Energy-Spectrum Efficient Content Distribution in Fog-RAN Using Rate-Splitting, Common Message Decoding, and 3D-Resource Matching

Md. Zoheb Hassan, Md. Jahangir Hossain, *Senior Member, IEEE*,
Julian Cheng, *Senior Member, IEEE*, and Victor C. M. Leung, *Fellow, IEEE*

Abstract

Multi-objective resource allocations are studied for a multi-level edge-caching enabled fog-radio access network. Notably, joint maximization of the energy-efficiency (EE) and spectrum-efficiency (SE) and interference management for the content distribution phase are investigated. In the envisioned system, the popular contents are cached at both the fog access point (F-AP) and fog user equipments (F-UEs), and the content-requesting devices are grouped into multiple device-clusters based on their locations. Using a rate-splitting with common message decoding based transmission strategy, each device-cluster is simultaneously served by a suitably selected F-UE and the F-AP over the same radio resource blocks. To maximize system EE and SE of the content distribution phase jointly, a multi-objective optimization problem (MOOP) is formulated. By using the ϵ -constraint method, the proposed MOOP is converted to an EE-SE trade-off optimization problem. Leveraging iterative function evaluation based power control and generalized 3D-resource matching, a novel algorithm is devised to obtain near-optimal Pareto front for the proposed MOOP. By inspecting solutions over the Pareto optimal front, a suitable operating EE-SE pair and the corresponding resource allocation variables are obtained. A low-complexity algorithm to obtain a suitable EE-SE trade-off is also devised. The conducted simulations confirm that the proposed algorithms achieve substantial improvement of system EE and SE over the baseline schemes.

I. INTRODUCTION

Leveraging distributed signal processing, resource allocation, and popular content caching at the network edge, fog-radio access network (F-RAN) presents a revolutionary paradigm to alleviate burden over fronthaul network in the beyond 5G architecture [1]. In a multi-level edge-caching (EC) enabled system, the content distribution phase is significantly important to ensure throughput and diverse latency requirements of the end-users. Particularly, in a multi-level EC

Md. Zoheb Hassan, Md. Jahangir Hossain and Julian Cheng are with School of Engineering, The University of British Columbia, Kelowna, BC, Canada (e-mail: zohassan@mail.ubc.ca, {jahangir.hossain, julian.cheng}@ubc.ca). Victor C. M. Leung is with the College of Computer Science, Shenzhen University, Shenzhen 518060, China, and also with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada (e-mail: vleung@ieee.org).

enabled F-RAN system, the distribution of contents exploits both cellular and device-to-device (D2D) communication links [2]. As a result, well-designed resource allocation plays a vital role for the content distribution phase in multi-level EC enabled F-RAN system [3], [4].

Interference among the content sharing links severely degrades the content distribution performance. The conventional approach of scheduling the content sharing links over orthogonal resources strictly limits number of the supported users in an EC enabled network, and thus, it is incapable to satisfy the increasing data demand of the overly congested 5G network. A promising way to improve the resource utilization for content distribution is to group the end-users in multiple clusters where the members of each cluster are scheduled to receive contents over the same resources. However, clustering inevitably increases the interference in the network and requires a complex interference management scheme. Meanwhile, to motivate user devices to participate in EC and satisfy end-users' data demands, energy-efficiency (EE) and spectrum-efficiency (SE) of the content distribution phase must be improved [5]. Nevertheless, joint maximization of EE and SE is challenging since we need to maximize two conflicting objectives, and interference management in a network with multiple clusters is an NP-hard problem. This work aims to address both of these challenges by designing novel resource allocations for the content distribution phase in multi-level EC enabled F-RAN system.

A. Related Work

In addition to improving throughput and latency, EC also improves the overall EE of a dense network. By proactively caching the popular contents at the network edge, EC significantly reduces the transmission over backhaul network, and thus, reduces backhaul energy consumption [6]. The enhancement of EE is attributed by optimizing both the content placement and content distribution phases [7]. Particularly, the proximity gain of an optimized D2D-assisted content distribution phase can be leveraged to reduce device energy consumption and improve the overall EE [8]. However, devices need to be provided with certain incentive for participating in caching and distribution of the popular contents. Using game theory, the authors in [9] designed optimized caching strategy and incentive mechanism to encourage the devices to participate in the EC system. The authors in [10] proposed optimized proactive caching strategy and transmission power control to offload the contents from the centralized content server via D2D links while reducing energy consumption of the offloading process. Clustering of the devices was also exploited to improve EE of the content distribution process. The authors in [11] proposed a novel D2D-clustering based content distribution framework for improving EE of the network. In

[12], novel unicast and multicast based D2D-assisted content distribution schemes were proposed for improving EE of network. Using the tools from stochastic geometry, the authors obtained optimal caching strategy for D2D-clustering and caching enabled network [13]. F-RAN with multi-level EC further improves the EE of the cellular network [14]. In [15]–[17], the authors studied multi-level EC enabled wireless network where mode selection, transmission power, and file combination were optimized to reduce energy consumption of the content distribution phase.

The works in [6]–[13], [15], [16] have two limitations. First, the authors solely concentrated on optimizing EE of the network. However, optimizing EE does not necessarily lead to an optimized SE, and therefore, it is more meaningful to optimize the trade-off between EE and SE of the network. Second, the authors did not explore mitigation of the interference among the content sharing links. In the most of these works, the interference was treated as a noise and consequently, the resultant system throughput is sub-optimal. Our proposed work overcomes the aforementioned limitations by jointly investigating the EE-SE trade-off optimization and interference management problems for a multi-level EC enabled F-RAN system. We also emphasize that the EE-SE trade-off optimization and interference management for various wireless networks were studied in a disjoint manner in [18]–[21]. Accordingly, our work is also different from [18]–[21].

Rate-splitting with common message decoding (RS-CMD) has recently been emerged as a robust and resource efficient transmission strategy for mitigating multi-user interference [22]. RS-CMD provides a flexibility of partial interference cancellation and treating partial interference as a noise, and thus, it encompasses the advantages of the well-known multi-user transmission strategies, namely, non-orthogonal multiple access, space-division multiple access, orthogonal multiple access, and multicasting [23]. The problem of SE optimization for RS-CMD system was extensively studied for various wireless networking scenarios, such as, multi-user broadcast system [24], cloud-radio access network [25], cooperative network with user relaying [26], and aerial network [27]. Meanwhile, both EE maximization and EE-SE trade-off optimization for RS-CMD systems were studied in [28]–[30]. Because of the well demonstrated success of RS-CMD in mitigating multi-user interference, in this work, RS-CMD is employed to distribute contents in multi-level EC enabled F-RAN. We emphasize that most of the aforementioned works mainly considered a downlink scenario with single transmitter and multiple receivers, and did not investigate the problem of association between receivers and transmitters. However, the role of association between transmitting and receiving nodes is crucial in a multi-level EC enabled F-RAN. Hence, unlike the prior works, we also study the problem of association between transmitting and receiving nodes to optimize the EE-SE trade-off of RS-CMD enabled system.

B. Contributions

In this work, we develop resource allocations to achieve Pareto optimal trade-off between system EE and SE for the content distribution phase of a multi-level EC enabled F-RAN system. In the envisioned system, the popular contents are cached at both the fog access point (F-AP) and at certain fog user equipments (F-UEs). As a result, the content-requesting (CR) devices can obtain their requested contents either from the F-UEs via D2D-networking or from F-AP. To enhance utilization of the radio resource blocks (RRBs) and manage interference in the network, we propose a novel framework by integrating clustering of the CR devices, RS-CMD in the device-clusters, and 3D-resource matching. We emphasize that the optimization of device-clustering and RS-CMD for interference management have been recently investigated for F-RAN system [31]. However, only D2D-aided content distribution was considered without utilizing multi-level EC, and a simplified device-clustering with two D2D links per cluster was adopted [31]. Essentially, the system models of the present work and [31] are entirely different. Moreover, to the best of our knowledge, this is the first work that optimizes the trade-off between system EE and SE for the RS-CMD integrated and multi-level EC enabled F-RAN system. Our specific contributions are summarized as follows.

- A multi-objective optimization problem (MOOP) is formulated to maximize the EE and SE of the system jointly. Notably, we optimize the power allocation of the transmitting nodes, association between F-UEs and device-clusters, and allocation of the RRBs among the device-clusters. To obtain the Pareto optimal front for the proposed MOOP, we devise the following three-step approach. First, we convert the MOOP to an EE-SE trade-off optimization problem by applying the ϵ -constraint method. Next, we determine the maximum achievable EE and SE of the system. Finally, we iteratively solve the EE-SE trade-off optimization problem to obtain a suitable trade-off between EE and SE of the system.
- To determine the maximum achievable EE and SE of the system, we iteratively solve two sub-problems. The first sub-problem obtains power allocations of the F-AP and F-UEs by using an iterative function evaluation (IFE) approach. Meanwhile, the second sub-problem determines a Pareto-efficient scheduling of RRBs and F-UEs among the device-clusters by solving a generalized 3D-resource matching problem. Two algorithms are proposed to determine the maximum achievable EE and SE of the system.
- To solve the EE-SE trade-off optimization problem iteratively, we develop a novel iterative rate-splitting and resource-matching (I-RSRM) algorithm. I-RSRM can obtain the overall

Fig. 1: Multi-level EC enabled F-RAN with 2 F-UEs, 4 RRBs, and 8 CR devices.

Pareto optimal front of the proposed MOOP by iteratively adjusting the trade-off between EE and SE. Each solution of I-RSRM leads to a certain trade-off between EE and SE of the system. By inspecting these solutions, the operating EE-SE pair is determined, and the required resource allocation variables to achieve the operating EE-SE pair are obtained.

- Despite optimality, I-RSRM has high computational complexity. To reduce the complexity of achieving optimal Pareto front, a low-complexity algorithm, referred as, decomposed rate-splitting and resource-matching (D-RSRM), is developed as well. Extensive simulations are conducted to verify the effectiveness of the proposed I-RSRM and D-RSRM algorithms.

The rest of the manuscript has following organizations. Section II provides overall system model and problem formulation. Maximum achievable EE and SE of the system are obtained in Section III. EE-SE trade-off optimization and the proposed algorithms are discussed in Section IV. Simulation results and concluding remarks are presented in Sections VI and VII, respectively.

II. SYSTEM MODEL

A. System Overview

Consider an F-RAN system having N orthogonal RRBs, one single antenna F-AP, L single antenna F-UEs, and M single antenna CR devices. Fig. 1 illustrates an example of such a network. In F-RAN, F-UEs are the smart UEs with signal processing and storage capability, and consequently, F-UEs also participate in caching the popular contents [3]. We assume that F-AP has the global cache (i.e., the complete set of popular contents) where F-UEs only share a local cache (i.e., a subset of popular contents). We further assume that each F-UE caches identical contents. Let $\mathcal{N} = \{1, 2, \dots, N\}$, $\mathcal{L} = \{1, 2, \dots, L\}$ and $\mathcal{M} = \{1, 2, \dots, M\}$ be the set of available orthogonal RRBs, F-UEs, and CR devices, respectively.

It is assumed that all the CR devices are associated with the same network operator. Hence, based on the locations of the CR devices, the network operator forms multiple non-overlapping device-clusters. To combat inter device-cluster interference in the content distribution phase, each device-cluster is scheduled with certain orthogonal RRBs. The CR devices in a given device-cluster can receive the requested contents either from F-AP or an F-UE. To reduce the signaling and coordination among the F-UEs, each device-cluster is allowed to be associated with maximum one F-UE. Meanwhile, an F-UE can serve multiple device-clusters over the orthogonal RRBs. Moreover, each device-cluster is decomposed into two non-overlapping sets of local cache-hit (LCH) and global cache-hit (GCH) CR devices. The LCH and GCH CR devices receive contents from the associated F-UE and F-AP, respectively, over the same RRBs. To combat the multi-user interference in a given device-cluster, 1-layer RS-CMD is exploited for both LCH and GCH CR devices. The described system runs on a slotted-time basis where the overall time duration is divided into equal and non-overlapping time-slots (TSs). We assume block fading channel where the channels remain constant in a given TS, and can independently vary in different TSs. We also assume that the CR devices can accurately perform successive interference cancellation (SIC), and the locations of CR devices and F-UEs remain fixed for a number of TSs. These assumptions are common in the existing literature of F-RAN [3], [4].

B. Distance Based Device-clustering

Let $\mathbf{x}_i = (x_i, y_i)$ be the 2D-location of the i -th CR device, $\forall i \in \mathcal{M}$. The essence of device-clustering is to cluster M CR devices into K non-overlapping clusters, denoted by $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_K\}$, as such the mean-square distance between the clustered devices and cluster center is minimized. Device-clustering is written as the following optimization problem:

$$\{\mathcal{S}_1, \dots, \mathcal{S}_K\} = \arg \min_{\mathcal{S}, \{\boldsymbol{\mu}\}} \sum_{k=1}^K \sum_{i \in \mathcal{S}_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 \text{ s.t. } \mathcal{S}_k \cap \mathcal{S}_{k' \neq k} = \emptyset, \forall k, k'; \cup_{k=1}^K \mathcal{S}_k = \mathcal{M} \quad (1)$$

where $\boldsymbol{\mu}_k$ is center of the k -th device-cluster, and it is defined as $\boldsymbol{\mu}_k = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \mathbf{x}_i$ with $|\cdot|$ as the cardinality of a set. Eq. (1) is solved by applying the well-known K -means clustering algorithm. K -means clustering is an unsupervised-learning method, and it can obtain local optimal clusters with linear time complexity. For the j -th device-cluster, obtained from solving (1), let $\mathcal{S}_j^{(L)}$ and $\mathcal{S}_j^{(H)}$ be the sets of LCH and GCH CR devices where $\mathcal{S}_j = \mathcal{S}_j^{(L)} \cup \mathcal{S}_j^{(H)}$ and $\mathcal{S}_j^{(L)} \cap \mathcal{S}_j^{(H)} = \emptyset$. To optimally partition $\mathcal{S}_j$ into $\mathcal{S}_j^{(L)}$ and $\mathcal{S}_j^{(H)}$ sets, computationally prohibitive exhaustive search over all possible partitions is required. Hence, we consider the following heuristic approach. Particularly, we denote by $\mathcal{S}_j(i)$ the i -th element of the j -th device-cluster, $D_{0, \mathcal{S}_j(i)}$ the distance

between F-AP and $\mathcal{S}_j(i)$, d_o a given threshold distance, $\mathcal{W}_{\mathcal{S}_j(i)}$ the content requested by $\mathcal{S}_j(i)$, and $\mathcal{W}_L$ the set of contents cached at the F-UEs. The set assignment rule for the i -th element of the j -th device-cluster, $\forall i$, is obtained as

$$\mathcal{S}_j(i) \in \begin{cases} \mathcal{S}_j^{(L)} & \text{if } D_{0,\mathcal{S}_j(i)} > d_o \quad \text{and} \quad \mathcal{W}_{\mathcal{S}_j(i)} \in \mathcal{W}_L \\ \mathcal{S}_j^{(H)} & \text{if } D_{0,\mathcal{S}_j(i)} \leq d_o \quad \text{or} \quad \mathcal{W}_{\mathcal{S}_j(i)} \notin \mathcal{W}_L \end{cases} \quad (2)$$

C. RS-CMD Strategy

We describe RS-CMD strategy for the j -th device-cluster. The LCH and GCH CR devices of the j -th device-cluster are denoted by $\mathcal{S}_j^{(L)} = \{\mathcal{S}_j^{(L)}(e)\}$ and $\mathcal{S}_j^{(H)} = \{\mathcal{S}_j^{(H)}(t)\}$, respectively, where $\mathcal{S}_j^{(L)}(e)$ and $\mathcal{S}_j^{(H)}(t)$ denote the indices of the e -th and t -th LCH and GCH CR devices of the j -th device-cluster, respectively. Here, $e = 1, 2, \dots, |\mathcal{S}_j^{(L)}|$ and $t = 1, 2, \dots, |\mathcal{S}_j^{(H)}|$. We consider that the l -th F-UE and the n -th RRB are assigned to the j -th device-cluster. Hence, the CR devices in the $\mathcal{S}_j^{(L)}$ and $\mathcal{S}_j^{(H)}$ sets receive data from the l -th F-UE and F-AP, respectively, over the n -th RRB. The l -th F-UE first splits the content requested by the e -th LCH CR device, denoted by $\mathcal{W}_{\mathcal{S}_j^{(L)}(e)}$, into two parts $\left\{ \mathcal{W}_{\mathcal{S}_j^{(L)}(e)}^{(c)}, \mathcal{W}_{\mathcal{S}_j^{(L)}(e)}^{(p)} \right\}$, $\forall e \in \mathcal{S}_j^{(L)}$. Subsequently, the l -th F-UE combines and encodes $\left\{ \mathcal{W}_{\mathcal{S}_j^{(L)}(e)}^{(c)} \right\}$ into a common message denoted by s_c . Simultaneously, the l -th F-UE encodes $\left\{ \mathcal{W}_{\mathcal{S}_j^{(L)}(e)}^{(p)} \right\}$, $\forall e \in \mathcal{S}_j^{(L)}$, into $|\mathcal{S}_j^{(L)}|$ private messages denoted by $\{s_e^{(p)}\}$. The signal transmitted by the l -F-UE over the n -th RRB is expressed as $X_{l,n} = \sqrt{P_{l,n}^{C,j}} s_c + \sum_{e=1}^{|\mathcal{S}_j^{(L)}|} \sqrt{P_{l,e,n}^{p,j}} s_e^{(p)}$ where $P_{l,n}^{C,j}$ and $\{P_{l,e,n}^{p,j}\}$ denote the transmission powers assigned for the common and private messages for the LCH CR devices in the j -th device-cluster, respectively. Similarly, F-AP encodes the data requested by the GCH devices into a common message, $\tilde{s}_c$, and $|\mathcal{S}_j^{(H)}|$ private messages, $\{\tilde{s}_t^{(p)}\}$. The signal transmitted by the F-AP over the n -th RRB is expressed as $\tilde{X}_n = \sqrt{Q_n^{C,j}} \tilde{s}_c + \sum_{t=1}^{|\mathcal{S}_j^{(H)}|} \sqrt{Q_{t,n}^{p,j}} \tilde{s}_t^{(p)}$ where $Q_n^{C,j}$ and $\{Q_{t,n}^{p,j}\}$ denote the transmission powers assigned for the common and private messages for the GCH CR devices in the j -th device-cluster, respectively. The received signals at the CR devices are expressed as

$$Y_e^{(n)} = h_{e,n}^{(l)} X_{l,n} + h_{e,n}^{(0)} \tilde{X}_n + n, \forall e \in \mathcal{S}_j^{(L)} \quad (3a)$$

$$\tilde{Y}_t^{(n)} = h_{t,n}^{(0)} \tilde{X}_n + h_{t,n}^{(l)} X_{l,n} + n, \forall t \in \mathcal{S}_j^{(H)} \quad (3b)$$

where $h_{e,n}^{(l)}$ (resp. $h_{t,n}^{(l)}$) is the channel gain between the l -th F-UE and the e -th CR device (resp. the t -th CR device) over the n -th RRB; $h_{e,n}^{(0)}$ (resp. $h_{t,n}^{(0)}$) is the channel gain between F-AP and

$$\gamma_{l,n}^{C,j} = \frac{P_{l,n}^{C,j} |h_{k_j,n}^{(l)}|^2}{\left(\sum_{e=1}^{|\mathcal{S}_j^{(L)}|} P_{l,e,n}^{p,j} \right) |h_{k_j,n}^{(l)}|^2 + \left(Q_n^{C,j} + \sum_{t=1}^{|\mathcal{S}_j^{(H)}|} Q_{t,n}^{p,j} \right) |h_{k_j,n}^{(0)}|^2 + \sigma^2}, \quad (4)$$

$$\tilde{\gamma}_n^{C,j} = \frac{Q_n^{C,j} |h_{\tilde{k}_j,n}^{(0)}|^2}{\left(\sum_{t=1}^{|\mathcal{S}_j^{(H)}|} Q_{t,n}^{p,j} \right) |h_{\tilde{k}_j,n}^{(0)}|^2 + \left(P_{l,n}^{C,j} + \sum_{e=1}^{|\mathcal{S}_j^{(L)}|} P_{l,e,n}^{p,j} \right) |h_{\tilde{k}_j,n}^{(l)}|^2 + \sigma^2}. \quad (5)$$

the e -th CR device (resp. the t -th CR device) over the n -th RRB; and n is the additive white Gaussian noise with variance σ^2 . Along the lines of [3], [4], the channel gains are assumed to be available at both CR devices and transmitting nodes (i.e. F-UE and F-AP).

The CR devices in the $\mathcal{S}_j^{(L)}$ and $\mathcal{S}_j^{(H)}$ sets first decode the common messages transmitted from the l -th F-UE and F-AP, respectively. To determine the scheduled common message rates from the l -th F-UE and F-AP, we denote $k_j = \arg \min_{e \in \mathcal{S}_j^{(L)}} \frac{|h_{e,n}^{(l)}|^2}{\left(Q_n^{C,j} + \sum_{t=1}^{|\mathcal{S}_j^{(H)}|} Q_{t,n}^{p,j} \right) |h_{e,n}^{(0)}|^2 + \sigma^2}$ and $\tilde{k}_j = \arg \min_{t \in \mathcal{S}_j^{(H)}} \frac{|h_{t,n}^{(0)}|^2}{\left(P_{l,n}^{C,j} + \sum_{e=1}^{|\mathcal{S}_j^{(L)}|} P_{l,e,n}^{p,j} \right) |h_{t,n}^{(l)}|^2 + \sigma^2}$. Consequently, for the common messages over the n -th RRB, the l -th F-UE and F-AP target the achievable data rate of the CR devices indexed by k_j and $\tilde{k}_j$, respectively. The common message data rates for the LCH and GCH CR devices are expressed as $R_{l,n}^{C,j} = \log_2 (1 + \gamma_{l,n}^{C,j})$ and $\tilde{R}_n^{C,j} = \log_2 (1 + \tilde{\gamma}_n^{C,j})$, respectively. Here, $\gamma_{l,n}^{C,j}$ and $\tilde{\gamma}_n^{C,j}$ are defined as (4) and (5), respectively, at the top of the current page.

After decoding the common messages, both the LCH and GCH CR devices decode their own private messages while canceling the interference from the decoded common messages by applying SIC, and treating the private messages of other CR devices in the system as noise. The LCH (GCH) CR devices also treat the interference from the common message of the GCH (LCH) CR devices as noise. Thus, the scheduled data rates for the private messages of the e -th LCH and t -th GCH CR devices, $\forall e \in \mathcal{S}_j^{(L)}$ and $\forall t \in \mathcal{S}_j^{(H)}$, are obtained as, respectively, $R_{l,e,n}^{p,j} = \log_2 (1 + \gamma_{l,e,n}^{p,j})$ and $\tilde{R}_{t,n}^{p,j} = \log_2 (1 + \tilde{\gamma}_{t,n}^{p,j})$. Here,

$$\gamma_{l,e,n}^{p,j} = \frac{P_{l,e,n}^{p,j} |h_{e,n}^{(l)}|^2}{\left(\sum_{\substack{\tilde{e}=1 \\ \tilde{e} \neq e}}^{|\mathcal{S}_j^{(L)}|} P_{l,\tilde{e},n}^{p,j} \right) |h_{e,n}^{(l)}|^2 + \left(Q_n^{C,j} + \sum_{t=1}^{|\mathcal{S}_j^{(H)}|} Q_{t,n}^{p,j} \right) |h_{e,n}^{(0)}|^2 + \sigma^2}, \quad (6)$$

$$\tilde{\gamma}_{t,n}^{p,j} = \frac{Q_{t,n}^{p,j} |h_{t,n}^{(0)}|^2}{\left(\sum_{\substack{\tilde{t}=1 \\ \tilde{t} \neq t}}^{|\mathcal{S}_j^{(H)}|} Q_{\tilde{t},n}^{p,j} \right) |h_{t,n}^{(0)}|^2 + \left(P_{l,n}^{C,j} + \sum_{e=1}^{|\mathcal{S}_j^{(L)}|} P_{l,e,n}^{p,j} \right) |h_{t,n}^{(l)}|^2 + \sigma^2}. \quad (7)$$

$$\begin{aligned}
\text{P0:} \quad & \max_{\mathbb{P} \geq 0, \mathbf{Q} \geq 0, \{\mathbf{a}, \mathbf{b}\} \in \{0,1\}, \mu \geq 0} \{ \text{EE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}), \text{SE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}) \} \\
\text{s.t.} \quad & \begin{cases} \text{C1: } \sum_{l=1}^L a_{l,j} = 1; \sum_{j=1}^K a_{l,j} \leq N_R, \forall l \in \mathcal{L}, j \in \{1, 2, \dots, K\} \\ \text{C2: } \sum_{j=1}^K \sum_{l=1}^L b_{n,l,j} = 1; \sum_{n=1}^N \sum_{l=1}^L b_{n,l,j} \geq 1, \forall n \in \mathcal{N}, l \in \mathcal{L}, j \in \{1, 2, \dots, K\} \\ \text{C3: } b_{n,l,j} \leq a_{l,j}, \forall n \in \mathcal{N}, l \in \mathcal{L}, j \in \{1, 2, \dots, K\} \\ \text{C4: } \mathcal{P}_l \leq P_{\max}^{(D)}, \forall l \in \mathcal{L}; \quad \mathcal{P}_{FAP} \leq P_{\max}^{FAP}. \end{cases} \quad (8)
\end{aligned}$$

$$\begin{aligned}
\text{P1:} \quad & \max_{\mathbb{P} \geq 0, \mathbf{Q} \geq 0, \{\mathbf{a}, \mathbf{b}\} \in \{0,1\}, \mu \geq 0} \text{SE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}) \\
\text{s.t.} \quad & \begin{cases} \text{EE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}) \geq \epsilon \\ \text{C1, C2, C3, C4.} \end{cases} \quad (9)
\end{aligned}$$

Using (4)-(7), we obtain the total scheduled data rates for the LCH and GCH CR devices in the j -th device-cluster, respectively, as $\mathcal{R}_{l,n,j}^{(L)} = R_{l,n}^{C,j} + \sum_{e \in \mathcal{S}_j^{(L)}} R_{l,e,n}^{p,j}$ and $\mathcal{R}_{n,j}^{(G)} = \tilde{R}_n^{C,j} + \sum_{t \in \mathcal{S}_j^{(H)}} \tilde{R}_{t,n}^{p,j}$. Therefore, the overall throughput of the j -th device-cluster, given that the l -th F-UE and the n -th RRB are assigned to the j -th device-cluster, is obtained as, $\mathcal{T}_j^{(l,n)} = \mathcal{R}_{l,n,j}^{(L)} + \mathcal{R}_{n,j}^{(G)}$.

D. Problem Formulation

We introduce two binary variables, $a_{l,j} \in \{0,1\}$ and $b_{n,l,j} \in \{0,1\}$ such that $a_{l,j} = 1$ if the l -th F-UE is associated with the j -th device-cluster and $a_{l,j} = 0$ otherwise; and $b_{n,l,j} = 1$ if the n -th RRB is assigned to both the l -th F-UE and the j -th device-cluster and $b_{n,l,j} = 0$ otherwise. Hence, the total power consumption of the l -th F-UE is obtained as $\mathcal{P}_l = P_c + \frac{1}{\varepsilon} \sum_{j=1}^K \sum_{n=1}^N a_{l,j} b_{n,l,j} \left(P_{l,n}^{C,j} + \sum_{e \in \mathcal{S}_j^{(L)}} P_{l,e,n}^{p,j} \right)$ where P_c is the constant circuitry power consumption and ε is the power amplifier efficiency. Meanwhile, the total power consumption of the F-AP is obtained as $\mathcal{P}_{FAP} = P_c + \frac{1}{\varepsilon} \sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \left(Q_n^{C,j} + \sum_{t \in \mathcal{S}_j^{(H)}} Q_{t,n}^{p,j} \right)$. Thereby, the system EE and SE are obtained as, respectively, $\text{EE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}) = \frac{\sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)}}{\sum_{l \in \mathcal{L}} \mathcal{P}_l + \mathcal{P}_{FAP}}$, and $\text{SE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}) = \sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)}$. We formulate an optimization problem, given by P0 at the top of current page, to maximize simultaneously the system EE and SE by jointly performing transmission power allocation of the F-AP and F-UEs, association between the F-UEs and device-clusters, and allocation of RRBs among the device-clusters. In P0, C1 implies that each device-cluster will be associated with only one F-UE whereas a given F-UE can be associated with maximum N_R device-clusters; C2 implies the orthogonal allocation of the RRBs among the device-clusters; C3 implies that the n -th RRB can be assigned to both the l -th F-UE and the j -th device-cluster if and only if the l -th F-UE is associated with the j -th device-cluster; and C4 provides the transmission power allocation constraints for F-UE and FAP with $P_{\max}^{(D)}$ and $P_{\max}^{FAP}$ as the maximum transmission power of the F-UE and F-AP, respectively.

P0 is a non-convex MOOP. The ϵ -constraint method can obtain Pareto-optimal solution for a non-convex MOOP [32]. Consequently, we employ the ϵ -constraint method to obtain a Pareto-optimal solution to P0. By introducing a lower-bound for the system EE, denoted by ϵ , a single objective formulation of P0 is obtained as P1 at the top of the previous page. Note that P1 is an EE-SE trade-off optimization problem since each solution to P1 provides a trade-off between system EE and SE. The following lemma justifies the efficiency of the problem formulation P1,

Lemma 1: *P1 provides a weakly Pareto-optimal solution to P0.*

Proof: The proof is provided in Appendix A.

It is obvious that different values of ϵ will lead to different solutions to P1, and Lemma 1 infers that each of these solutions will be weakly Pareto-optimal to P0. Hence, to solve P0 optimally, we need to obtain overall set of the Pareto-optimal solutions by solving P1 for different values of ϵ and select the solution providing a suitable EE-SE trade-off. However, to do so, the maximum achievable EE and SE of the system are required [19]. In what follows, we first obtain the maximum achievable EE and SE of the system, and afterwards, we solve P1 by varying ϵ .

III. EE-OPTIMAL AND SE-OPTIMAL SOLUTIONS

To obtain maximum EE and SE of the system, we formulate two optimization problems as

$$\text{P2.1 : } \max_{\mathbb{P} \geq \mathbf{0}, \mathbb{Q} \geq \mathbf{0}, \{\mathbf{a}, \mathbf{b}\} \in \{0,1\}} \text{EE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}) \quad \text{s.t.} \quad \text{C1, C2, C3, C4} \quad (10)$$

$$\text{P2.2 : } \max_{\mathbb{P} \geq \mathbf{0}, \mathbb{Q} \geq \mathbf{0}, \{\mathbf{a}, \mathbf{b}\} \in \{0,1\}} \text{SE}(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b}) \quad \text{s.t.} \quad \text{C1, C2, C3, C4.} \quad (11)$$

According to the well-known Dinkelbach theorem, if $(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b})$ is a solution to P2.1 and η is the maximum EE of the system, the following condition must be satisfied.

$$\sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)} - \eta \left(\sum_{l \in \mathcal{L}} \mathcal{P}_l + \mathcal{P}_{FAP} \right) = 0. \quad (12)$$

Hence, solving P2.1 is equivalent to finding a resource allocation tuple as such the objective function of the following parameterized optimization problem approaches zero

$$\text{P2.3 : } \max_{\substack{\mathbb{P} \geq \mathbf{0}, \mathbb{Q} \geq \mathbf{0}, \\ \{\mathbf{a}, \mathbf{b}\} \in \{0,1\}}} \sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)} - \eta \left(\sum_{l \in \mathcal{L}} \mathcal{P}_l + \mathcal{P}_{FAP} \right) \quad \text{s.t.} \quad \text{C1, C2, C3, C4.} \quad (13)$$

However, in P2.3, η also depends on the resource allocation variables as depicted from (12). Hence, we first solve P2.3 for a given η , and then, we update η as such (12) is satisfied. Moreover,

$$\mathcal{L} = \sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)} - \eta \left(\sum_{l \in \mathcal{L}} \mathcal{P}_l + \mathcal{P}_{FAP} \right) - \sum_{l \in \mathcal{L}} \lambda_l (\mathcal{P}_l - P_{max}^{(D)}) - \nu (\mathcal{P}_{FAP} - P_{max}^{FAP}). \quad (15)$$

the solutions to P2.2 and P2.3 converge when $\eta = 0$. Thereby, we focus on solving P2.3. Since P2.3 is NP-hard, we propose a modular approach to solve P2.3 with reasonable computational complexity. Particularly, we decompose P2.3 into two sub-problems for performing power control and resource assignment in an iterative manner. The detailed analysis is given as follows.

A. F-AP and F-UE power allocation

Considering that the assignment variables $\{\mathbf{a}, \mathbf{b}\}$ are given, P2.3 is simplified as

$$\text{P2.4 : } \max_{\mathbb{P} \geq \mathbf{0}, \mathbb{Q} \geq \mathbf{0}} \sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)} - \eta \left(\sum_{l \in \mathcal{L}} \mathcal{P}_l + \mathcal{P}_{FAP} \right) \quad \text{s.t.} \quad \text{C3, C4.} \quad (14)$$

Although P2.4 is a non-convex power allocation problem, high quality approximate optimal solutions to P2.4 can be obtained through Lagrangian relaxation technique [33]. The Lagrangian function corresponding to P2.4 is expressed as (15) at the top of the current page. Here, $\{\lambda_l\}$ and ν are the Lagrangian multipliers corresponding to C3 and C4 constraints, respectively. Using Lagrangian dual-decomposition, the power allocation problem for the j -th device-cluster, $\forall j = 1, 2, \dots, K$, is expressed as

$$\text{P2.5 : } \left\{ P_{l,n}^{C,j}, P_{l,e,n}^{p,j}, Q_n^{C,j}, Q_{t,n}^{p,j} \right\} = \arg \max_{\mathbb{P} \geq \mathbf{0}, \mathbb{Q} \geq \mathbf{0}} \mathcal{L}_j \quad (16)$$

where $\mathcal{L}_j$ is defined as (17) at the top of the next page. To efficiently solve P2.5, it is decomposed into inner and outer layer problems. Using an iterative approach, the inner layer problem updates the power allocation variables and the outer layer problem updates the Lagrangian multipliers.

1) *Power allocation update:* Let the Lagrangian multipliers be given. Capitalizing the IFE based power control method in [17], the near-optimal power allocations of the j -th device cluster, $\forall j = 1, 2, \dots, K$, are obtained in the following proposition.

Proposition 1: Let $\left\{ \hat{P}_{l,n}^{C,j}, \hat{P}_{l,e,n}^{p,j}, \hat{Q}_n^{C,j}, \hat{Q}_{t,n}^{p,j} \right\}$ be the given power allocations in the k -th iteration. A converged power allocation is obtained by updating the power levels at the $(k+1)$ -th iteration, $\forall k$, according to (18)-(21) given at the top of the next page. Here, the values of $\gamma_{l,n}^{C,j}$, $\tilde{\gamma}_n^{C,j}$, $\gamma_{l,e,n}^{p,j}$, and $\tilde{\gamma}_{t,n}^{p,j}$ are obtained by plugging $\left\{ \hat{P}_{l,n}^{C,j}, \hat{P}_{l,e,n}^{p,j}, \hat{Q}_n^{C,j}, \hat{Q}_{t,n}^{p,j} \right\}$ to (4)-(7).

Proof: The proof is provided in Appendix B.

$$\mathcal{L}_j = \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{l,j,n} \left(\mathcal{T}_j^{(l,n)} - \frac{\eta + \lambda_l}{\varepsilon} \left(P_{l,n}^{C,j} + \sum_{e \in \mathcal{S}_j^{(L)}} P_{l,e,n}^{p,j} \right) - \frac{\eta + \nu}{\varepsilon} \left(Q_n^{C,j} + \sum_{t \in \mathcal{S}_j^{(H)}} Q_{t,n}^{p,j} \right) \right). \quad (17)$$

$$P_{l,n}^{C,j} = \frac{\gamma_{l,n}^{C,j} / (1 + \gamma_{l,n}^{C,j})}{\frac{\eta/\varepsilon + \lambda_l/\varepsilon}{1/\ln 2} + \frac{(\tilde{\gamma}_n^{C,j})^2 |h_{k,j,n}^{(l)}|^2}{(1 + \tilde{\gamma}_n^{C,j}) \hat{Q}_n^{C,j} |h_{k,j,n}^{(0)}|^2} + \sum_{t \in \mathcal{S}_j^{(H)}} \frac{(\tilde{\gamma}_{t,n}^{p,j})^2 |h_{t,n}^{(l)}|^2}{(1 + \tilde{\gamma}_{t,n}^{p,j}) \hat{Q}_{t,n}^{p,j} |h_{t,n}^{(0)}|^2}}, \quad (18)$$

$$Q_n^{C,j} = \frac{\tilde{\gamma}_n^{C,j} / (1 + \tilde{\gamma}_n^{C,j})}{\frac{\eta/\varepsilon + \nu/\varepsilon}{1/\ln 2} + \frac{(\gamma_{l,n}^{C,j})^2 |h_{k,j,n}^{(0)}|^2}{(1 + \gamma_{l,n}^{C,j}) \hat{P}_{l,n}^{C,j} |h_{k,j,n}^{(l)}|^2} + \sum_{e \in \mathcal{S}_j^{(L)}} \frac{(\gamma_{l,e,n}^{p,j})^2 |h_{e,n}^{(0)}|^2}{(1 + \gamma_{l,e,n}^{p,j}) \hat{P}_{l,e,n}^{p,j} |h_{e,n}^{(l)}|^2}}, \quad (19)$$

$$P_{l,e,m}^{p,j} = \frac{\gamma_{l,e,n}^{p,j} / (1 + \gamma_{l,e,n}^{p,j})}{\frac{\eta + \lambda_l}{\varepsilon/\ln 2} + \frac{(\gamma_{l,n}^{C,j})^2}{(1 + \gamma_{l,n}^{C,j}) \hat{P}_{l,n}^{C,j}} + \sum_{\substack{e' \in \mathcal{S}_j^{(L)} \\ e' \neq e}} \frac{(\gamma_{l,e',n}^{p,j})^2}{(1 + \gamma_{l,e',n}^{p,j}) \hat{P}_{l,e',n}^{p,j}} + \frac{(\tilde{\gamma}_n^{C,j})^2 |h_{k,j,n}^{(l)}|^2}{(1 + \tilde{\gamma}_n^{C,j}) \hat{Q}_n^{C,j} |h_{k,j,n}^{(0)}|^2} + \sum_{t \in \mathcal{S}_j^{(H)}} \frac{(\tilde{\gamma}_{t,n}^{p,j})^2 |h_{t,n}^{(l)}|^2}{(1 + \tilde{\gamma}_{t,n}^{p,j}) \hat{Q}_{t,n}^{p,j} |h_{t,n}^{(0)}|^2}}, \quad (20)$$

$$Q_{t,n}^{p,j} = \frac{\tilde{\gamma}_{t,n}^{p,j} / (1 + \tilde{\gamma}_{t,n}^{p,j})}{\frac{\eta + \nu}{\varepsilon/\ln 2} + \frac{(\tilde{\gamma}_n^{C,j})^2}{(1 + \tilde{\gamma}_n^{C,j}) \hat{Q}_n^{C,j}} + \sum_{\substack{t' \in \mathcal{S}_j^{(H)} \\ t' \neq t}} \frac{(\tilde{\gamma}_{t',n}^{p,j})^2}{(1 + \tilde{\gamma}_{t',n}^{p,j}) \hat{Q}_{t',n}^{p,j}} + \frac{(\gamma_{l,n}^{C,j})^2 |h_{k,j,n}^{(0)}|^2}{(1 + \gamma_{l,n}^{C,j}) \hat{P}_{l,n}^{C,j} |h_{k,j,n}^{(l)}|^2} + \sum_{e \in \mathcal{S}_j^{(L)}} \frac{(\gamma_{l,e,n}^{p,j})^2 |h_{e,n}^{(0)}|^2}{(1 + \gamma_{l,e,n}^{p,j}) \hat{P}_{l,e,n}^{p,j} |h_{e,n}^{(l)}|^2}}. \quad (21)$$

2) *Lagrangian multiplier update:* Using (18)-(21), the power consumption of the l -th F-UE and F-AP are calculated as $\hat{\mathcal{P}}_l$ and $\hat{\mathcal{P}}_{FAP}$, respectively. Leveraging the well known sub-gradient method, at the $(k+1)$ -th iteration, the Lagrangian multipliers are updated as

$$\lambda_l[k+1] = \max \left[\lambda_l[k] + \zeta_1(k) \left(\hat{\mathcal{P}}_l - P_{\max}^{(D)} \right), 0 \right], \forall l \in \mathcal{L} \quad (22a)$$

$$\nu[k+1] = \max \left[\nu[k] + \zeta_2(k) \left(\hat{\mathcal{P}}_{FAP} - P_{\max}^{FAP} \right), 0 \right] \quad (22b)$$

where $\{\zeta_1(k), \zeta_2(k)\}$ are the step sizes at the k -th iteration, and for convergence, the step sizes are selected as such $\sum_{k=1}^{\infty} \zeta_i(k) = \infty$ and $\sum_{k=1}^{\infty} \zeta_i^2(k) < \infty$, $i = 1, 2$ are satisfied.

3) *Algorithm development:* We propose Algorithm 1 to obtain power allocation in the j -th device-cluster. The convergence and implementation of Algorithm 1 are discussed as follows

Proposition 2: *Algorithm 1 converges to a stable and local optimal solution to P2.4.*

Proof: The proof is provided in Appendix C.

Implementation: For a given device-cluster, the required computational complexity of Algorithm 1 is $\mathcal{O}(LN)$. Using the following steps, the power allocations in the device-clusters can

be determined in a decentralized manner. Particularly, at the beginning of the TS, the CR devices broadcast pilots to both F-AP and F-UE. By decoding these pilots, F-AP and F-UE estimate the received channel gains at the CR devices, and share these channel gains with each other. Next, both F-AP and F-UE initialize power allocations over the assigned RRBs, and share these power levels with each other. Subsequently, by exploiting the received channel gains and power levels, both F-AP and F-UE identify the CR devices whose rates are targeted during transmission of the common message(s). Finally, by using the Steps 5-7 of Algorithm 1, both F-AP and F-UE update their corresponding power levels and Lagrangian multipliers. The last three steps are iteratively repeated until the objective function value of P2.4 converges. After the convergence, F-UE informs the F-AP about its total scheduled data rate and transmission power.

Algorithm 1 Power Allocation Algorithm for the j -th Device-cluster

- 1: **Input:** Assignment variables $\{a_{l,j}\}$ and $\{b_{n,l,j}\}$ and EE η .
 - 2: **Initialize** $\hat{P}_{l,n}^{C,j} = \hat{Q}_n^{C,j} = \hat{P}_{l,e,n}^{p,j} = \hat{Q}_{t,n}^{C,j} = P_0; \forall l \in \mathcal{L}, n \in \mathcal{N}, e \in \mathcal{S}_j^{(L)}, t \in \mathcal{S}_j^{(H)}$.
 - 3: **Initialize** Lagrangian multipliers $\lambda_l, \forall l \in \mathcal{L}$ and ν ; iteration index $k = k + 1$.
 - 4: **repeat**
 - 5: Update $\{P_{l,n}^{C,j}, P_{l,e,n}^{p,j}, Q_n^{C,j}, Q_{t,n}^{p,j}\}, \forall l \in \mathcal{L}, n \in \mathcal{N}, e \in \mathcal{S}_j^{(L)}, t \in \mathcal{S}_j^{(H)}$, by using (18)-(21).
 - 6: Update $\{\lambda_l\}$ and ν by using (22a) and (22b), respectively.
 - 7: Set $\hat{P}_{l,n}^{C,j} = P_{l,n}^{C,j}, \hat{Q}_n^{C,j} = Q_n^{C,j}, \hat{P}_{l,e,n}^{p,j} = P_{l,e,n}^{p,j}$, and $\hat{Q}_{t,n}^{C,j} = Q_{t,n}^{C,j}, \forall l \in \mathcal{L}, n \in \mathcal{N}, e \in \mathcal{S}_j^{(L)}, t \in \mathcal{S}_j^{(H)}$; $k = k + 1$
 - 8: **until** $k > k_{max}$ or Convergence of the objective value function of P2.4.
-

B. 3D-Resource Matching

1) *Algorithm development:* For a given set of power allocation variables, P2.3 is written as

$$\text{P2.6 : } \max_{\{a,b\} \in \{0,1\}} \sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \hat{U}_{l,n}^{(j)} \quad \text{s.t.} \quad \text{C1, C2} \quad (23)$$

where $\hat{U}_{l,n}^{(j)}$ is a function of the power allocations obtained from Algorithm 1, and it is given by

$$\hat{U}_{l,n}^{(j)} = \mathcal{T}_j^{(l,n)} \left(P_{l,n}^{C,j}, P_{l,e,n}^{p,j}, Q_n^{C,j}, Q_{t,n}^{p,j} \right) - \frac{\eta}{\varepsilon} \left(P_{l,n}^{C,j} + \sum_{e \in \mathcal{S}_j^{(L)}} P_{l,e,n}^{p,j} + Q_n^{C,j} + \sum_{t \in \mathcal{S}_j^{(H)}} Q_{t,n}^{p,j} \right). \quad (24)$$

To solve P2.6, we model the problem of assigning F-UEs and RRBs among the device-clusters as a 3D-resource matching problem. The rules of matching and preference-orders of the device-clusters, F-UEs, and RRBs are defined as follows.

Definition 1: A 3D-matching is defined as a function $\Phi(\cdot, \cdot)$ that maps elements of the following sets, $\mathcal{S} \times \mathcal{L}$, $\mathcal{S} \times \mathcal{N}$, and $\mathcal{L} \times \mathcal{N}$, into the elements of following sets $\mathcal{N}$, $\mathcal{L}$, and $\mathcal{S}$, respectively. For all $\{j, l, n\} \in \{\mathcal{S}, \mathcal{L}, \mathcal{N}\}$, such a mapping satisfies the following conditions.

- 1) $\Phi(j, n) \in \mathcal{L}$ and $\Phi(j, n) = \Phi(j, \tilde{n})$, $\forall n, \tilde{n} \in \mathcal{N}, n \neq \tilde{n}$.
- 2) $\Phi(l, n) \in \mathcal{S}$, $\Phi(l, n) \neq \Phi(\tilde{l}, \tilde{n})$, $\forall n \neq \tilde{n}, n, \tilde{n} \in \mathcal{N}$ and $\forall l \neq \tilde{l}, l, \tilde{l} \in \mathcal{L}$.
- 3) $|\Phi(l, n)| \leq 1$ and $|\cup_{n \in \mathcal{N}_{sc}} \Phi(l, n)| \leq N_R$.
- 4) $\Phi(j, l) \subset \mathcal{N}$, $\Phi(j, l) \cap \Phi(\tilde{j}, l) = \emptyset$, $\Phi(j, l) \cap \Phi(\tilde{j}, \tilde{l}) = \emptyset$, $\forall j, \tilde{j} \in \mathcal{S}, l, \tilde{l} \in \mathcal{L}, j \neq \tilde{j}, l \neq \tilde{l}$.
- 5) $\Phi(j, l) = \emptyset \iff \nexists n, \Phi(j, n) = l$
- 6) $\Phi(j, n) = l \iff \Phi(l, n) = j \iff \Phi(j, l) = n$.

Condition 1) implies that if the n -th and $\tilde{n}$ -th RRBs are matched with the j -th device-cluster, then both the (j, n) and $(j, \tilde{n})$ pairs will be matched with the same F-UE. Condition 2) implies that if the n -th and $\tilde{n}$ -th RRBs are matched with the l -th and $\tilde{l}$ -th F-UEs, respectively, then the (l, n) and $(\tilde{l}, \tilde{n})$ pairs must be matched with different device-clusters. Condition 3) implies that the (l, n) pair can be matched with maximum one device-cluster, and the total number of matched device-clusters with the l -th F-UE over all the RRBs can not exceed N_R . Condition 4) implies that two different device-clusters will always be matched with orthogonal sets of RRBs. Finally, Conditions 5) and 6) imply that the j -th device-cluster, the l -th F-UE, and the n -th RRB will be matched if and only if any of these two elements are matched with the third element.

Definition 2: To maximize the objective function value of P2.6, the j -th device-cluster prefers the (l, n) pair over the (l', n') pair if the following condition is satisfied

$$(l, n) \succ_j (l', n') \iff \hat{U}_{l,n}^{(j)} > \hat{U}_{l',n'}^{(j)} \quad \forall l, l' \in \mathcal{L}, \forall n, n' \in \mathcal{N}. \quad (25)$$

The l -th F-UE prefers the (j, n) pair over the (j', n') pair if the following condition is satisfied

$$(j, n) \succ_l (j', n') \iff \hat{U}_{l,n}^{(j)} > \hat{U}_{l,n'}^{(j')} \quad \forall j, j' \in \mathcal{S}, \forall n, n' \in \mathcal{N}. \quad (26)$$

The n -th RRB prefers the (j, l) pair over the (j', l') pair if the following condition is satisfied

$$(j, l) \succ_n (j', l') \iff \hat{U}_{l,n}^{(j)} > \hat{U}_{l',n}^{(j')} \quad \forall j, j' \in \mathcal{S}, \forall l, l' \in \mathcal{L}. \quad (27)$$

In this matching problem, as the players are matched with their preferred pairs according to (25)-(27), either the system SE is increased or the power consumption of the system is reduced. However, a global optimal 3D-matching is intractable. We obtain a Pareto-efficient 3D-matching where each device-cluster is satisfied with its matched resources, and no device-cluster can

unilaterally improve its utility. The developed matching has three different phases as follows.

F-UE matching (FM) phase: In FM phase, the device-clusters are matched with suitable F-UEs as such a single device-cluster is matched with only one F-UE and any F-UE is matched with maximum N_R device-clusters. In this phase, two different sets, namely, UM_D and UM_R are maintained for the unmatched device-clusters and F-UEs capable of accepting device-clusters, respectively. At each iteration of the FM phase, an *eligibility score*, D_j , is calculated for $\forall j \in UM_D$, and it provides the eligibility of the unmatched device-clusters to be matched with their most preferred F-UE in the UM_R set. The unmatched device-cluster having maximum eligibility score is matched with its most preferred F-UE in the UM_R set based on (25), and such a device-cluster is removed from the UM_D set. The UM_R set is updated as well. The FM phase is continued until all the device-clusters matched with certain F-UEs.

RRB matching (RM) phase: In RM phase, the remaining RRBs, which are not matched with any device-cluster, are allocated among the device-clusters. The RM phase is a two-sided many-to-one matching problem where the matched device-cluster and F-UEs in the FM phase constitute one side, and the unmatched RRBs constitute other side. Each unmatched RRB is matched with its most preferred device-cluster and F-UE pair according to (27). The RM phase is continued until all the RRBs are matched.

RRB swapping (RS) phase: In RS phase, a set of RRBs between any two device-clusters are swapped given the overall utility of the both device-clusters is improved. To explain the RS operation, we assume that $\Phi(j, l) = n$ and $\Phi(\tilde{j}, \tilde{l}) = \tilde{n}$. If $\hat{U}_{l, \tilde{n}}^{(j)} + \hat{U}_{\tilde{l}, n}^{(\tilde{j})} > \hat{U}_{l, n}^{(j)} + \hat{U}_{\tilde{l}, \tilde{n}}^{(\tilde{j})}$ is satisfied, the $(n, \tilde{n})$ -th RS are swapped between the j -th and $\tilde{j}$ -th device-clusters. The RS phase is continued until there is no swappable device-cluster pair in the system.

The overall steps of the proposed 3D-resource matching are summarized in Algorithm 2. In Algorithm 2, L_j^* denotes the assigned F-UE to the j -th device-cluster, $\mathcal{N}_j^*$ denotes set of the assigned RRBs to the j -th device-cluster, K_l is the number of device-clusters assigned to the l -th F-UE, and UM_N is the set of the unmatched RRBs.

2) *Property Analysis of Algorithm 2:* In what follows, we analyze the properties of Algorithm 2 in terms of Pareto-efficiency, local optimality, and computational complexity.

Proposition 3: *The final matching obtained by Algorithm 2, Φ^* , is Pareto-efficient.*

Proof: The proof is provided in Appendix D.

We obtain solution to P2.6 from Φ^* as follows. Particularly, $\forall j \in \mathcal{S}$, $a_{l,j} = 1$ and $b_{n,l,j} = 1$ if $\Phi^*(j, n) = l$, $\Phi^*(l, n) = j$, $\Phi^*(j, l) = n$, and $a_{l,j} = 0$ and $b_{n,l,j} = 0$, otherwise. The following proposition establishes the local optimality of the obtained solutions.

Proposition 4: For the given power allocations, Φ^* provides a local-optimal solution to P2.6.

Proof: The proof is provided in Appendix E.

Algorithm 2 3D-Resource Matching Algorithm

- 1: **Input:** $\widehat{U}_{l,n}^{(j)}, \forall j \in \mathcal{S}, l \in \mathcal{L}, n \in \mathcal{N}$.
 - 2: **Initialize:** $\mathbf{UM}_D = \mathcal{S}, \mathbf{UM}_N = \mathcal{N}, L_j^* = \emptyset, \mathcal{N}_j^* = \emptyset, K_l = 0, \forall j \in \mathcal{S}, l \in \mathcal{L}$.
 - 3: **while** $\mathbf{UM}_D \neq \emptyset$ **do** (Beginning of FM phase)
 - 4: Determine $\mathbf{UM}_R = \{l \in \mathcal{L} | K_l < N_R\}$.
 - 5: For each $j \in \mathbf{UM}_D$, determine

$$(l_j^*, n_j^*) = \arg \max_{l \in \mathbf{UM}_R, n \in \mathbf{UM}_N} \widehat{U}_{l,n}^{(j)} \quad \text{and} \quad D_j = \widehat{U}_{l_j^*, n_j^*} - \max_{\substack{l \neq l_j^*, l \in \mathbf{UM}_R, \\ n \neq n_j^*, n \in \mathbf{UM}_N}} \widehat{U}_{l,n}^{(j)}$$
 - 6: Select $\hat{j} = \arg \max_{j \in \mathbf{UM}_D} D_j$; assign $L_{\hat{j}}^* \leftarrow l_{\hat{j}}^*$ and $\mathcal{N}_{\hat{j}}^* \leftarrow \mathcal{N}_{\hat{j}}^* \cup \{n_{\hat{j}}^*\}$.
 - 7: Update $\mathbf{UM}_D \leftarrow \mathbf{UM}_D \setminus \{\hat{j}\}$, $\mathbf{UM}_N \leftarrow \mathbf{UM}_N \setminus \{n_{\hat{j}}^*\}$, and $K_{l_{\hat{j}}^*} = K_{l_{\hat{j}}^*} + 1$.
 - 8: **end while** (End of FM phase)
 - 9: **repeat** $\forall n \in \mathbf{UM}_N$ (Beginning of RM phase)
 - 10: Find $j' = \arg \max_{j \in \mathcal{S}} \widehat{U}_{L_{j'}^*, n}^{(j)}, \mathcal{N}_{j'}^* \leftarrow \mathcal{N}_{j'}^* \cup \{n\}$, and $\mathbf{UM}_N \leftarrow \mathbf{UM}_N \setminus \{n\}$.
 - 11: **until** $\mathbf{UM}_N = \emptyset$ (End of RM phase)
 - 12: **repeat** $\forall j \in \mathcal{S}$ and $\forall \tilde{j} \in \mathcal{S} \setminus \{j\}$ (Start of RS phase)
 - 13: Find $\tilde{n}_{\tilde{j}} = \arg \max_{n \in \mathcal{N}_{\tilde{j}}^*} \widehat{U}_{L_j^*, n}^{(j)}$.
 - 14: **if** $\exists n \in \mathcal{N}_{\tilde{j}}^*$ such that the j -th and $\tilde{j}$ -th device-clusters are swappable over the $(n, \tilde{n})$ -th RRBs **then**
 - 15: $\mathcal{N}_{\tilde{j}}^* \leftarrow \mathcal{N}_{\tilde{j}}^* \setminus \{n\} \cup \{\tilde{n}\}$ and $\mathcal{N}_j^* \leftarrow \mathcal{N}_j^* \setminus \{\tilde{n}\} \cup \{n\}$.
 - 16: **end if**
 - 17: **until** The system does not have any swappable device-cluster pair (End of RS phase)
 - 18: **Output:** Final matching Φ^* .
-

Algorithm 2 is implemented at the F-AP. At first, F-AP collects the scheduled rates and transmission powers of all the F-UEs over control channels, and based on this information, F-AP constructs $\widehat{U}_{l,n}^{(j)}$ of (24), $\forall j, l, n$. Thereafter, F-AP executes Algorithm 2 and obtains the assignment of F-UEs and RRBs among the device-clusters. The computational complexity of Algorithm 2 is explained as follows. Particularly, in each iteration of the FM phase in Algorithm 2, a single device-cluster is matched with a F-UE and RRB pair. In each iteration of the FM phase, the metric $\{D_j\}$ is calculated for all the unmatched device-clusters, and the set $\mathbf{UM}_R$ is computed for the F-UEs. Therefore, the required computation of the FM phase is $\sum_{j=1}^{|S|} j + L = \frac{|S|(|S|+1)}{2} + |S|L$. For each unmatched RRB in the RM phase, $\mathcal{O}(|S|)$ operation is needed to execute the Step (10) of the Algorithm 2. Since the total number of unmatched RRBs in the RM phase is $N - |S|$, the total number of required computations is $(N - |S|)|S|$. Finally, for performing RS phase, each device-cluster has to visit the rest of the $|S| - 1$ device-clusters. Therefore, in the worst-case,

total required computations to execute the Steps (14) and (15) in Algorithm 2 is $|S|(|S| - 1)N$. By combining the required computations of the FM, RM, and RS phases and considering the fact that $|S| = K$, the worst-case computational complexity of Algorithm 2 is obtained as $\mathcal{O}(|S|^2N + |S|L) = \mathcal{O}(K^2N + KL)$.

C. Algorithms for Maximizing EE and SE

Capitalizing the aforementioned development on IFE based power control and 3D-resource matching, we develop Algorithm 3 and Algorithm 4 to determine maximum achievable EE and SE of the system, respectively. Both of the proposed algorithms have provable convergence. In both Algorithm 3 and Algorithm 4, the Step 3 is implemented at the network edge by exploiting the signal processing capabilities of F-UEs and the cooperation between F-UEs and F-AP, and the Step 4 is implemented at the F-AP. Hence, the implementation of these algorithms is in accordance with F-RAN principle where some network controls are pushed to the network edge.

Algorithm 3 EE-optimal Resource Allocation Algorithm

- 1: **Initialize:** EE $\eta^{(1)} = \eta_o$, assignment variables $\{\mathbf{a}, \mathbf{b}\}$, iteration index $d = 1$.
 - 2: **repeat**
 - 3: Update the power allocations, $\{\mathbf{P}, \mathbf{Q}\}$, for all the device-clusters by using the value of $\eta^{(d)}$ in Algorithm 1.
 - 4: Update the assignment variables, $\{\mathbf{a}, \mathbf{b}\}$, by using Algorithm 2 while plugging the updated power allocation variables of the previous step in (24).
 - 5: Update $\eta^{(d+1)} = \frac{\sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)}}{\sum_{l \in \mathcal{L}} \mathcal{P}_l + \mathcal{P}_{FAP}}$, $d = d + 1$.
 - 6: **until** $|\eta^{(d+1)} - \eta^{(d)}| \leq \text{Tolerance}$ or Maximum number of iterations is reached.
-

Algorithm 4 SE-optimal Resource Allocation Algorithm

- 1: **Initialize:** Assignment variables $\{\mathbf{a}, \mathbf{b}\}$, iteration index $d = 1$, and set $\eta = 0$.
 - 2: **repeat**
 - 3: Update the power allocations, $\{\mathbf{P}, \mathbf{Q}\}$, for all the device-clusters by using Algorithm 1.
 - 4: Update the assignment variables, $\{\mathbf{a}, \mathbf{b}\}$, by using Algorithm 2 while plugging the updated power allocation variables of the previous step in (24), $d = d + 1$.
 - 5: **until** SE $(\mathbf{P}, \mathbf{Q}, \mathbf{a}, \mathbf{b})$ converges or Maximum number of iterations is reached.
-

IV. EE-SE TRADE-OFF OPTIMIZATION AND PROPOSED ALGORITHMS

A. Algorithm for Optimal Pareto Front

P1 can be solved by applying a Lagrangian optimization technique. Let μ be the Lagrangian multiplier correspond to C5 constraint in P1. The partial Lagrangian of P1 is written as

$$\mathcal{L}^{(2)} = (1 + \mu) \sum_{j=1}^K \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{n,l,j} \mathcal{T}_j^{(l,n)} - \mu \epsilon \left(\sum_{l \in \mathcal{L}} \mathcal{P}_l + \mathcal{P}_{FAP} \right) \quad (28)$$

For a given value of ϵ , a local optimal solution to P1 can be obtained as

$$\text{P3.1} : \{P^*(\epsilon), Q^*(\epsilon), a^*(\epsilon), b^*(\epsilon)\} = \arg \min_{\mu \geq 0} \max_{\substack{P \geq 0, Q \geq 0, \\ \{a, b\} \in \{0, 1\}}} \mathcal{L}^{(2)} \quad \text{s.t.} \quad \text{C1, C2, C3, C4.} \quad (29)$$

Note that, P3.1 and P2.3 have a similar structure except some scalar multipliers. Hence, for the given values of μ and ϵ , Algorithm 1 and Algorithm 2 are applied to obtain $\{P^*(\epsilon), Q^*(\epsilon)\}$ and $\{a^*(\epsilon), b^*(\epsilon)\}$, respectively. To avoid redundancy, the detailed analysis is omitted. Meanwhile, at the optimality of P3.1, $\mu^* > 0$ will be satisfied; otherwise, P3.1 will converge to an SE maximization problem. According to the complementary slackness condition, an optimal value of μ^* must be obtained as such the C5 constraint becomes active. Since P3.1 provides a (local) optimal solution to P1, such a solution is also weakly Pareto optimal to P0. The overall set of the Pareto optimal solutions to P0 is known as the Pareto optimal front of P0. To obtain the Pareto optimal front of P0, we introduce the following proposition.

Proposition 5: Let $\{P_{SE}, Q_{SE}, a_{SE}, b_{SE}\}$ and $\{P_{EE}, Q_{EE}, a_{EE}, b_{EE}\}$ are the SE-optimal and EE-optimal resource allocations. The Pareto optimal front of P0 can be achieved by solving P3.1 while varying ϵ in the range $EE(P_{SE}, Q_{SE}, a_{SE}, b_{SE}) \leq \epsilon \leq EE(P_{EE}, Q_{EE}, a_{EE}, b_{EE})$.

Proof: The proof is provided in Appendix F.

Algorithm 5 Iterative Rate Splitting and Resource Matching (I-RSRM) Algorithm

- 1: Determine EE-optimal and SE-optimal resource allocations, denoted by, $(P_{SE}, Q_{SE}, a_{SE}, b_{SE})$ and $(P_{EE}, Q_{EE}, a_{EE}, b_{EE})$, respectively, by using Algorithm 1 and Algorithm 2.
 - 2: **Initialize:** $\epsilon^{(1)} = EE(P_{SE}, Q_{SE}, a_{SE}, b_{SE})$, iteration index $m = 1$, step size Δ .
 - 3: **while** $EE(P_{EE}, Q_{EE}, a_{EE}, b_{EE}) > \epsilon^{(m)}$ **do**
 - 4: Initialize μ , Assignment variables $\{a, b\}$.
 - 5: **repeat**
 - 6: Update power allocation variables using Algorithm 1 while replacing $\frac{\eta}{\epsilon}$ and $1/\ln 2$ in (18)-(21) with $\frac{\mu\epsilon^{(m)}}{\epsilon}$ and $(1 + \mu)/\ln 2$, respectively.
 - 7: Update the assignment variables, $\{a, b\}$, by plugging the updated power allocation variables of the previous step in Algorithm 2.
 - 8: Update μ using a bi-section method as such constraint C5 in P1 becomes active.
 - 9: **until** $\mathcal{L}^{(2)}$ converges or Maximum number of iterations is reached
 - 10: Record the resource allocations $(P^*(\epsilon^{(m)}), Q^*(\epsilon^{(m)}), a^*(\epsilon^{(m)}), b^*(\epsilon^{(m)}))$. Update $\epsilon^{(m+1)} = \epsilon^{(m)} + \Delta$, $m = m + 1$.
 - 11: **end while**
 - 12: **Output:** Set of Pareto-optimal resource allocations $\{(P^*(\epsilon^{(m)}), Q^*(\epsilon^{(m)}), a^*(\epsilon^{(m)}), b^*(\epsilon^{(m)}))\}$, $\forall m = 1, 2, \dots, N_F$.
-

Based on *Proposition 5*, we propose an I-RSRM algorithm. The overall steps of I-RSRM algorithm are summarized in Algorithm 5. By tuning the value of step size Δ , I-RSRM obtains a near-optimal Pareto optimal front of P0. The total number of the solutions obtained by I-RSRM algorithm is N_F , and each of these solutions provides a certain trade-off between EE and SE of the system. The procedure of selecting a suitable operating EE-SE pair is explained as follows.

The decision of the system to trade-off certain portion of its optimal EE (or optimal SE) can be perceived as a cost paid by the system to achieve certain profit in terms of SE (or EE) gain. We consider that the system allows maximum α percentage loss of its optimal EE or SE. In other words, the system can either trade-off maximum α percentage of the optimal EE or the optimal SE. Here, α is a system defined EE-SE trade-off control parameter. We are interested in the overall profit obtained by the system while paying the cost of the α percentage loss of its optimal EE and α percentage loss of its optimal SE. To this end, we define EE^{**} and SE^{**} by the achievable EE and SE given the system is willing to trade-off α percentage of optimal EE and SE, respectively. Let, EE^* and SE^* be the maximum values of EE and SE in the Pareto optimal front, respectively. We formulate the following two optimization problems.

$$EE^{**} = \max_{\{(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}^*(\epsilon^{(m)}), \mathbf{b}^*(\epsilon^{(m)}))\}} EE(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}^*(\epsilon^{(m)}), \mathbf{b}^*(\epsilon^{(m)})) \quad (30)$$

s.t. $SE(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}^*(\epsilon^{(m)}), \mathbf{b}^*(\epsilon^{(m)})) \geq \left(1 - \frac{\alpha}{100}\right) SE^*.$

$$SE^{**} = \max_{\{(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}^*(\epsilon^{(m)}), \mathbf{b}^*(\epsilon^{(m)}))\}} SE(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}^*(\epsilon^{(m)}), \mathbf{b}^*(\epsilon^{(m)})) \quad (31)$$

s.t. $EE(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}^*(\epsilon^{(m)}), \mathbf{b}^*(\epsilon^{(m)})) \geq \left(1 - \frac{\alpha}{100}\right) EE^*.$

In both (30) and (31), the resource allocation variables, $\{(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}^*(\epsilon^{(m)}), \mathbf{b}^*(\epsilon^{(m)}))\}$, are taken from the output of the I-RSRM algorithm. Next, we define two factors as follows.

$$\chi_{EE} = \frac{EE^{**} - \widehat{EE}_{SE=SE^*}}{\widehat{EE}_{SE=SE^*}} \times 100 \quad (32a)$$

$$\chi_{SE} = \frac{SE^{**} - \widehat{SE}_{EE=EE^*}}{\widehat{SE}_{EE=EE^*}} \times 100 \quad (32b)$$

where $\widehat{EE}_{SE=SE^*}$ is the achievable EE when the system operates with its maximum SE, SE^* , and $\widehat{SE}_{EE=EE^*}$ is the achievable SE when the system operates with its maximum EE, EE^* . Here, χ_{EE} provides the percentage gain of EE when the system allows maximum α percentage loss of its optimal SE, and χ_{SE} provides the percentage gain of SE when the system allows

maximum α percentage loss of its optimal EE. We define the following two metrics.

$$\mathcal{NP}_1 = w_1(\chi_{EE} - \alpha) + w_2 \frac{EE^{**} - (1 - \frac{\alpha}{100}) EE^*}{(1 - \frac{\alpha}{100}) EE^*} + w_3 \frac{(1 - \frac{\alpha}{100}) SE^* - SE^{**}}{SE^{**}} \quad (33a)$$

$$\mathcal{NP}_2 = w_1(\chi_{SE} - \alpha) + w_2 \frac{(1 - \frac{\alpha}{100}) EE^* - EE^{**}}{EE^{**}} + w_3 \frac{SE^{**} - (1 - \frac{\alpha}{100}) SE^*}{(1 - \frac{\alpha}{100}) SE^*} \quad (33b)$$

where $\{w_1, w_2, w_3\}$ is a set of network determined weight-factors. The first term in (33a) and (33b) provides the net profit (or loss) obtained from trading-off the α percentage of optimal SE and EE, respectively; the second term in (33a) and (33b) provides the profit (loss) obtained from selecting a strategy of higher (lower) EE; and the third term in (33a) and (33b) provides the profit (loss) obtained from selecting a strategy of higher (lower) SE. Eqs. (33a) and (33b) provide the overall profit metrics obtained by the system for trading-off the α percentage of the optimal SE and EE to improve the achievable EE and SE, respectively. If $\mathcal{NP}_1 > \mathcal{NP}_2$ holds, the operating EE-SE pair of the system is selected as $(EE^{**}, (1 - \frac{\alpha}{100}) SE^*)$. Conversely, if $\mathcal{NP}_2 > \mathcal{NP}_1$ holds, the operating EE-SE pair of the system is selected as $((1 - \frac{\alpha}{100}) EE^*, SE^{**})$. After selecting an operating EE-SE pair, the remaining step is to choose the resource allocation variables that can achieve the selected EE-SE pair. By inspecting the outputs of the I-RSRM algorithm, the resource allocation variables attaining the operating EE-SE pair are readily obtained.

Algorithm 6 Distributed Rate Splitting and Resource Matching (D-RSRM) Algorithm

- 1: **Initialize:** Equal power allocations of the F-UEs and F-AP among all the RRBs and $\eta = \eta_o$;
 - 2: Determine assignment variables $(\mathbf{a}, \mathbf{b})$ by using Algorithm 2.
 - 3: By plugging $(\mathbf{a}, \mathbf{b})$ to Algorithms 3 and 4, determine EE-optimal and SE-optimal power allocations, respectively. Denote SE-optimal and EE-optimal resource allocations by $(\mathbf{P}_{SE}, \mathbf{Q}_{SE}, \mathbf{a}, \mathbf{b})$ and $(\mathbf{P}_{EE}, \mathbf{Q}_{EE}, \mathbf{a}, \mathbf{b})$, respectively.
 - 4: **Initialize:** $\epsilon^{(1)} = EE(\mathbf{P}_{SE}, \mathbf{Q}_{SE}, \mathbf{a}, \mathbf{b})$, iteration index $m = 1$, step size Δ .
 - 5: **while** $EE(\mathbf{P}_{EE}, \mathbf{Q}_{EE}, \mathbf{a}, \mathbf{b}) > \epsilon^{(m)}$ **do**
 - 6: Initialize μ .
 - 7: **repeat**
 - 8: Update power allocation variables using Algorithm 1 while replacing $\frac{\eta}{\epsilon}$ and $1/\ln 2$ in (18)-(21) with $\frac{\mu\epsilon^{(m)}}{\epsilon}$ and $(1 + \mu)/\ln 2$, respectively.
 - 9: Update μ using a bi-section method as such constraint C5 in P1 becomes active.
 - 10: **until** $\mathcal{L}^{(2)}$ converges or Maximum number of iterations is reached
 - 11: Record the resource allocations $(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}, \mathbf{b})$. Update $\epsilon^{(m+1)} = \epsilon^{(m)} + \Delta$, $m = m + 1$.
 - 12: **end while**
 - 13: **Output:** Set of resource allocations $\{(\mathbf{P}^*(\epsilon^{(m)}), \mathbf{Q}^*(\epsilon^{(m)}), \mathbf{a}, \mathbf{b})\}, \forall m = 1, 2, \dots, N_F$.
-

B. Low-Complexity Algorithm Development

For the optimality, at each iteration of I-RSRM algorithm, $\mathcal{O}(KLN)$ information exchanges are required between the F-UEs and F-AP. For a large network, such a high number of information exchange can be prohibitive. To reduce the number of information exchanges, we propose D-RSRM algorithm. The overall steps of D-RSRM are summarized in Algorithm 6. Unlike I-RSRM, D-RSRM decouples the power allocations and resource assignment. Particularly, D-RSRM initially determines the assignment of the F-UEs and RRBs among the device-clusters by considering a uniform power allocation, and keeps such assignments fixed for the remaining steps. Essentially, the total information exchanges are substantially reduced in D-RSRM. Similar to I-RSRM, D-RSRM provides a set of resource allocation variables where each solution has a certain EE-SE trade-off. By applying a similar procedure described in the previous sub-section, first the operating EE-SE pair is determined. Thereafter, by inspecting the outputs of D-RSRM algorithm, the resource allocation variables attaining the operating EE-SE pair are selected.

C. Computational Complexity Analysis

I-RSRM: We first determine total number of the required operations to obtain a single Pareto-optimal point. Using the reported computational complexity of Algorithms 1 and 2 and assuming Δ_{max} iterations are required for convergence of the Steps 3-11 of Algorithm 5, total $\mathcal{O}(\Delta_{max}(KLN + K^2N + KL)) \approx \mathcal{O}(\Delta_{max}(KLN + K^2N))$ operations are required to obtain a single Pareto-optimal point. Since the aforementioned procedure is repeated N_F times, the overall computational complexity of the proposed I-RSRM algorithm is $\mathcal{O}(N_F\Delta_{max}(KLN + K^2N))$.

D-RSRM: The major computational complexity of D-RSRM algorithm is contributed by the Step 2 and the Steps 5-12 of Algorithm 6. For Step 2, Algorithm 6 requires $\mathcal{O}(K^2N + KL)$ operations. Since only power allocations are used in the Steps 5-12 of Algorithm 6, the required number of operations for these steps is $\mathcal{O}(\Delta_{max}KLN)$. Therefore, the overall computational complexity of the proposed D-RSRM algorithm to obtain N_F Pareto-optimal points is obtained as $\mathcal{O}(N_F\Delta_{max}KLN + K^2N)$. Evidently, D-RSRM brings substantial reduction of the computational complexity, especially for the system with large number of F-UEs and RRBs.

V. SIMULATION RESULTS

For simulations, we consider a $300m \times 300m$ square F-RAN cell with one F-AP in the center, 28 CR devices, 10 F-UEs, and 36 RRBs. The CR devices are grouped into 7 device-clusters where each device-cluster contains 4 CR devices. The center and radius of the device-clusters, and the coordinates of the CR devices at the device-clusters are randomly generated at each

simulation instance. Meanwhile, locations of the F-UEs are randomly generated on a circle of $75m$ radius. In each device-cluster, the CR devices are uniformly divided into LCH and GCH device sets. The channel coefficients of are generated by considering Rayleigh distributed fading and the 3GPP path loss model, given by, $131.1 + 42.8 \log_{10}(d)$ dB (where d is the distance in kilometers). Without further specification, we consider $P_{\max}^{(D)} = P_{\max}^{\text{FAP}} = 33$ dBm, $P_c = 0.1$ Watt, $\varepsilon = 0.75$, $\sigma^2 = -174$ dBm, and $\{w_i\} = 1, \forall i \in \{1, 2, 3\}$. For the sake of performance comparison, we consider the following baseline schemes. For all these baseline schemes, the trade-off between EE and SE is optimized by employing the method described in Section IV. A

- **Downlink multicasting:** Here, the F-UEs and F-AP transmit contents to the LCH and GCH CR devices, respectively, using a multicasting technique. Particularly, in a given device-cluster, F-UE and F-AP first combine the messages for LCH and GCH CR devices, respectively, and multicast such combined messages by using two common rates. Meanwhile, Algorithm 2 is used for scheduling the F-UEs and RRBs among the device-clusters.
- **I-RSRM with iterative resource assignment (I-RSRM/IRA):** Here, the power allocations of the F-UEs and F-AP are determined according to Algorithm 1, and an alternating optimization based resource assignment, given in Algorithm 3 of [31], is used to determine the scheduling of F-UEs and RRBs among the device-clusters.
- **Conventional unicast content distribution:** Each device-cluster is allocated with certain RRBs, and the CR devices receive contents only from F-AP. Specifically, at each device-cluster, F-AP transmits unicast messages over the allocated RRBs and the CR devices decode their intended messages while treating interference from other devices' messages as noise. An IFE based power control of [17] is applied for allocating F-AP's power over the RRBs, and a many-to-one matching is employed to allocate the RRBs among the device-clusters.

A. Effect of EE-SE Trade-off Control Parameter

Figs. 2a and 2b illustrate the EE and SE of the proposed I-RSRM algorithm for different values of the EE-SE trade-off control parameter α . For all the considered device-clusters in Figs. 2a and 2b, when α approaches small or large value, such as $\alpha = 10\%$ or $\alpha = 60\%$, a large system EE is achieved, and when α approaches moderate value, such as $\alpha = 30\%$, a large system SE is achieved. Moreover, for a given range of α , the system EE and SE depict opposite behavior, i.e., increase of system SE results in decrease of system EE and vice versa. Such an observation is intuitive as EE and SE of the system can not be maximized simultaneously. An explanation of achieving large EE for both small and large values of α is provided as follows.

- *Small α* : When $\alpha \rightarrow 0$, the solutions to (30) and (31) approach the corner points of the developed Pareto-optimal front. Particularly, the EE obtained by trading-off α percentage of optimal SE approaches the achievable EE of the system operating with maximum SE, and the SE obtained by trading-off α percentage of optimal EE approaches the achievable SE of the system operating with maximum EE. Essentially, as $\alpha \rightarrow 0$, $\mathbf{EE}^{**} \rightarrow \widehat{\mathbf{EE}}_{SE=SE^*}$ and $\mathbf{SE}^{**} \rightarrow \widehat{\mathbf{SE}}_{EE=EE^*}$, and consequently, $\chi_{EE} \rightarrow 0$ and $\chi_{SE} \rightarrow 0$. We can illustrate that in the considered simulation setup, the relative EE gap between the two corner points of the developed Pareto optimal front is much larger than relative the SE gap. Hence, for small α , $\frac{\mathbf{EE}^* - \mathbf{EE}^{**}}{\mathbf{SE}^* - \mathbf{SE}^{**}} \gg 1$ and $\mathcal{NP}_2 > \mathcal{NP}_1$ are usually satisfied. Recall, when $\mathcal{NP}_2 > \mathcal{NP}_1$ holds, the achievable EE and SE of the system are given by $(1 - \frac{\alpha}{100}) \mathbf{EE}^*$ and $\mathbf{SE}^{**}$, respectively. Since α is small, the system EE approaches the optimal EE, and owing to the conflicting behavior of EE and SE, SE becomes small. Accordingly, when α is small, the considered system example obtains large EE and small SE.
- *Large α* : When α is large, the SE and EE constraints of (30) and (31) are relaxed. As a result, $\mathbf{EE}^{**} \rightarrow \mathbf{EE}^*$ and $\mathbf{SE}^{**} \rightarrow \mathbf{SE}^*$ are satisfied. Therefore, eqs. (33a) and (33b) can be modified as $\mathcal{NP}_1 \rightarrow \chi_{EE} - \alpha + \frac{(\alpha/100)^2}{1-\alpha/100}$ and $\mathcal{NP}_2 \rightarrow \chi_{SE} - \alpha + \frac{(\alpha/100)^2}{1-\alpha/100}$, respectively. Since the relative EE gap between the two corner points of the developed Pareto optimal front is much larger than the relative SE gap, $\chi_{EE} > \chi_{SE}$ and $\mathcal{NP}_1 > \mathcal{NP}_2$ are satisfied. Recall, when $\mathcal{NP}_1 > \mathcal{NP}_2$ holds, the achievable EE of the system is given by $\mathbf{EE}^{**}$, and as mentioned earlier that for a large α , $\mathbf{EE}^{**}$ approaches the optimal system EE. Consequently, the considered system example also achieves large EE as α becomes large.

Meanwhile, between the aforementioned two extreme cases, certain α does exist for which system chooses a strategy that obtains a near-optimal SE. Fig. 2a illustrates that as α takes a moderate value, i.e., $\alpha = 30\%$, for all the device-clusters, the considered system example achieves larger SE compared to the scenarios when $\alpha = 10\%$ and $\alpha = 60\%$. Note that α is a system defined control parameter. To determine a suitable value of α , one can first conduct an offline simulation where system EE and SE are evaluated for different values of α . Subsequently, by inspecting these results, the suitable value of α is selected so that the required design criteria are satisfied.

B. EE and SE Comparison Among the Proposed and Baseline Schemes

Figs. 3a and 3b compare the achievable SE and EE of the proposed and baseline schemes for $\alpha = 30\%$ and $\alpha = 60\%$, respectively. The achievable SE and EE of the EE-optimal and SE-optimal schemes are also depicted in Figs. 3a and 3b, respectively. The SE-optimal and EE-

Fig. 2: Impact of EE-SE trade-off control parameter, α , on the EE and SE of the proposed I-RSRM algorithm considering $L = 3$ F-UEs, $N = 36$ RRBs, and $N_R = 3$.

Fig. 3: EE and SE comparison among the proposed and baseline schemes for different number of device-clusters considering $L = 3$ F-UEs, $N = 36$ RRBs, and $N_R = 3$.

optimal schemes provide the two corner points of the Pareto-optimal front developed by I-RSRM algorithm. Figs. 3a and 3b depict that for all the device-clusters, the relative EE gap between the EE-optimal and SE-optimal schemes is much wider than the relative SE gap between EE-optimal and SE-optimal schemes. This confirms the previously mentioned fact, that is, the relative EE gap between the two corner points of the developed Pareto optimal front is much larger than the relative SE gap. Fig. 3a depicts that I-RSRM with $\alpha = 30\%$ achieves approximately 70% of the optimal SE which is intuitively expected. Particularly, by trading-off 30% of optimal SE, I-RSRM tremendously improves the achievable system EE. For instance, for $K = 5$ device-clusters, I-RSRM with $\alpha = 30\%$ improves the system EE by 3.57 times compared to the SE-optimal scheme. Moreover, I-RSRM with $\alpha = 30\%$ achieves improved SE compared to the EE-optimal scheme as well. In Fig. 3a, for $K = 5$ device-clusters, I-RSRM with $\alpha = 30\%$ achieves 65.76%

improved SE compared to the EE-optimal scheme. Finally, Fig. 3b depicts that for all the device-clusters, I-RSRM with $\alpha = 60\%$ achieves a near-optimal EE, and it is consistent with Fig. 2b. We observe that the conventional EE-optimal and SE-optimal schemes do not provide both SE and EE guarantee at the same time. In contrast, by tuning α , the proposed I-RSRM algorithm not only improves the trade-off between EE and SE but also can approach the solution provided by the EE-optimal and SE-optimal schemes. Essentially, the proposed I-RSRM algorithm is more amenable compared to the conventional EE-optimal and SE-optimal schemes.

Figs. 3a and 3b also confirm the substantial SE and EE gain achieved by the proposed I-RSRM algorithm. Fig. 3a depicts that for $K = 5$ device-clusters, I-RSRM with $\alpha = 30\%$ achieves 17.45% and 2.4803 times improved SE compared to downlink multicasting and unicasting schemes, respectively. Similarly, Fig. 3b depicts that for $K = 5$ device-clusters, I-RSRM with $\alpha = 60\%$ achieves 15.12% and 28.4 times improved achievable EE compared to downlink multicasting and unicasting schemes, respectively. In a downlink multicasting scheme, only the minimum throughput of the LCH and GCH CR devices in a given device-cluster is targeted. Therefore, although downlink multicasting overcomes the multi-user interference in a given device-cluster, the achievable throughput of such a scheme is substantially reduced, especially when the channel condition of the device's with minimum throughput is deteriorated. Meanwhile, in each device-cluster of a unicasting system, the cooperation between F-UEs and the CR devices is entirely eliminated while the multi-user interference is completely treated as a noise. Thus, the achievable throughput of a unicasting system is also significantly reduced. Overall, both the downlink multicasting and unicasting schemes suffer from considerable EE and SE loss compared to the I-RSRM algorithm. Figs. 3a and 3b also illustrate that the SE and EE gap between the proposed I-RSRM and D-RSRM algorithms is small when the number of device-clusters in the network is small. However, as the number of device-clusters in the network is increased, the EE gap between the proposed I-RSRM and D-RSRM algorithms is noticeably increased as well.

Figs. 4a and 4b illustrate variation of the SE and EE of the proposed I-RSRM algorithm for different number of F-UEs, respectively. As the number of F-UEs in the network is increased, the diversity gain obtained from the opportunistic assignment between device-clusters and F-UEs is enhanced. Such a fact leads to the improvement of both SE and EE of the system. As expected, for all the F-UEs, the SE (EE) of the proposed I-RSRM algorithm resides between SE (EE) of SE-optimal and EE-optimal schemes. Moreover, Figs. 4a and 4b depict that I-RSRM achieves maximum SE and EE for $\alpha = 30\%$ and $\alpha = 60\%$ which is consistent with the observation made from Figs. 2a and 2b, respectively. For $L = 4$ F-UEs, compared to SE-optimal

(a) SE vs. number of F-UEs for $\alpha = 10\%, 30\%, 60\%$. (b) EE vs. number of F-UEs for $\alpha = 10\%, 30\%, 60\%$.

Fig. 4: EE and SE of the proposed I-RSRM algorithm for different number of F-UEs considering $K = 3$ device-clusters, $N = 36$ RRBs, and $N_R = 3$.

(a) SE vs. N_R for $\alpha = 10\%, 30\%, 60\%$. (b) EE vs. N_R for $\alpha = 10\%, 30\%, 60\%$.

Fig. 5: EE and SE of the proposed I-RSRM and D-RSRM algorithms for different values of N_R considering $K = 3$ device-clusters, $L = 2$ F-UEs, and $N = 36$ RRBs.

scheme, I-RSRM with $\alpha = 10\%, 30\%$, and 60% achieves 49.32%, 71.16%, and 44.53% of optimal SE, respectively, and improves the achievable EE by 6.13, 3.33, and 6.56 times, respectively. Similarly, For $L = 4$ F-UEs, compared to EE-optimal scheme, I-RSRM with $\alpha = 60\%, 30\%$, and 10% achieves 99.11%, 50.35%, and 92.52% of optimal EE, respectively, and improves the achievable SE by 1.03, 1.65, and 1.14 times, respectively. Indeed, the proposed I-RSRM algorithm improves the trade-off between EE and SE of both EE-optimal and SE-optimal schemes.

Figs. 5a and 5b compare SE and EE of the proposed I-RSRM and D-RSRM algorithms for different N_R , respectively. As N_R is increased, the number of supported device-clusters per F-UE is increased. On one hand, the opportunity of assigning device-clusters with their most suitable F-UEs is increased, and on the other hand, the number of active F-UEs is decreased. Essentially, both SE and EE of the network are increased with N_R . However, Figs. 5a and 5b depict that

Fig. 6: EE and SE comparison among I-RSRM, D-RSRM, and I-RSRM/IRA algorithms for different number of RRBs considering $K = 3$ device-clusters, $L = 2$ F-UEs, and $N_R = 3$.

Fig. 7: EE vs. SE trade-off comparison among the proposed and baseline schemes considering $K = 3$ device-clusters, $L = 3$ F-UEs, $N_R = 3$, and $N = 36$ RRBs.

for a sufficiently large N_R , both SE and EE become saturated, i.e., further increase of N_R does not improve SE and EE. Figs. 5a and 5b also illustrate that for a given value of N_R , the SE and EE gap between the proposed I-RSRM and D-RSRM algorithm is small. From Fig. 5a, at $N_R = 3$, the SE gap between I-RSRM and D-RSRM algorithms is 6.7%, 3.3%, and 6.3% for $\alpha = 10\%$, 30%, and 60%, respectively. From Fig. 5b, at $N_R = 3$, the EE gap between I-RSRM and D-RSRM algorithms is 6.5%, 16.8%, and 8.1% for $\alpha = 10\%$, 30%, and 60%, respectively. Recall, D-RSRM has much lower implementation and computational complexity than I-RSRM. Hence, for the low-complexity systems, D-RSRM is a suitable alternative to I-RSRM algorithm.

Figs. 6a and 6b compare SE and EE of I-RSRM, D-RSRM, and I-RSRM/IRA algorithms for different RRBs. As the number of RRBs is increased, both SE and EE of the system is improved at the cost of increased computational complexity. As expected, the performance gap

between the proposed I-RSRM and D-RSRM algorithms is small. Note that I-RSRM, D-RSRM, and I-RSRM/IRA algorithms utilize the same power control method described in the proposed Algorithm 1. However, in I-RSRM/IRA, the scheduling of RRBs and assignment of F-UEs among the device-clusters are performed according to Algorithm 3 of [31]. Since Algorithm 3 of [31] is sub-optimal, we observe a noticeable performance gap between the proposed I-RSRM and D-RSRM algorithms and I-RSRM/IRA algorithm, especially for large RRBs. From Fig. 6a, at $N = 36$ RRBs, the proposed I-RSRM and D-RSRM algorithms with $\alpha = 30\%$ improve the SE of I-RSRM/IRA by 52.32% and 45.24%, respectively. From Fig. 6b, at $N = 36$ RRBs, the proposed I-RSRM and D-RSRM algorithms with $\alpha = 60\%$ improve the EE of I-RSRM/IRA by 112.96% and 95.41%, respectively. Such an observation clearly demonstrates that the our proposed 3D-resource matching is highly efficient to improve both SE and EE of the system.

C. EE (resp. SE) Loss-rate vs. SE (resp. EE) Improvement

Figs. 7a and 7b demonstrate EE and SE loss-rate vs. SE and EE improvement for the proposed and baseline schemes, respectively. Each point on Figs. 7a and 7b provides the achievable EE and SE obtained by trading-off certain percentage of optimal SE and EE, respectively. We observe the following three observations. First, as the SE loss-factor and EE loss-factor become asymptotically large, the achievable EE and SE of I-RSRM algorithm approach the EE and SE upper bounds obtained by the EE-optimal and SE-optimal schemes, respectively. Second, for I-RSRM algorithm, the improvement of EE by trading-off optimal SE is much larger than the improvement of SE by trading-off optimal EE. For instance, when the SE loss-factor is increased from 20% to 50%, the EE of I-RSRM algorithm is improved by 162.09% as depicted from Fig. 7a. In contrast, when the EE loss-factor is increased from 20% to 50%, the SE of I-RSRM algorithm is improved by 29.25% as depicted from Fig. 7b. Finally, we observe that for a given SE and EE loss factor, the proposed I-RSRM and D-RSRM algorithms achieve substantially large EE and SE gain compared to the baseline schemes. For instance, Fig. 7a depicts that when the system is willing to trade-off 30% of optimal SE, I-RSRM (resp. D-RSRM) achieves 60.25%, 60.87% and 139% (resp. 43.71%, 44.27%, and 123%) improved EE compared to the downlink multicasting, I-RSRM/IRA, and unicasting schemes, respectively. Similarly, Fig. 7b depicts that when the system is willing to trade-off 40% of optimal EE, I-RSRM (resp. D-RSRM) achieves 20.47%, 38.27% and 197% (resp. 12.51%, 29.13%, and 178%) improved SE compared to the downlink multicasting, I-RSRM/IRA, and unicasting schemes, respectively. The aforementioned observations suggest that compared to the baseline schemes, both of the proposed I-RSRM and D-RSRM algorithms significantly improve the trade-off between EE and SE of the system.

VI. CONCLUSION

We investigated resource allocations to simultaneously maximize system EE and SE of the content distribution phase of RS-CMD integrated and multi-level EC enabled F-RAN. By exploiting clustering of the CR devices and the ϵ -constraint method, an EE-SE trade-off optimization was conducted while jointly performing the transmission power allocation of F-AP and F-UEs, assignment of the F-UEs among the device-clusters, and scheduling of the RRBs among the device-clusters. An IFE based power control method and 3D-resource matching were developed. Based on such analyses, two different algorithms were proposed to obtain the near-optimal Pareto front of EE-SE trade-off optimization problem, and select a suitable operating EE-SE pair. Simulation results confirmed the following two observations: (i) By tuning a suitable EE-SE trade-off control parameter, the proposed algorithm substantially improves the trade-off between EE and SE compared to the conventional EE-optimal and SE-optimal algorithms, and (ii) the proposed algorithms yield significant EE and SE gain compared to the baseline schemes.

APPENDIX A

We assume that the tuple $\mathbf{z}^* = (\mathbf{P}^*, \mathbf{Q}^*, \mathbf{a}^*, \mathbf{b}^*)$ is an optimal solution to P1. For a proof by contradiction, we consider that $\mathbf{z}^*$ is not weakly Pareto-optimal to P0. Hence, another tuple $\tilde{\mathbf{z}} \triangleq (\hat{\mathbf{P}}, \hat{\mathbf{Q}}, \hat{\mathbf{a}}, \hat{\mathbf{b}})$ will exist for which $\text{EE}(\tilde{\mathbf{z}}) > \text{EE}(\mathbf{z}^*)$ and $\text{SE}(\tilde{\mathbf{z}}) > \text{SE}(\mathbf{z}^*)$ will be satisfied. Moreover, since $\mathbf{z}^*$ is a solution to P1, we obtain $\text{SE}(\tilde{\mathbf{z}}) > \text{SE}(\mathbf{z}^*) \geq \epsilon$. Hence, $\tilde{\mathbf{z}}$ is also a feasible solution to P1, and it improves objective function of P1 compared to $\mathbf{z}^*$. Essentially, $\mathbf{z}^*$ can not be an optimal solution to P1 which is contradictory to our initial assumption. Hence, an optimal solution to P1 is weakly Pareto-optimal to P0. ■

APPENDIX B

According to the Karush-Khun-Tucker (KKT) optimality conditions, the first order derivative of $\mathcal{L}_j$ with respect to the power allocation variables will be zero. By setting $\frac{\partial \mathcal{L}_j}{\partial P_{l,n}^{C,j}} = 0$, we obtain

$$\frac{1}{1 + \gamma_{l,n}^{C,j}} \frac{\partial \gamma_{l,n}^{C,j}}{\partial P_{l,n}^{C,j}} + \frac{1}{1 + \tilde{\gamma}_n^{C,j}} \frac{\partial \tilde{\gamma}_n^{C,j}}{\partial P_{l,n}^{C,j}} + \sum_{e \in S_j^{(L)}} \frac{1}{1 + \gamma_{l,e,n}^{P,j}} \frac{\partial \gamma_{l,e,n}^{P,j}}{\partial P_{l,n}^{C,j}} + \sum_{t \in S_j^{(H)}} \frac{1}{1 + \tilde{\gamma}_{t,n}^{P,j}} \frac{\partial \tilde{\gamma}_{t,n}^{P,j}}{\partial P_{l,n}^{C,j}} = \frac{\eta + \lambda_l}{\epsilon / \ln 2}. \quad (34)$$

By plugging (4)-(7) to (34) and performing some algebraic manipulations, we obtain

$$P_{l,n}^{C,j} = \frac{\gamma_{l,n}^{C,j} / (1 + \gamma_{l,n}^{C,j})}{\frac{\eta/\varepsilon + \lambda_l/\varepsilon}{1/\ln 2} + \frac{(\tilde{\gamma}_n^{C,j})^2 |h_{k_j,n}^{(l)}|^2}{(1 + \tilde{\gamma}_n^{C,j}) Q_n^{C,j} |h_{k_j,n}^{(0)}|^2} + \sum_{t \in \mathcal{S}_j^{(H)}} \frac{(\tilde{\gamma}_{t,n}^{p,j})^2 |h_{t,n}^{(l)}|^2}{(1 + \tilde{\gamma}_{t,n}^{p,j}) Q_{t,n}^{p,j} |h_{t,n}^{(0)}|^2}}. \quad (35)$$

Eq. (35) provides the optimality condition for $P_{l,n}^{C,j}$. Similarly, the optimality conditions for the other power allocation variables, namely, $Q_n^{C,j}$, $P_{l,e,n}^{p,j}$, and $Q_{l,n}^{p,j}$, are respectively obtained as

$$Q_n^{C,j} = \frac{\tilde{\gamma}_n^{C,j} / (1 + \tilde{\gamma}_n^{C,j})}{\frac{\eta/\varepsilon + \nu/\varepsilon}{1/\ln 2} + \frac{(\gamma_{l,n}^{C,j})^2 |h_{k_j,n}^{(0)}|^2}{(1 + \gamma_{l,n}^{C,j}) P_{l,n}^{C,j} |h_{k_j,n}^{(l)}|^2} + \sum_{e \in \mathcal{S}_j^{(L)}} \frac{(\gamma_{l,e,n}^{p,j})^2 |h_{e,n}^{(0)}|^2}{(1 + \gamma_{l,e,n}^{p,j}) P_{l,e,n}^{p,j} |h_{e,n}^{(l)}|^2}} \quad (36)$$

$$P_{l,e,n}^{p,j} = \frac{\gamma_{l,e,n}^{p,j} / (1 + \gamma_{l,e,n}^{p,j})}{\frac{\eta + \lambda_l}{\varepsilon/\ln 2} + \frac{(\gamma_{l,n}^{C,j})^2}{(1 + \gamma_{l,n}^{C,j}) P_{l,n}^{C,j}} + \sum_{\substack{e' \in \mathcal{S}_j^{(L)} \\ e' \neq e}} \frac{(\gamma_{l,e',n}^{p,j})^2}{(1 + \gamma_{l,e',n}^{p,j}) P_{l,e',n}^{p,j}} + \frac{(\tilde{\gamma}_n^{C,j})^2 |h_{k_j,n}^{(l)}|^2}{(1 + \tilde{\gamma}_n^{C,j}) Q_n^{C,j} |h_{k_j,n}^{(0)}|^2} + \sum_{t \in \mathcal{S}_j^{(H)}} \frac{(\tilde{\gamma}_{t,n}^{p,j})^2 |h_{t,n}^{(l)}|^2}{(1 + \tilde{\gamma}_{t,n}^{p,j}) Q_{t,n}^{p,j} |h_{t,n}^{(0)}|^2}} \quad (37)$$

$$Q_{t,n}^{p,j} = \frac{\tilde{\gamma}_{t,n}^{p,j} / (1 + \tilde{\gamma}_{t,n}^{p,j})}{\frac{\eta + \nu}{\varepsilon/\ln 2} + \frac{(\tilde{\gamma}_n^{C,j})^2}{(1 + \tilde{\gamma}_n^{C,j}) Q_n^{C,j}} + \sum_{\substack{t' \in \mathcal{S}_j^{(H)} \\ t' \neq t}} \frac{(\tilde{\gamma}_{t',n}^{p,j})^2}{(1 + \tilde{\gamma}_{t',n}^{p,j}) Q_{t',n}^{p,j}} + \frac{(\gamma_{l,n}^{C,j})^2 |h_{k_j,n}^{(0)}|^2}{(1 + \gamma_{l,n}^{C,j}) P_{l,n}^{C,j} |h_{k_j,n}^{(l)}|^2} + \sum_{e \in \mathcal{S}_j^{(L)}} \frac{(\gamma_{l,e,n}^{p,j})^2 |h_{e,n}^{(0)}|^2}{(1 + \gamma_{l,e,n}^{p,j}) P_{l,e,n}^{p,j} |h_{e,n}^{(l)}|^2}} \quad (38)$$

Since the closed-form power allocation expressions are intractable, we use (35)-(38) to iteratively update the power allocation variables. To this end, we denote $\{\hat{P}_{l,n}^{C,j}, \hat{P}_{l,e,n}^{p,j}, \hat{Q}_n^{C,j}, \hat{Q}_{t,n}^{p,j}\}$ as the given power allocations. Thereupon, by replacing $\{P_{l,n}^{C,j}, P_{l,e,n}^{p,j}, Q_n^{C,j}, Q_{t,n}^{p,j}\}$ of (35)-(38) with $\{\hat{P}_{l,n}^{C,j}, \hat{P}_{l,e,n}^{p,j}, \hat{Q}_n^{C,j}, \hat{Q}_{t,n}^{p,j}\}$, we obtain the power allocation update equations of *Proposition 1*. ■

APPENDIX C

The convergence of IFE based power allocation for high signal-to-inference-plus-noise-ratio (SINR) regime was justified in [17]. We present an alternative proof by using game theory. Specifically, we describe power allocation in the j -th device-cluster, $\forall j$, as a non-cooperative power control game (NCPCG) where the transmitted common messages and private messages are the selfish and rational (virtual) players. The NCPCG is defined by the tuple $\mathcal{G} = (\{I\}, \{\Gamma_I\}, \{\Pi_I\})$. Here, $\{I\}$ is the set of the players in the game and the considered game has total $|\mathcal{S}_j^{(H)}| + |\mathcal{S}_j^{(L)}| + 2$ players. Particularly, $I = 1$ denotes the transmitted common message for the LCH CR devices; $I = 2$ denotes the transmitted common message for the GCH CR devices; $I = 3, \dots, |\mathcal{S}_j^{(L)}| + 2$

$$\Pi_1(\Gamma_1|\mathbf{P}_{-1}) = \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{l,j,n} \left(R_{l,n}^{C,j} + \tilde{R}_n^{C,j} + \sum_{t \in \mathcal{S}_j^{(H)}} \tilde{R}_{t,n}^{P,j} - \frac{\eta + \lambda_l}{\varepsilon} P_{l,n}^{C,j} \right) \quad (39)$$

$$\Pi_2(\Gamma_2|\mathbf{P}_{-2}) = \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{l,j,n} \left(\tilde{R}_n^{C,j} + R_{l,n}^{C,j} + \sum_{e \in \mathcal{S}_j^{(L)}} R_{l,e,n}^{bmP,j} - \frac{\eta + \nu}{\varepsilon} Q_n^{C,j} \right) \quad (40)$$

$$\begin{aligned} \Pi_I(\Gamma_I|\mathbf{P}_{-I}) = & \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{l,j,n} \left(R_{l,I-2,n}^{P,j} + \sum_{\substack{e' \in \mathcal{S}_j^{(L)} \\ e' \neq I-2}} R_{l,e',n}^{P,j} + R_{l,n}^{C,j} + \tilde{R}_n^{C,j} \right. \\ & \left. + \sum_{t \in \mathcal{S}_j^{(H)}} R_{t,n}^{P,j} - \frac{\eta + \lambda_l}{\varepsilon} P_{l,I-2,n}^{P,j} \right), I = 3, \dots, |\mathcal{S}_j^{(L)}| + 2 \end{aligned} \quad (41)$$

$$\begin{aligned} \Pi_I(\Gamma_I|\mathbf{P}_{-I}) = & \sum_{l=1}^L \sum_{n=1}^N a_{l,j} b_{l,j,n} \left(\tilde{R}_{I-|\mathcal{S}_j^{(L)}|-2,n}^{P,j} + \sum_{\substack{t' \in \mathcal{S}_j^{(H)} \\ t' \neq I-|\mathcal{S}_j^{(L)}|-2}} \tilde{R}_{t',n}^{P,j} + R_{l,n}^{C,j} + \tilde{R}_n^{C,j} \right. \\ & \left. + \sum_{e \in \mathcal{S}_j^{(L)}} R_{l,e,n}^{P,j} - \frac{\eta + \nu}{\varepsilon} Q_{I-|\mathcal{S}_j^{(L)}|-2,n}^{P,j} \right), I = |\mathcal{S}_j^{(L)}| + 3, \dots, |\mathcal{S}_j^{(L)}| + |\mathcal{S}_j^{(H)}| + 2. \end{aligned} \quad (42)$$

denotes the transmitted private messages for the LCH CR devices; and $I = |\mathcal{S}_j^{(L)}| + 3, \dots, |\mathcal{S}_j^{(L)}| + |\mathcal{S}_j^{(H)}| + 2$ denotes the transmitted private messages for the GCH CR devices. Γ_I is the strategy space set for the I -th player. Particularly, $\Gamma_1 = \{P_{l,n}^{C,j} | P_{l,n}^{C,j} \geq 0\}$; $\Gamma_2 = \{Q_n^{C,j} | Q_n^{C,j} \geq 0\}$; for $I = 3, \dots, |\mathcal{S}_j^{(L)}| + 2$, $\Gamma_I = \{P_{l,e,n}^{C,j} | P_{l,e,n}^{C,j} \geq 0\}$ where $e = I - 2$; and for $I = |\mathcal{S}_j^{(L)}| + 3, \dots, |\mathcal{S}_j^{(L)}| + |\mathcal{S}_j^{(H)}| + 2$, $\Gamma_I = \{Q_{t,n}^{C,j} | Q_{t,n}^{C,j} \geq 0\}$ where $t = I - |\mathcal{S}_j^{(L)}| - 2$. Π_I is the utility function for the I -th player and for different I , utility functions are defined as (39)-(42) at the top of current page. In (39)-(42), $\mathbf{P}_{-I}$ denotes the power allocation strategy of all the players in the game except the I -th player. The utility function of the I -th player, $\forall I$, is continuous and differentiable in the strategy space. For each utility function, the first term provides the reward for improving the achievable data rate and the remaining terms provide the reward (penalty) for creating less (high) interference in the network. We can show that the utility function of the I -th player, $\forall I$, behaves similar to a quasi-concave function. Essentially, the considered NCPCG must exhibit a Nash-equilibrium (NE) power allocation strategy [34, Theorem 3.2]. Under an

NE power allocation strategy, no player can further improve its utility by choosing an alternative strategy, and thus, power allocation converges to a stable point. Note that for the quasi-concave utility functions, NE is achieved at the maximizing points of the players' utility functions. It can be readily verified that power allocation update equations in *Proposition 1* are obtained at the maximizing points of the utility functions in (39)-(42). Obviously, the power allocations in *Proposition 1* must converge to a stable point regardless of the high SINR assumption. Moreover, the stable power allocations obtained from Algorithm 1 satisfies the KKT conditions of P2.4. Accordingly, Algorithm 1 converges to a stable and local optimal solution to P2.4. ■

APPENDIX D

Assume that Φ^* is not Pareto-efficient. Consequently, Φ^* will contain at least two device-clusters, namely $\mathcal{S}_j$ and $\mathcal{S}_{j'}$ where $\Phi^*(l, n) = j$ and $\Phi^*(l', n') = j'$, and both device-clusters want to improve their current utility by exchanging their matched resource(s) with each other. Hence, we can have the following scenarios: **(S1)** $(l', n') \leftarrow_j (l, n)$ and $(l, n) \leftarrow_{j'} (l', n')$, **(S2)** $(l', n) \leftarrow_j (l, n)$ and $(l, n') \leftarrow_{j'} (l', n')$, and **(S3)** $(l, n') \leftarrow_j (l, n)$ and $(l', n) \leftarrow_{j'} (l', n')$. Here, $\Phi' \leftarrow_j \Phi^*$ and $\Phi'' \leftarrow_{j'} \Phi^*$ imply that the j -th and the j' -th device-clusters want to switch their resource assignment matching from Φ^* to Φ' and Φ'' , respectively. For proving *Proposition 3*, it is suffice to show that Φ^* does not contain two device-clusters for which **S1-S3** hold true. In Φ^* , the j -th and the j' -th device-clusters are matched with the l -th and the l' -th F-UEs at the FM phase, respectively. However, the n -th and the n' -th RRBs can be matched with the j -th and the j' -th device-clusters, respectively, in either the FM phase, the RM phase, or the RS phase.

Let the n -th and the n' -th RRBs be matched with the j -th and the j' -th device-clusters, respectively, in the FM phase. We assume that the j -th and j' -th device-clusters select their preferred pairs in two consecutive iterations of the FM phase, and the j -th device-cluster has a larger eligibility score than the j' -th device-cluster. We consider the following three cases: (i) Both the j -th and j' -th device-clusters prefer to be matched with the (l, n) pair; (ii) the j -th and j' -th device-clusters prefer to be matched with the (l, n) and (l, n') pairs, respectively; and (ii) the j -th and j' -th device-clusters prefer to be matched with the (l, n) and (l', n) pairs, respectively. Capitalizing the Steps 5 and 6 of Algorithm 2, for these three cases, we obtain, respectively,

$$\widehat{U}_{l,n}^{(j)} - \widehat{U}_{l',n'}^{(j)} > \widehat{U}_{l,n}^{(j')} - \widehat{U}_{l',n'}^{(j')} \implies \widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')} > \widehat{U}_{l',n'}^{(j)} + \widehat{U}_{l,n}^{(j')} \quad (43a)$$

$$\widehat{U}_{l,n}^{(j)} - \widehat{U}_{l',n}^{(j)} > \widehat{U}_{l,n'}^{(j')} - \widehat{U}_{l',n'}^{(j')} \implies \widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')} > \widehat{U}_{l',n}^{(j)} + \widehat{U}_{l,n'}^{(j')} \quad (43b)$$

$$\widehat{U}_{l,n}^{(j)} - \widehat{U}_{l,n'}^{(j)} > \widehat{U}_{l',n}^{(j')} - \widehat{U}_{l',n'}^{(j')} \implies \widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')} > \widehat{U}_{l,n'}^{(j)} + \widehat{U}_{l',n}^{(j')} \quad (43c)$$

However, **S1**, **S2**, and **S3** imply that $\widehat{U}_{l',n'}^{(j)} + \widehat{U}_{l,n}^{(j')} > \widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')}$, $\widehat{U}_{l',n}^{(j)} + \widehat{U}_{l,n'}^{(j')} > \widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')}$, and $\widehat{U}_{l,n'}^{(j)} + \widehat{U}_{l',n}^{(j')} > \widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')}$, respectively. Essentially, **S1**, **S2**, and **S3** are in contradiction with (43a), (43b), and (43c), respectively. Thus, when the j -th and j' -th device-clusters are chosen to select their preferred pairs in two consecutive iterations of the FM phase, exchanging their matched resource(s) is not beneficial. We can readily extend such a conclusion for any two device-clusters that are not necessarily chosen to select their preferred pairs in two consecutive iterations of the FM phase. Therefore, **S1-S3** do not hold true for the FM phase.

Next, we assume that the n -th and the n' -th RRBs are matched with the j -th and the j' -th device-clusters, respectively, either at the RM or RS phase. We only need to show that **S3** does not hold true in both scenarios. Particularly, if the considered RRBs are matched with the device-clusters at the RM phase, as per the Step 10 of Algorithm 2, we obtain $j = \arg \max_{j \in \mathcal{S}} \widehat{U}_{l,n}^{(j)}$ and $j' = \arg \max_{j' \in \mathcal{S}} \widehat{U}_{l',n'}^{(j')}$. Obviously, $\widehat{U}_{l,n}^{(j)} > \widehat{U}_{l',n}^{(j')}$ and $\widehat{U}_{l',n'}^{(j')} > \widehat{U}_{l,n'}^{(j)}$, and consequently, $\widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')} > \widehat{U}_{l,n'}^{(j)} + \widehat{U}_{l',n}^{(j')}$. On the other hand, if the (n, n') -th RRBs are swapped between the j -th and j' -th device clusters at the RS phase, as per the Steps (14) and (15) of Algorithm 2, $\widehat{U}_{l,n}^{(j)} + \widehat{U}_{l',n'}^{(j')} > \widehat{U}_{l,n'}^{(j)} + \widehat{U}_{l',n}^{(j')}$ must be satisfied. Therefore, **S3** does not hold true for both RM and RS phases.

The aforementioned analysis depicts that both the j -th and j' -th device-clusters can not improve their overall utility by partially or completely exchanging their matched resources. Therefore, no two device-clusters in Φ^* has an incentive to exchange their matched resources since such an exchange will not improve the social welfare and will decrease the utility of at least one device-cluster involved in the exchange process. In other words, each device-cluster has to be satisfied with its matched resources. Hence, Φ^* is Pareto-efficient. ■

APPENDIX E

Suppose, Φ^* does not provide a local optimal solution to P2.2, and another matching, Φ^{**} , exists which provides an improved overall utility compared to Φ^* . However, it is already proved that Φ^* is Pareto-efficient. Essentially, it is not possible to improve the overall utility of the device-clusters by exchanging the already matched resources among themselves. In other words, an improved matching from Φ^* can not be derived. Hence, for the given RRB prices, Φ^* must converge to a local-optimal solution to P2.2. ■

APPENDIX F

Since P3.1 provides a (local) optimal solution to P1, according to *Lemma 1*, a solution to P3.1 is also weakly Pareto optimal to P0. We denote set of the weakly Pareto optimal solutions obtained

from I-RSRM algorithm as $\mathcal{F} = \{f_1, f_2, \dots, f_{N_F}\}$. Notably, $f_i = \{f_i^{(1)}, f_i^{(2)}\}$ where $f_i^{(1)}$ and $f_i^{(2)}$ denote weakly Pareto optimal SE and EE obtained from the i -th solution of the I-RSRM algorithm and $\mathcal{F} \subseteq \Psi$ where Ψ is the Pareto optimal front of P0. We emphasize that no solution in the feasible space of P0 can dominate $(\mathbf{SE}(P_{SE}, Q_{SE}, \mathbf{a}_{SE}, \mathbf{b}_{SE}), \mathbf{EE}(P_{SE}, Q_{SE}, \mathbf{a}_{SE}, \mathbf{b}_{SE}))$ and $(\mathbf{SE}(P_{EE}, Q_{EE}, \mathbf{a}_{EE}, \mathbf{b}_{EE}), \mathbf{EE}(P_{EE}, Q_{EE}, \mathbf{a}_{EE}, \mathbf{b}_{EE}))$. Consequently, they will be the two corner points of the Pareto optimal front of P0. Meanwhile, in $\mathcal{F}$, f_1 and f_{N_F} are obtained by setting $\epsilon = \mathbf{EE}(P_{SE}, Q_{SE}, \mathbf{a}_{SE}, \mathbf{b}_{SE})$ and $\epsilon = \mathbf{EE}(P_{EE}, Q_{EE}, \mathbf{a}_{EE}, \mathbf{b}_{EE})$, respectively. Thus, the corner points of $\mathcal{F}$ coincide with the corner points of the Pareto optimal front. For the i -th non-corner point of $\mathcal{F}$, $\forall i$, $\mathbf{SE}(P_{EE}, Q_{EE}, \mathbf{a}_{EE}, \mathbf{b}_{EE}) < f_i^{(1)} < \mathbf{SE}(P_{SE}, Q_{SE}, \mathbf{a}_{SE}, \mathbf{b}_{SE})$ and $\mathbf{EE}(P_{SE}, Q_{SE}, \mathbf{a}_{SE}, \mathbf{b}_{SE}) < f_i^{(2)} < \mathbf{EE}(P_{EE}, Q_{EE}, \mathbf{a}_{EE}, \mathbf{b}_{EE})$ are satisfied. Since each solution in $\mathcal{F}$ is non-dominated, $\forall i$, $f_{(i+1)}^{(1)} < f_i^{(1)}$ and $f_{(i)}^{(2)} < f_{(i+1)}^{(2)}$ are satisfied. Without loss of generality, let consider $\hat{f} = \theta f_i + (1 - \theta) f_{i+1}$ where $\theta \in (0, 1)$. We can readily demonstrate that $f_{(i+1)}^{(1)} < \hat{f}^{(1)} < f_i^{(1)}$ and $f_{(i)}^{(2)} < \hat{f}^{(2)} < f_{(i+1)}^{(2)}$ are satisfied. Therefore, $\hat{f}$ is non-dominated and it lies between the corner points of $\mathcal{F}$. Furthermore, EE and SE of $\hat{f}$ can be achieved by using the resource allocations of f_i and f_{i+1} in a time-sharing manner. Hence, $\hat{f}$ is also a weakly Pareto optimal to P0. Therefore, by taking linear combination of the two consecutive points of $\mathcal{F}$, we obtain $\mathcal{F}^{(1)}$ as a new set of weakly Pareto optimal solutions to P0 where $|\mathcal{F}^{(1)}| - |\mathcal{F}| = N_F - 1$. In this way, the set of achievable weakly Pareto optimal solutions can be evolved as $\mathcal{F} \rightarrow \mathcal{F}^{(1)} \rightarrow \mathcal{F}^{(2)} \rightarrow \dots \rightarrow \mathcal{F}^{(final)}$. In this sequence, the number of weakly Pareto optimal solutions is increased at each evolution. As $|\mathcal{F}^{(final)}|$ becomes sufficiently large, further evolution will not create distinct Pareto optimal EE and SE. Essentially, the aforementioned sequence must converge, i.e., $\mathcal{F}^{(final)} \approx \Psi$. In practice, such an $\mathcal{F}^{(final)}$ set can be obtained by setting $\Delta = \lim_{N_F \rightarrow \infty} \frac{\mathbf{EE}(P_{EE}, Q_{EE}, \mathbf{a}_{EE}, \mathbf{b}_{EE}) - \mathbf{EE}(P_{SE}, Q_{SE}, \mathbf{a}_{SE}, \mathbf{b}_{SE})}{N_F}$ in the proposed I-RSRM algorithm. Accordingly, the Pareto optimal front P0 can be obtained by solving P3.1 while varying ϵ in the range $\mathbf{EE}(P_{SE}, Q_{SE}, \mathbf{a}_{SE}, \mathbf{b}_{SE}) \leq \epsilon \leq \mathbf{EE}(P_{EE}, Q_{EE}, \mathbf{a}_{EE}, \mathbf{b}_{EE})$. ■

REFERENCES

- [1] Z. Zhao *et al.*, "Federated-learning-enabled intelligent fog radio access networks: Fundamental theory, key techniques, and future trends," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 22-28, Apr. 2020.
- [2] R. Karasik, O. Simeone, and S. Shamai, "How much can D2D communication reduce content delivery latency in fog networks with edge caching?" *IEEE Trans. on Commun.*, vol. 68, no. 4, pp. 2308-2323, Apr. 2020.
- [3] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in NOMA-based fog radio access networks," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 110-115, June 2018.
- [4] I. Randrianantenaina, M. Kaneko, H. Dahrouj, H. ElSawy, and M.-S. Alouini, "Interference management in NOMA-based fog-radio access networks via Scheduling and power allocation," *IEEE Trans. Commun.* (accepted for publication).

- [5] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003-5009, Sept. 2016.
- [6] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907-922, Apr. 2016.
- [7] Z. Zhao *et al.*, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207-1221, May 2016.
- [8] M. Sheng *et al.*, "Enhancement for content delivery with proximity communications in caching enabled wireless networks: Architecture and challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 70-76, Aug. 2016.
- [9] L. Shi, L. Zhao, G. Zheng, Z. Han, and Y. Ye, "Incentive design for cache-enabled D2D underlaid cellular networks using Stackelberg game," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 765-779, Jan. 2019.
- [10] B. Chen, C. Yang, and A. F. Molisch, "Cache-enabled device-to-device communications: Offloading gain and energy cost," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4519-4536, Jul. 2017.
- [11] Y. Shen *et al.*, "Device-to-device-assisted communications in cellular networks: An energy efficient approach in downlink video sharing scenario," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1575-1587, Feb. 2016.
- [12] Y. Xu, X. Li, and J. Zhang, "Device-to-device content delivery in cellular networks: Multicast or unicast," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4401-4414, May 2018.
- [13] R. Amer, H. Elsaywy, M. M. Butt, E. A. Jorswieck, M. Bennis, and N. Marchetti, "Optimizing joint probabilistic caching and communication for clustered D2D networks," [Online]. Available: <https://arxiv.org/pdf/1810.05510.pdf>.
- [14] Z. Piao, M. Peng, Y. Liu and M. Daneshmand, "Recent advances of edge cache in radio access networks for Internet of Things: Techniques, performances, and challenges," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 1010-1028, Feb. 2019.
- [15] K. Kanevay, N. Aboutoraby, S. Sorour, and M. C. Reedy, "Cross-layer offloading in fog-RANs using device cooperation and network coding" *IEEE Global Communication Workshops*, Abu Dhabi, Dec. 2018, pp. 1-6.
- [16] Y. Liu, Y. Wang, R. Sun, S. Meng, and R. Su, "Energy efficient downlink resource allocation for D2D-assisted cellular networks with mobile edge caching," *IEEE Access*, vol. 7, pp. 2053-2067, Jan. 2019.
- [17] A. Douik, H. Dahrouj, O. Amin, B. AIOuibi, T. Y. A.-Naffouri, and M.-S. Alouini, "Mode selection and power allocation in multi-level cache-enabled networks," *IEEE Communication Lett.* (accepted for publication).
- [18] W. Yu, L. Musavian, and Q. Ni, "Statistical delay QoS driven energy efficiency and effective capacity tradeoff for uplink multi-user multi-carrier systems," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3494-3507, Aug. 2017.
- [19] A. Mesodiakaki *et al.*, "Energy and spectrum efficient user association in millimeter wave backhaul small cell networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2 pp. 1810-1821, Feb. 2017.
- [20] F. Fang, J. Cheng, and Z. Ding, "Joint energy efficient subchannel and power optimization for a downlink NOMA heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2 pp. 1810-1821, Feb. 2019.
- [21] K. Singh, K. Wang, S. Biswas, Z. Ding, F. Khan, and T. Ratnarajah, "Resource-optimization in full-duplex non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4312-4325, Sept. 2019.
- [22] Y. Mao, B. Clerckx, and V. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. and Network.*, no. 133, pp. 1-54, May 2018.
- [23] B. Clerckx *et al.*, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 349-353, Mar. 2020.
- [24] H. Joudeh and B. Clerckx, "Robust transmission in downlink multiuser MISO systems: A rate-splitting approach," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6227-6242, Dec. 2016.
- [25] D. Yu, J. Kim, and S.-H. Park, "An efficient rate-splitting multiple access scheme for the downlink of C-RAN systems," *IEEE Commun. Lett.*, vol. 8, no. 6, pp. 1555-1558, Dec. 2019.

- [26] Y. Mao, B. Clerckx, J. Zhang, V. O.K. Li, and M. Arafah, "Max-min fairness of K -user cooperative rate-splitting in MISO broadcast channel with user relaying," [Online]. Available: <https://arxiv.org/pdf/1910.07843.pdf>.
- [27] W. Jaafar, S. Naser, S. Muhaidat, P.C. Sofotasios, H. Yanikomeroglu, "Multiple access in aerial networks: From orthogonal and non-orthogonal to rate-splitting,"[Online]. Available: <https://arxiv.org/pdf/2005.13122.pdf>.
- [28] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi- antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754-8770, Dec. 2019.
- [29] A. A. Ahmad, B. Matthiesen, A. Sezgin, and E. Jorswieck, "Energy efficiency in C-RAN using rate splitting and common message decoding," *IEEE Int. Conf. Commun. Wrkshps.*, Dublin, June 2020, pp. 1-6.
- [30] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff,"[Online]. Available: <https://arxiv.org/pdf/2001.03206.pdf>.
- [31] M. Z. Hassan *et al.*, "Decentralized content dissemination in fog radio access network using unsupervised learning empowered rate-splitting framework," [Online]. Available: <https://arxiv.org/pdf/2007.03827.pdf>.
- [32] K.-M. Miettinen, *Nonlinear Multi-objective Optimization*, Kluwer Academic Publishers, Boston, 1999.
- [33] Z.-Q. Luo, and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 1, pp. 57-73, Feb. 2008.
- [34] Z. Han *et al.*, *Game Theory in Wireless and Communication Networks: Theory, Models and Applications*, Cambridge, UK: Cambridge University Press, 2011.