Resource Allocation in Cache-Enabled CRAN with Non-Orthogonal Multiple Access

Jingjing Zhao*, Yuanwei Liu†, Toktam Mahmoodi*, Kok Keong Chai†, Yue Chen†, and Zhu Han‡
* King's College London, London, UK
† Queen Mary University of London, London, UK
‡ University of Houston, Houton, Tx, USA

***Abstract*—This paper studies the application of non-orthogonal multiple access (NOMA) to cache-enabled cloud radio access network (CRAN) with mixed multicast and unicast transmission. Users requesting the same content are grouped together and served with a cluster of remote radio heads (RRHs) using distributed beamforming. In addition, the user with better channel condition in each group is allowed to request an extra unicast content via the NOMA protocol. Each RRH has a local cache which enables it to acquire the requested contents either from the local cache or from the central processor via the fronthaul link. Taking the maximum fronthaul capacity into consideration, we investigate the subchannel (SC) allocation problem to both RRHs and multicast groups to improve the weighted network sum rate. The optimal solution requires exhaustive search, which become prohibitively complicated as the number of RRHs and groups increases. To tackle this problem effectively, we formulate this problem as a three-sided matching problem among SCs, RRHs and multicast groups, and propose a novel low-complexity matching algorithm. We prove mathematically that the proposed algorithm converges to a stable matching within limited number of iterations. Numerical results unveil that the proposed algorithm closely approaches the optimal solution and outperforms the conventional orthogonal multiple access (OMA)-based CRAN.**

I. Introduction

Non-orthogonal multiple access (NOMA), as a promising candidate in the fifth-generation (5G) networks for tackling the massive connectivity and high data speed challenges [1], has recently received a great deal of attentions. It has been shown that NOMA offers considerable performance gains compared to conventional orthogonal multiple access (OMA) [2]–[4]. In particular, a general downlink NOMA scenario was investigated in [2], where one BS was capable of communicating with several spatially random deployed users. In [3], the authors exploited the resource allocation problems of multiple carrier NOMA systems, where the BS was capable of working in full-duplex mode to simultaneously communicate uplink and downlink NOMA users. On the standpoint of tackling the physical layer security (PLS) issue, protection zones and artificial noise aided techniques were invoked for enhancing PLS of NOMA in stochastic geometry based large-scale networks [4].

Recently, cloud radio access network (CRAN) is recognized to curtail the capital and operating expenditures, which provides a cost-effective way to achieve network densification. The low-power and low-complexity remote radio heads (RRHs) replace conventional BSs to compress and forward the received signals from the central processor (CP) to mobile users via the fronthaul links. Meanwhile, current wireless services is experiencing a transfer from traditional *connection-centric* communications to the emerging *content-centric* communications, such as video streaming and mobile applications download [5]. A main feature of content-centric communication is that the same contents are requested by multiple users at the same time. To address this paradigm shift, *multicasting* and *caching* are developed as two main techniques to exploit the content diversity. In previous work [6], [7], the content-centric transmission design based on CRAN architectures for efficient content delivery has been investigated. In [6], the joint design of content-centric BS clustering and multicast beamforming in the cache-enabled CRAN was investigated, with the aim of minimizing the total network cost. In [7], the energy-efficient transmission design in an orthogonal frequency-division multiple access (OFDMA)-based CRAN with cache-enabled RRHs was studied. Yet, we note that [6], [7] focused only on the OMA schemes in CRAN, where orthogonality was achieved in the spatial or frequency domain. However, the spectrum efficiency can be further improved by incorporating NOMA into CRAN. To the best of our knowledge, CRAN-NOMA systems have not been investigated in the literature yet.

This paper is to consider the application of NOMA to a cache-enabled CRAN architecture with mixed multicasting and unicasting traffic, where each cluster of RRHs transmits two types of data streams, one for multicasting and one for unicasting, to their serving users. The idea of joint transmission of multicast and unicast contents is motivated by the observation that in a multicast group, some users are with better channel conditions while the others with worse ones. As such, unicasting transmission can be superimposed with multicasting following the NOMA principle, where the difference of channel conditions can be used to improve the performance of unicasting, while maintaining the reliability of multicasting. Aiming at finding an effective algorithm, we recognize that the subchannel (SC) allocation problem to both RRHs and multicast groups can be regarded as a three-sided matching process in which SCs, RRHs, and multicast groups are three sets of players to be matched together to achieve the maximum sum rate. This enables us to adopt matching theory to solve our problem in a adaptive and low-complexity way [8], [9]. The main contribution of this paper is summarized as the following: 1) We apply NOMA into cache-enabled CRAN to enable mixed unicasting and multicasting transmission; 2)

By formulating the three-sided matching problem, where SCs, RRHs and multicast groups are matched with each other to form different matching triples, we develop a low-complexity three-sided matching algorithm.

II. Network Model

Consider the downlink transmission of a cache-enabled CRAN with N single antenna RRHs denoted by $\mathcal{N} = \{1, ..., N\}$, M single antenna users denoted by $\mathcal{M} = \{1, ..., M\}$, and K SCs denoted by $\mathcal{K} = \{1, ..., K\}$. Each RRH is connected to the CP via a limited-capacity fronthaul link. The CP can access a database that contains a set of C contents, denoted by $\mathcal{C} = \{1, ..., C\}$, with equal size. Each RRH can store up to $U \leq C$ contents in its cache. Let $i_{n,c} = 1$ if content c is cached at RRH n, and $i_{n,c} = 0$ otherwise. Users requesting the same content are grouped together, denoted as $\mathcal{G} = \{1, ..., G\}$, and served using multicast transmission. We make the assumption that at most two users requesting the same content are grouped together and served using multicast transmission[1]. In each group, the user with better channel condition can request an extra unicast content from its serving RRHs via the NOMA protocol. Let $j_{g,c,mu} = 1$ if group g requests multicast content c, and $j_{g,c,mu} = 0$ otherwise. Similarly, let $j_{g,c,un} = 1$ if group g requests unicast content c, and $j_{g,c,un} = 0$ otherwise. Each multicast group g is served by a cluster of RRHs cooperatively during each frame, and can request at most one multicast content and one unicast content. Each RRH in a cluster acquires the requested contents either from its local cache or from the database in the CP through the fronthaul link.

Let $\alpha_{g,k}$ denote SC assignment to multicast groups, i.e., $\alpha_{g,k} = 1$ if multicast group g is assigned to SC k. $\alpha_{g,k} = 0$ otherwise. Similarly let $\theta_{n,k}$ denote SC assignment to RRHs, i.e., $\theta_{n,k} = 1$ if SC k is assigned to RRH n; $\theta_{n,k} = 0$ otherwise. We assume that each multicast group can occupy no more than one SC and one SC can be allocated to no more than one multicast group, i.e., $\sum_{k=1}^{K} \alpha_{g,k} \leq 1, \forall g \in \mathcal{G}$ and $\sum_{g=1}^{G} \alpha_{g,k} \leq 1, \forall k \in \mathcal{K}$. The detailed system description can be found in Fig. 1. In this figure, good user (GU) denotes the user with better channel condition, while bad user (BU) denotes the user with wore channel condition.

The NOMA protocol requires the superposed coding technique at the RRH side and successive interference cancellation (SIC) techniques at the users. Assuming the set of RRHs occupying SC k is denoted by $\mathcal{S}_k = \{n \in \mathcal{N} | \gamma_{n,k} = 1\}$, the RRHs in $\mathcal{S}_k$ cooperatively send the data to group g assigned to SC k based on the distributed beamforming transmission strategy. To be more specific, both small scale Rayleigh fading and large scale path loss are considered in our channel model. Therefore, the channel gain between RRH n and the GU of group g on SC k can be expressed as $f_{n,k,g,1} = \frac{|\hat{f}_{n,k,g,1}|^2}{1+d^{\alpha}_{n,g,1}}$, where $\hat{f}_{n,k,g,1} \sim \mathcal{CN}(0,1)$ represents the Rayleigh fading between RRH n and the GU of group g on SC k, and α is the path loss exponent. It is worth noting that we use the bounded path loss model in the channel gain expression to avoid the singularity issue when $d_{n,g,1} \rightarrow 0$. Similarly, we can get the channel gain between RRH n and the BU of group g on SC k, which is represented by $f_{n,k,g,2}$. We assume equal power allocation of RRHs on SCs, i.e., $p_{n,k} = P_n/K$, where P_n is the total transmit power at each RRH n. For multicast group g, the vector $\{a_{n,g,mu}, a_{n,g,un}\}$ represents the power allocation coefficients of RRH n to the multicast and unicast content, which is also assumed to be fixed. RRHs in the set $\mathcal{S}_k$ send both the required multicast and unicast contents to the group occupying SC k. Particularly, with the application of the NOMA protocol, RRH $n \in \mathcal{S}_k$ transmits the following signal to group g occupying k:

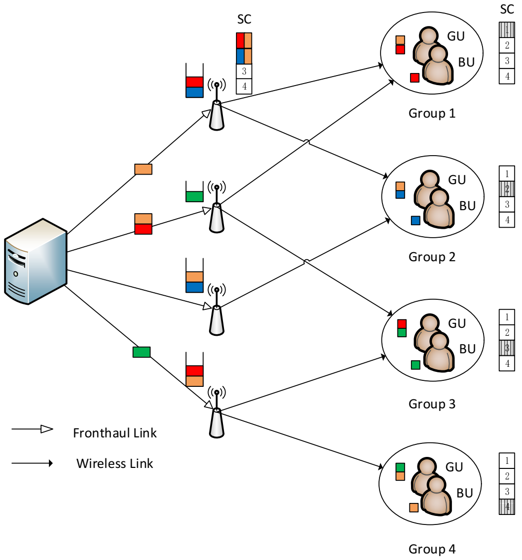

Fig. 1: Cache-Enabled CRAN with NOMA

$$x_{n,g,k} = w\left(\sqrt{p_{n,k}a_{n,g,mu}}s_{g,mu} + \sqrt{p_{n,k}a_{n,g,un}}s_{g,un}\right), \quad (1)$$

where $s_{g,mu}$, $s_{g,un}$ are the required multicast and unicast contents, respectively. w is the distributed beamforming coefficient which is designed to artificially create the difference between the users' effective channel gains. Particularly, w is designed to improve the effective channel gain of GU 1, i.e., $w = \frac{\hat{f}^*_{n,k,g,1}/\sqrt{1+d^{\alpha}_{n,g,1}}}{\sqrt{\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}}}$. As a result, the NOMA principle can be applied even if the GU and BU in a group have similar channel conditions.

By using the design of the coefficient w, the observation at the GU of group g with the transmission RRHs $\mathcal{S}_k$ is given by

$$z_{k,g,1} = \sum_{n\in\mathcal{S}_k} \frac{f_{n,k,g,1}}{\sqrt{\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}}}(\sqrt{p_{n,k}a_{n,g,mu}}s_{g,mu} + \sqrt{p_{n,k}a_{n,g,un}}s_{g,un}) + \zeta_{k,g,1}, \quad (2)$$

where $\zeta_{k,g,1}$ is the additive Gaussian noise (AWGN) with variance σ^2. Like traditional NOMA networks, the GU can first successively subtracts the multicast content, and then decode the unicast one. Therefore, the received

[1] Note that the work can be extended to the scenario where each group contains more than two users.

signal-to-interference-plus-noise ratio (SINR) at the GU for the multicast content is given by $\gamma_{k,g,1} = \frac{t_{k,g,1}}{q_{k,g,1}+\sigma^2}$, where $t_{k,g,1} = \frac{\left(\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}\sqrt{p_{n,k}a_{n,g,mu}}\right)^2}{\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}}$ and $q_{k,g,1} = \frac{\left(\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}\sqrt{p_{n,k}a_{n,g,un}}\right)^2}{\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}}$.

After the multicast content is decoded successfully, the GU first removes this message from its observation and then detects the unicast content with the following signal-to-noise ratio (SNR): $\beta_{k,g,1} = q_{k,g,1}/\sigma^2$. Actually, in reality, SIC decoding in the first step may not be successful and thus error is carried over to the next level coding, which is called SIC error propagation. Therefore, the received SNR at the GU for the unicast content is given by $\beta_{k,g,1} = \frac{q_{k,g,1}}{\epsilon t_{k,g,1}+\sigma^2}$, where $\epsilon \in [0,1]$ denotes the fraction of NOMA interference due to SIC error propogation. The observation at the BU of group g is given by

$$z_{k,g,2} = \sum_{n\in\mathcal{S}_k} \frac{h_{n,k,g,2}}{\sqrt{\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}}}(\sqrt{p_{n,k}a_{n,g,mu}}s_{g,mu} + \sqrt{p_{n,k}a_{n,g,un}}s_{g,un}) + \zeta_{k,g,2}, \tag{3}$$

where $h_{n,k,g,2} = \frac{\hat{f}^*_{n,k,g,1}\hat{f}_{n,k,g,2}}{\sqrt{\left(1+d^{\alpha}_{n,g,1}\right)\left(1+d^{\alpha}_{n,g,2}\right)}}$. The BU detects the multicast content by treating the unicast content as noise. Therefore, the received SINR at the BU for the multicast content is given by $\gamma_{k,g,2} = \frac{t_{k,g,2}}{q_{k,g,2}+\sigma^2}$, where $t_{k,g,2} = \frac{\left(\sum_{n\in\mathcal{S}_k} |h_{n,k,g,2}|\sqrt{p_{n,k}a_{n,g,mu}}\right)^2}{\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}}$ and $q_{k,g,2} = \frac{\left(\sum_{n\in\mathcal{S}_k} |h_{n,k,g,2}|\sqrt{p_{n,k}a_{n,g,un}}\right)^2}{\sum_{n\in\mathcal{S}_k} f_{n,k,g,1}}$.

Based on Shannon formula, the data rates at the GU for the multicast content and the unicast content can be expressed as $R_{k,g,1} = \log_2(1+\gamma_{k,g,1})$ and $R'_{k,g,1} = \log_2(1+\beta_{k,g,1})$, respectively. Similarly, the data rate at the BU for the multicast content is given by $R_{k,g,2} = \log_2(1+\gamma_{k,g,2})$. The weighted sum rate of multicast group g on SC k is denoted by $R_{k,g} = w_{k,g,1}R_{k,g,1}+w'_{k,g,1}R'_{k,g,1}+w_{k,g,2}R_{k,g,2}$, where $w_{k,g,1}$, $w'_{k,g,1}$ and $w_{k,g,2}$ are weighted factors indicating the priorities in resource allocation, which is specified in the media access control layer to achieve a certain notion of fairness. The interference cancellation is successful if the GU's received SINR for the multicast signal is larger or equal to the received SINR of the BU for the multicast signal [2]. Therefore, the condition of our SIC decoding order is expressed as $\gamma_{k,g,1} \geq \gamma_{k,g,2}$. For each RRH n, if content c has been cached in its local storage, it can access the content directly without costing fronthaul. Otherwise, it needs to fetch this content from the CP via the fronthaul link. Depending on the popularity profile, the same content c may be requested by several users, thus the fronthaul link rate is at least equal to the maximum data rate to transmit the content to users over all SCs. Moreover, the fronthaul data rate is restricted by the fronthaul capacity R_n^{max}. As such, we have the following constraint:

$$\sum_{c=1}^{C}(1-i_{n,c})\max_{g\in\mathcal{G}}\left\{\sum_{k=1}^{K}\gamma_{n,k}\alpha_{g,k}\bar{R}_g\right\} \leq R_n^{max}, \tag{4}$$

where $\bar{R}_g = \max\left\{j_{g,c,mu}R_{k,g,1}, j_{g,c,mu}R_{k,g,2}, j_{g,c,un}R'_{k,g,1}\right\}$.

III. Problem Formulation and Proposed Optimization Method

A. Problem Formulation

We aim to maximize the total wireless data rate in the cache-enabled C-RAN network subject to the maximum fronthaul capacity, by optimizing the user-SC assignment vector $\alpha = \{\alpha_{g,k}, \forall g, k\}$, and the RRH assignment vector $\theta = \{\theta_{n,k}, \forall n, k\}$. The formulated problem can be expressed as the following:

$$\max_{\alpha,\theta}\sum_{k=1}^{K}\sum_{g=1}^{G}R_{k,g}, \tag{5a}$$

$$s.t. \;\; \gamma_{k,g,1} \geq \gamma_{k,g,2}, \;\; \forall k, g, \tag{5b}$$

$$\sum_{c=1}^{C}(1-i_{n,c})\max_{g\in\mathcal{G}}\left\{\sum_{k=1}^{K}\gamma_{n,k}\alpha_{g,k}\bar{R}_g\right\} \leq R_n^{max}, \tag{5c}$$

$$\alpha_{g,k} \in \{0,1\}, \;\; \forall g, k, \tag{5d}$$

$$\theta_{n,k} \in \{0,1\}, \;\; \forall n, k, \tag{5e}$$

$$\sum_{k=1}^{K}\alpha_{g,k} \leq 1, \;\; \forall g, \quad \sum_{g=1}^{G}\alpha_{g,k} \leq 1, \;\; \forall k, \tag{5f}$$

$$\sum_{n=1}^{N}\theta_{n,k} \leq q_r, \;\; \forall k, \quad \sum_{k=1}^{K}\theta_{n,k} \leq q_s, \;\; \forall n. \tag{5g}$$

Constraint (5b) guarantees the optimal SIC decoding order. Constraint (5c) gives the restriction on the maximum fronthaul capacity. Constraint (5d) and (5e) show that the values of $\alpha_{g,k}$ and $\theta_{n,k}$ should be either 0 or 1, respectively. According to (5f) and (5g), each multicast group can only be paired with one SC and vice versa, each SC can be assigned to no more than q_r RRHs, and one RRH can occupy no more than q_s SCs in terms of user fairness.

Due to the $\log_2(\cdot)$ functions and the interference terms in the expressions of $R_{k,g,1}$, $R'_{k,g,1}$ and $R_{k,g,2}$, problem (5) is a nonlinear optimisation problem and it is not trivial to convert it to a convex optimisation problem. There is no systematic and computational efficient approach to solve this problem optimally. Therefore, we associate this problem as equivalent to a three-sided matching problem and develop a low-complexity algorithm to find a near-optimal solution.

B. Three-Sided Matching Game

In this subsection, we present the three-sided matching problem among multicast groups, RRHs and SCs. Different from the traditional two-sided matching problem [9], [10], the mutual relationship between the multicast groups, RRHs and the SCs is more complicated to be handled in the three-sided matching problem.

To better describe the three-sided matching relationship among RRHs, SCs, and multicast groups, we first introduce some notations and definitions of the matching game in the following:

Definition 1. *$T = (n, k, g)$ is defined as a matching triple, which implies that RRH n transmits contents to multicast group g on SC k. In other words, g, n and k are matched together.*

Since one multicast group can occupy no more than one SC, and one SC can be allocated to at most one multicast group, it is a one-to-one matching between multicast groups and SCs. To simplify the proposed three-sided matching problem to a traditional two-sided matching one [11], we set SCs and multicast groups as a SG unit, i.e., $(k, g) \in \mathcal{SG}$, then the matching between SG units and RRHs can be regarded as a two-sided matching game which is defined as the following

Definition 2. *In the many-to-many matching model between SG units and RRHs, a* matching Ω *is a function from the set $\mathcal{SG} \cup \mathcal{N}$ into the set of all subsets of $\mathcal{SG} \cup \mathcal{N}$ such that 1) $|\Omega(k,g)| \leq q_r, \forall k \in \mathcal{K}, g \in \mathcal{G}$; 2) $|\Omega(n)| \leq q_s, \forall n \in \mathcal{N}$; and 3) $(k,g) \in \Omega(n)$ if and only if $n \in \Omega(k,g)$.*

Condition 1) implies that the each SG unit is matched with at most q_r RRHs, and each RRH is matched with at most q_s SG units, where q_r and q_s are called *quota* in the matching model.

Remark 1. *The matching between SG units and RRHs is a many-to-many matching game with externalities.*

Proof. In the proposed matching model between SG units and RRHs, one SG unit can request contents from multiple RRHs, and one RRH can transmit to multiple SG units. Therefore, the matching process between SG units and RRHs can be regarded as a many-to-many matching. Furthermore, we observe that the achievable data rate of a multicast group depends not only on the individual RRH matched with, but also the inner-relationship among the set of matched RRHs due to the distributed beamforming transmission. Thus, this is a many-to-many matching game with *externalities*, also known as *peer effects*. □

The preference list of each SG unit and RRH is in the descending order with respect to the rate of each potential matching triple. For any SG unit (k, g), its preference $\succ_{(k,g)}$ over the set of RRHs can be described as follows. For any two subsets of RRHs $\mathcal{T}$, $\mathcal{T}' \subseteq \mathcal{N}, \mathcal{T} \neq \mathcal{T}'$:

$$\mathcal{T} \succ_{(k,g)} \mathcal{T}' \Longleftrightarrow R_{k,g}(\mathcal{T}) > R_{k,g}(\mathcal{T}') \tag{6}$$

indicates that the SG unit (k, g) prefers $\mathcal{T}$ to $\mathcal{T}'$ only when (k, g) can achieve a higher rate over $\mathcal{T}$ than over $\mathcal{T}'$. Similarly, for any RRH $n \in \mathcal{N}$, its preference $\succ_n$ over the SG unit can be described as follows. For any two SG units (k, g) and (k, g'), and any two matchings Ω, Ω', with $\Omega(n) = (k, g)$, $\Omega'(n) = (k, g')$:

$$((k,g), \Omega) \succ_n ((k,g'), \Omega') \Longleftrightarrow R_{k,g}(n, \Omega) > R_{k,g'}(n, \Omega') \tag{7}$$

implies that RRH n prefers SG unit (k, g) to (k, g') only when n can get a higher rate from (k, g) than to (k, g').

Due to the existence of *externalities*, each players' preference list is dynamic with different matching states, which is different from the traditional matching games where players have fixed preference lists [8], [9], [12]. Therefore, different from the traditional deferred acceptance algorithm solution [9], the *swap operation* is considered in order to enable players to switch their current matches while keeping other players' matches the same, and thus to improve the system sum rate. To better describe the interdependencies between the players' preference, we first define the concept of *swap matching* as follows:

Definition 3. *A swap matching between RRH n and n' with their current matches $(k, g) \in \Omega(n)$ and $(k', g') \in \Omega(n')$ is defined as*

$$\begin{aligned}\Omega_{n(k,g)}^{n'(k',g')} =& \Omega \setminus \{(n,k,g), (n',k',g')\} \cup \\ & \{(n,k',g'), (n',k,g)\}.\end{aligned} \tag{8}$$

Note that the players involved in a swap operation are allowed to be unmatched, and therefore allowing all the players to be active. Based on the concept of swap operation, we define *swap-blocking* pair as the following:

Definition 4. *RRH (n, n') is a swap-blocking pair if and only if the following constraints are satisfied*

1) $\forall x \in \{n, n', \Omega(n), \Omega(n')\}$, $\left(\Omega_{n(k,g)}^{n'(k',g')}(x), \Omega_{n(k,g)}^{n'(k',g')}\right) \succeq_x (\Omega(x), \Omega)$,

2) $\exists x \in \{n, n', \Omega(n), \Omega(n')\}$, $\left(\Omega_{n(k,g)}^{n'(k',g')}(x), \Omega_{n(k,g)}^{n'(k',g')}\right) \succ_x (\Omega(x), \Omega)$.

Condition 1) implies that the data rates of all the involved players should not be reduced after the swap operation between a swap-blocking pair. Condition 2) indicates that at least one of the players' data rate is increased after the swap operation. This avoids looping between equivalent matchings where the utilities of all involved agents are indifferent.

C. Proposed RRH-SC-Group Three-Sided Matching Algorithm

The specific details of the proposed three-sided matching algorithm are described in Algorithm 1 and Algorithm 2, where Algorithm 1 does the initialization and Algorithm 2 enables the swap operations among RRHs to further improve the SC allocation results. In three-sided matching algorithm for initialization (TSMA-I), each RRH n proposes to its most preferred SG unit, then each SG unit accepts the most preferred q_r RRHs and rejects the others. This process continues until all the RRHs or all the SG units are fully matched, or the preference list of all RRHs are empty. Based on the initial matching state obtained by TSMA-I, three-sided matching algorithm with swap operations (TSMA-S) focuses on the swap operations between the RRHs. Each RRH keeps searching for all the other RRHs to check whether there exists a swap-blocking pair. The swap-matching process continues until there exists no swap-blocking pair. Note that we set the flag $\mathcal{SR}_{n,n'}$ to record the times that RRH n and n' swap their paired SG units. Each RRH n can at most swap with another RRH n' twice, which prevents flip flop and ensures convergence.

In the following, we analyze the properties of TSMA-S in terms of stability, convergence, and complexity.

Algorithm 1 Three-Sided Matching Algorithm - Initialization (TSMA-I)

```
Form $K \times G$ SG units from $\mathcal{K} \times \mathcal{G}$;
Calculate $R_{k,g}(n), \forall k \in \mathcal{K}, g \in \mathcal{G}, n \in \mathcal{N}$, and $R_{k,g}(\mathcal{T}), \forall k \in \mathcal{K}, g \in \mathcal{G}, \mathcal{T} \subseteq \mathcal{N}$;
Construct the preference lists of RRHs on SG units, and that of SG units on subsets of RRHs, represented by $\mathcal{RLIST}_n, n \in \mathcal{N}$ and $\mathcal{SLIST}_{(k,g)}, k \in \mathcal{K}, g \in \mathcal{G}$, respectively;
Construct the lists of RRHs and SG units that are not fully matched, i.e., $\mathcal{RNFMATCH}$ and $\mathcal{SNFMATCH}$;
repeat
    for $\forall n \in \mathcal{RNFMATCH}$ do
        RRH $n$ proposes to its most preferred SG unit that has never rejected it before;
    end for
    for $\forall (k,g) \in \mathcal{SG}$ do
        SG unit $(k,g)$ keeps the most preferred $q_r$ RRHs, and rejects the others;
        Remove $(k,g)$ from the preference lists of RRHs that have sent proposals;
    end for
until $\mathcal{RNFMATCH} = \emptyset$ or $\mathcal{SNFMATCH} = \emptyset$ or $\mathcal{RLIST} = \emptyset$
```

Algorithm 2 Three-Sided Matching Algorithm - Swap (TSMA-S)

```
– Step 1: Initialization
Obtain the initial matching state: $\Omega_0$ by TSMA-I;
Initialize the number of swapping requests that RRH $n$ sends to $n'$, i.e., $\mathcal{SR}_{n,n'} = 0$;
– Step 2: Swap-matching process:
For each RRH $n$, it searches for another RRH $n'$ to check whether it is a swap-blocking pair;
if $(n, n')$ forms a swap-blocking pair along with $(k,g) = \Omega(n)$, and $(k',g') = \Omega(n')$, as well as $\mathcal{SR}_{n,n'} < 2$ then
    Update the current matching state to $\Omega_{n,(k,g)}^{n'(k'g')}$;
    $\mathcal{SR}_{n,n'} = \mathcal{SR}_{n,n'} + 1$;
else
    Keep the current matching state;
end if
Repeat Step 2 until there is no swap-blocking pair.
```

1) Stability: It has been proved in [13] that the problem of deciding whether a three-sided matching algorithm can reach stability is NP-complete. Thus, there may exist scenarios that a three-sided matching problem has no stable result. However, in our work, we transfer the original three-sided matching problem among RRHs, SCs and multicast groups to the two-sided matching problem between RRHs and SG units, and such the stability can be analyzed in a traditional two-sided exchange-stable way, which is defined as the following:

Definition 5. *A matching Ω between RRHs and SG units is two-sided exchange-stable if there does not exist a swap-blocking pair.*

The *two-sided exchange stability* is a distinct notion of stability which well handle the "externalities" among RRHs. Based on Definition 5, we have the following observation:

Lemma 1. *The final matching Ω^* of TSMA-S is a two-sided exchange-stable one between RRHs and SG units.*

Proof. Assume that there exists a swap-blocking pair (n, n') in the final matching Ω^* satisfying conditions 1) and 2) in Definition 4. According to TSMA-S, the algorithm does not terminate until all the swap-blocking pairs are eliminated. In other words, Ω^* is not the final matching, which causes conflict. Therefore, there does not exist a swap-blocking pair in the final matching, and thus we can conclude that TSMA-S reaches a two-sided exchange stability between RRHs and SG units in the end of the algorithm. □

2) Convergence: Regarding the convergence of TSMA-S, we have the following conclusion:

Theorem 1. *TSMA-S converges to a two-sided exchange-stable matching between RRHs and SG units within limited number of iterations.*

Proof. The convergence of TSMA-S depends mainly on the swap-matching process. According to Definition 4, after each swap operation between RRH n and n' along with their matches $\Omega(n)$ and $\Omega(n')$, none of the participating players' rate can be decreased, and at least one of the players' rate is increased. Since RRHs and spectrum resource are limited, the users' sum rate has an upper bound. Therefore, there exists a swap operation after which no swap-blocking pair can further improve users' sum rate, and then the TSMA-S converges to the final matching Ω^*. □

3) Complexity: The complexity of TSMA-S mainly focuses on the swap-matching process, for which the complexity is upper bounded by $\mathcal{O}(N^2)$. Since we restrict that each RRH n can at most swap with another RRH n' twice, the number of swap operations is upper bounded by $2 \times \binom{N}{2}$. Therefore, the complexity of TSMA-S is upper bounded by $\mathcal{O}(N^2)$.

IV. Numerical Results

In this section, we investigate the performance of the proposed SC allocation algorithm through simulations. We consider a CRAN scenario where the number of RRHs is 5, and the number of SCs is the same as that of the multicast groups. Following the topology model in [7], we assume that one RRH is located in the center of a square region with the side length of 100 m, while the others are located on the vertices. The users are uniformly and independently distributed within a larger square region with the side length of 200 m, excluding an inner circle of 5 m around each RRH. The center of the user square region coincides with that of the RRH square region. The maximum transmit power of each RRH is assumed to be 24 dBm. The noise power spectral density is -174 dBm/Hz. There are $C = 50$ distinct contents and the users' requests follow a Zipf distribution [6], which implies that the probability that a user requests content $c \in \mathcal{C}$ is given by $\phi_c = c^{-\nu} / \sum_{u=1}^{C} u^{-\nu}$. ν is the skewness parameter and set by $\nu = 0.9$. Each RRH caches $U = 10$ contents if not specified otherwise. The caching strategy follows the *popularity-aware* rule, which implies that each RRH caches the most popular contents until its storage is full. In this case, all the RRHs cache the same set of contents if they have the same storage size. We compare the proposed solution with three benchmarks: 1) **Benchmark 1: Optimal solution.** We adopt the exhaustive searching method to find the optimal solution for the formulated RRH-SC-Group pairing problem; 2) **Benchmark 2: Single RRH selection.** Only one RRH is selected on each SC to transmit to the multicast groups, which loses the distributed beamforming transmission gain. This is achieved by restricting $q_r = 1$ in problem (5). The new formulated problem is then solved using TSMA-S by changing the matching between SG units and RRHs to a many-to-one matching one; 3) **Benchmark 3: Conventional OMA-based CRAN.** In this scheme, we consider the conventional C-RAN network where each RRH transmits to each multicast group with only the multicast content.

Fig. 2 demonstrates the weighted sum rate versus different numbers of multicast groups in the network. As can be seen from Fig. 2 that the weighted sum rate increases monotonically with the number of multicast groups due to the exploitation of multi-user diversity gain. It is worth noting that with the increase of the number of multicast groups in the networks, the increasing rate of sum rate becomes smaller due to the limited number of RRHs and the restriction of maximum number of multicast groups can be paired with each RRH, .i.e., q_s. It is also observed that TSMA-S achieves much higher sum rate compared to Benchmark 2 and Benchmark 3. Meanwhile, the performance of TSMA-S is close to the optimal solution obtained from the exhaustive search. Recall the complexity of TSMA-S, which is much lower than the exhaustive search, unequivocally substantiates the plausibility of TSMA-S.

Fig. 3 plots the weighted sum rate versus different fronthaul capacity, averaged over random user locations and content request profiles. It can be seen from the figure that higher fronthaul capacity contributes to a higher weighted sum rate, especially when the fronthaul capacity is small. This is because higher transmission rate between RRH and users is allowed with larger fronthaul capacity. Besides, the sum rate stops increasing when the fronthaul capacity gets large enough. This is due to the fact that, from the beginning from a fixed point, the fronthaul capacity is able to support all the potential data transmissions from RRHs to users.

V. Conclusions

In this paper, the application of NOMA to cahche-enabled CRAN was studied. With the maximum fronthaul capacity constraint, SC allocation to both RRHs and multicast groups was investigated to improve the network sum rate. By formulating the problem as a three-sided matching problem among RRHs, SCs and multicast groups, a low-complexity matching algorithm was proposed to obtain the near-optimal result. Both analytical and simulation results were demonstrated to show that the NOMA-enabled CRAN yielded a significant improvement in network sum rate compared to OMA cases.

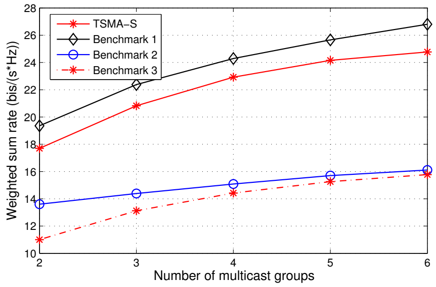

Fig. 2: Weighted sum rate with different number of multicast groups, with $R_n^{max} = 20$ Mbps, $\forall n \in \mathcal{N}$.

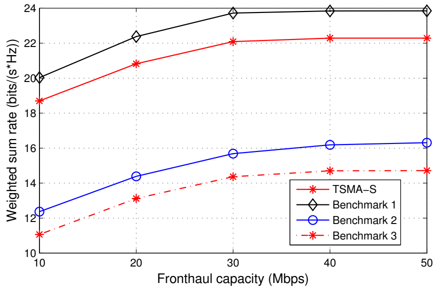

Fig. 3: Weighted sum rate with different fronthaul capacity, with $G = 3$.

References

[1] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[2] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[3] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[4] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[5] C. V. N. Index, "Global mobile data traffic forecast update, 2012-2017," *Cisco white paper*, Feb. 2015.

[6] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sept. 2016.

[7] R. G. Stephen and R. Zhang, "Green OFDMA resource allocation in cache-enabled CRAN," *arXiv preprint arXiv:1612.04065*, Dec. 2016.

[8] Y. Gu, Y. Zhang, M. Pan, and Z. Han, "Matching and cheating in device to device communications underlying cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2156–2166, Oct. 2015.

[9] A. E. Roth and M. A. O. Sotomayor, *Two-sided matching: A study in game-theoretic modeling and analysis*. Cambridge University Press, 1992, no. 18.

[10] D. F. Manlove, *Algorithmics of matching under preferences*. World Scientific, 2013, vol. 2.

[11] B. Di, S. Bayat, L. Song, Y. Li, and Z. Han, "Joint user pairing, subchannel, and power allocation in full-duplex multi-user OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8260–8272, Dec. 2016.

[12] H. Zhang, Y. Xiao, S. Bu, D. Niyato, R. Yu, and Z. Han, "Computing resource allocation in three-tier iot fog networks: a joint optimization approach combining stackelberg game and matching," *accepted by Int. of Things J.*, Jan. 2017.

[13] A. Subramanian, "A new approach to stable matching problems," *SIAM J. Comput.*, vol. 23, no. 4, pp. 671–700, Aug. 1994.