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

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

Abstract—Multi-objective resource allocation is studied for edge-caching enabled fog-radio access network. Notably, joint maximization of the energy-efficiency (EE) and spectrum-efficiency (SE) and interference management are investigated for distributing contents from the cache-enabled fog access points (F-APs) and cloud base station (CBS) to the user devices (UDs). In our envisioned system, the UDs are grouped into multiple non-overlapping device-clusters based on their locations. A rate-splitting with common message decoding based transmission strategy is applied to enable UDs of each device-cluster to receive data from a suitably selected F-AP and CBS over the same radio resource blocks. To maximize system EE and SE jointly, a multi-objective optimization problem (MOOP) is formulated and it is solved in three stages. At first, by employing the ε-constraint method, the MOOP is converted to an EE-SE trade-off optimization problem. Then, by leveraging iterative function evaluation based power control and generalized 3D-resource matching, the EE-SE trade-off optimization problem is solved and a novel resource allocation algorithm is proposed to obtain near-optimal Pareto-front for the proposed MOOP. To reduce the complexity of obtaining near-optimal Pareto-front, a sub-optimal resource allocation algorithm is proposed as well. Finally, a low-complexity algorithm is devised to select a suitable operating EE-SE pair from the obtained Pareto-front. The conducted simulations demonstrate that the proposed resource allocation schemes achieve substantial improvement of system EE and SE over the benchmark schemes.

Index Terms—Energy efficiency, fog radio access network, interference management, rate-splitting multiple access.

I. INTRODUCTION

EVERAGING centralized cloud processing based resource allocation, distributed signal processing, and popular content caching at the network edge, fog-radio access network (F-RAN) presents a revolutionary paradigm to satisfy the beyond 5G (B5G) performance requirements [1]. In edge-caching (EC) enabled F-RAN, the content distribution phase is significantly important to ensure throughput and diverse latency requirements of the user devices (UDs). Particularly, EC enabled F-RAN jointly exploits both the cache-enabled fog access points (F-APs) and centralized cloud base station (CBS) to deliver contents to the UDs [2]. Hence, well-designed resource allocation plays a vital role for improving the content distribution performance of EC enabled F-RAN.

Because of the proliferation of mobile devices and various applications in the B5G network, there is a dire need of solution to efficiently deliver high-volume and latency-sensitive enhanced mobile broadband (eMBB) data, generated from 3D video streaming and extended reality, to the UDs. Using the joint edge-cloud processing, EC enabled F-RAN presents a promising architecture to deliver eMBB data to the UDs with reduced latency [3]. However, the advantage of F-RAN enabled eMBB data delivery is confronted by the increased interference in the dense network and scarcity of the spectrum resources. In particular, interference severely reduces the spectrum-efficiency (SE) which is detrimental to the eMBB data. The conventional approach of interference management by scheduling the active links over orthogonal spectrum resources strictly limits number of the supported UDs in EC enabled F-RAN, and thus, it is incapable to satisfy the data demand of the overly congested B5G network. A promising way to improve the spectrum resource utilization is to group the UDs into multiple device-clusters where the members of each cluster are scheduled to receive data over the same spectrum resources. However, clustering inevitably increases the interference in the network and requires a complex interference management scheme. In addition, the number of F-APs in F-RAN needs to be increased to satisfy enhanced data-rate requirements of a large number of UDs. Hence, from a network operation point of view, maximization of the system energy-efficiency (EE) is also imperative to ensure...
sustainable operation of F-RAN. Indeed, the B5G network is expected to have 10 times enhancement of EE than 5G network [4]. Accordingly, F-RAN enabled eMBB data delivery calls for a novel solution that maximizes system EE and SE jointly, and provides efficient interference management with improved spectrum resource utilization. However, developing such a solution is non-trivial. This is because on one hand, joint maximization of EE and SE requires to maximize two conflicting objectives, and on the other hand, interference management in a network with multiple device-clusters is a well-known NP-hard problem. To address these challenges, we focus on designing novel resource allocation scheme while aiming at joint maximization of system EE and SE and interference management for the content distribution phase of EC enabled F-RAN.

A. Related Works

EC significantly reduces the backhaul energy consumption and leads to improved EE [5]. In EC enabled system, the enhancement of EE is attributed by optimizing both the content placement and content distribution phases [6]. A large number of recent works have investigated on-device caching and improvement of the system EE by capitalizing the proximity gain of an optimized device-to-device (D2D) assisted content distribution phase. Using game theory, the authors in [8] designed optimized caching strategy and incentive mechanism to encourage the devices to participate in the EC system. The authors in [9] 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. The authors in [10] proposed a novel D2D-clustering based content distribution framework for improving EE of the network. In [11], 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 enabled network [12]. Besides device-caching and D2D assisted content distribution, multi-level EC was also utilized to improve EE of the system. In [13]–[15], 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. Although the aforementioned works depicted the advantages of EC for improving system EE, these works did not explore interference management for delivering the cached contents. In the most of these works, interference was treated as a noise and consequently, the resultant EE is sub-optimal. Hence, it is imperative to investigate the benefit of interference management for improving EE of EC enabled system.

Resource allocation is envisaged to play the pivotal role for interference management in EC enabled F-RAN. In the contemporary literature, non-orthogonal multiple access (NOMA) has received strong reputation to multiplex multiple UDs over the same spectrum resource(s), and consequently, NOMA has also been applied to manage interference in F-RAN. By optimizing power control and spectrum resource-UD matching, the authors in [16]–[18] developed several novel resource allocation schemes to manage interference and maximize SE of NOMA-based F-RAN. The existing literature also focused on improving EE of F-RAN through resource allocation and interference management. The authors in [19] developed cross-tier interference aware pricing mechanism for two-tier F-RAN, and developed a game theoretic resource allocation scheme. Considering the cost of content caching at the F-APs, the authors in [20] developed optimized beamforming vectors and UDF-AP associations to maximize a cost-aware EE of F-RAN. The optimized beamforming vectors and UDF-AP associations were also developed to maximize EE of F-RAN while considering outdated channel information at the cloud [21]. Note that the aforementioned works solely concentrated on optimizing either SE or EE of F-RAN in an individual manner. However, optimization of only SE is insufficient to characterize energy saving in the network, and optimization of only EE does not necessarily lead to an optimized SE. Hence, to ensure both greenness for network operation and improved SE, as required by B5G network, optimization of the trade-off between EE and SE of F-RAN is inherently important.

Rate-splitting with common message decoding (RS-CMD) has recently been emerged as a robust and resource efficient transmission strategy for mitigating interference [22]. RS-CMD enables decoding/canceling partial interference and treating partial interference as a noise, and thus, it appears as a more generalized multi-user transmission strategy than NOMA [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]. Moreover, both EE maximization and EE-SE trade-off optimization for RS-CMD systems were studied in [28]–[30]. Although the aforementioned works applied RS-CMD in multi-antenna setting, RS-CMD is also advantageous for interference management in single-antenna network [31]. Motivated by the success of RS-CMD in interference management, in this work, RS-CMD is employed for the content distribution phase of EC enabled F-RAN. Note that, for an RS-CMD integrated system, the aforementioned works investigated a downlink scenario with single transmitter and multiple receivers and did not consider the problem of association between receivers and transmitters. However, the association between UDs and F-APs is crucial in EC enabled F-RAN as mentioned in [3], [17]. Consequently, in this work, we optimize the EE-SE trade-off of RS-CMD integrated F-RAN while taking the problem of association between UDs and F-APs into account.

B. Contributions and Paper Organization

In this work, a novel cluster-based resource allocation framework is developed for EC enabled F-RAN. In our envisioned system, the UDs are clustered based on their locations. We employ RS-CMD based transmission strategy to enable UDs of each device-cluster to receive data from a suitably selected cache-enabled F-AP and CBS over the same radio resource blocks (RRBs). As a result, our envisioned system enhances utilization of the limited RRBs in the
content distribution phase of EC enabled F-RAN. However, the considered strategy leads to increased interference that affects both system EE and SE. To this end, our developed resource allocation framework aims to efficiently manage interference in the content distribution phase of EC enabled F-RAN while jointly maximizing system EE and SE. To the best of the authors’ knowledge, this is the first work that optimizes the trade-off between system EE and SE for the RS-CMD integrated and EC enabled F-RAN. The specific contributions of this work 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 F-APs and CBS, association between F-APs and device-clusters, and allocation of the RRBs among the device-clusters. To obtain the near-optimal Pareto-front of the proposed MOOP, it is converted to an EE-SE trade-off optimization problem by applying the $\epsilon$-constraint method. Since the EE-SE trade-off optimization problem is NP-hard, a three stage solution is devised.

- At the first stage, we determine the maximum achievable system EE and SE by decomposing the resource allocation into two sub-problems. The first sub-problem obtains near-optimal power allocations of the CBS and F-APs by applying an iterative function evaluation (IFE) approach. The second sub-problem obtains a Pareto-efficient matching of RRBs and F-APs with the device-clusters by solving a generalized 3D-resource matching problem. Based on a block-alternating accent method, two convergent algorithms of polynomial computational complexity are proposed to determine the maximum achievable EE and SE of the system.

- At the second stage, we propose a novel iterative rate-splitting and resource-matching (I-RSRM) algorithm to solve the EE-SE trade-off optimization problem iteratively. I-RSRM obtains the near-optimal Pareto-front of the proposed MOOP by iteratively adjusting the trade-off between EE and SE. To reduce the complexity of achieving near-optimal Pareto-front of the proposed MOOP, a sub-optimal resource allocation algorithm, referred as, decomposed rate-splitting and resource-matching (D-RSRM), is proposed as well.

- At the final stage, a low-complexity algorithm is developed to select an operating EE-SE pair from the solutions obtained by I-RSRM and D-RSRM algorithms. The developed algorithm provides a suitable trade-off between system EE and SE while adjusting a single parameter. Extensive simulations are conducted to verify the superiority of the proposed near-optimal and sub-optimal resource allocation algorithms over several benchmark schemes.

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 V and VI, respectively.

---

**II. System Model**

**A. System Overview**

Consider a downlink F-RAN with one CBS, $L$ single antenna F-APs, $N$ orthogonal RRBs, and $M$ single antenna UDs. Fig. 1 illustrates an example of such a network. Here, F-APs are connected with the CBS via fronthaul links. As per the contemporary F-RAN literature [16], [17], CBS proactively places data of the UDs at the cache-memory of F-APs in offline (i.e., before the communication takes place). As such, in online the F-APs can directly transmit data to the UDs without requiring data to be delivered from CBS over fronthaul links. The considered system is a standard F-RAN architecture [2], and it is benefited from both the cloud processing empowered resource allocation and local signal processing/encoding operations of F-APs. For simplicity, we assume that each F-AP caches identical contents. Let $\mathcal{N} = \{1, 2, \ldots, N\}$, $\mathcal{L} = \{1, 2, \ldots, L\}$ and $\mathcal{M} = \{1, 2, \ldots, M\}$ be the set of available orthogonal RRBs, F-APs, and UDs, respectively.

Based on the locations of UDs, multiple non-overlapping device-clusters are formed as such the mean-square distance between the clustered UDs and cluster center is minimized. Let, $M$ UDs be grouped into $K$ device-clusters, denoted by, $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K$ where $\mathcal{S}_k \cap \mathcal{S}_{k'} = \emptyset$, $\forall k, k'$; $\bigcup_{k=1}^K \mathcal{S}_k = \mathcal{M}$. Since the F-APs are cost-efficient devices, they can cache only a subset of the UDs’ data in their storage. To this end, each device-cluster is further decomposed into two non-overlapping sets of cache-hit and cache-miss UDs. The requested data of the cache-hit UDs is available at the F-APs, and hence, they obtain data by directly accessing the local cache. In contrast, the requested data of the cache-miss UDs are not available at the F-APs, and as a result, they can obtain the data only from CBS. Note that our proposed resource allocation framework is valid for any device-clustering algorithm. Hence, in the ensuing analysis, we consider that the device-clusters are given, and we focus on the resource allocations among the device-clusters.

In a given device-cluster, both the cache-hit and cache-miss UDs receive data from a certain F-AP and CBS, respectively, over the same RRBs. Hence, the system EE and SE depends on the mitigation of the following interference in the content distribution phase,\(^1\) namely, (i) inter-device-cluster interference, (ii) interference between F-AP and CBS in each device-cluster, (iii) intra-device-cluster interference, such as, \(^1\)Interference from the malicious and illegal devices also affects the system reliability. However, mitigation of such interference is beyond the scope of this work, and it will be considered in the future extension of this work.
interference among the cache-hit UDs and interference among the cache-miss UDs in each device-cluster. To combat inter-device-cluster interference, each device-cluster is allocated with certain orthogonal RRBs. To combat interference between F-AP and CBS in a given device-cluster, joint F-AP and CBS power allocation over the shared RRBs is leveraged. Finally, to combat intra-device-cluster interference, 1-layer RS-CMD is exploited for both the cache-hit and cache-miss UDs in each device-cluster. Along the lines of [3], [17], each device-cluster is allowed to be associated with maximum one F-AP. Meanwhile, an F-AP can serve multiple device-clusters over the orthogonal RRBs.

For analytical tractability, the following assumptions are made. A1: 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. Data transmission from all the F-APs at each TS is synchronized by the CBS. A2: UDs can accurately perform successive interference cancellation (SIC), and the locations of UDs remain fixed for a number of TSs. A3: F-AP and CBS has accurate knowledge of the local and global channel state information, respectively, of both cache-hit and cache-miss UDs in the system. A4: Using the cloud processing, all the resource allocation tasks are executed at the CBS in a centralized manner. Note that the aforementioned assumptions are common in the contemporary literature of F-RAN [2], [16], [17]. Because of these assumptions, the proposed work provides the maximum theoretical performance gain for EC enabled F-RAN by jointly optimizing power allocation of the F-APs and CBS, association between F-APs and device-clusters, and allocation of the RRBs among the device-clusters.

B. RS-CMD Based Transmission Strategy

We describe RS-CMD based transmission strategy for the j-th device-cluster, $S_j$. The cache-hit and cache-miss UDs of the j-th device-cluster are denoted by $S_{j}^{(L)}$ and $S_{j}^{(H)}$, respectively, where $S_{j}^{(L)} \cap S_{j}^{(H)} = \emptyset$ and $S_{j}^{(L)} \cup S_{j}^{(H)} = S_j$. Here, $S_{j}^{(L)} = \{1, 2, \cdots , |S_{j}^{(L)}|\}$ and $S_{j}^{(H)} = \{1, 2, \cdots , |S_{j}^{(H)}|\}$ where $|\cdot|$ refers to the cardinality of a set. We consider that the l-th F-AP and the n-th RRB are assigned to the j-th device-cluster. Hence, the UDs in the $S_{j}^{(L)}$ and $S_{j}^{(H)}$ sets receive data from the l-th F-AP and CBS, respectively, over the n-th RRB. The l-th F-AP first splits the data of the e-th UD into two parts, denoted by, $\{W_{e}^{(0)}, W_{e}^{(1)}\}$, $\forall e \in S_{j}^{(L)}$. Subsequently, the l-th F-AP combines and encodes $W_{e}^{(0)}$, $\forall e \in S_{j}^{(L)}$ into a common message denoted by $s_e$. Simultaneously, the l-th F-AP encodes $W_{e}^{(1)}$, $\forall e \in S_{j}^{(L)}$, into $|S_{j}^{(L)}|$ private messages denoted by $\{s_{e}^{(p)}\}$. The signal transmitted by the l-FAP over the n-th RRB is expressed as $X_{l,n} = \sqrt{P_{l,n}^{C,j}} s_e + \sum_{e=1}^{\mid S_{j}^{(L)} \mid} \sqrt{P_{l,n,n,e}^{p,j}} s_{e}^{(p)}$ where $P_{l,n}^{C,j}$ and $\{P_{l,n,n,e}^{p,j}\}$ denote the transmission powers assigned for the common and private messages for the cache-hit UDs in the j-th device-cluster, respectively. Similarly, CBS encodes the data of the cache-miss UDs into a common message, denoted by, $\tilde{s}_c$, and $|S_{j}^{(H)}|$ private messages, denoted by, $\{s_{i}^{(p)}\}$. The signal transmitted by the CBS over the n-th RRB is expressed as $\tilde{X}_n = \sqrt{Q_{n,j}^{C,j}} \tilde{s}_c + \sum_{i=1}^{\mid S_{j}^{(H)} \mid} \sqrt{Q_{n,j,n}^{p,j}} s_{i}^{(p)}$ where $Q_{n,j}^{C,j}$ and $\{Q_{n,j,n}^{p,j}\}$ denote the transmission powers assigned for the common and private messages for the cache-miss UDs in the j-th device-cluster, respectively. The received signals at the UDs are expressed as

$$Y_{e}^{(n)} = h_{e,n}^{(0)} X_{l,n} + n, \forall e \in S_{j}^{(L)} \quad (1a)$$

$$Y_{t}^{(n)} = h_{t,n}^{(0)} \tilde{X}_n + n, \forall t \in S_{j}^{(H)} \quad (1b)$$

where $h_{e,n}^{(0)}$ (resp. $h_{t,n}^{(0)}$) is the channel gain between the l-th F-AP and the e-th UD (resp. the t-th UD) over the n-th RRB; $h_{e,n}^{(0)}$ (resp. $h_{t,n}^{(0)}$) is the channel gain between CBS and the e-th UD (resp. the t-th UD) over the n-th RRB; and $n$ is the additive white Gaussian noise with variance $\sigma^2$.

The UDs in the $S_{j}^{(L)}$ and $S_{j}^{(H)}$ sets first decode the common messages transmitted from the l-th F-AP and CBS, respectively. All the UDs of the $S_{j}^{(L)}$ and $S_{j}^{(H)}$ sets should be able to decode their corresponding common messages. We denote $k_j = \arg \min_{e \in S_{j}^{(L)}} \frac{\mid h_{e,n}^{(0)} \mid^2}{Q_{n,j}^{C,j} + \sum_{e=1}^{\mid S_{j}^{(L)} \mid} \sqrt{Q_{n,j,n}^{p,j}} |s_{e}^{(p)}| + \sigma^2}$ and $\tilde{k}_j = \arg \min_{t \in S_{j}^{(H)}} \frac{\mid h_{t,n}^{(0)} \mid^2}{P_{l,n,n}^{p,j} + \sum_{t=1}^{\mid S_{j}^{(H)} \mid} \sqrt{P_{l,n,n}^{p,j}} |s_{i}^{(p)}| + \sigma^2}$. Consequently, for the common messages over the n-th RRB, the l-th F-AP and CBS target the achievable data rate of the UDs indexed by $k_j$ and $\tilde{k}_j$, respectively. The common message data rates for the cache-hit and cache-miss UDs are expressed as $R_{l,n}^{C,j} = \log_2 \left( 1 + \frac{\gamma_{l,n}^{C,j}}{1} \right)$ and $\tilde{R}_{l,n}^{C,j} = \log_2 \left( 1 + \frac{\gamma_{l,n}^{C,j}}{1} \right)$, respectively. Here, $\gamma_{l,n}^{C,j}$ and $\gamma_{l,n}^{C,j}$ are defined as (2) and (3), at the bottom of the next page.

After decoding the common messages, both the cache-hit and cache-miss UDs decode their own private messages while canceling the interference from the decoded common messages by applying SIC, and treating the private messages of other UDs in the system as noise. The cache-hit (cache-miss) UDs also treat the interference from the common message of the cache-miss (cache-hit) UDs as noise. Thus, the scheduled data rates for the private messages of the e-th cache-hit and t-th cache-miss UDs, $\forall e \in S_{j}^{(L)}$ and $\forall t \in S_{j}^{(H)}$, are obtained as, respectively, $R_{l,n,e}^{p,j} = \log_2 \left( 1 + \frac{\gamma_{l,n,e}}{1} \right)$ and $\tilde{R}_{l,n,t}^{p,j} = \log_2 \left( 1 + \frac{\gamma_{l,n,t}}{1} \right)$. Here, $\gamma_{l,n,e}$ and $\gamma_{l,n,t}$ are given as (4) and (5) at the bottom of the next page. Using (2)-(5), we obtain the total scheduled data rates for the cache-hit and cache-miss UDs in the j-th device-cluster, respectively, as $R_{l,n}^{C,j} = R_{l,n}^{C,j} + \sum_{e=1}^{\mid S_{j}^{(L)} \mid} R_{l,n,e}^{p,j}$ and $\tilde{R}_{l,n,t}^{C,j} = \tilde{R}_{l,n,t}^{C,j} + \sum_{t=1}^{\mid S_{j}^{(H)} \mid} \tilde{R}_{l,n,t}^{p,j}$. 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, $T_j^{(l,n)} = R_{l,n,j}^{C,j} + \tilde{R}_{l,n,t}^{C,j}$.

C. Problem Formulation

We introduce two binary variables, $a_{i,j} \in \{0, 1\}$ and $b_{n,t,j} \in \{0, 1\}$ such that $a_{i,j} = 1$ if the l-th F-AP is associated with the j-th device-cluster and $a_{i,j} = 0$ otherwise; and $b_{n,t,j} = 1$ if the n-th RRB is assigned to both the l-th F-AP
and the \( j \)-th device-cluster and \( b_{n,l,j} = 0 \) otherwise. Hence, the total power consumption of the \( l \)-th F-AP is obtained as
\[
P_l = P_c + \frac{1}{\varepsilon} \sum_{j=1}^{K} \sum_{n=1}^{N} \alpha_{l,j} b_{n,l,j} (P_{C,j}^{l,n} + \sum_{e \in S^{(l)}} P_{p,j}^{l,c,n})
\]
where \( P_c \) is the constant circuitry power consumption and \( \varepsilon \) is the power amplifier efficiency. Meanwhile, the total power consumption of the CBS is obtained as
\[
P_{CBS} = P_c + \frac{1}{\varepsilon} \sum_{j=1}^{K} \sum_{n=1}^{N} \alpha_{l,j} b_{n,l,j} (Q_{n,j}^{C} + \sum_{e \in S^{(l)}} Q_{T,j}^{p})
\]
Thereby, the system EE and SE are obtained as, respectively, \( \sum_{n,l,j} P_{C,j}^{l,n} b_{n,l,j} \) and \( \sum_{n,l,j} P_{C,j}^{l,n} b_{n,l,j} \), and \( \sum_{n,l,j} P_{p,j}^{l,c,n} b_{n,l,j} \), and \( \sum_{n,l,j} P_{p,j}^{l,c,n} b_{n,l,j} \). We formulate an optimization problem, given by \( P_0 \) at the bottom of the page, to maximize both the system EE and SE by jointly performing transmission power allocation of the F-APs and CBS, association between the F-APs and device-clusters, and allocation of RRBs among the device-clusters.

In \( P_0 \), \( C_1 \) implies that each device-cluster will be associated with only one F-AP whereas a given F-AP can be associated with maximum \( N_R \) device-clusters; \( C_2 \) implies the orthogonal allocation of the RRBs among the device-clusters; \( C_3 \) implies that the \( n \)-th RRB can be assigned to the \( l \)-th F-AP and the \( j \)-th device-cluster if and only if the \( l \)-th F-AP is associated with the \( j \)-th device-cluster; and \( C_4 \) provides the transmission power allocation constraints for F-AP and CBS with \( P_{\text{FAP}}^{\text{max}} \) and \( P_{\text{CBS}}^{\text{max}} \) as the maximum transmission power of the F-AP and CBS, respectively.

\( P_0 \) is a non-convex MOOP. The \( \varepsilon \)-constraint method can obtain Pareto-optimal solution for a non-convex MOOP [32]. Consequently, we employ the \( \varepsilon \)-constraint method to obtain a Pareto-optimal solution to \( P_0 \). By introducing a lower-bound for the system EE, denoted by \( \varepsilon \), a single objective formulation of \( P_0 \) is obtained as \( P_1 \) at the bottom of the page. Note that \( P_1 \) is an EE-SE trade-off optimization problem since each solution to \( P_1 \) provides a trade-off between system EE and SE. The following lemma justifies the efficiency of the problem formulation \( P_1 \).

*Lemma 1*: \( P_1 \) provides a weakly Pareto-optimal solution to \( P_0 \).

*Proof*: The proof is omitted, and it can be found in [34, Appendix A].

It is obvious that different values of \( \varepsilon \) will lead to different solutions to \( P_1 \), and as per *Lemma 1*, each of these solutions will be weakly Pareto-optimal to \( P_0 \). Hence, to solve \( P_0 \) optimally, we need to obtain overall set of the Pareto-optimal solutions by solving \( P_1 \) for different values of \( \varepsilon \) 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 [35]. In what follows, we first obtain the maximum achievable EE and SE of the system, and afterwards, we solve \( P_1 \) by varying \( \varepsilon \).
III. EE-OPTIMAL AND SE-OPTIMAL SOLUTIONS

To determine the maximum achievable EE and SE of the system by exploiting the considered resource allocation strategy, we formulate the following two optimization problems.

\[
P_2.1 : \max_{P \geq 0, Q \geq 0, \{a,b\} \in \{0,1\}} \text{EE} (P, Q, a, b) \\
\text{s.t. } C1,C2,C3,C4
\]

(8)

\[
P_2.2 : \max_{P \geq 0, Q \geq 0, \{a,b\} \in \{0,1\}} \text{SE} (P, Q, a, b) \\
\text{s.t. } C1,C2,C3,C4.
\]

(9)

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

\[
\max_{(P, Q, a, b) \mid C1,C2,C3,C4} \left( \sum_{j=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{n,l,j} T_{j}^{(l,n)} \right) - \eta \left( \sum_{l \in L} P_{l} + P_{CBS} \right) = 0. \tag{10}
\]

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.

\[
P_2.3 : \max_{P \geq 0, Q \geq 0} \sum_{j=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{n,l,j} T_{j}^{(l,n)} - \eta \left( \sum_{l \in L} P_{l} + P_{CBS} \right) \\
\text{s.t. } C1,C2,C3,C4.
\]

(11)

However, in P2.3, \(\eta\) also depends on the resource allocation variables as depicted from (10). Hence, we first solve P2.3 for a given \(\eta\), and then, we update \(\eta\) as such (10) is satisfied. Moreover, 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. CBS and F-AP Power allocation

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

\[
P_2.4 : \max_{P \geq 0, Q \geq 0} \sum_{j=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{n,l,j} T_{j}^{(l,n)} - \eta \left( \sum_{l \in L} P_{l} + P_{CBS} \right) \\
\text{s.t. } C3,C4. \tag{12}
\]

Although P2.4 is NP-hard, a high quality approximate optimal solution can be obtained by applying Lagrangian relaxation technique [33]. Therefore, to solve P2.4, we formulate a Lagrangian dual-optimization problem and solve it by applying the well-known Karush-Kuhn-Tucker (KKT) condition. The Lagrangian function corresponding to P2.4 is expressed as (13) at the bottom of the next page. Here, \(\{\lambda_i\}\) and \(\nu\) 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, \(\forall j = 1, 2, \ldots, K\), is expressed as

\[
P_2.5 : \left\{ \begin{array}{l} \tilde{P}_{C,j}, \tilde{P}_{C,j}, \tilde{Q}_{C,j}, \tilde{Q}_{C,j} \\ \tilde{P}_{F,j}, \tilde{P}_{F,j}, \tilde{Q}_{F,j}, \tilde{Q}_{F,j} \end{array} \right\} = \arg \max_{P \geq 0, Q \geq 0} L_j \tag{14}
\]

where \(L_j\) is defined as (15) at the bottom of the next page. To apply the KKT condition, P2.5 is decomposed into inner and outer layer problems. The inner and outer layer problems iteratively update the power allocation variables and the Lagrangian multipliers, respectively.

1) Power Allocation Update: Let the Lagrangian multipliers be given. Capitalizing the IFE based power control method in [15], the near-optimal power allocations of the \(j\)-th device-cluster, \(\forall j = 1, 2, \ldots, K\), are obtained in the following proposition.

**Proposition 1**: Let \(\{\tilde{P}_{C,j}, \tilde{P}_{F,j}, \tilde{Q}_{C,j}, \tilde{Q}_{F,j}\}\) 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 (16)-(19) given at the bottom of the next page. The values of \(\gamma_{t,n}^{j}, \gamma_{l,e,n}^{j}, \gamma_{t,n}^{j}, \gamma_{l,e,n}^{j}\) and \(\tilde{P}_{C,j}, \tilde{P}_{F,j}\) are obtained by plugging \(\{\tilde{P}_{C,j}, \tilde{P}_{F,j}, \tilde{Q}_{C,j}, \tilde{Q}_{F,j}\}\) to (2)-(5).

**Proof**: The proof is omitted, and it can be found in [34, Appendix B].

2) Lagrangian Multiplier Update: Using (16)-(19), the power consumption of the \(l\)-th F-AP and CBS are calculated as \(P_l\) and \(P_{CBS}\), respectively. Leveraging the well-known sub-gradient method, at the \((k+1)\)-th iteration, the Lagrangian multipliers are updated as

\[
\lambda_k = \max \left[ \lambda_{k-1} + \zeta_{k}(k) \left( \tilde{P}_{l} - P_{FAP}^{\text{max}} \right) , 0 \right], \forall l \in L \tag{20a}
\]

\[
\nu_k = \max \left[ \nu_{k-1} + \zeta_{2}(k) \left( \tilde{P}_{CBS} - P_{CBS}^{\text{max}} \right) , 0 \right] \tag{20b}
\]

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 \(\sum_{k=1}^{\infty} \zeta_{i}^{2}(k) < \infty, i = 1, 2\) satisfied.

3) Algorithm Development: We propose Algorithm 1 to obtain power allocations in the \(j\)-th device-cluster. For a given device-cluster, Algorithm 1 requires \(O(LN)\) computational complexity to obtain a converged solution to P2.4. To demonstrate efficiency of Algorithm 1, we introduce the following proposition.

**Proposition 2**: Algorithm 1 converges to a stable and near-optimal solution to P2.4.
B. 3D-Resource Matching

1) Algorithm Development: For a given set of power allocation variables, (2.3) is written as

\[
P_2.3: \quad \max_{\{a,b\} \in \{0,1\}} \sum_{j=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{n,l,j} \tilde{U}^{(j)}_{l,n}, \quad \text{s.t.} \quad C1, \ C2 \ (21)
\]

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

\[
\tilde{U}^{(j)}_{l,n} = T_{j}^{(l,n)} \left( P_{c,j}^{C_{l,n}}, P_{b,j}^{P_{l,e,n}}, Q_n^{C_{l,n}}, Q_n^{P_{l,e,n}} \right) - \frac{\eta}{\varepsilon} \left( P_{c,j}^{C_{l,n}} + \sum_{e \in S_{j}^{(L)}} P_{b,j}^{P_{l,e,n}} + Q_n^{C_{l,n}} + \sum_{t \in S_{j}^{(H)}} Q_n^{P_{l,e,n}} \right). \tag{22}
\]

To solve P2.6, we describe the problem of assigning F-APs and RRBS among the device-clusters as a 3D-resource matching problem. The rules of matching and preference-orders of the device-clusters, F-APs, 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, \( S \times L \), \( S \times N \), and \( L \times N \), into the elements of following sets \( N \), \( L \), and \( S \), respectively. For all \( \{j,l,n\} \in \{S,L,N\} \), such a mapping satisfies the following conditions.

\[
\mathcal{L} = \sum_{j=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{l,n,j} \tilde{T}^{(l,n,j)} - \eta \left( \sum_{l \in L} P_{l} + P_{FAP} \right) - \sum_{l \in L} \lambda_l \left( P_{l} - P_{{\text{max}}^{D}} \right) - \eta \left( \sum_{l \in L} P_{l} - P_{{\text{max}}^{D}} \right). \tag{13}
\]

\[
\mathcal{L}_j = \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{l,n,j} \left( T_{j}^{(l,n,j)} - \frac{\eta}{\varepsilon} \frac{\lambda_l}{\lambda_j} \left( P_{c,j}^{C_{l,n}} + \sum_{e \in S_{j}^{(L)}} P_{b,j}^{P_{l,e,n}} + Q_n^{C_{l,n}} + \sum_{t \in S_{j}^{(H)}} Q_n^{P_{l,e,n}} \right) \right). \tag{15}
\]

\[
P_{c,j}^{C_{l,n}} = \frac{\gamma_{c,j}^{C_{l,n}} / (1 + \gamma_{c,j}^{C_{l,n}})}{1/\ln 2 + \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} + \sum_{t \in S_{j}^{(H)}} \left( \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} \right)^2 + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2}}, \tag{16}
\]

\[
Q_n^{C_{l,n}} = \frac{\gamma_{c,j}^{C_{l,n}} / (1 + \gamma_{c,j}^{C_{l,n}})}{1/\ln 2 + \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} + \sum_{e \in S_{j}^{(L)}} \left( \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} \right)^2 + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2}}, \tag{17}
\]

\[
P_{b,j}^{P_{l,e,n}} = \frac{\gamma_{p,j}^{P_{l,e,n}} / (1 + \gamma_{p,j}^{P_{l,e,n}})}{1/\ln 2 + \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} + \sum_{e \in S_{j}^{(L)}} \left( \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} \right)^2 + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2}}, \tag{18}
\]

\[
Q_n^{P_{l,e,n}} = \frac{\gamma_{p,j}^{P_{l,e,n}} / (1 + \gamma_{p,j}^{P_{l,e,n}})}{1/\ln 2 + \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} + \sum_{e \in S_{j}^{(L)}} \left( \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2} \right)^2 + \frac{\gamma_{p,j}^{P_{l,e,n}}}{1/\ln 2} \frac{\gamma_{c,j}^{C_{l,n}}}{1/\ln 2}}. \tag{19}
\]
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 \((\tilde{j}, \tilde{n})\) pairs will be matched with the same F-AP. Condition 2) implies that if the n-th and \( \tilde{n} \)-th RRBs are matched with the l-th and l-th F-APs, 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-AP over all the RRBs cannot 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-AP, 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

\[
(\ell, n) \succ_j (l', n') \iff \tilde{U}^{(j)}_{l,n} > \tilde{U}^{(j')}_{l',n}, \forall l, l' \in \mathcal{L}, \forall n, n' \in \mathcal{N}. \quad (23)
\]

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

\[
(j, n) \succ_l (j', n') \iff \tilde{U}^{(j)}_{l,n} > \tilde{U}^{(j')}_{l,n}, \forall j, j' \in \mathcal{S}, \forall n, n' \in \mathcal{N}. \quad (24)
\]

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

\[
(j, l) \succ_n (j', l') \iff \tilde{U}^{(j)}_{l,n} > \tilde{U}^{(j')}_{l',n}, \forall j, j' \in \mathcal{S}, \forall l, l' \in \mathcal{L}. \quad (25)
\]

In this matching problem, as the players are matched with their preferred pairs according to (23)-(25), either the system SE is increased or the power consumption of the system is reduced. However, for this matching problem, finding a stable 3D-matching is NP-hard and computationally intractable [36]. Consequently, we develop a heuristic algorithm to obtain matching among the device-clusters, F-APs, and RRBs. The overall matching has the following three phases.

F-AP Matching (FM) Phase: In FM phase, the device-clusters are matched with suitable F-APs as such a single device-cluster is matched with only one F-AP and any F-AP is matched with maximum \( N_R \) device-clusters. In this phase, two different sets, namely, \( \text{UM}_D \) and \( \text{UM}_R \) are maintained for the unmatched device-clusters and the F-APs capable of accepting device-clusters, respectively. At each iteration of the FM phase, an eligibility score, \( D_j \), is calculated for \( \forall j \in \text{UM}_D \), and it provides the eligibility of the unmatched device-clusters to be matched with their most preferred F-AP in the \( \text{UM}_R \) set. The unmatched device-cluster having maximum eligibility score is matched with its most preferred F-AP in the \( \text{UM}_R \) set based on (23), and such a device-cluster is removed from the \( \text{UM}_D \) set. The \( \text{UM}_R \) set is updated as well. The FM phase is continued until all the device-clusters are matched with certain F-APs.

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-APs 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-AP pair according to (25). 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 \( \Phi_2(l, n) = n \) and \( \Phi_2(j, \tilde{n}) = \tilde{n} \). If \( \tilde{U}^{(j)}_{l,n} + \tilde{U}^{(\tilde{j})}_{l,n} > \tilde{U}^{(j)}_{l,n} + \tilde{U}^{(\tilde{j})}_{l,n} \) is satisfied, the \((n, \tilde{n})\)-th RS is 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-AP to the \( j \)-th device-cluster, \( N_R^\star \) denotes set of the assigned RRBs to the \( j \)-th device-cluster, \( K_i \) is the number of device-clusters associated with the \( l \)-th F-AP, and \( \text{UM}_N \) is the set of the unmatched RRBs.

2) Property Analysis of Algorithm 2: Although the proposed Algorithm 2 is heuristic, it provides a single-sided Pareto-efficient matching for the device-clusters. To this end, we first introduce the following definitions.

Definition 3: Two device-clusters will be Pareto-improvement pair if by partially or completely exchanging their matched resources with each other, their overall utility is strictly improved.

Definition 4: The matching obtained by Algorithm 2 will be Pareto-efficient for the device-clusters if it does not contain any Pareto-improvement pair.

Proposition 3: The final matching obtained by Algorithm 2, \( \Phi^\star \), is Pareto-efficient for the device-clusters.

Proof: The proof is omitted, and it can be found in [34, Appendix D].

Consequently, despite being heuristic, Algorithm 2 is efficient. We obtain solution to P2.6 from \( \Phi^\star \) as follows. Particularly, \( \forall j \in \mathcal{S}, a_{j,j} = 1 \) and \( b_{n,l,j} = 1 \) if \( \Phi^\star(j, n) = l, \Phi^\star(l, n) = j, \Phi^\star(j, l) = n, \) and \( a_{j,j} = 0 \) and \( b_{n,l,j} = 0 \), otherwise. The following proposition establishes the near-optimality of the derived solutions.

Proposition 4: For the given power allocations, Algorithm 2 provides a near-optimal solution to P2.6.

Proof: The proof is omitted, and it can be found in [34, Appendix E].

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 an F-AP 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 \( \text{UM}_R \) is computed for the F-APs. Therefore, the required computation of the FM phase is \( \sum_{j=1}^{|\mathcal{S}|} j + L = \frac{|\mathcal{S}|(|\mathcal{S}|+1)}{2} + |\mathcal{S}|L \). For each unmatched RRB in the RM phase, \( O(|\mathcal{S}|) \) operation is needed to execute the Step (10) of the Algorithm 2. Since the total number of unmatched RRBs in the RM phase is \( N - |\mathcal{S}| \), the total number of required
Algorithm 2 3D-Resource Matching Algorithm
1: Input: \( \bar{U}_{ij} \), \( \forall j \in r, l \in L, n \in N \).
2: Initialize: \( \text{UM}_D = S, \text{UM}_N = N, L_j^* = \emptyset, N_j^* = \emptyset, K_1 = 0, \forall j \in S, l \in L. \)
3: while \( \text{UM}_D \neq \emptyset \) do (Beginning of FM phase)
   4: Determine \( \text{UM}_R = \{ l \in L | K_l < N_R \} \).
   5: For each \( j \in \text{UM}_D \), determine
      \[
      (l_j^*, n_j^*) = \arg \max_{l \in \text{UM}_R, n \in \text{UM}_N} \bar{U}_{ij} \\
      \text{and} \]
      \[
      D_j = \bar{U}_{l_j^*, n_j^*} - \max_{l \neq l_j^*, n \in \text{UM}_N} \bar{U}_{ij} 
      \]
   6: Select \( \tilde{j} = \arg \max_{j \in \text{UM}_D} D_j \); assign \( L_j^* \leftarrow l_j^* \) and \( N_j^* \leftarrow N_j^* \cup \{ n_j^* \} \).
   7: Update \( \text{UM}_D \leftarrow \text{UM}_D \setminus \{ S_j \} \), \( \text{UM}_N \leftarrow \text{UM}_N \setminus \{ n_j^* \} \), and \( K_1 = K_2 + 1 \).
8: end while (End of FM phase)
9: repeat \( \forall n \in \text{UM}_N \) (Beginning of RM phase)
   10: Find \( j' = \arg \max_{j \in S} \{ \tilde{j} \} \), \( N_j^* \leftarrow N_j^* \cup \{ n_j^* \} \), and \( \text{UM}_N \leftarrow \text{UM}_N \setminus \{ n_j^* \} \).
   11: until \( \text{UM}_N = \emptyset \) (End of RM phase)
   12: repeat \( \forall j \in S \) and \( \forall \tilde{j} \in S \setminus \{ j \} \) (Start of RS phase)
   13: Find \( n_j^* = \arg \max_{n \in N_j^*} \bar{U}_{l_j^*, n} \).
   14: if \( \exists n \in N_j^* \) such that the \( j \)-th and \( \tilde{j} \)-th device-clusters are swappable over the \( (n, \tilde{n}) \)-th RRBs then
      \[
      N_j^* \leftarrow N_j^* \setminus \{ n \} \cup \{ \tilde{n} \} \text{ and } N_j^* \leftarrow N_j^* \setminus \{ \tilde{n} \} \cup \{ n \}. 
      \]
   15: end if
   16: until The system does not have any swappable device-cluster pair (End of RS phase)
18: Output: Final matching \( \Phi^* \).

Algorithm 3 EE-Optimal Resource Allocation Algorithm
1: Initialize: \( \text{EE } \eta^{(1)} = \eta_0, \) assignment variables \{a, b\}, iteration index \( d = 1 \).
2: repeat
   3: Update the power allocations, \( \{ P, Q \} \), for all the device-clusters by using the value of \( \eta^{(d)} \) in Algorithm 1.
   4: Update the assignment variables, \{a, b\}, by using Algorithm 2 while plugging the updated power allocation variables of the previous step in (22).
   5: Update \( \eta^{(d+1)} = \frac{\sum_{l=1}^{K} \sum_{j=1}^{N} a_{l,j} b_{n,l,j} \phi_{l,j}(n)}{\sum_{l \in L} P_l + P_{C_{BS}}} \), \( d = d + 1 \).
   6: until \( \left| \eta^{(d+1)} - \eta^{(d)} \right| \leq \text{Tolerance or Maximum number of iterations is reached.} \)

Algorithm 4 SE-Optimal Resource Allocation Algorithm
1: Initialize: Assignment variables \{a, b\}, iteration index \( d = 1 \), and set \( \eta = 0 \).
2: repeat
   3: Update the power allocations, \( \{ P, Q \} \), for all the device-clusters by using Algorithm 1.
   4: Update the assignment variables, \{a, b\}, by using Algorithm 2 while plugging the updated power allocation variables of the previous step in (22), \( d = d + 1 \).
   5: until \( \text{SE } (P, Q, a, b) \) converges or Maximum number of iterations is reached.

Proof: The proof is omitted, and it can be found in [34, Appendix F].

The computational complexity of Algorithm 3 and Algorithm 4 is explored as follows. Let the convergence of Algorithm 3 and Algorithm 4 requires \( \Delta_{\text{max}}^{(1)} \) iterations. Using the reported computational complexity of Algorithms 1 and 2, the worst-case computational complexity of Algorithm 3 and Algorithm 4 is obtained as \( O\left( \Delta_{\text{max}}^{(1)} (KLN + K^2N + KL) \right) \approx O\left( \Delta_{\text{max}}^{(1)} (KLN + K^2N) \right) \).

IV. EE-SE Trade-Off Optimization and Proposed Algorithms

A. Near-Optimal Algorithm Design

We obtain the Pareto-front of P0 by solving P1. P1 can be solved by applying a Lagrangian optimization technique. Let \( \mu \) be the Lagrangian multiplier correspond to C5 constraint in P1. The partial Lagrangian of P0 is written as

\[
\mu = \max_{P, Q} \min_{a, b} \left( 1 + \mu \right) \sum_{j=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{n,l,j} \phi_{l,j}(n) - \mu \left( \sum_{l \in L} P_l + P_{C_{BS}} \right) \quad (26)
\]

For each value of \( \epsilon \), a local optimal solution to P1 can be obtained as

\[
P_{3.1} = \{ P^*(\epsilon), Q^*(\epsilon), a^*(\epsilon), b^*(\epsilon) \} = \arg \min_{\mu \geq 0, \phi_{l,j}(n)} \max_{\{a, b\} \in (0, 1)} \left( 1 + \mu \right) \sum_{j=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} a_{l,j} b_{n,l,j} \phi_{l,j}(n) - \mu \left( \sum_{l \in L} P_l + P_{C_{BS}} \right) \quad (27)
\]

C. Algorithms for Maximizing EE and SE

Capitalizing the 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 Algorithm 3 and Algorithm 4 iteratively update solutions to P2.4 and P2.6 by adopting a block-alternating accent method. Since both P2.4 and P2.6 are near-optimally solved, Algorithm 3 and Algorithm 4 must converge as confirmed by the following proposition.

Proposition 5: Algorithm 3 and Algorithm 4 converge to near-optimal system EE and SE, respectively.
Note that, P3.1 and P2.3 have a similar structure except some scalar multipliers. Hence, for the given values of $\mu$ and $\epsilon$, 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 $\mu^*$ 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 (weakly) Pareto optimal solutions to P0 is defined as the Pareto-front of P0. To obtain the Pareto-front of P0, we introduce the following proposition.

**Proposition 6:** 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 allocation variables. The near-optimal Pareto-front of P0 can be achieved by solving P3.1 while varying $\epsilon$ 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 omitted, and it can be found in [34, Appendix G].

Based on Proposition 6, we propose an I-RSRM algorithm. The overall steps of I-RSRM algorithm are summarized in Algorithm 5. 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. However, since P0 is a non-convex and NP-hard multi-objective optimization problem, the global optimality of the Pareto-front obtained by I-RSRM algorithm is not guaranteed. Nevertheless, by selecting a small step size $\Delta$ and increasing $N_F$, I-RSRM can obtain a good approximation of the optimal Pareto-front. Accordingly, I-RSRM provides a near-optimal Pareto-front for the multi-objective resource optimization problem, P0.

**B. Low-Complexity Algorithm Design**

Despite near-optimality, I-RSRM algorithm requires significant large number of iterations, and number of iterations increases with the size of network. Consequently, the convergence rate of I-RSRM algorithm may not affordable in large scale network. To reduce the number of iterations, 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-APs and RRBs among the device-clusters by considering a uniform power allocation, and keeps such assignments fixed for the remaining steps. Essentially, the total number of iterations is 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. However, the Pareto-front obtained by D-RSRM is sub-optimal.

**C. Operating EE-SE Pair Selection Algorithm**

In this section, we explain the procedure of selecting a suitable operating EE-SE pair from the Pareto-front obtained by the proposed I-RSRM and D-RSRM algorithms. Note that 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) improvement. Assume that the system allows maximum $\alpha$ percentage loss of its optimal EE or SE. Thus, the system can either trade-off maximum $\alpha$ percentage of the optimal EE or the optimal SE. Here, $\alpha$ is a system defined EE-SE trade-off control parameter. We need to decide whether it is beneficial to trade-off $\alpha$ percentage of optimal EE (resp. optimal SE) for having a gain in the achievable SE (resp. achievable EE) of the system. To this end, we denote $EE^{**}$ and $SE^{**}$ by the achievable EE and SE given the system is willing to trade-off $\alpha$ percentage of optimal EE and SE, respectively. Let, $EE^*$ and $SE^*$ be the maximum EE and SE in the available Pareto-front, respectively. To determine $EE^{**}$ and $SE^{**}$, we formulate two optimization problems, given by, (28) and (29) at the bottom of the next page.

![Algorithm 5 Iterative Rate Splitting and Resource Matching (I-RSRM) Algorithm](Image)

In both (28) and (29), the resource allocation variables are taken from the output of the proposed I-RSRM (D-RSRM) algorithm. Hence, both (28) and (29) are solved by performing simple search over the EE and SE pairs in the obtained Pareto-front. Thereby, we obtain two possible operating EE-SE pairs, such as, (i) $EE^{**}, (1 - \frac{\alpha}{100}) SE^{**}$ obtained by trading-off $\alpha$ percentage of optimal SE and (ii) $(1 - \frac{\alpha}{100}) EE^*, SE^{**}$ obtained by trading-off $\alpha$ percentage of optimal EE. For the ease of description, we denote $(EE^*, (1 - \frac{\alpha}{100}) SE^{**})$ and $(1 - \frac{\alpha}{100}) EE^*, SE^{**}$ as EE-SE pair-A and EE-SE pair-B,
Algorithm 6 Distributed Rate Splitting and Resource Matching (D-RSRM) Algorithm
1: Initialize: Equal power allocations of the F-APs and CBS among all the RRBs and \( \eta = \eta_0 \);
2: Determine assignment variables \((a, b)\) by using Algorithm 2.
3: By plugging \((a, b)\) to Algorithms 3 and 4, determine EE-optimal and SE-optimal power allocations, respectively. Denote SE-optimal and EE-optimal resource allocations by \((P_{SE}, Q_{SE}, a, b)\) and \((P_{EE}, Q_{EE}, a, b)\), respectively.
4: Initialize: \(\epsilon^{(1)} = EE(P_{SE}, Q_{SE}, a, b)\), iteration index \(m = 1\), step size \(\Delta\).
5: while \(EE(P_{EE}, Q_{EE}, a, b) > \epsilon^{(m)}\) do
6: Initialize \(\mu\).
7: repeat
8: Update power allocation variables using Algorithm 1 while replacing \(\tau\) and \(1/\ln 2\) in (16)-(19) with \((a/m)\) and \((1 + \mu)/\ln 2\), respectively.
9: Update \(\mu\) 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 \((P^{*}(\epsilon^{(m)}), Q^{*}(\epsilon^{(m)}), a, b)\). Update \(\epsilon^{(m+1)} = \epsilon^{(m)} + \Delta, m = m + 1\).
12: end while
13: Output: Set of resource allocations \(\{(P^{*}(\epsilon^{(m)}), Q^{*}(\epsilon^{(m)}), a, b)\}, \forall m = 1, 2, \cdots, N_F\).

respectively. Our target is to select either EE-SE pair-A or EE-SE pair-B as such the benefit of trade-off (B.O.T) is maximized. To calculate the B.O.T metric, we require the following information: (1) the net (relative) EE gain of EE-SE pair-A for trading-off \(\alpha\) percentage of optimal SE, and the net (relative) SE gain of EE-SE pair-B for trading-off \(\alpha\) percentage of optimal EE; (2) the relative gain (or loss) of EE for selecting EE-SE pair-A (resp. EE-SE pair-B) instead of EE-SE pair-B (resp. EE-SE pair-A), and (3) the relative gain (or loss) of SE for selecting EE-SE pair-A (resp. EE-SE pair-B) instead of EE-SE pair-B (resp. EE-SE pair-A). By linearly combining the aforementioned gains, we obtain the B.O.T metrics for both EE-SE pair-A and EE-SE pair-B. Then, we select the EE-SE pair having the maximum B.O.T as the operating EE-SE pair. The B.O.T metrics for EE-SE pair-A and EE-SE pair B are calculated as (30a) and (30b), respectively, at the bottom of the page.

In (30a) and (30b), \(\{w_1, w_2, w_3\}\) is a set of network determined weight-factors. In (30a) and (30b), \(\chi_{EE}\) and \(\chi_{SE}\) are defined as, respectively,

\[
\chi_{EE} = \frac{EE^{**} - \widetilde{EE}_{SE=SE^*}}{EE_{SE=SE^*}} \times 100
\]

\[
\chi_{SE} = \frac{SE^{**} - \widetilde{SE}_{EE=EE^*}}{SE_{EE=EE^*}} \times 100
\]

where \(\widetilde{EE}_{SE=SE^*}\) is the achievable EE when the system operates with its maximum SE, \(SE^*\), and \(\widetilde{SE}_{EE=EE^*}\) is the achievable SE when the system operates with its maximum EE, \(EE^*\). Here, \(\chi_{EE}\) provides the percentage gain of EE when the system allows maximum \(\alpha\) percentage loss of its optimal SE, and \(\chi_{SE}\) provides the percentage gain of SE when the system allows maximum \(\alpha\) percentage loss of its optimal EE. Obviously, if \(BOT_1 > BOT_2\) holds, it is more beneficial to trade-off \(\alpha\) percentage of the optimal SE and the operating EE-SE pair of the system is selected as \((EE^{**}, (1 - \frac{\alpha}{100}) SE^*)\). Conversely, if \(BOT_2 > BOT_1\) holds, it is more beneficial to trade-off \(\alpha\) percentage of the optimal EE and the operating EE-SE pair of the system is selected as \((1 - \frac{\alpha}{100}) EE^*, SE^{**}\). The overall steps of selecting a suitable EE-SE pair and the corresponding resource allocation variables are summarized in Algorithm 7.

Note that the optimality of Algorithm 7’s output depends on the optimality of the available Pareto-front. Accordingly, by increasing the number of solutions of I-RSRM and D-RSRM algorithms, the near-optimality of both the obtained Pareto-front and Algorithm 7’s output can be improved.

D. Implementation and Computational Complexity Analysis

Implementation: Leveraging the reliable frontnal links between CBS and F-APs and the computation resources of the CBS, both I-RSRM and D-RSRM algorithms are executed at the CBS. In each TS, CBS first collects the global CSI of the
Algorithm 7 Operating EE-SE Pair Selection Algorithm

1: **Input:** Pareto-front achieved by I-RSRM/D-RSRM algorithm and the EE-SE trade-off control parameter $\alpha$.
2: Calculate $EE^*$ and $SE^*$ by solving (28) and (29), respectively.
3: Calculate $BOT_1$ and $BOT_2$ by using (30a) and (30b), respectively.
4: **if** $BOT_1 > BOT_2$ **then**
5: Select $(EE^*, (1 - \frac{\alpha}{100}) SE^*)$ as the operating EE-SE pair.
6: **else**
7: Select $((1 - \frac{\alpha}{100}) EE^*, SE^*)$ as the operating EE-SE pair.
8: **end if**
9: By inspecting the outputs of the I-RSRM/D-RSRM algorithm, obtain the resource allocation variables correspond to the operating EE-SE pair.

Network.

V. Simulation Results

In this section, the performance of the proposed resource allocation algorithms is evaluated by simulations. For simulations, we consider a $300m \times 300m$ square F-RAN cell with one CBS in the center, 28 UD, 10 F-APs, and 36 RRBs. The USs are grouped into 7 device-clusters where cluster-1, cluster-2, cluster-3, cluster-4, cluster-5, cluster-6, and cluster-7 contain 5 USs, 3 USs, 3 USs, 5 USs, 4 USs, 3 USs, and 5 USs, respectively. The F-APs are deployed on a circle of $35m$ radius. The minimum distance from the F-APs to USs and the minimum distance from the CBS to USs are set as $40m$ and $75m$, respectively. An example of the considered simulation setting is provided in [34, Fig. 2]. In each device cluster, the maximum number of USs is considered as cache-miss USs and remaining USs are considered as cache-hit USs where $M'$ is the total number of USs in the device-cluster. In the ensuing simulation results, each simulation experiment is repeated 25 times with randomly chosen coordinates of the USs at the device-clusters and locations of F-APs, and the average result is presented. The channel coefficients 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) [37]. Similar to [38], we consider $P^{F\text{AP}} = P^{CBS} = 33$ dBm, $P_{c} = 0.2$ Watt, $\epsilon = 0.7$, $\sigma^2 = -174$ dBm. Finally, for selecting a suitable EE-SE pair, we consider Algorithm 7, we consider $\{w_i\} = 1, \forall i \in \{1, 2, 3\}$, in (30a) and (30b). We consider the following benchmark schemes. For all the benchmark schemes, the operating EE-SE pair is selected by employing the method described in Section V. C.

- **Downlink multicasting:** Downlink multicasting approach entirely eliminates the intra-device cluster interference, such as, interference among the cache-hit USs and interference among the cache-miss USs of each device-cluster. To apply multicasting in a given device-cluster, both the F-AP and CBS first combine the messages of the cache-hit USs and cache-hit USs, respectively. Then, the F-AP and CBS transmit the combined message over the RRBs using two rates as such the transmitted message from the F-AP and CBS are decodable to all the cache-hit and cache-miss USs, respectively. To mitigate interference between F-AP and CBS in the device-clusters, transmission power of both F-AP and CBS are jointly optimized by using Algorithm 1. To mitigate inter device-cluster interference, Algorithm 2 is used to obtain the associations among device-clusters, F-APs, and RRBs.

- **Unicast and treat interference as noise (Unicast/TIN):** Unicast/TIN eliminates the interference between F-AP and CBS of the considered system model by allowing each device-cluster to be solely supported via CBS. However, this approach is subject to both inter device-cluster and intra device-cluster interference. To eliminate inter device-cluster interference, a many-to-one matching algorithm is applied to allocate orthogonal RRBs among the device-clusters. To mitigate intra device-cluster...
interference, an iterative function evaluation based power control is applied to allocate CBS’s power over the RRBs and UD.

- **I-RSRM with iterative resource assignment (I-RSRM/IRA):** I-RSRM/IRA applies similar resource allocation framework of the proposed I-RSRM algorithm. The only difference between I-RSRM and I-RSRM/IRA is that instead of 3D-resource matching, I-RSRM/IRA applies an alternating optimization technique, given in Algorithm 3 of [39], to determine the associations among device-clusters, F-APs, and RRBs.

**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 $\alpha$. For all the considered device-clusters in Figs. 2a and 2b, when $\alpha$ approaches small or large value, such as $\alpha = 10\%$ or $\alpha = 60\%$, a large system EE is achieved, and when $\alpha$ approaches moderate value, such as $\alpha = 30\%$, a large system SE is achieved. Moreover, for a given range of $\alpha$, 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 $\alpha$ is provided as follows.

- **Small $\alpha$:** When $\alpha \rightarrow 0$, the solutions to (28) and (29) approach the corner points of the developed Pareto-front. Particularly, the EE obtained by trading-off $\alpha$ percentage of optimal SE approaches the achievable EE of the system operating with maximum SE, and the SE obtained by trading-off $\alpha$ percentage of optimal EE approaches the achievable SE of the system operating with maximum EE. Essentially, as $\alpha \rightarrow 0$, $EE^\star \rightarrow \bar{EE}_{SE=SE^\star}$ and $SE^\star \rightarrow \bar{SE}_{EE=EE^\star}$, and consequently, $\chi_{EE} \rightarrow 0$ and $\chi_{SE} \rightarrow 0$ are satisfied. We can illustrate that in the considered simulation setup, the relative EE gap between the two corner points of the developed Pareto-front is much larger than relative the SE gap. Hence, for small $\alpha$, $EE^\star \rightarrow \bar{EE}_{SE=SE^\star}$ and $SE^\star \rightarrow \bar{SE}_{EE=EE^\star}$ are usually satisfied. Recall, when $BOT_2 > BOT_1$ holds, the achievable EE and SE of the system are given by $(1 - \frac{\alpha}{100}) EE^\star$ and $SE^\star$, respectively. Since $\alpha$ is small, the system EE approaches the optimal EE, and owing to the conflicting behavior of EE and SE, SE becomes small. Accordingly, when $\alpha$ is small, the considered system example obtains large EE and small SE.

- **Large $\alpha$:** When $\alpha$ is large, the SE and EE constraints of (28) and (29) are relaxed. As a result, $EE^\star \rightarrow EE^\star$ and $SE^\star \rightarrow SE^\star$ are satisfied. Therefore, eqs. (30a) and (30b) can be modified as $BO_{T_1} \rightarrow \chi_{EE} - \alpha + (\frac{\alpha}{100})^2 \frac{1}{1-\frac{\alpha}{100}}$ and $BO_{T_2} \rightarrow \chi_{SE} - \alpha + (\frac{\alpha}{100})^2 \frac{1}{1-\frac{\alpha}{100}}$, respectively. Since the relative EE gap between the two corner points of the developed Pareto-front is much larger than the relative SE gap, $\chi_{EE} > \chi_{SE}$ and $BO_{T_1} > BO_{T_2}$ are satisfied. Recall, when $BO_{T_2} > BO_{T_1}$ holds, the achievable EE of the system is given by $EE^\star$, and as mentioned earlier that for a large $\alpha$, $EE^\star$ approaches the optimal system EE. Consequently, the considered system example also achieves large EE as $\alpha$ becomes large.

Meanwhile, between the aforementioned two extreme cases, certain $\alpha$ does exist for which system chooses a strategy that obtains a near-optimal SE. Fig. 2a illustrates that as $\alpha$ takes a moderate value, i.e., $\alpha = 30\%$, for all the device-clusters, the considered system example achieves larger EE compared to the scenarios when $\alpha = 10\%$ and $\alpha = 60\%$. Note that $\alpha$ is a system defined control parameter. To determine a suitable value of $\alpha$, one can first conduct an offline simulation where the EE and SE are evaluated for different values of $\alpha$. Subsequently, by inspecting these results, the suitable value of $\alpha$ 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 plot the achievable SE and EE of the proposed schemes for different device-clusters while considering $\alpha = 30\%$ and $\alpha = 60\%$, respectively. SE and EE of the EE-optimal and SE-optimal schemes are also depicted in Figs. 3a and 3b, respectively. Performance of the EE-optimal and SE-optimal schemes is obtained from Algorithm 3 and Algorithm 4, respectively. The SE-optimal and EE-optimal schemes provide the two corner points of the Pareto-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-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 6.09 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 28.25% improved SE compared to the EE-optimal scheme. Finally, Fig. 3b depicts that for all the device-clusters, I-RSRM with $\alpha = 60\%$ approaches EE-optimal scheme, 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 simultaneously. In contrast, by tuning $\alpha$, 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. Hence, 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. Note that both downlink multicasting and unicast/TIN approaches provide certain form of interference management for the considered system model, and consequently, they are the viable benchmark schemes for proposed resource allocation algorithms. Moreover, both of these approaches can be obtained as special cases of the proposed I-RSRM algorithm. Essentially, I-RSRM outperforms both downlink multicasting and unicast/TIN approaches. Fig. 3a depicts that for $K = 5$ device-clusters, I-RSRM achieves 22.36% and 23.65% improved SE compared to downlink multicasting and unicast/TIN approaches, respectively. Similarly, Fig. 3b depicts that for $K = 5$ device-clusters, I-RSRM achieves 88.76% and 121.93% improved system EE compared to downlink multicasting and unicast/TIN approaches, respectively. In a downlink multicasting approach, to avoid intra device-cluster interference, only the minimum throughput of the cache-hit and cache-miss UDs in each device-cluster is targeted. The achievable throughput of such an approach is substantially reduced as the channel condition of the UD’s with minimum throughput is deteriorated. Note that both downlink multicasting and I-RSRM use the same algorithm for 3D-resource matching. Essentially, compared to downlink multicasting, I-RSRM achieves performance gain by exploiting an optimized RS-CMD technique. On the other hand, compared to unicast/TIN, I-RSRM achieves performance gain by exploiting the simultaneous transmission from both F-APs and CBSs and optimized power allocations to effectively suppress interference in the network. Overall, although the proposed I-RSRM algorithm allows more interference in the system than both downlink multicasting and unicast/TIN approaches, due to an optimized resource allocation, I-RSRM achieves considerable EE and SE gain over both downlink multicasting and unicast/TIN approaches. Fig. 3a also depicts that the proposed low-complexity D-RSRM algorithm achieves 95% to 97% of SE of I-RSRM algorithm. However, Fig. 3b depicts that 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.

Figs. 4a and 4b illustrate variation of the SE and EE of the proposed I-RSRM algorithm for different number of F-APs, respectively. As the number of F-APs in the network is increased, the diversity gain obtained from the opportunistic assignment between device-clusters and F-APs is enhanced. Such a fact leads to the improvement of both SE and EE of the system. As expected, for all the F-APs, 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-APs, compared to SE-optimal scheme, I-RSRM with $\alpha = 10\%, 30\%$, and $60\%$ achieves 61.10%, 69.92%, and 53.90% of optimal SE, respectively, and improves the achievable EE by 8.24, 6.49, and 9.01 times, respectively. Similarly, For $L = 4$ F-APs, compared to EE-optimal scheme, I-RSRM with $\alpha = 60\%, 30\%$, and $10\%$ achieves 99.92%, 72%, and 91.31% of optimal EE, respectively, and improves the achievable SE by 1.004, 1.3, 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.
Fig. 4. EE and SE of the proposed I-RSRM algorithm for different number of F-APs considering $K = 3$ device-clusters, $N = 36$ RRBs, and $N_R = 3$.

Fig. 5. EE and SE of the proposed D-RSRM algorithm for different number of F-APs considering $K = 3$ device-clusters, $N = 36$ RRBs, and $N_R = 3$.

Figs. 5a and 5b plot SE and EE of the proposed D-RSRM algorithm for different number of F-APs, respectively. As expected, both SE and EE of the D-RSRM algorithm are improved as the number of F-APs is increased. However, compared to I-RSRM, D-RSRM experiences higher SE and EE loss compared to SE-optimal and EE-optimal schemes, respectively. For example, for $L = 4$ F-APs, I-RSRM can achieve up to 69.92% of optimal SE and 99.92% of optimal EE of the system. However, for the same system setting, D-RSRM can achieve up to 67.55% of optimal SE and 84.49% of optimal EE of the system. In addition, compared to I-RSRM, D-RSRM has inferior EE-SE trade-off performance. For example, for $L = 4$ F-APs, I-RSRM can achieve up to 99.92% of EE of EE-optimal scheme while improving SE by 1.004 times. However, for the same system setting, D-RSRM can achieve up to 84.49% of EE of EE-optimal scheme while sacrificing 11.68% loss of SE. From the aforementioned discussion, it is evident that the gap between maximum achievable SE of I-RSRM and D-RSRM algorithms is small. Moreover, D-RSRM has much lower computational complexity than I-RSRM. Hence, D-RSRM is a suitable alternative to I-RSRM for large size system given the system prioritizes SE. However, when the system prioritizes the improvement of the trade-off between EE and SE, I-RSRM is preferred.

Figs. 6a and 6b compare SE and EE of I-RSRM, D-RSRM, and I-RSRM/IRA algorithms for different RRBs. As expected, as the number of RRBs is increased, both SE and EE of all the considered schemes in Figs. 6a and 6b are increased. Note that unlike I-RSRM and D-RSRM, I-RSRM/IRA uses an alternating optimization technique given in [39, Algorithm 3] to obtain the assignments of RRBs and F-APs among the device-clusters. However, despite convergence, such an
alternating optimization technique does not guarantee near-optimal solution to P2.6. Consequently, I-RSRM/IRA obtains an inferior matching among the device-clusters, F-APs, and RRBs, and such an inferior matching leads to the noticeable SE/EE loss compared to the proposed algorithms, especially for large RRBs. From Fig. 6a, at \( N = 30 \) RRBs, the proposed I-RSRM and D-RSRM algorithms with \( \alpha = 30\% \) improve the achievable SE of I-RSRM/IRA by 22.69% and 16.17%, respectively. From Fig. 6b, at \( N = 30 \) RRBs, the proposed I-RSRM and D-RSRM algorithms with \( \alpha = 60\% \) improve the achievable EE of I-RSRM/IRA by 63.78% and 38.98%, respectively. Such an observation clearly demonstrates that the 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 benchmark 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 make 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 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 137.59% 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 14.57% 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 benchmark schemes. For instance, from Fig. 7a, when the system is willing to trade-off 30% of optimal SE, I-RSRM (resp. D-RSRM) achieves 131.45%, 131.67% and 134.61% (resp. 72.26%, 72.43%, and 74.61%) improved EE compared to the downlink multicasting, unicast/TIN, and I-RSRM/IRA schemes, respectively. Similarly, from Fig. 7b, when the system is willing to trade-off 30% of optimal EE, I-RSRM (resp. D-RSRM) achieves 37.32%, 36.36% and 38.76% (resp. 24.03%, 23.37%, and 25.33%) improved SE compared to the downlink multicasting, unicast/TIN, and I-RSRM/IRA schemes, respectively. The aforementioned observations suggest that compared to the benchmark 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 proposed cluster-based resource allocations to simultaneously maximize system EE and SE of the content distribution phase of RS-CMD integrated and EC enabled F-RAN. By adopting the \( c \)-constraint method, an EE-SE trade-off optimization was conducted while jointly performing the transmission power allocation of CBS and F-APs, assignment of the F-APs 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, near-optimal and sub-optimal algorithms were proposed to obtain the 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 near-optimal algorithm substantially improves the trade-off between EE and SE compared to the conventional EE-optimal and SE-optimal schemes, and (ii) the proposed near-optimal and sub-optimal algorithms yield significant EE and SE gain compared to the benchmark schemes.

REFERENCES

[1] Z. Zhao, C. Feng, H. H. Yang, and X. Luo, “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] M. S. Al-Abiad, M. J. Hossain, and S. Sorour, “Cross-layer cloud offloading with quality of service guarantees in fog-RANs,” IEEE Trans. Commun., vol. 67, no. 12, pp. 8435–8449, Dec. 2019.

[3] M. Kaneko, I. Randrianantenaina, H. Dahrouj, H. Elsawy, and M.-S. Alouini, “On the opportunities and challenges of NOMA-based fog radio access networks: An overview,” IEEE Access, vol. 8, pp. 205467–205476, 2020.

[4] S. Chen, Y.-C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, “Vision, requirements, and technology trend of 6G: How to tackle the challenges of system coverage, capacity, user data-rate and movement speed,” IEEE Wireless Commun., vol. 27, no. 2, pp. 218–228, Apr. 2020.

[5] 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.
Y. Shen, C. Jiang, T. Q. S. Quek, and Y. Ren, “Device-to-device-assisted 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.

M. Sheng, C. Xu, J. Liu, J. Song, X. Ma, and J. Li, “Enhancement for content delivery with proximity communications in caching enabled wireless networks: Architecture and challenges,” IEEE Commun. Mag., vol. 54, no. 4, pp. 70–76, Aug. 2016.

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.

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. 4549–4563, Jul. 2017.

Y. Shen, C. Jiang, T. Q. S. Quek, and Y. Ren, “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.

Y. Xu, X. Li, and J. Zhang, “Device-to-device content delivery in cellular networks: Multicast or unicast,” IEEE Trans. Veh. Technol., vol. 6, no. 5, pp. 4401–4414, May 2018.

R. Amer, H. Elsawy, M. M. Butt, E. A. Jorswieck, M. Bennis, and N. Marchetti, “Optimizing joint probabilistic caching and communication for clustered D2D networks,” 2018, arXiv:1810.05510. [Online]. Available: http://arxiv.org/abs/1810.05510

K. Kaneva, N. Aboutorab, S. Sorour, and M. C. Reed, “Cross-layer offloading in fog-RANs using device cooperation and network coding,” in Proc. IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

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.

A. Douik, H. Dahrourj, O. Amin, B. Alouqui, T. Y. Al-Naffouri, and M.-S. Alouini, “Mode selection and power allocation in multi-level cache-enabled networks,” IEEE Commun. Lett., vol. 24, no. 8, pp. 1789–1793, Aug. 2020.

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, Jun. 2018.

I. Randrianantenaiana, M. Kaneko, H. Dahrourj, N. Elsawy, and M.-S. Alouini, “Interference management in NOMA-based fog-radio access networks via scheduling and power allocation,” IEEE Trans. Veh. Technol., vol. 68, no. 8, pp. 5056–5071, Aug. 2020.

B. Liu, C. Liu, M. Peng, Y. Liu, and S. Yan, “Resource allocation for non-orthogonal multiple access-enabled fog radio access networks,” IEEE Trans. Wireless Commun., vol. 19, no. 6, pp. 3867–3878, Jun. 2020.

X. Wen, H. Zhang, H. Zhang, and F. Fang, “Interference pricing resource allocation and user-subchannel matching for NOMA hierarchy fog networks,” IEEE J. Sel. Topics Signal Process., vol. 13, no. 3, pp. 467–479, Jun. 2019.

Z. Yan, M. Peng, and M. Daneshmand, “Cost-aware resource allocation for optimization of energy efficiency in fog radio access networks,” IEEE J. Sel. Areas Commun., vol. 36, no. 11, pp. 2581–2590, Nov. 2018.

T. H. L. Dinh, M. Kaneko, E. H. Fukuda, and L. Boukhatem, “Energy efficient resource allocation optimization in fog radio access networks with outdated channel knowledge,” IEEE Trans. Green Commun. Netw., early access, Oct. 29, 2020, doi: 10.1109/TGCN.2020.3034638.

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. Netw., vol. 2018, no. 1, pp. 1–54, May 2018.

B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, “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.

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.

D. Yu, J. Kim, and S.-H. Park, “An efficient rate-splitting multiple access scheme for the downlink of C-RAN systems,” IEEE Wireless Commun. Lett., vol. 8, no. 6, pp. 1555–1558, Dec. 2019.

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,” 2019, arXiv:1910.07843. [Online]. Available: http://arxiv.org/abs/1910.07843

W. Jaafar, S. Naser, S. Muhamad, P. C. Softosiates, and H. Yanikomeroglu, “Multiple access in aerial networks: From orthogonal and non-orthogonal to rate-splitting,” 2020, arXiv:2005.13122. [Online]. Available: http://arxiv.org/abs/2005.13122

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.

A. A. Ahmad, B. Matthiasen, A. Sezgin, and E. Jorswieck, “Energy efficiency in C-RAN using rate splitting and common message decoding,” in Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops), Dublin, Ireland, Jun. 2020, pp. 1–6.

G. Zhou, Y. Mao, and B. Clerckx, “Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff,” 2020, arXiv:2001.03206. [Online]. Available: http://arxiv.org/abs/2001.03206

Z. Yang, M. Chen, W. Saad, and M. Shikh-Bahaei, “Optimization of rate allocation and power control for rate splitting multiple access (RSMA),” 2019, arXiv:1903.08068. [Online]. Available: http://arxiv.org/abs/1903.08068

K.-M. Miettinen, Nonlinear Multi-Objective Optimization, Boston, MA: Academic, 1999

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

M. Z. Hassan et al., “Energy-spectrum efficient content distribution in fog-RAN using rate-splitting, common message decoding, and 3D-resource matching,” TechRxiv, 2021. [Online]. Available: https://doi.org/10.36227/techrxiv.12933761.v2

M. Z. Hassan, M. J. Hossain, J. Cheng, and V. C. M. Leung, “Device-clustering and rate-splitting enabled device-to-device cooperation framework in fog radio access network,” 2020, arXiv:2007.03827. [Online]. Available: http://arxiv.org/abs/2007.03827

Md. Zoheb Hassan (Student Member, IEEE) received the Ph.D. degree from The University of British Columbia, Vancouver, BC, Canada, in 2019. He is currently a Post-Doctoral Research Fellow with the School of Engineering, The University of British Columbia, Okanagan Campus, Kelowna, BC, Canada. His research interests include wireless optical communications, optimization, and resource allocation in wireless communication networks, and digital communications over fading channels. He serves/served as a member for the Technical Program Committee of IEEE IWWCMC 2018, IEEE ICC 2019, and IEEE ICC 2020. He was a recipient of the Four-Year Doctoral Fellowship from The University of British Columbia in 2014.
Md. Jahangir Hossain (Senior Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, the M.A.Sc. degree from the University of Victoria, Victoria, BC, Canada, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, BC, Canada. He served as a Lecturer with BUET. He was a Research Fellow with McGill University, Montreal, QC, Canada, and a Research Intern position with Communication Technology Laboratory, Intel, Inc., Hillsboro, OR, USA. He is currently working as an Associate Professor with the School of Engineering, UBC Okanagan campus, Kelowna, BC, Canada. His research interests include designing spectrally and power-efficient modulation schemes, quality of service issues, and resource allocation in wireless networks and optical wireless communications. He regularly serves as a member for the Technical Program Committee of the IEEE International Conference on Communications (ICC) and IEEE Global Telecommunications Conference (Globecom). He is serving as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Julian Cheng (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 1995, the M.Sc.(Eng.) degree in mathematics and engineering from Queens University, Kingston, ON, Canada, in 1997, and the Ph.D. degree in electrical engineering from the University of Alberta, Edmonton, AB, Canada, in 2003. He was with Bell Northern Research and NORTEL Networks. He is currently a Full Professor with the School of Engineering, Faculty of Applied Science, The University of British Columbia, Kelowna, BC, Canada. His current research interests include machine learning and deep learning for wireless communications, optical wireless technology, and quantum communications. He is a Registered Professional Engineer with the Province of British Columbia, Canada. He currently serves as the President for the Canadian Society of Information Theory and the Secretary for the Radio Communications Technical Committee of IEEE Communications Society. He was the Co-Chair of the 12th Canadian Workshop on Information Theory in 2011, the 28th Biennial Symposium on Communications in 2016, and the sixth EAI International Conference on Game Theory for Networks (GameNets 2016). He currently serves as an Area Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS. He was a past Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE COMMUNICATIONS LETTERS, and IEEE ACCESS. He served as a Guest Editor for a Special Issue of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS on Optical Wireless Communications.

Victor C. M. Leung (Life Fellow, IEEE) is currently a Distinguished Professor of Computer Science and Software Engineering with Shenzhen University, China. He is also an Emeritus Professor of Electrical and Computer Engineering and the Director of the Laboratory for Wireless Networks and Mobile Systems with the University of British Columbia (UBC), Canada. His research interests include wireless networks and mobile systems, and he has published widely in these areas. He is a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. He received the 1977 APEBC Gold Medal, the NSERC Postgraduate Scholarships from 1977 to 1981, the IEEE Vancouver Section Centennial Award, the 2011 UBC Killam Research Prize, the 2017 Canadian Award for Telecommunications Research, the 2018 IEEE TGC1 C Distinguished Technical Achievement Recognition Award, and the 2018 ACM MSWiM Reginald Fessenden Award. He coauthored article that won the 2017 IEEE ComSoc Fred W. Ellersick Prize, the 2017 IEEE Systems Journal Best Paper Award, the 2018 IEEE CSIM Best Journal Paper Award, and the 2019 IEEE TGC1 C Best Journal Paper Award. He is serving on the Editorial Board for the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE ACCESS, IEEE NETWORK, and several other journals. He is named in the current Clarivate Analytics list of “Highly Cited Researchers”.