth and energy-aware resource allocation for cloud radio access networks,'' *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6487–6500, Oct. 2018.

[28] D. Zeng, J. Zhang, L. Gu, S. Guo, and J. Luo, ''Energy-efficient coordinated multipoint scheduling in green cloud radio access network,'' *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9922–9930, Oct. 2018.

[29] Q. Liu, T. Han, and N. Ansari, ''Energy-efficient on-demand resource provisioning in cloud radio access networks,'' *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 1142–1151, Dec. 2019.

[30] K. Lin, W. Wang, Y. Zhang, and L. Peng, ''Green spectrum assignment in secure cloud radio network with cluster formation,'' *IEEE Trans. Sustain. Comput.*, vol. 4, no. 2, pp. 191–203, Apr. 2019.

[31] M. Khan, Z. H. Fakhri, and H. S. Al-Raweshidy, ''Semistatic cell differentiation and integration with dynamic BBU-RRH mapping in cloud radio access network,'' *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 1, pp. 289–303, Mar. 2018.

[32] J. Yao and N. Ansari, ''QoS-aware joint BBU-RRH mapping and user association in cloud-RANs,'' *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 4, pp. 881–889, Dec. 2018.

[33] N. Li, Z. Yao, Y. Tu, and Y. Chen, ''Cooperative optimization for OFDMA resource allocation in multi-RRH millimeter-wave CRAN,'' *IEEE Access*, vol. 8, pp. 164035–164044, 2020.

[34] L. You and D. Yuan, ''User-centric performance optimization with remote radio head cooperation in C-RAN,'' *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 340–353, Jan. 2020.

[35] J. Ye and Y. Zhang, ''Pricing-based resource allocation in virtualized cloud radio access networks,'' *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 7096–7107, Jul. 2019.

[36] C. Fang, P. Li, and K. Feng, ''Joint interference cancellation and resource allocation for full-duplex cloud radio access networks,'' *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3019–3033, Jun. 2019.

[37] Z. Lin and Y. Liu, ''Joint uplink and downlink transmissions in user-centric OFDMA cloud-RAN,'' *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7776–7788, Aug. 2019.

[38] M. Labana and W. Hamouda, ''Unsupervised deep learning for power allocation in CRAN,'' in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

[39] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini, ''Joint scheduling and beamforming via cloud-radio access networks coordination,'' in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–5.

[40] W. Tang and S. Feng, ''User selection and power minimization in full-duplex cloud radio access networks,'' *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2426–2438, May 2019.

[41] S. D'Oro, M. A. Marotta, C. B. Both, L. DaSilva, and S. Palazzo, ''Power-efficient resource allocation in C-RANs with SINR constraints and deadlines,'' *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6099–6113, Jun. 2019.

[42] J. Dong, Q. Yang, B. Li, and K. S. Kwak, ''Efficient virtual machine scheduling for downlink joint transmission of comp in C-RAN,'' in *Proc. Int. Conf. Wirel. Commun. Signal Process. (WCSP)*, 2015, pp. 1–5.

[43] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, ''Dynamic resource allocation for cloud-RAN in LTE with real-time BBU/RRH assignment,'' in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[44] Z. Iftikhar, S. Jangsher, H. K. Qureshi, and M. Aloqaily, ''Resource efficient allocation and RRH placement for backhaul of moving small cells,'' *IEEE Access*, vol. 7, pp. 47379–47389, 2019.

[45] M. Bigdeli, S. Farahmand, B. Abolhassani, and H. H. Nguyen, ''Globally optimal resource allocation and time scheduling in downlink cognitive CRAN favoring big data requests,'' *IEEE Access*, vol. 10, pp. 27504–27521, 2022.

[46] *IEEE Standard for Information Technology—Telecommunications and information Exchange Between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, Standard 802.11- 2016 (Revision IEEE Std 802.11-2012), pp. 2321–2582, 2016.

[47] 3GPP. (2021). *5G; NR; Physical Layer Procedures for Control (3GPP TS 38.213 Version 16.6.0 Release 16)*. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[48] C. Ranaweera, E. Wong, A. Nirmalathas, C. Jayasundara, and C. Lim, ''5G C-RAN with optical fronthaul: An analysis from a deployment perspective,'' *J. Lightw. Technol.*, vol. 36, no. 11, pp. 2059–2068, Dec. 13, 2017.

[49] T. A. Weiss and F. K. Jondral, ''Spectrum pooling: An innovative strategy for the enhancement of spectrum efficiency,'' *IEEE Commun. Mag.*, vol. 42, no. 3, pp. S8–14, Mar. 2004.

[50] M. Bigdeli and B. Abolhassani, ''A novel cooperative spectrum sharing algorithm based on optimal cognitive radio user selection,'' *Int. J. Commun., Netw. Syst. Sci.*, vol. 5, no. 1, pp. 7–16, 2012.

[51] M. Awais, A. Ahmed, M. Naeem, M. Iqbal, W. Ejaz, A. Anpalagan, and H. S. Kim, ''Efficient joint user association and resource allocation for cloud radio access networks,'' *IEEE Access*, vol. 5, pp. 1439–1448, 2017.

[52] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini, ''Distributed hybrid scheduling in multi-cloud networks using conflict graphs,'' *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 209–224, Jan. 2018.

[53] T. Erlebach and F. C. Spieksma, ''Simple algorithms for a weighted interval selection problem,'' in *Proc. Int. Symp. Algorithms Comput.* Cham, Switzerland: Springer, 2000, pp. 228–240.

[54] H. Chetto and M. Chetto, ''Some results of the earliest deadline scheduling algorithm,'' *IEEE Trans. Softw. Eng.*, vol. 15, no. 10, pp. 1261–1269, Oct. 1989.

[55] J. H. Anderson and A. Srinivasan, ''Early-release fair scheduling,'' in *Proc. 12th Euromicro Conf. Real-Time Syst. Euromicro RTS*, Jun. 2000, pp. 35–43.

[56] Y. Xu, F. Yin, W. Xu, J. Lin, and S. Cui, ''Wireless traffic prediction with scalable Gaussian process: Framework, algorithms, and verification,'' *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1291–1306, Jun. 2019.

[57] *5G; NR; Physical Channels and Modulation*, document TS 38.211, version 16.6.0, Release 16, 3GPP, 2021.

[58] S. Sun, T. S. Rappaport, S. Rangan, T. A. Thomas, A. Ghosh, I. Z. Kovacs, I. Rodriguez, O. Koymen, A. Partyka, and J. Jarvelainen, ''Propagation path loss models for 5G urban micro- and macro-cellular scenarios,'' in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–6.

[59] W. Ning, X. Huang, K. Yang, F. Wu, and S. Leng, ''Reinforcement learning enabled cooperative spectrum sensing in cognitive radio networks,'' *J. Commun. Netw.*, vol. 22, no. 1, pp. 12–22, Feb. 2020.

[60] S. Pudlewski, N. Cen, Z. Guan, and T. Melodia, ''Video transmission over lossy wireless networks: A cross-layer perspective,'' *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 6–21, Feb. 2015.

[61] D. Casini, A. Biondi, and G. Buttazzo, ''Handling transients of dynamic real-time workload under EDF scheduling,'' *IEEE Trans. Comput.*, vol. 68, no. 6, pp. 820–835, Jun. 2019.

[62] M. Chen, C. Hwang, and H. Wang, ''Analysis of the queue service probability for the EDF scheduling algorithm,'' in *Proc. 30th Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Mar. 2016, pp. 960–963.

[63] K. Wang and Y. Cen, ''Real-time partitioned scheduling in cloud-RAN with hard deadline constraint,'' in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[64] I.-H. Hou and R. Singh, ''Scheduling of access points for multiple live video streams,'' in *Proc. 14th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jul. 2013, pp. 267–270.

[65] F. C. R. Spieksma, ''On the approximability of an interval scheduling problem,'' *J. Scheduling*, vol. 2, no. 5, pp. 215–227, Sep. 1999.

[66] O. Göbel, ''Online resource allocation on stochastic input models,'' Ph.D. dissertation, Mathematik, Informatik und Naturwissenschaften, Universitätsbibliothek der RWTH Aachen, Aachen, Germany, 2016.

MOHAMMAD BIGDELI received the B.Sc. degree in electrical engineering and computer engineering and the M.Sc. degree (Hons.) in telecommunications engineering from the Iran University of Science and Technology, Tehran, Iran, in 2009 and 2012, respectively, where he is currently pursuing the Ph.D. degree with the School of Electrical Engineering. His current research interests include big data, information theory, channel coding, signal processing for communications, wireless networking, and cooperative communications. He has served as a reviewer for several IEEE journals and major conferences.

BAHMAN ABOLHASSANI was born in Tehran, Iran. He received the B.Sc. degree from the Iran University of Science and Technology (IUST), Tehran, and the M.Sc. and Ph.D. degrees from the University of Saskatchewan, Saskatoon, SK, Canada, all in electrical engineering. He was an Instrumentation Engineer with the College of Water and Power Technology, Iranian Ministry of Energy, for three years. Then, he worked as a Communication System Engineer in a number of private and government companies. He joined the School of Electrical Engineering, IUST, where he is currently an Associate Professor. He served as the Dean of the School of Electrical Engineering and an Associate Dean for Research. He also served as a Sessional Lecturer at the University of Saskatchewan. His research interests include the fields of wireless communication systems, network planning, spread spectrum, cognitive radio networks, resource allocation, VANETs, and optimization of large systems.

SHAHROKH FARAHMAND was born in Tehran, Iran, in 1980. He received the B.Sc. degree in electrical engineering from the Sharif University of Technology, in 2003. Then, he pursued his graduate studies in United States where he obtained his M.Sc. degree in 2006 and Ph.D. degree in 2011 both from University of Minnesota (UMN) at Twin-Cities in the field of communications and signal processing. From 2011 to 2014, he was with the Iran Research Organization for Science and Technology (IROST), where he held a research faculty position. Since 2018, he has been with the Electrical Engineering Department, Iran University of Science and Technology (IUST), where he is currently an Assistant Professor. His general research interests include applications of statistical signal processing, optimization, and machine learning in communications and networking. His current focus is on Internet of things (IoT), intelligent reflecting surfaces (IRS), massive MIMO, and ultra-wideband impulse radio (UWB-IR).

CHINTHA TELLAMBURA (Fellow, IEEE) received the B.Sc. degree (Hons.) in electrical and electronic engineering from the University of Moratuwa, Sri Lanka, the M.Sc. degree in electrical engineering from the King's College, University of London, and the Ph.D. degree in electrical engineering from the University of Victoria, Canada.

He was with Monash University, Australia, from 1997 to 2002. He is a Professor with the Department of Electrical and Computer Engineering, University of Alberta. He has authored or coauthored over 600 journals and conference papers, demonstrating his expertise in the field. His exceptional scholarly contributions have earned him an impressive H-index of 81 according to Google Scholar. He has made significant contributions to various areas of research, including future wireless networks, machine learning for wireless networks, and signal processing. Recognizing his outstanding accomplishments, he was elected as an IEEE Fellow, in 2011, for his noteworthy contributions to physical layer wireless communication theory. In 2017, he was further honored as a fellow of the Canadian Academy of Engineering, a testament to his exceptional achievements. His dedication and expertise have been acknowledged through prestigious awards, including the Best Paper Awards in the Communication Theory Symposium, in 2012, the IEEE International Conference on Communications (ICC) held in Canada, in 2017, and another ICC in France. Moreover, he has been honored with the esteemed McCalla Professorship and the Killam Annual Professorship by the University of Alberta, further underscoring his significant impact on academia. He has also played a vital role in editorial responsibilities within the IEEE community. He served as an Editor for IEEE Transactions on Communications, from 1999 to 2011, and IEEE Transactions on Wireless Communications, from 2001 to 2007. In the latter role, he held the position of Area Editor of *Wireless Communications Systems and Theory*, from 2007 to 2012, contributing to the advancement of the field through his editorial expertise.