Resource Allocation in NOMA Systems for Centralized and Distributed Antennas with Mixed Traffic using Matching Theory

Marie-Joséphа Youssef, Joumana Farah, Charbel Abdel Nour, Catherine Douillard

Resource Allocation in NOMA Systems for Centralized and Distributed Antennas with Mixed Traffic using Matching Theory

Marie-Josepha Youssef, *Student Member, IEEE,* Joumana Farah, *Member, IEEE.*
Charbel Abdel Nour, *Senior Member, IEEE,* and
Catherine Douillard, *Senior Member, IEEE.*

Abstract

In this paper, we study the traffic-aware resource allocation problem for a system with mixed traffic types. The considered framework encompasses real-time (RT) users having strict QoS requirements (in terms of amount of data and latency), and best-effort (BE) users for which the system tries to strike a balance between throughput and fairness. The resource allocation problem is studied in different contexts: orthogonal and non-orthogonal multiple access (OMA and NOMA respectively) in either centralized or distributed antenna systems (CAS and DAS respectively). Following the formulation of the resource optimization problem, we propose a low complexity suboptimal solution based on matching theory for each system context. We also propose an iterative approach to determine the number of subbands per antenna for the DAS contexts. The proposed techniques aim at guaranteeing the requirements of RT users while maximizing the utility function of BE users. Simulation results show that the proposed allocation method based on matching theory greatly outperforms a previously proposed greedy approach, especially in terms of RT users satisfaction.

Index Terms

Mixed traffic types, latency, resource allocation, matching theory, NOMA, DAS.

I. Introduction

The explosive growth in connectivity and information sharing brought by the proliferation of IoT applications has been paving the way towards the 5^{th} generation of cellular networks. 5G systems are expected to fulfill a certain set of diverse requirements [1]. In addition to increasing the achieved data rates, they are expected to accommodate a massive number of connected

M. J. Youssef, C. Abdel Nour and C. Douillard are with IMT Atlantique, LabSTICC, UBL, F-29238 Brest, France, (e-mail: marie-josepha.youssef@imt-atlantique.fr; charbel.abdelnour@imt-atlantique.fr; catherine.douillard@imt-atlantique.fr). J. Farah is with the Department of Electricity and Electronics, Faculty of Engineering, Lebanese University, Roumieh, Lebanon (e-mail: joumanafarah@ul.edu.lb).

This work has been funded with support from the Lebanese University and IMT-Atlantique.

devices deployed to enable different applications. These span various sectors (e.g. autonomous vehicles, automated control, e-health, virtual reality, etc.) and should co-exist with traditional applications (e.g. file download, web browsing, etc.). However, the new envisioned applications have very different requirements, compared to traditional services, in terms of data rate, latency and reliability. As a result, 5G systems must adopt new technologies to cope with mixed or heterogeneous traffic models.

Resource allocation for mixed traffic types was previously investigated in the literature. In [2], the authors adopted utility theory for a system consisting of real-time (RT) and best-effort (BE) users, and proposed a heuristic algorithm to solve the resource allocation problem, based on Lagrange multipliers. In [3] and [4], after partitioning users among different classes based on their requirements, the priority of each user was calculated using fuzzy logic before scheduling the most urgent ones. In [5], the authors proposed a heuristic to perform QoS-based scheduling for small-cell users. They also developed an admission control algorithm to enhance the scheduling policy. Network coordination was employed to enhance the performance of RT users in [6] and minimize the amount of resources needed by RT users, increasing their availability for BE users. This minimization was also the target of [7] where a scalable transmission time interval (TTI) was adapted to the data and latency requirements of the users.

With the exception of [8] and [9], all previous studies on resource allocation for mixed traffic employed orthogonal multiple access (OMA) to enable different users to simultaneously access the spectrum. This orthogonality aims at limiting inter-user interference. Although OMA benefits from both, good system level performance and a simplified receiver design, it suffers from several drawbacks. First, the number of admitted users in OMA systems is limited by the number of available frequency subbands. Also, OMA restricts the allocation of each subband to one user only. This results in a poor overall spectral efficiency, as was also noted by [10], especially if the allocated resource exceeds the requirements of the scheduled user or if the latter suffers from poor channel conditions. Since 5G systems are expected to provide massive connectivity for users with very heterogeneous requirements, OMA is becoming a limiting factor for system design. Hence, there is a need to diverge towards new radio access technologies with better support for the changing needs of connected devices.

From an information-theoretical point of view, it is well-known that non-orthogonal user multiplexing using superposition coding at the transmitter and proper decoding techniques at the receiver not only outperforms orthogonal multiplexing, but is also optimal in the sense of

achieving the capacity region of the downlink broadcast channel [11]. As a result, non-orthogonal multiple access (NOMA) emerged as a promising multiple access technology for 5G systems [12]–[14]. NOMA allows multiple users to be scheduled on the same time-frequency resource by multiplexing them in the power domain. At the receiver side, successive interference cancellation (SIC) is performed to retrieve superimposed signals. By allowing multiple users to access the same resource, NOMA enhances spectral efficiency and increases the number of admitted users which is necessary to achieve massive connectivity, rendering NOMA a promising solution to support mixed traffic systems. In fact, in a mixed traffic system where some users have rate requirements and others aim to maximize theirs, NOMA enables the sharing of one subband between two users of the two categories. That way, if the rate requirement of a user is low, the system spectral efficiency is not penalized as in an OMA system as another user can benefit from the same subband. Moreover, NOMA enables the spectrum to be overloaded which ensures the accommodation and the satisfaction of a higher number of users when compared to OMA scheduling.

Resource allocation for NOMA systems has been extensively studied with different performance measures. To name a few, the weighted sum rate of a NOMA system was maximized in [15]; however the proposed method has exponential complexity. Maximizing system fairness was the target of [16], minimizing the used power subject to rate requirements was the target of the works in [17] and [18]. Considering a millimeter wave system with mixed traffic, [9] proposed an algorithm for user grouping, then determined the optimal power allocation to maximize the spectral efficiency of BE users while serving RT users with their rate requirements. However, [9] restricted RT users to be scheduled as second users in NOMA, i.e. users not performing SIC, which decreases the probability of satisfying their needs.

In addition to NOMA, distributed antenna systems (DAS) and their evolution to cloud radio access networks (C-RAN) were recently introduced as promising network architectures. By using multiple remote radio heads (RRHs) coordinated by a central controller, DAS enable higher capacities and increased coverage. To further enhance system performance, the combination of NOMA and DAS or C-RAN was given some attention in recent literature. In [19] and [20], the authors used NOMA in the transmission from the central controller to various RRHs. They proposed a power allocation scheme between the RRHs as well as an algorithm that finds the optimal number of BSs in order to guarantee the cell-edge user requirement in [19] or maximize the energy efficiency (EE) in [20]. [21] considered an uplink setting, where the RRHs cooperate

to remove the interference brought by NOMA. In [22], the outage probability of a downlink two-user C-RAN system was derived using stochastic geometry. In [23], several joint subcarrier, RRH and power allocation techniques were proposed for reducing the total transmit power in each cell, using proper combinations of NOMA with DAS. However, none of these works investigated the use of NOMA and DAS to accommodate mixed traffic systems.

To achieve the full-potential of NOMA in the DAS settings, RRH and sub-channel assignment, as well as power allocation must be optimized jointly. However, this results in a mixed-integer optimization problem which is NP-hard and for which the optimal solution is found by exhaustive search. That said, exhaustive search has a prohibitive complexity for practical systems. Therefore, suboptimal but more efficient resource allocation techniques are preferred in practice.

In a previous work [8], we proposed a greedy algorithm to perform resource allocation for a mixed traffic system consisting of RT and BE users. In the current work, we study the resource allocation problem with a focus on antenna and subband assignment under different system configurations. To tackle the assignment problem in an efficient manner, the antenna and subband assignment problem is cast as a matching game. The primary goal of the proposed solution is to ensure the satisfaction of RT users through guaranteeing their rate requirements. Then, when possible, the introduced technique must maximize both the data rates and fairness of BE users. The study is conducted in different system settings combining DAS or centralized antenna systems (CAS) with different signaling technologies, namely OMA and NOMA, to compare their performance in the mixed traffic context.

Matching theory-based algorithms for resource allocation have recently gained significant attention. In [24], the authors considered a hybrid C-RAN system with D2D communications and adopted matching theory to perform the subband-RRH assignment. However, to simplify the problem, they supposed that the user-RRH association is done beforehand and restricted each user to be assigned to one antenna and access one subband only. Similarly, in the context of NOMA, [25] developed a user pairing technique based on matching theory. In [26] and [27], an algorithm based on matching theory was proposed to perform subband allocation for users and D2D pairs respectively. However, to the best of our knowledge, no previous work considered the application of matching theory to solve the subband assignment problem for mixed traffic in a NOMA-DAS system. More specifically, none of the previously proposed matching theory-based algorithms for NOMA systems ensured that rate requirements are met. Furthermore, in the DAS settings, the restrictions made by previous algorithms on antenna and subband assignment

are unrealistic in practice, and are introduced only to simplify the resource allocation problem. Indeed, the work in [24] assumed that the assignment of users to the distributed antennas is done beforehand. In addition, almost all previous studies in this context restricted each user to be assigned to one antenna and to access one subband only, while others [28] considered that the spectrum consists of one subband to bypass the subband assignment step.

Contrary to previous works, the method proposed in this paper is the first to provide a joint solution for the assignment problem at hand while allowing each user to access any number of RRHs and subbands simultaneously. Our contributions can be summarized as follows:

- After formulating the performance measures for both RT and BE users, we propose an optimization problem to conduct resource allocation in the considered framework. To ensure successful SIC decoding at the receiver side, the optimization problem takes into account the NOMA power multiplexing constraints for users scheduled on each subband. These constraints state that the signal that is to be decoded first must have a higher power level than the other received signals, so that it is detectable at the receiver side. Note that the power multiplexing constraints were mostly neglected in previous works dealing with matching theory to simplify the analysis.
- We consider different system settings, namely OMA-CAS, OMA-DAS, NOMA-CAS and NOMA-DAS for performance comparison. More precisely, OMA-CAS (NOMA-CAS resp.) corresponds to a CAS setting employing OMA signaling (NOMA signaling resp.), whereas OMA-DAS (NOMA-DAS resp.) corresponds to a DAS setting employing OMA signaling (NOMA signaling resp.). For each scenario, we formulate the channel allocation problem as a one-to-many matching game. We then propose algorithms based on the deferred acceptance (DA) [29] method to solve the channel allocation in each of the considered settings. To the best of our knowledge, no previous study has considered the use of matching theory to resolve the mixed traffic resource allocation problem, combining DAS and NOMA. It should also be noted that none of the previous works applying matching theory to solve the resource allocation problem for a NOMA system incorporated rate requirements into their analysis, which is not the case for the current work. Moreover, the proposed algorithms can be easily applied to solve other problems with different objectives in the different considered system settings. This can be done by modifying the preference relations of the users and, in the case of systems with one user type only, slightly adapting the different algorithms.
- For the NOMA-CAS and NOMA-DAS system settings, a hybrid NOMA system is devised

using matching theory, where subbands are either allocated to single users or user-pairs in such a way to optimize system performance. Moreover, an algorithm that overcomes the need for a swapping phase to deal with the interdependencies between users' preferences is introduced.

- We prove that the proposed algorithm, for each of the considered system settings, converges to a stable solution within a limited number of iterations.
- For the DAS setting, no *a priori* information about the assignment of subbands or users to the different RRHs is assumed as was done in most previous works in the DAS context, such as [24], [30], [31]. Moreover, contrary to previous works, users are not restricted to be assigned to one predefined antenna and one subband only. Instead, an iterative approach is proposed to determine this assignment and its convergence is proved.
- Through simulation results, the proposed method is shown to achieve a better performance than a previous method introduced in [8], especially in terms of the percentage of RT users that meet their QoS requirements.

The rest of this paper is organized as follows. In section II, the system model is described. Sections III, IV, V present the proposed algorithms to solve the resource allocation problem in OMA-CAS, OMA-DAS, NOMA (CAS and DAS) respectively. The properties of the proposed algorithms are analyzed in Section VI. Finally, simulation results are presented in Section VII, before drawing the conclusion in Section VIII.

II. System Model and Problem Formulation

A. System Description

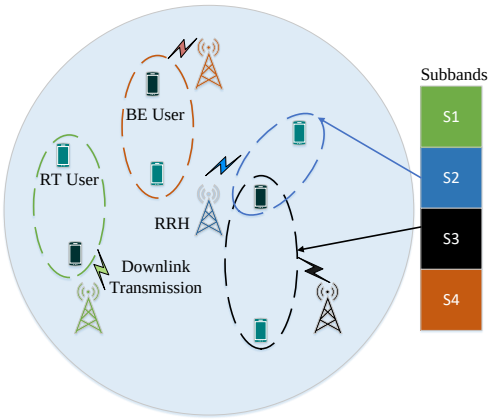


Fig. 1: System Model

Consider a downlink system as shown in Fig. 1 with K single-antenna users uniformly deployed over a cell. The total system bandwidth B is divided into S subbands, leading to a bandwidth of $B_c = B/S$ per subband. In this work, different system configurations are studied. More precisely, we consider both CAS and DAS settings: CAS consists of one antenna located at the cell center, with a power budget P_{CAS}, whereas in DAS, A single-antenna RRHs are uniformly deployed over the cell. Each antenna has a total power budget of P_{DAS}. For multiple access, OMA as well as NOMA, which enables up to N_s users to be non-orthogonally multiplexed over a subband s, are considered. Hence, the different studied system configurations are OMA-CAS, OMA-DAS, NOMA-CAS and NOMA-DAS. The sets of users, subbands and RRHs will be respectively denoted by $\mathcal{K}, \mathcal{S}$ and $\mathcal{A}$.

When DAS is considered, a subband can only be assigned to one antenna during a scheduling slot to limit intra-cell interference. When NOMA is adopted for multiple access, the messages of up to N_s users are superposed and transmitted over subband s. This results in co-channel interference between the collocated users. Therefore, user k applies SIC [32] before demodulating its own signal, resulting in the following achieved rate:

$$R^t_{k,s,a} = B_c \log_2 \left(1 + \frac{p^t_{k,s,a}(h^t_{k,s,a})^2}{\sum\limits_{k' \in \mathcal{I}^t_{k,s,a}} p^t_{k',s,a}(h^t_{k,s,a})^2 + N_0 B_c} \right). \tag{1}$$

In (1), $p^t_{k,s,a}$ and $h^t_{k,s,a}$ are respectively the transmit power and the channel gain of user k over subband s when assigned to antenna a at timeslot t. N_0 is the noise power spectral density. The first term in the denominator reflects the interference experienced by user k from users in $I^t_{k,s,a} = \{(k' \in \mathcal{S}_s \setminus \{k\}) \cap \left(h^t_{k',s,a} > h^t_{k,s,a}\right)\}$, i.e. users scheduled on subband s and having a higher channel gain than k on s, when the latter is assigned to antenna a.

SIC results in a significant complexity increase at the receiver side; therefore, in this study, the maximum value of N_s is restricted to 2, $\forall s \in \mathcal{S}$.

B. User Characteristics

In this work, we differentiate between two user classes characterized by different requirements.

1) BE users: This category includes users running delay-tolerant, rate-demanding applications such as file download or web browsing. The goal of these users is to maximize both achieved data rates and system fairness. The performance measure for BE users is therefore chosen to be:

$$M^t_{BE}(\boldsymbol{x^t}, \boldsymbol{p^t}) = \sum_{k=1}^{K_{BE}} \sum_{a=1}^{A} \sum_{s=1}^{S} x^t_{k_{BE},s,a} R^t_{k_{BE},s,a} f\left(T^t_{k_{BE}}\right), \tag{2}$$

where $x^t_{k_{BE},s,a} = 1$ if k_{BE} is scheduled on s when the latter is assigned to antenna a, and 0 otherwise. $\boldsymbol{p^t}$ is the power allocation vector and f is a measure of system fairness that depends on the average data rate $T^t_{k_{BE}}$ of each BE user until the beginning of timeslot t. $T^t_{k_{BE}}$ is updated at the beginning of each timeslot according to:

$$T^t_{k_{BE}} = \left(1 - \frac{1}{t_c}\right) T^{t-1}_{k_{BE}} + \frac{1}{t_c} R^{t-1}_{k_{BE}}. \tag{3}$$

In (3), t_c is the averaging window and $R^{t-1}_{k_{BE}}$ is the total rate of user k_{BE} at timeslot $(t-1)$.

The expression of (2) is a generic form that can enclose a wide range of specific performance metrics. A common trait of these metrics is the combination of the achieved rate and the system fairness in the scheduling decision. If the expression of (2) did not include the achieved throughput term, the scheduler would optimize system fairness, while penalizing the achieved sum rate. In contrast, if this expression did not include the fairness measure, the system throughput would be maximized by only scheduling users with a high channel gain, hence achieving a low system fairness. Therefore, to optimize performance, both the achieved throughput and the fairness measure need to be accounted for in the expression of (2). The maximum of (2) for BE users is reached when the product between their achieved rates and the fairness between them is maximized. Hence, by adopting this measure, a tradeoff between the maximization of the achieved rates and that of user fairness is reached. This tradeoff can be efficiently reached by the well known proportional fairness (PF) scheduler [12], known to achieve the best balance between rate and system fairness. In fact, the PF metric is one of the expressions embodied by (2), and will be adopted later on in the proposed solutions.

2) RT users: This category includes users running latency-constrained applications. While some applications require only a small rate (e.g. autonomous cars or sensor applications), others are more bandwidth-hungry (e.g. virtual reality). Therefore, RT users are characterized by a strict latency limit $L_{k_{RT}}$ (expressed as an integer number of timeslots) as well as a requested amount of data bits $D^{\text{req}}_{k_{RT}}$. Their satisfaction depends upon receiving $D^{\text{req}}_{k_{RT}}$ prior to their latency limit. In this work, $D^{\text{req}}_{k_{RT}}$ is equally divided among the timeslots preceding the latency limit. Hence, from the start of the scheduling period till the end of timeslot t, each user k_{RT} needs to be allocated a number of bits equal to:

$$D^{\text{req},t}_{k_{RT}} = t D^{\text{req}}_{k_{RT}} / L_{k_{RT}}. \tag{4}$$

$D_{k_{RT}}^{\text{req},t}$ is an increasing function of the current timeslot index t and $D_{k_{RT}}^{\text{req}}$ and decreasing in $L_{k_{RT}}$. Hence, its value increases when the latency limit is small, the total required number of bits is large and when the current timeslot index approaches the latency limit.

Adopting $D_{k_{RT}}^{\text{req},t}$ as the number of required bits at timeslot t leads to the following required rate in timeslot t:

$$R_{k_{RT}}^{\text{req},t} = \frac{D_{k_{RT}}^{\text{req},t} - D_{k_{RT}}^{\text{ach},t-1}}{\tau}, \tag{5}$$

where τ is the timeslot duration and $D_{k_{RT}}^{\text{ach},t-1}$ denotes the received number of bits by k_{RT} at the end of the previous timeslot $(t-1)$.

Let $\mathbb{I}_{k_{RT}}^{t}(\boldsymbol{x^t}, \boldsymbol{p^t})$ be a measure of the satisfaction of k_{RT} defined by:

$$\mathbb{I}_{k_{RT}}^{t}(\boldsymbol{x^t}, \boldsymbol{p^t}) = \begin{cases} 1 & \text{if } \sum_{a=1}^{A} \sum_{s=1}^{S} x_{k_{RT},s,a}^{t} R_{k_{RT},s,a}^{t} \geq R_{k_{RT}}^{\text{req},t}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In (6), $\mathbb{I}_{k_{RT}}^{t}(\boldsymbol{x^t}, \boldsymbol{p^t}) = 1$, and hence user k_{RT} is satisfied if the current resource and power allocation scheme allows it to achieve a sum rate that is at least equal to its requested rate $R_{k_{RT}}^{\text{req},t}$, which is calculated to allow user k_{RT} to reach its requested number of data bits before its latency limit. In the opposite case, k_{RT} is not satisfied with the current allocation which reflects in having $\mathbb{I}_{k_{RT}}^{t}(\boldsymbol{x^t}, \boldsymbol{p^t}) = 0$.

Having (6) at hand, we propose to formulate the optimization function for all RT users as:

$$M_{RT}^{t}(\boldsymbol{x^t}, \boldsymbol{p^t}) = \sum_{k=1}^{K_{RT}} \mathbb{I}_{k_{RT}}^{t}(\boldsymbol{x^t}, \boldsymbol{p^t}). \tag{7}$$

Using the above formulation, (7) measures the number of RT users having received their requested data rate at each timeslot t. Hence, at the end of the latency period of all RT users, (7) finds the number of satisfied RT users, i.e. users having received the totality of requested data bits.

For concision purposes, the timeslot index t will be dropped in the following when there is no confusion. Table I contains the main notation used throughout this paper.

C. Optimization Problem

Having defined the performance measures to be maximized for both user types, the following optimization problem must be solved at each time slot:

$$\max_{\boldsymbol{a},\boldsymbol{p}} \quad (2), (7) \tag{8}$$

TABLE I: Notation Table

K	Total number of users	K_{RT}, K_{BE}	Number of RT and BE users resp.
S	Number of subbands	$\mathcal{S}_s$	Set of users scheduled on subband s
A	Number of distributed antennas	$\boldsymbol{N} = \{N_a, a \in \mathcal{A}\}$	Vector containing the number of subbands assigned to each antenna
N_0	Noise power spectral density	B_c	Subband bandwidth
P_{CAS}, P_{DAS}	Power per antenna in the CAS and DAS settings resp.	P_a	Power per subband on antenna a
t	Timeslot index	$\mathcal{K}_{\text{active}}$	Set of active users
$p^t_{k,s,a}$	Power allocated to user k on subband s and antenna a at timeslot t	$h^t_{k,s,a}$	Channel gain of user k on subband s and antenna a at timeslot t
$R^t_{k,s,a}$	Rate achieved by user k on subband s and antenna a at timeslot t	T^t_k	Average rate achieved by user k before reaching timeslot t
M^t_{RT}, M^t_{BE}	Satisfaction measure of RT and BE users resp. at timeslot t	$\mathcal{U}^t_{RT}(.,s)$, $\mathcal{U}^t_{BE}(.,s)$	Utility of RT and BE users resp. on subband s at timeslot t
L_k, D^{req}_k	Latency limit and number of requested data bits of user k	$R^{\text{req},t}_k$, $D^{\text{req},t}_k$	Required rate and required number of bits of user k at timeslot t
Ψ	Matching outcome	$\mathcal{PL}$	Preference list
$\mathcal{M}(s)$, $\mathcal{AS}(s)$	Matching set and applying user set of subband s resp.	$v^{k,a}_p$	Proposing virtual user relative to user k on antenna a

$$\text{such that} \sum_{a\in\mathcal{A}} x_{k,s,a} \leq 1,\ \forall (k,s) \in \mathcal{K}\times\mathcal{S} \tag{8a}$$

$$\sum_{k\in\mathcal{K}} x_{k,s,a} \leq 2,\ \forall (s,a) \in \mathcal{S}\times\mathcal{A} \tag{8b}$$

$$\sum_{k\in\mathcal{K}}\sum_{s\in\mathcal{S}} P_{s,a} x_{k,s,a} \leq P,\ \forall a \in \mathcal{A} \tag{8c}$$

$$p_{k_1,(s,a)} < p_{k_2,(s,a)}\ \forall (s,a), \in \mathcal{S}\times\mathcal{A} \tag{8d}$$

$$x_{k,s,a} \in \{0,1\}. \tag{8e}$$

Constraint (8a) restricts each subband to be assigned to one antenna only in the DAS case while (8b) limits the maximum number of users per subband to 2. (8c) is the power budget per antenna, where $P = P_{CAS}$ (resp. $P = P_{DAS}$) in the CAS setting (resp. DAS setting). Denoting by k_1 and k_2 the users scheduled on (s,a) s.t. $h_{k_1,s,a} > h_{k_2,s,a}$, k_2 must be allocated a higher power value than user k_1 as expressed in (8d) to guarantee SIC stability [8], [33], i.e. successful decoding at the user side. Indeed, as shown in [34], the power of the weak user must be strictly greater than that of the strong user. In the opposite case, the outage probabilities of the users will be always one. Note that (8) is formulated for the general case of a NOMA-DAS system, the other system configurations being special cases of it.

The optimization problem in (8) has two objectives. Since the applications of RT users are

time sensitive and can be considered as "more urgent" than BE applications, satisfying RT users is given a higher priority in the proposed solutions. Moreover, (8) is a mixed-integer multi-objective optimization problem for which an optimal solution is computationally intractable. If equal power repartition between subbands assigned to the same antenna is assumed, solving (8) consists of finding the optimal subband and antenna assignment. Therefore, we invoke the two-sided matching theory framework to obtain a suboptimal solution for the formulated problem.

III. Matching Theory in the OMA-CAS Context

A. Definition

To develop a low-complexity solution for (8), the subband assignment problem can be modeled as a two-sided one-to-many matching game. In this model, the set of subbands $\mathcal{S}$ and the set of users $\mathcal{K} = \mathcal{K}_{RT} \cup \mathcal{K}_{BE}$ form two disjoint sets of selfish and rational agents. In this first context, a subband s can be assigned to at most one user while a user k can be matched with more than one subband. If a user k is scheduled on subband s, (k, s) forms a *matching pair*. Note that, since only a single central base station is involved in the current CAS context, the antenna index a will be dropped from all involved variables.

A one-to-many matching Ψ is defined as a mapping from the set $\mathcal{K} \cup \mathcal{S}$ into the set of all subsets of $\mathcal{K} \cup \mathcal{S}$ such that for each $k \in \mathcal{K}$ and $s \in \mathcal{S}$:

1) $\Psi(k) \subseteq \mathcal{S}, \forall k \in \mathcal{K}$
2) $\Psi(s) \subseteq \mathcal{K}, \forall s \in \mathcal{S}$
3) $|\Psi(s)| = 1, \forall s \in \mathcal{S}$
4) $s \in \Psi(k) \Leftrightarrow k \in \Psi(s)$

To reach a final matching Ψ, each player p builds a preference relation $\succ_p$ over the players from the other set. Using these predefined preference relations, players dynamically interact with each other to reach a stable matching. Ψ is stable when there are no user k and subband s that are not matched together, but prefer each other over their partners $\Psi(k)$ and $\Psi(s)$, respectively. Hence, the subband assignment problem can be represented by the tuple $(\mathcal{S}, \mathcal{K}, \succ_s, \succ_k)$.

B. Preference Lists

To decide on the outcome of the game, each user k builds a preference list $\mathcal{PL}_k$ over the set of subbands. $\mathcal{PL}_k$ is sorted in a descending order based on the channel gain experienced by user k over all subbands in the set $\mathcal{S}$. In other words, $\mathcal{PL}_k$ is based on Definition 1.

Definition 1. Assuming equal inter-subband power allocation, users base their preferences on the channel gains over different subbands. Put differently, user k prefers subband s_i over s_j, i.e. $s_i \succ_k s_j$, where $s_i, s_j \in \mathcal{S}$, if $h_{k,s_i} > h_{k,s_j}$.

Similarly each subband s_i bases its preferences over the set of users based on Definition 2.

Definition 2. Subbands must account for the heterogeneity of users while building their preference lists by always preferring RT users over BE users, since they have the highest priority. Therefore, we define the preference relation of subband s_i as:

$$k_{RT} \succ_{s_i} k_{BE}, \forall s_i \in \mathcal{S}, k_{RT} \in \mathcal{K}_{RT} \text{ and } k_{BE} \in \mathcal{K}_{BE}.$$

In addition to preferring RT users, subbands must also be able to separately prioritize among the set of RT users and that of BE users. Therefore, at timeslot t, the following utility metric, inspired by the proposed metrics of [4], [5], is introduced for RT users:

$$\mathcal{U}^t_{RT}(k^l_{RT}, s_i) = \frac{R^t_{k^l_{RT},s_i}}{R^t_{k^l_{RT}}} \times \frac{t}{L_{k^l_{RT}}}. \tag{9}$$

In (9), $R^t_{k^l_{RT},s_i}$ denotes the achievable rate by k^l_{RT} over subband s_i if matched together at timeslot t and $R^t_{k^l_{RT}}$ is the rate already achieved by k^l_{RT} before reaching timeslot t. (9) is proportional to the achieved rate of user k_{RT} over subband s_i, and to the timeslot index t. Therefore, the users with the highest priorities are those who would benefit the most from subband s_i in terms of rate, and who have most approached their latency limit. Moreover, (9) is inversely proportional to $R^t_{k^l_{RT}}$ as well as to the latency limit of user k_{RT}. By considering t, $\mathcal{U}^t_{RT}(k^l_{RT}, s_i)$ grows larger as t increases. Also, by accounting for the latency limit $L_{k^t_{RT}}$, $\mathcal{U}^t_{RT}(k^l_{RT}, s_i)$ is increased for a more stringent latency requirement. In addition to considering the time and latency limits, (9) captures the transmission rate of k^t_{RT} if it were scheduled on s_i. Hence, this enables users with a better channel quality to have a higher priority during scheduling, therefore increasing spectral efficiency. Last, by also accounting for the achieved data rate of k^t_{RT} prior to reaching timeslot t, (9) achieves a certain fairness between RT users by prioritizing users that have not previously achieved a large enough throughput.

Each subband s_i bases its preferences over RT users according to the following definition.

Definition 3. At timeslot t, subband s_i prefers k^l_{RT} over k^m_{RT}, i.e. $k^l_{RT} \succ_{s_i} k^m_{RT}$, if $\mathcal{U}^t_{RT}\left(k^l_{RT}, s_i\right) > \mathcal{U}^t_{RT}\left(k^m_{RT}, s_i\right)$.

Similarly, subbands must differentiate between BE users, and the utility metric should strike a tradeoff between fairness and rate maximization. Hence, the PF scheduler metric [12] is adopted:

$$\mathcal{U}_{BE}^{t}\left(k_{BE}^{l}, s_i\right) = \frac{R_{k_{BE}^{l}, s_i}^{t}}{T_{k_{BE}^{l}}^{t}}. \tag{10}$$

It should be noted that (10) aims to maximize the performance measure for BE users formulated in (2), as the fairness measure $f(T_{k_{BE}}^{t})$ in (2) is represented by the weight $1/T_{k_{BE}}^{t}$ in (10).

Each subband s_i bases its preferences over BE users at timeslot t according to Definition 4.

Definition 4. Subband s_i prefers k_{BE}^{l} over k_{BE}^{m}, i.e. $k_{BE}^{l} \succ_{s_i} k_{BE}^{m}$, if $\mathcal{U}_{BE}^{t}\left(k_{BE}^{l}, s_i\right) > \mathcal{U}_{BE}^{t}\left(k_{BE}^{m}, s_i\right)$.

Having defined the preference relations for both sets of players, the proposed algorithm to solve the formulated matching game is described next.

C. Proposed OMA-CAS DA Algorithm

Since users have different priority levels, the classical DA algorithm [35] cannot be directly used to solve the considered matching game. That is why, in Algorithm 1, a priority-aware version of the DA algorithm suitable for the studied context in the OMA-CAS setting is proposed.

Initially, the set of active users $\mathcal{K}_{\text{active}}$ is built; it includes all RT users that have not yet received their requested data rate and all BE users. Each user $k \in \mathcal{K}_{\text{active}}$ builds its preference list $\mathcal{PL}_k$, whereas each subband s_i initializes its matching set $\mathcal{M}(s_i)$ consisting of the user to which it is assigned throughout the algorithm. At the first iteration of the algorithm, $\mathcal{M}(s_i) = \emptyset, \forall s_i \in \mathcal{S}$. In the first phase of the algorithm, each user k applies to its most preferred subband, i.e. the very first element in $\mathcal{PL}_k$. Each subband s_i receiving proposals forms an applying user set $\mathcal{AS}(s_i)$, to which it adds all proposing users as well as the user to which it was matched at the previous iteration of the algorithm. Note that $\mathcal{AS}(s_i)$ can be empty if none of the users apply to s_i and $\mathcal{M}(s_i) = \emptyset$. Having a system with heterogeneous users, subband s_i must prioritize RT users in the decision phase. Therefore, for each subband s_i receiving user proposals, i.e. having $\mathcal{AS}(s_i) \neq \emptyset$, if RT users are among the applicants, the most preferred RT user according to (9), k_{RT}^{*}, is accepted by s_i. All users in $\mathcal{AS}(s_i) \setminus \{k_{RT}^{*}\}$ are rejected and $\mathcal{K}_{\text{active}}$ is updated to reflect the resulting rate requirement changes. However, if a subband receives applications from BE users only, it accepts the most preferred BE user according to (10) and rejects all others. At the end of the second phase of the algorithm, every user removes the subband that it proposed to

at the current iteration from its preference list. This process continues until the preference lists of all active users are empty. Upon termination, the optimal matching result is obtained.

IV. Matching Theory in the OMA-DAS Context

A. DAS Matching Game Model and Algorithm

In this section, we aim to apply matching theory to solve the subband allocation problem in the DAS context. However, the DAS layout brings a new dimension into the resource allocation problem: antenna association. In addition to the user-subband assignment of the CAS context, it is also necessary to decide on the serving antenna for each subband assigned to a user. This new dimension complicates the problem considerably and renders the application of matching theory challenging. In fact, since we have to associate each user with a subband and an antenna, we are faced with a three-dimensional matching problem for which a stable solution is not guaranteed to exist [36].

Most previous studies on resource allocation in distributed settings make some assumptions with the aim of making the problem tractable. For example, [28] focused on antenna selection and power allocation to maximize the EE of a DAS setting, where the spectrum was assumed to consist of one subband only. However, in practical systems, this assumption does not hold. Maximizing the EE was also the purpose of [31], where the authors associated the user with the antenna providing the best average SINR, before proceeding with the subband and power allocation steps. Although the antenna selection scheme may seem logical, it might result in an antenna being associated a large number of users. This decreases the power available to each user on that antenna and some users may benefit from being assigned to other, less congested antennas. Moreover, in [31], a user was constrained to be associated with one antenna only, and all RRHs have access to the whole spectrum which increases the interference. In [30], the subband and power allocation steps were separated. For the subband assignment, the number of subbands per antenna was estimated based on the average path-loss of the users, and the actual subband assignment was performed with the aim of maximizing proportional fairness.

In the current work, a user is not restricted to be assigned to a unique antenna. Moreover, a subband can be assigned to one antenna only to limit interference. In addition to that, no *a priori* information about the distribution of subbands among RRHs is assumed. To overcome these challenges, the concept of virtual users is introduced, in which each user is duplicated into A virtual users, A being the number of antennas in the cell. This leads to a total of $K \times A$ virtual

users with each virtual user representing the potential association of a real user and a serving antenna. This transformation recovers the two-dimensional aspect of the resource allocation problem, which makes it possible to find a solution using matching theory.

As in the CAS case, the sets of players in the matching game as well as the preference lists must be defined. In the DAS context, the sets participating in the matching game are the set of virtual users $\mathcal{KV}$, consisting of RT and BE virtual users, and the set of subbands $\mathcal{S}$. Virtual users and subbands also build their preference relations according to Definition 1 and 2 respectively. However, the algorithm conceived for the CAS case is revisited.

First, allowing each active virtual user to propose to its most preferred subband might result in a real user being allocated more than one subband at each iteration. Although inconsistent with the nature of the matching game, this variation was compared to a second one that restricts each real user to apply through one virtual user only. Simulations showed that the second variation yields better results. Therefore, at each iteration, among virtual users pertaining to the same real user, only one is allowed to propose to its most preferred subband. This virtual user is selected to guarantee the best performance among virtual users relative to the same real user. Put differently, each real user must aim to be assigned to the antenna guaranteeing the best performance. However, the choice should not only take into consideration the channel gains of the users. In fact, the power levels of the subbands generally differ between antennas, depending on the respective congestion levels of the antennas. Consequently, the power level per subband on each antenna should also be considered in the decision process. Assuming equal inter-subband power allocation on each antenna, the power allocated to each subband assigned to antenna a (hence to each virtual user associated with a) is given by: $P_a = \frac{P_{DAS}}{N_a}$. N_a is the number of subbands assigned to antenna a, the derivation of which will be detailed in section IV-B. Then, each real user chooses the proposing virtual user according to Definition 5.

Definition 5. Proposing virtual user $v_p^{k,a}$ representing the association of real user k with antenna a is the one satisfying:

$$v_p^{k,a} = \underset{\substack{v^{k,a'},\\ a'=1,\cdots,A}}{\operatorname{argmax}} \left(P_{a'} \times h^2_{v^{k,a'},s^*_{v^{k,a'}}} \right). \tag{11}$$

In (11), $v^{k,a'}$ represents the virtual user associated to antenna a' and relating to real user k. $s^*_{v^{k,a'}}$ is the preferred subband of virtual user $v^{k,a'}$, i.e. the very first one in its preference list $\mathcal{PL}_{v^{k,a'}}$. Choosing the proposing user following (11) ensures that real users are matched with

their best subbands and antennas, in terms of rate maximization.

Algorithm 1 summarizes the proposed DA resource allocation algorithm in the OMA-DAS context, when $N_a, \forall a \in \mathcal{A}$, is known.

Algorithm 1 Priority-Aware OMA DA Algorithm

Input: $\mathcal{K}_{RT}, \mathcal{K}_{BE}, \boldsymbol{H_{RT}}, \boldsymbol{H_{BE}}, \boldsymbol{R}^{\text{req}}_{\boldsymbol{RT}}, \boldsymbol{L_{RT}}, \boldsymbol{T_{BE}}, t, \boldsymbol{N}$

Output: $\boldsymbol{A_{RT}}, \boldsymbol{A_{BE}}, \boldsymbol{R_{RT}}, \boldsymbol{R_{BE}}$ // $\boldsymbol{A}_{RT}$ (resp. $\boldsymbol{A}_{BE}$) is the assignment matrix of RT users (resp. BE users) to subbands and antennas while $\boldsymbol{R}_{RT}$ (resp. $\boldsymbol{R}_{BE}$) denotes their achieved rates over each (subband, antenna) pair

Initialization:

if CAS setting **then**
  $\mathcal{K}^{RT}_{\text{active}} = \left\{k_{RT} \in \mathcal{K}_{RT} / R^{\text{req}}_{k_{RT}} > 0\right\}, \mathcal{K}_{\text{active}} = \mathcal{K}_{BE} \cup \mathcal{K}^{RT}_{\text{active}}$.
else if DAS setting **then**
  Construct virtual user sets $\mathcal{K}^{v}_{RT}$ and $\mathcal{K}^{v}_{BE}$.
  $\mathcal{K}^{RT,v}_{\text{active}} = \left\{v^{k_{RT},a}, a = 1, \ldots, A \ / R^{\text{req}}_{k_{RT}} > 0\right\}$, $\mathcal{K}_{\text{active}} = \mathcal{K}^{v}_{BE} \cup \mathcal{K}^{RT,v}_{\text{active}}$.
end if
Build the preference lists $\mathcal{PL}_k$ of users in $\mathcal{K}_{\text{active}}$. Set $\mathcal{M}(s_i) = \emptyset, \forall s_i \in \mathcal{S}$.

Phase 1: Active users apply to subbands

if CAS setting **then**
  Each user $k \in \mathcal{K}_{\text{active}}$ proposes to the first subband in $\mathcal{PL}_k$.
else if DAS setting **then**
  Each real user $k \in \mathcal{K}_{\text{active}}$ chooses its proposing virtual user $v^{k,a}_p$ and proposes to the first subband in $\mathcal{PL}_{v^{k,a}_p}$.
end if
Find the applicant set $\mathcal{AS}(s_i)$ for each subband $s_i \in \mathcal{S}$, $\mathcal{AS}(s_i) = \mathcal{AS}(s_i) \cup \mathcal{M}(s_i), \ \forall s_i \in \mathcal{S}$.

Phase 2: Subbands make decisions

if $\mathcal{AS}(s_i) \neq \emptyset$ and $\mathcal{AS}(s_i) \cap \mathcal{K}^{RT}_{\text{active}} \neq \emptyset$ **then**
  $\mathcal{M}(s_i) = \{k^*_{RT}\}$, where $k^*_{RT} = \underset{k^l_{RT} \in \mathcal{AS}(s_i) \cap \mathcal{K}^{RT}_{\text{active}}}{\text{argmax}} \mathcal{U}^t_{RT}(k^l_{RT}, s_i)$.
  Update $R^{\text{req}}_{k^*_{RT}}$, $\mathcal{K}^{RT}_{\text{active}}, \mathcal{K}^{RT,v}_{\text{active}}$ and $\mathcal{K}_{\text{active}}$.
else if $\mathcal{AS}(s_i) \neq \emptyset$ **then**
  $\mathcal{M}(s_i) = \{k^*_{BE}\}$, where $k^*_{BE} = \underset{k^l_{BE} \in \mathcal{AS}(s_i)}{\text{argmax}} \mathcal{U}^t_{BE}(k^l_{BE}, s_i)$.
end if

Phase 3: Preference lists update

Each user k that proposed to $s_i, \forall s_i \in \mathcal{S}$, removes s_i from $\mathcal{PL}_k$.

Repeat Phases 1, 2 and 3 until $\mathcal{PL}_k = \emptyset, \ \forall k \in \mathcal{K}_{\text{active}}$

B. Estimation of the number of subbands per antenna

To find $N_a, \ \forall a \in \mathcal{A}$, we first use $w_{k,a}$, the large-scale fading parameter between user k and antenna a, as was done in [30]:

$$N_a = \left\lfloor \frac{S \times \sum_{k=1}^{K} w_{k,a}}{\sum_{a=1}^{A} \sum_{k=1}^{K} w_{k,a}} \right\rfloor, a \in \mathcal{A}. \tag{12}$$

However, contrary to [30], in our work, this estimation is only used for initial power approximation and does not dictate the number of assigned subbands to each antenna a in each timeslot.

Using this initial estimation of the number of subbands per antenna, Algorithm 2 provides the final assignment of users and subbands to antennas. In each iteration of Algorithm 2, each antenna performs equal-power distribution using its estimated number of assigned subbands. Then, subband assignment is determined using Algorithm 1. The number of subbands per antenna is then updated, as well as $P_a,\ \forall a \in \mathcal{A}$. Algorithm 2 converges when the number of subbands per antenna remains unchanged between two successive iterations.

Algorithm 2 OMA-DAS Resource Allocation

Input: $\boldsymbol{N}, \mathcal{K}_{RT}, \mathcal{K}_{BE}, \boldsymbol{H_{RT}}, \boldsymbol{H_{BE}}, \boldsymbol{R}^{\text{req}}_{\boldsymbol{RT}}, \boldsymbol{L_{RT}}, \boldsymbol{T_{BE}}, t$
Output: $\boldsymbol{A_{RT}}, \boldsymbol{A_{BE}}, \boldsymbol{R_{RT}}, \boldsymbol{R_{BE}}, \boldsymbol{N}$ // $\boldsymbol{N}$ is the number of subbands per antenna

Repeat:
  $N^{\text{old}} = N$.
  $P_a = P_{DAS}/N_a$.
  Find the assignment of users and subbands to antennas according to Algorithm 1.
  Using the resulting $\boldsymbol{A_{RT}}$ and $\boldsymbol{A_{BE}}$, re-calculate $\boldsymbol{N} \in \mathbb{N}^{A\times 1}$ and $\boldsymbol{P} \in \mathbb{R}_{+}^{A\times 1}$ as well as $\boldsymbol{R_{RT}}$ and $\boldsymbol{R_{BE}}$.
Until convergence

V. Matching Theory in the NOMA Context

Solving the resource allocation problem in the NOMA context using matching theory is not straightforward. On the one hand, the power multiplexing constraints, neglected in previous works like [26], must be respected to guarantee SIC stability. On the other hand, applying the methods proposed in previous works like [26] and [37] does not guarantee the rate requirements of RT users. Nor does applying the same algorithms proposed in sections III and IV, while allowing multiple users to be scheduled on the same subband. Moreover, because of the interdependencies between users' preferences, due to the inter-user interference between paired users on NOMA subbands, the outcome of these algorithms is not guaranteed to be optimal. That is why, in this section, we generalize the resource allocation techniques proposed in the previous sections to encompass the NOMA case, starting with the NOMA-CAS context.

The proposed solution to the resource allocation problem in the NOMA-CAS context is divided into two stages. The first one, consisting of the assignment of single RT users and NOMA BE users, aims at maximizing the number of satisfied RT users, as well as boosting the performance of BE users when possible. The second stage, where user pairings are performed on the subbands allocated to RT users, aims at further enhancing system performance.

A. Assignment of Subbands to Single RT Users and NOMA BE Users

To ensure that rate requirements of RT users are met, the allocation process starts by scheduling OMA RT users, as done in section III. The preference relations of RT users and subbands are formulated in the same way as in section III. If all RT users reach their rate requirements and free subbands remain in the system, BE users are scheduled on these subbands directly via NOMA.

NOMA Matching Game for BE Users: In [26], matching theory was used to solve the subband allocation problem in NOMA and a technique based on the DA algorithm was proposed. However, interdependencies between users' preferences exist in the NOMA case because of the interference brought by non-orthogonally scheduling different users on the same subband. This is why previous work as in [27] and [37] performed a swapping step at the end of the matching step to further enhance system performance. This step allows different users to swap their assigned subbands, conditioned by the approval of all involved players, and generally requires a significant number of additional iterations.

In this study, we follow a different direction by allowing single users as well as pairs of users to make proposals. By doing so, interdependencies between users are directly taken into account, without the need for an additional swap phase. Also, thanks to this new idea, a hybrid-NOMA system (where subbands are either allocated to sole or paired users) is enabled, which achieves a better performance than non-hybrid NOMA [38].

Similarly to the previous sections, the sets of players in the NOMA-CAS case are subbands and users. However, the user set, denoted by $\mathcal{US}$, now consists of both single users and pairs of users. It contains $|K_{BE}| + P(|K_{BE}|, 2)$ elements, where $|K_{BE}|$ accounts for single users and $P(|K_{BE}|, 2)$ for user pairs. Each user combination in $\mathcal{US}$ complies with the following definition to form its preference list.

Definition 6. Each user combination $us_n \in \mathcal{US}$ divides the available power per subband, P_s, among its members according to the fractional transmit power allocation (FTPA) [13] (Note that if $|us_n| = 1$, P_s is entirely allocated to the sole user in us_n). Then, us_n bases its preferences on (13), where R_{us_n,s_i} is the rate achieved by users in the set us_n on subband s_i.

$$s_i \succ_{us_n} s_j \text{ if } R_{us_n,s_i} > R_{us_n,s_j}. \tag{13}$$

For subbands, the utility achieved by scheduling us_n is defined in (14), with $\mathcal{U}^t_{BE}(k, s_i)$ given by (10).

$$\mathcal{U}^t_{BE}(us_n, s_i) = \sum_{k \in us_n} \mathcal{U}^t_{BE}(k, s_i), \tag{14}$$

Then, subbands base their preferences on the following relation:

$$us_n \succ_{s_i} us_m \quad \text{if} \quad \mathcal{U}^t_{BE}(us_n, s_i) > \mathcal{U}^t_{BE}(us_m, s_i). \tag{15}$$

Phases 1 and 2 of Algorithm 3 describe the steps involved in the first stage of the resource allocation technique in the NOMA-CAS case.

B. NOMA Pairing on Subbands Assigned to RT Users

Since the first stage of the proposed solution in the NOMA-CAS context schedules RT users based on OMA and assuming equal inter-subband power repartition, some scheduled RT users exceed their required rates, while others may not have satisfied their requirements (because of system congestion or bad channel states). Consequently, in this second stage of the allocation process, NOMA pairing is performed on the subbands assigned to RT users in the first stage.

This second stage starts by finding the amount of power that satisfied RT users can spare, without jeopardizing their satisfaction. Therefore, for each RT user having exceeded its required rate, the following optimization problem is solved:

$$\min_{\boldsymbol{p}_{k_{RT}}} \sum_{s \in \mathcal{S}_{k_{RT}}} p_{k_{RT},s} \tag{16}$$

$$\text{such that} \sum_{s \in \mathcal{S}_{k_{RT}}} R_{k_{RT},s} = R^{\text{req}}_{k_{RT}} \tag{16a}$$

$$0 \leq p_{k_{RT},s,a_s} \leq P_s. \tag{16b}$$

In (16), $\mathcal{S}_{k_{RT}}$ is the set of subbands assigned to user k_{RT} in the first stage of the allocation process. Solving the above optimization problem leads to the well-known waterfilling solution:

$$p_{k_{RT},s} = \left[\frac{\lambda_{k_{RT}} B_c}{\log(2)} - \frac{N_0 B_c}{H^2_{k_{RT},s,a_s}} \right]^{P_s}_0. \tag{17}$$

In (17), $\lambda_{k_{RT}}$ is the Lagrange multiplier given by:

$$\lambda_{k_{RT}} = 2^{\frac{1}{S_{k_{RT}}} \left[\frac{R^{\text{req}}_{k_{RT}}}{B_c} - \sum\limits_{s \in \mathcal{S}_{k_{RT}}} \log_2 \left(\frac{H^2_{k_{RT},s}}{\log(2) N_0} \right) \right]}. \tag{18}$$

The required rate on $s \in \mathcal{S}_{k_{RT}}$ is given by:

$$R^{\text{req}}_{k_{RT},s} = B_c \log_2 \left(1 + \frac{p_{k_{RT},s} H^2_{k_{RT},s}}{N_0 B_c} \right). \tag{19}$$

This rate must be guaranteed regardless of the pairing order on s. Note that if the rate $R_{k_{RT},s}$ obtained on s at the end of the first allocation stage is lower than $R^{\text{req}}_{k_{RT},s}$, s is removed from

$\mathcal{S}_{RT}$, the set of subbands available for NOMA pairing (given by step 7 of Algorithm 3). In the opposite case, s can be shared with another user. To guarantee $R^{\text{req}}_{k_{RT},s}$, two separate cases are considered concerning the channel gain of the candidate user for pairing, k':

1) $h_{k_{RT},s} > h_{k',s}$*:* In this case, k_{RT} is paired as first user on s with a required rate given by (19). However, to guarantee SIC stability, $p_{k_{RT},s} < p_{k',s}$ must hold. This translates into considering s for NOMA pairing if $p_{k_{RT},s} < P_s/2$. If this condition is verified, the available power for k' on s is: $p^{2,\text{av}}_s = P_s - p_{k_{RT},s}$.

2) $h_{k_{RT},s} < h_{k',s}$*:* In this case, k_{RT} is paired as second user on s. To guarantee $R^{\text{req}}_{k_{req},s}$, k_{RT} needs to be assigned a power equal to:

$$p^{2,\text{req}}_{k_{RT},s} = \frac{(a-1)\left(P_s h^2_{k_{RT},s} + N_0 B_c\right)}{a h^2_{k_{RT},s}}, \tag{20}$$

where $a = 2^{R^{\text{req}}_{k_{RT},s}/B_c}$. SIC stability is guaranteed if $p^{2,\text{req}}_{k_{RT},s} > P_s/2$ which is warranted if $a \geq 2$. Hence, if being scheduled as second user, the required rate of k_{RT} to ensure SIC stability is chosen to be $R^{\text{req}}_{k_{RT},s} = \max\left(R^{\text{req}}_{k_{RT},s}, B_c\right)$. In the case where $R^{\text{req}}_{k_{RT},s}$ takes the value of B_c, $p^{2,\text{req}}_{k_{RT},s}$ is recalculated to reflect the change. Then, the power available for a potential first user k' is given by: $p^{1,\text{av}}_s = P_s - p^{2,\text{req}}_{k_{RT},s}$.

Having determined the amounts of power available for NOMA pairing, the matching algorithm proposed for the second allocation stage can now be described.

The active users in this second stage are the unsatisfied RT users and the BE users, hence $\mathcal{K}_{\text{active}} = \{\mathcal{K}_{BE} \cup k_{RT} \in \mathcal{K}_{RT} / R^{\text{req}}_{k_{RT}} > 0\}$. First, the achievable rate of each candidate user $k \in \mathcal{K}_{\text{active}}$ over subband $s \in \mathcal{S}_{\text{rem}}$ (where k_{RT} is scheduled) is calculated using (21). Then, each active user builds its preference list according to the decreasing order of achievable rates.

$$R^{\text{av}}_{k,s} = \begin{cases} B_c \log_2\left(1 + \dfrac{p^{1,\text{av}}_s h^2_{k,s}}{N_0 B_c}\right), \text{ if } h_{k,s} > h_{k_{RT},s}, \\ B_c \log_2\left(1 + \dfrac{p^{2,\text{av}}_s h^2_{k,s}}{p_{k_{RT},s} h^2_{k,s} + N_0 B_c}\right), \text{ otherwise.} \end{cases} \tag{21}$$

Since only one additional user is to be scheduled on each subband allocated to RT users, Algorithm 1 can be used to solve the second stage of the allocation process. However, instead of considering all subbands in set $\mathcal{S}$ as done in Algorithm 1, only those assigned to RT users in the previous stage must be considered (i.e. subbands in $\mathcal{S}_{RT}$).

C. Matching Technique for NOMA-CAS

The complete algorithm used to solve the allocation problem in the NOMA-CAS setting is given in Algorithm 3.

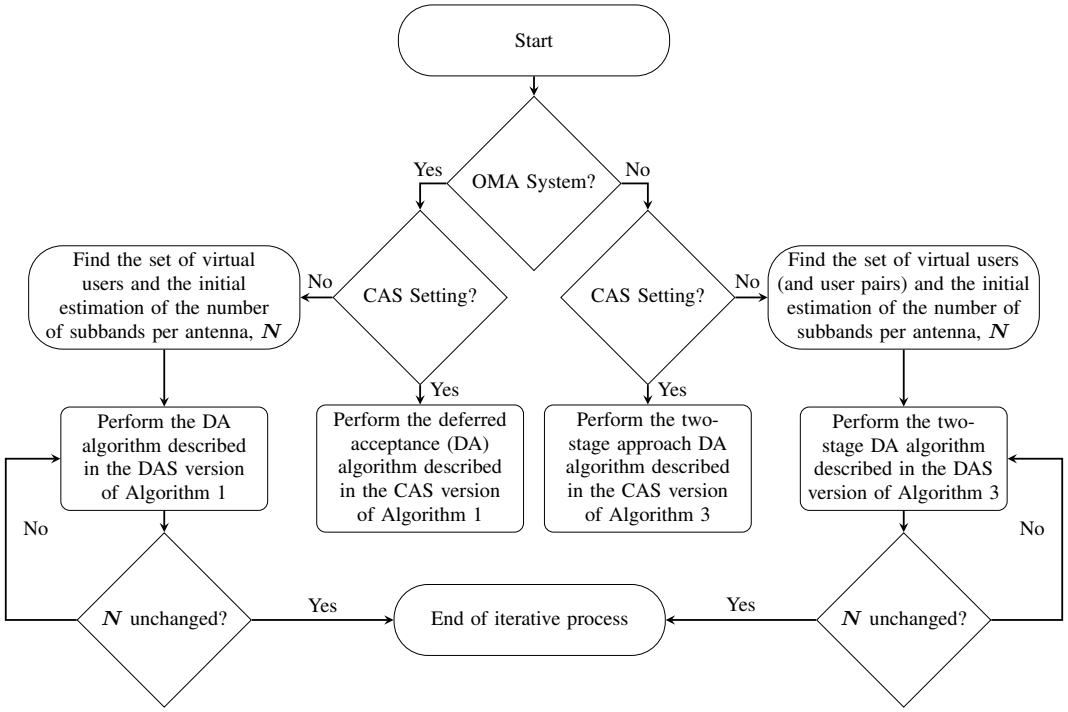


Fig. 2: Flowchart of the proposed matching technique in the studied settings

D. Matching Technique in the NOMA-DAS Context

Resource allocation through matching theory in the NOMA-DAS case is conducted by incorporating the concepts from Section IV and subsections V-A, V-B and V-C. The proposed method is detailed in Algorithm 3.

To summarize, the flowchart in Fig. 2 shows how the matching technique is used in the different system settings.

VI. Analysis of Stability, Convergence and Complexity

To analyze the properties of the proposed matching technique in all studied system settings (i.e. OMA-CAS, OMA-DAS, NOMA-CAS, NOMA-DAS), we separately consider the following two parts: 1) The matching algorithm, 2) The iterative approach proposed to find the number of subbands per antenna in the DAS case.

A. Properties of the Matching Technique

Before discussing the stability property of the matching technique, the definition of *blocking pair* [29] is first recalled.

Algorithm 3 Priority-Aware NOMA DA Algorithm

Input: $\mathcal{K}_{RT}, \mathcal{K}_{BE}, \boldsymbol{H_{RT}}, \boldsymbol{H_{BE}}, \boldsymbol{R}^{\text{req}}_{\boldsymbol{RT}}, \boldsymbol{L_{RT}}, \boldsymbol{T_{BE}}, t$
Output: $\boldsymbol{A_{RT}}, \boldsymbol{A_{BE}}, \boldsymbol{R_{RT}}, \boldsymbol{R_{BE}}$

Initialization:

$\mathcal{K}^{RT}_{\text{active}} = \left\{k_{RT} \in \mathcal{K}_{RT} / R^{\text{req}}_{k_{RT}} > 0\right\}$ in the CAS setting and $\mathcal{K}^{RT}_{\text{active}} = \left\{v^{k_{RT},a}, a = 1, \ldots, A \ / R^{\text{req}}_{k_{RT}} > 0\right\}$ in the DAS setting.

Build the preference lists $\mathcal{PL}_k$ of users in $\mathcal{K}^{RT}_{\text{active}}$.

Set $\mathcal{M}(s_i) = \emptyset, \ \forall s_i \in \mathcal{S}$.

$P_a = P_{DAS}/N_a, \ \forall a \in \mathcal{A}$. // Only in DAS setting

Repeat steps 5 to 27 in the DAS setting

$N^{\text{old}} = N$.

Phase 1: Scheduling RT users

Perform phases 1, 2 and 3 from Algorithm 1.

Repeat Phase 1 until $\mathcal{PL}_k = \emptyset, \ \forall k \in \mathcal{K}^{RT}_{\text{active}} \ || \ \mathcal{K}^{RT}_{\text{active}} = \emptyset$

$\mathcal{S}_{RT} \leftarrow$ Subbands used by RT users,
$\mathcal{S}_{\text{rem}} \leftarrow \mathcal{S} \setminus \mathcal{S}_{RT}$.

Phase 2: Scheduling BE users

if $\mathcal{S}_{\text{rem}} \neq \emptyset$ **then**

Initialization:

Construct user set $\mathcal{US}$ consisting of both single users and user pairs in the CAS setting (resp. virtual user set in the DAS setting consisting of virtual users and user sets).

Build the preference lists $\mathcal{PL}_{us_n}$ of user sets $us_n \in \mathcal{US}$.

Set $\mathcal{M}(s_i) = \emptyset, \ \forall s_i \in \mathcal{S}_{\text{rem}}$.

Phase 2.1: BE users and BE pairs apply to subbands

Each $us_n \in \mathcal{US}$ proposes to the first subband in $\mathcal{PL}_{us_n}$ in the CAS setting (In the DAS setting, each real user chooses its proposing virtual user us^v_n and proposes to the first subband in $\mathcal{PL}_{us^v_n}$).

Construct the applicant set $\mathcal{AS}(s_i)$ for each subband $s_i \in \mathcal{S}_{\text{rem}}$, $\mathcal{AS}(s_i) = \mathcal{AS}(s_i) \cup \mathcal{M}(s_i), \ \forall s_i \in \mathcal{S}_{\text{rem}}$.

Phase 2.2: Subbands make decisions

if $\mathcal{AS}(s_i) \neq \emptyset$ **then**

$\mathcal{M}(s_i) = \{us^*_n\}, \ \text{where } us^*_n = \underset{us_n \in \mathcal{A}(s_i)}{\text{argmax}} \ \mathcal{U}^t_{BE}(us_n, s_i)$.

end if

Phase 2.3: Preference lists update

Each user set us_n that proposed to $s_i, \ \forall s_i \in \mathcal{S}_{\text{rem}}$, removes s_i from $\mathcal{PL}_{us_n}$.

Repeat Phases 2.1, 2.2 and 2.3 until $\mathcal{PL}_{us_n} = \emptyset, \ \forall \ us_n \in \mathcal{US}$

end if

Using the resulting $\boldsymbol{A_{RT}}$ and $\boldsymbol{A_{BE}}$, re-calculate $\boldsymbol{N} \in \mathbb{N}^{A \times 1}$ and $\boldsymbol{P} \in \mathbb{R}^{A \times 1}_{+}$ as well as $\boldsymbol{R_{RT}}$ and $\boldsymbol{R_{BE}}$.

Phase 3: NOMA pairing on subbands assigned to RT users

if $\mathcal{S}_{RT} \neq \emptyset$ **then**

Find RT users that exceed their required rates $\mathcal{K}^{\text{excess}}_{\text{RT}}$.

Solve (16) for $k \in \mathcal{K}^{\text{excess}}_{\text{RT}}$.

Find the available power on $s \in \mathcal{S}_{RT}$.

$\mathcal{K}_{\text{active}} = \{\mathcal{K}_{BE} \cup k_{RT} \in \mathcal{K}_{RT} / R^{\text{req}}_{k_{RT}} > 0\}$.

Build preference lists for users in $\mathcal{K}_{\text{active}}$.

Use Algorithm 1 to schedule additional users on subbands belonging to $\mathcal{S}_{RT}$.

end if

Until convergence

Definition 7. Given a matching Ψ and a pair (k_n, s_i), where $k_n \in \mathcal{US} \cup \mathcal{K}_{RT}$ and $s_i \in \mathcal{S}$, with $k_n \notin \Psi(s_i)$ and $s_i \notin \Psi(k_n)$, (k_n, s_i) forms a blocking pair if:

1) $k_n \succ_{s_i} \Psi(s_i)$,
2) $s_i \succ_{k_n} s_j$, where $s_j \in \Psi(k_n)$.

With the definition above, it is now possible to define the concept of stability and prove that the proposed matching technique does indeed converge towards a stable matching.

Definition 8. If there is no blocking pair $(k_n, s_i) \in \Psi$, matching Ψ is considered *stable*.

Theorem 1. The proposed matching technique in all system settings is guaranteed to converge to a stable matching Ψ^*.

Proof. See Appendix A. ■

Theorem 2. The matching technique is guaranteed to converge after a limited number of iterations for all studied system settings.

Proof. See Appendix B. ■

Theorem 3. The maximum complexity of the proposed matching technique is $\mathcal{O}\left((K_{RT} + |\mathcal{US}|)AS^2\right)$, and is achieved in the NOMA-DAS setting.

Proof. See Appendix C. ■

B. Properties of the Iterative Approach

Theorem 4. The iterative approach introduced in Algorithm 2 is guaranteed to converge within a limited number of iterations.

Proof. See Appendix D. ■

Concerning the complexity of Algorithm 2, the number of iterations cannot be given in a closed form expression because it is not certain at which round the algorithm converges to the final solution. However, an upper bound on the complexity can be derived. Since system performance increases after each iteration, if ΔP denotes the performance gain yielded by the iterative approach and δ_{min} is the minimum increase in performance at each iteration, an upper bound on the complexity of this method is given by $\mathcal{O}(\frac{\Delta P}{\delta_{min}})$.

VII. Numerical Results

The performance of the proposed subband assignment technique based on matching theory, denoted by "MM", is evaluated in this section through simulations. The performance of MM is tested in the OMA-CAS, OMA-DAS, NOMA-CAS and NOMA-DAS settings. A variation of the MM method, denoted by MM-FA, in the DAS settings is also tested. MM-FA adopts the approach of [30] and determines the number of subbands per antenna, $N_a, \forall a \in \mathcal{A}$, at the beginning of the resource allocation algorithm based on the average path-loss experienced by all users. For fair comparison, each antenna a is assigned the N_a subbands having the highest average channel gain for all users. The performance achieved by the greedy method, denoted by "GM", that was formerly proposed in [8] is also shown for comparison.

The parameters used in the simulations are summarized in Table II [39]. When a CAS case is considered, one antenna is assumed to be located at the center of the cell. In the case of DAS, $A = 4$ RRHs are assumed to be deployed in the cell. In addition to the centrally located antenna, the other 3 antennas are equally distanced and positioned on a circle of radius $2R_d/3$, R_d being the cell radius, with an angular separation of $120°$. The power budget per cell is 40 W; when DAS is considered, this budget is equally partitioned between antennas leading to 10 W per antenna. The number of RT users K_{RT}, is varied between 5 and 30 to depict different system congestion levels. To reflect different RT application requirements, RT users are partitioned among 3 classes (C1, C2 and C3). Users in all 3 classes request 10^5 bits. However, a user $k_{RT} \in$ C1 has a latency limit $L_{k_{RT}} = 6$ timeslots, whereas if $k_{RT} \in$ C2 (resp. $k_{RT} \in$ C3), $L_{k_{RT}} = 10$ timeslots (resp. $L_{k_{RT}} = 15$ timeslots). For all K_{RT} values, it is assumed that 20% of the users belong to each of C1 and C2 while the remaining 60% users belong to C3. The number of BE users K_{BE} is maintained at 20 throughout the simulations.

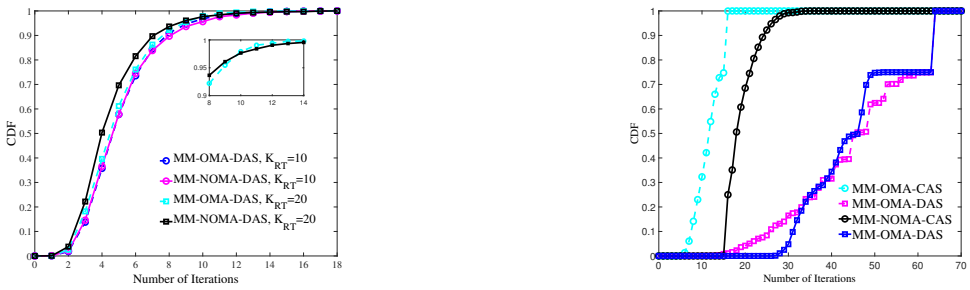
TABLE II: Simulation Parameters

Cell Radius R_d	500 m	Overall Transmission Bandwidth	10 MHz
Number of subbands	16	Cell Power Budget	40 Watts (46 dBm)
Number of RT users in the cell	5, 10, 15, 20, 25, 30	Number of BE users in the cell	20
Distance Dependent Path Loss	$128.1 + 37.6 \log_{10}(d)$(dB), d in Km	Receiver Noise Density	4.10^{-18} mW/Hz

A. Convergence of the Proposed Method

First, the convergence of the iterative method that aims to find the number of subbands per antenna is observed. Fig. 3a plots the cumulative distribution function (CDF) of the number of

iterations needed for Algorithm 2 to converge for both the OMA and NOMA settings, when the number of RT users takes the values 10 and 20. The CDF shows that the proposed method converges within a small number of iterations (90% of the cases converge within 8 iterations) for both the OMA and the NOMA settings, as well as for both $K_{RT} = 10$ and $K_{RT} = 20$ users. The convergence of the NOMA settings is slightly slower than its OMA counterpart. Fig. 3a shows that the convergence of the setting with $K_{RT} = 10$ users is slower than the one with 20 RT users. In fact, when $K_{RT} = 10$, BE users have a higher chance of being scheduled. Hence, all 30 users (10 RT users and 20 BE users) contribute to deciding on the assignment of subbands to the RRHs. However, when $K_{RT} = 20$, the system is more congested and BE users have a harder time getting resources. Therefore, the subband to antennas assignment is mostly decided upon by the 20 RT users in this case, which explains the faster convergence.



(a) (b)

Fig. 3: CDF of the number of iterations needed: (a) to find the number of subbands per antenna, (b) for the matching method with $K_{RT} = 10$

In Fig. 3b, the convergence of the MM technique is shown for the different settings, when $K_{RT} = 10$ users. As expected, the OMA settings converge faster than the NOMA ones. However, it can be seen that the maximum number of iterations needed for MM to converge is $AS = 64$ in the DAS settings, which is a relatively small number of iterations.

B. Performance of the MM Technique

In Fig. 4, the performance of the proposed technique in terms of RT users satisfaction is evaluated. It can be noted that, until $K_{RT} = 15$ users, MM and GM perform similarly regardless of the considered scenario. However, as the cell becomes more congested with a larger number of RT users, MM outperforms GM in all its variations. More concretely, when $K_{RT} = 30$, GM achieves almost no satisfaction for RT users. However, MM-OMA-CAS (resp. MM-OMA-DAS) outperforms its GM equivalent by almost 28% (resp. 62 %). Also, in the NOMA cases, MM-NOMA-CAS (resp. MM-NOMA-DAS) outperforms its GM equivalent by almost 30% (resp.

63%). Fig. 4 also shows the gains achieved by the iterative method to find the number of subbands per antenna introduced in Algorithm 2. For example, when $K_{RT} = 30$ users, MM-OMA-DAS (resp. MM-NOMA-DAS) outperforms MM-FA-OMA-DAS (resp. MM-FA-NOMA-DAS) by almost 30% (resp. 26%). Additionally, the results show the gain achieved by using a DAS setting, in comparison to CAS, as it can increase the performance by more than 30%.

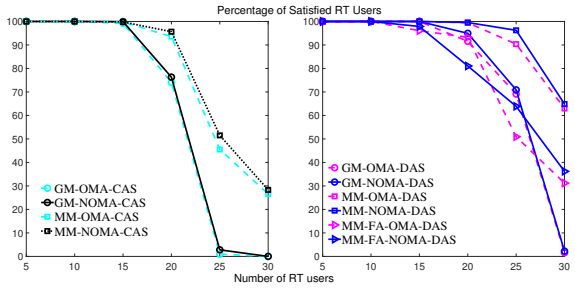


Fig. 4: Percentage of satisfied RT users for the CAS settings (left) and the DAS settings (right)

Having observed the performance enhancement brought by the DAS settings, hereinafter, only the performance of MM-NOMA-CAS will be compared to the other methods as it has the best behavior in the CAS setting. Also, MM-FA-OMA-DAS will be dropped since it is outperformed by its NOMA version.

In Fig. 5, the percentage of satisfied RT users per class is shown. As previously stated, GM and MM both satisfy all RT users when $K_{RT} \leq 15$ users. However, when $K_{RT} \geq 20$, the performance of GM degrades and the satisfaction of RT users belonging to the strictest class, C1 (with a latency of 6 timeslots), is mostly affected. In fact, for $K_{RT} = 25$ for example, the GM technique achieves close to 10% satisfaction for users belonging to C1 in the case of a DAS system, while a CAS system cannot satisfy any user in C1. On the other hand, MM-OMA-DAS and MM-NOMA-DAS achieve both almost 96% satisfaction while the CAS versions achieve almost 70% satisfaction. For the more relaxed classes, the percentage of satisfied RT users with the GM method is higher than for C1, since their requirements are more relaxed. Hence, even if the GM algorithm achieves an acceptable global percentage of satisfaction for $K_{RT} = 20$ and 25, this performance results from the satisfaction of the users in the most relaxed classes. This is not the case for the MM algorithm which prioritizes users having more strict requirements. Therefore, it achieves an acceptable performance in all classes.

Remark. In Fig. 5, for C1, we notice that MM-OMA-DAS outperforms MM-NOMA-DAS for a system consisting of 30 RT users. This results from the fact that at timeslot t, for NOMA pairing, all additional power allocated to a satisfied RT user k_{RT} is taken away from it to accommodate

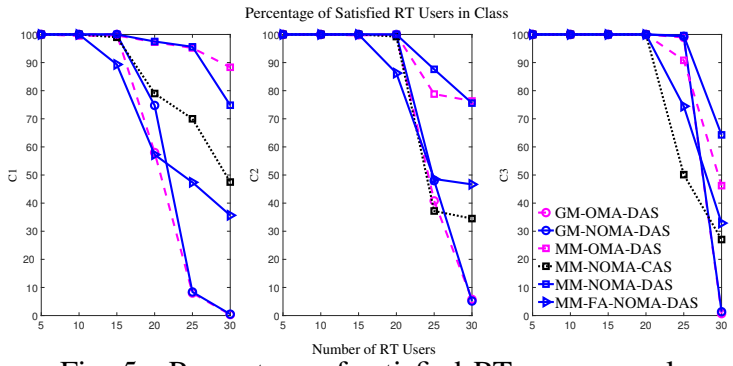
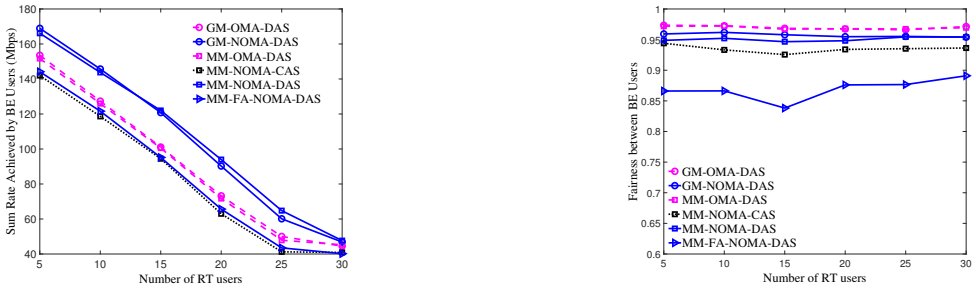


Fig. 5: Percentage of satisfied RT users per class

more users via NOMA. Hence, the rate achieved by k_{RT} at t is equal to its required rate, not more. At a subsequent timeslot t' ($t < t' \leq L_{k_{RT}}$), k_{RT} might not be able to reach its required rate (for example, because of a bad channel state). However, if k_{RT} was allowed to exceed its required rate at timeslot t, the required rate at t' would be reduced and hence k_{RT} could have been satisfied. A solution for this problem might be to allow RT users to keep a small amount of additional power during the NOMA pairing step.

Fig. 6a shows the achieved rate of BE users as K_{RT} increases. As expected, the sum rate of BE users decreases for all methods as K_{RT} grows, as less resources are available for BE users. Both methods (GM and MM) perform similarly when it comes to the sum rate achieved by BE users. For example, for the NOMA-DAS case, GM-NOMA-DAS achieves almost 1 Mbps gain over MM-NOMA-DAS when $K_{RT} = 5$ or 10 users. However, for $K_{RT} = 20$ or 25 users, MM-NOMA-DAS achieves almost 3 Mbps gain over GM-NOMA-DAS. Therefore, while significantly enhancing the satisfaction of RT users, MM does not degrade the sum rate of BE users. Moreover, MM-NOMA-DAS greatly outperforms MM-FA-NOMA-DAS. Hence, the complexity added by the use of Algorithm 2 is justified by the enhanced performance for both RT and BE users.



(a) (b)

Fig. 6: (a) Achieved rate for BE users as K_{RT} increases, (b) System Fairness in terms of K_{RT}

In Fig. 6b, we show the fairness achieved by the different methods in terms of K_{RT}. System

fairness is assessed through Jain's fairness index [40]: $J = \frac{\left(\sum_{k\in\mathcal{K}_{BE}} R_k\right)^2}{K_{BE}\sum_{k\in\mathcal{K}_{BE}} R_k^2}$, where R_k is the total rate achieved by user k. Jain's fairness index ranges between 0 and 1 with the maximum achieved in the case of optimal fairness. It can be seen that MM-NOMA-DAS outperforms its FA counterpart. Putting MM-FA apart, Fig. 6b shows that all considered methods have a good performance in terms of fairness with a Jain index higher than 0.9, with an advantage for the DAS settings. Therefore, system fairness is not a deciding factor in the evaluation of the different methods. It should be noted that although the OMA versions slightly outperform the NOMA ones, NOMA increases the minimum individual rate of BE users. Hence, the slightly decreased fairness is due to some users having slightly more rate that others in the NOMA setting.

To show the adequacy of the proposed metrics in (9) and (10), their performance is compared against the unified metric proposed in [2], given by (22), where $\alpha_k = 1$ for BE users and $\alpha_k = 1 + t$ for RT users. $\theta_k(t)$ is the normalized delay.

$$\mathcal{UM}_{k,s}(t) = \alpha_k R_{k,s}^t \exp(\theta_k(t)). \tag{22}$$

To compare, a NOMA-DAS setting is considered and Algorithm 3 is employed. However, the preference relation of the subbands is modified depending on the tested metric. The performance of the unified metric is denoted by MM-NOMA-DAS-Metric 2 in the simulations.

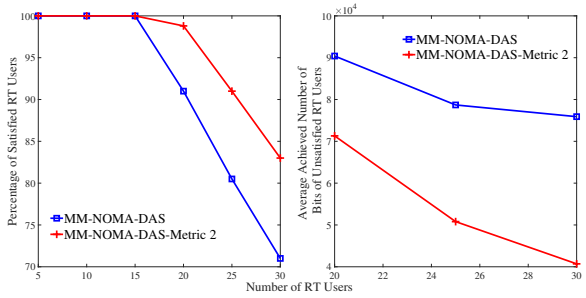


Fig. 7: Percentage of Satisfied RT Users as K_{RT} increases (left); Average Number of Bits Received by Unsatisfied RT Users as K_{RT} increases (right)

Fig. 7 compares the performance of the two metrics for RT users. In terms of the percentage of satisfied RT users, Fig. 7 (left) shows that both metrics achieve a similar performance for $K_{RT} \leq 15$ users. As K_{RT} increases, (22) outperforms (9) by up to 10% for $K_{RT} = 30$. That is because (22) privileges users with a high rate. In contrast, (9) seeks to achieve a high fairness between RT users by accounting for the received rate before timeslot t. In other words, even if it were unable to satisfy all users, (9) aims at approaching most RT users to their requirements,

as shown in Fig. 7 (right). Indeed, Fig. 7 (right) shows that (9) increases the amount of received data bits of unsatisfied RT users by up to 3.6×10^4 bits, when compared to (22).

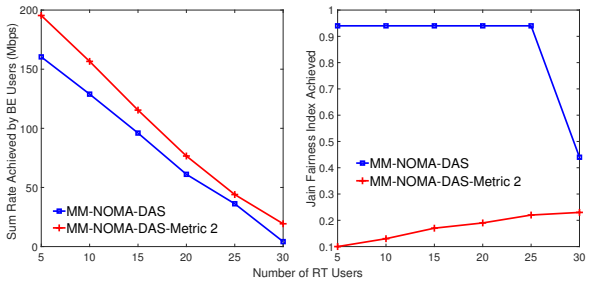


Fig. 8: Achieved Sum Rate of BE Users in terms of K_{RT} (left); Achieved Jain Fairness Index in terms of K_{RT} (right)

Fig. 8 compares the performance of both metrics for BE users. As expected, (22) achieves a higher sum rate for BE users than (10). This is due to the fact that in (22), only the achievable rate of BE users is taken into account, without any fairness consideration. This is not the case of (10), which seeks to achieve a tradeoff between rate and fairness maximization. The superior performance of (10) in terms of fairness is shown in Fig. 8 (right).

Fig. 7 and 8 show the different tradeoffs existing between the used metrics and (22). In the current paper, our goal is to formulate a matching theory based solution for the resource allocation problem in CAS, DAS, OMA and NOMA settings, without necessarily focusing on the optimal scheduling metric. However, to find new metrics reaping the best of these compared ones, a possible future work could consider the formulation, analysis and comparison between multiple metrics. Then, the new metrics can be readily plugged into our proposed algorithms.

On a final note, the performance of the proposed matching-based technique was compared with the optimal solution found by exhaustive search in the NOMA-DAS setting and denoted by ES-NOMA-DAS. For moderate values of the system parameters ($S = 2$, and 4 subbands, $K_{RT} = 2$ users and K_{BE} ranging from 2 to 6 users), the matching-based technique was able to achieve more than 90% of the performance of ES-NOMA-DAS, albeit with a much lower complexity. More precisely, MM-NOMA-DAS requires 2.78% and $5.36 \times 10^{-4}\%$ of the complexity of ES-NOMA-DAS when $K_{BE} = 6$ users, for 2 and 4 subbands, respectively. Moreover, Fig. 9 shows a comparison of the Pareto Frontier (the achieved data rate of BE users vs. the minimum percentage of satisfied RT users) between ES-NOMA-DAS and MM-NOMA-DAS, for $K_{RT} = K_{BE} = 4$ users and $S = 2$ subbands. It can be clearly seen that the total throughput achieved by BE users decreases when the minimum percentage of satisfied RT users increases. This is due to RT users

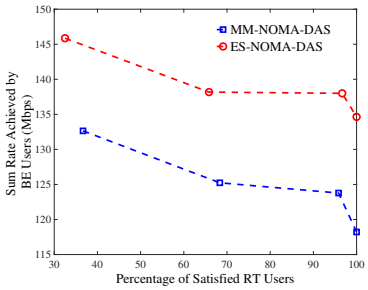


Fig. 9: Pareto Frontier comparison between MM-NOMA-DAS and ES-NOMA-DAS

requiring more resources for a higher percentage of satisfied RT users, leaving fewer resources for BE users. Fig. 9 also shows that the slope of decrease of the exhaustive search method and the matching technique is relatively the same. Moreover, the matching technique achieves 90% of the performance of ES-NOMA-DAS on average, with a much lower complexity. As a conclusion, depending on the system requirements regarding the satisfaction of RT users, a different performance of BE users can be achieved.

VIII. Conclusion

In this paper, subband allocation for mixed traffic types was studied for different system settings. For all settings, a new subband allocation method based on matching theory was proposed in such a way to take into account the particularities of each traffic type. Additionally, an iterative approach to determine the number of subbands per antenna for the DAS cases was introduced. The proposed techniques aim at maximizing the satisfaction of RT users while preserving a good performance for BE users. Simulation results showed that the proposed method based on matching theory outperforms the previously introduced greedy method by up to 63% in terms of RT users satisfaction when $K_{RT} = 30$ users. Moreover, the use of DAS in combination with MM was shown to increase satisfaction by more than 30%, compared to its CAS counterpart.

Appendix A
Proof of Theorem 1

Suppose that there exists a pair $(k_n, s_i) \notin \Psi^*$ such that (k_n, s_i) forms a blocking pair. According to phases 1 and 2 of Algorithm 1 (as well as Algorithm 3), k_n has already proposed to s_i and was rejected at a certain iteration q, meaning that $\Psi^q(s_i) \succ_{s_i} k_n$, where Ψ^q is the matching state at iteration q. Since $\Psi^*(s_i) \succ_{s_i} \Psi^q(s_i)$, s_i is matched to its final partner $\Psi^*(s_i)$ which it prefers to k_n. Hence, (k_n, s_i) does not form a blocking pair and the matching Ψ^* is stable. It should be noted that a stable matching is guaranteed to exist since the problem is modeled as a one-to-many matching game [29].

APPENDIX B
PROOF OF THEOREM 2

At the beginning of the algorithm, each active user k builds a preference list $\mathcal{PL}_k$ over the S subbands. Therefore, $\mathcal{PL}_k,\ \forall k \in \mathcal{K}$, has initially S elements, hence has a finite size. At each iteration of the algorithm, after subbands make decisions regarding accepted users (or user combinations in the NOMA case), each active user removes the subband it proposed to at the current iteration from its preference list. Hence, as the number of iterations increases, the preference lists of active users become smaller. When the maximum number of iterations is reached, the preference lists of active users become empty and the algorithm converges. Next, the maximum number of iterations needed by each setting is evaluated.

1) OMA-CAS: Each user can propose to, at most, S subbands, leading to a maximum number of iterations of S for OMA-CAS .
2) OMA-DAS: The DAS context involves duplicating each user A times. Hence, the system consists of $A \times K$ virtual users, each having preferences over the S subbands. In addition to that, during each iteration, only one of the duplicated users is allowed to propose to its favorite subband. Therefore, at each iteration, K entries of the preference lists are removed, which leads to the maximum number of iterations being upper bounded by $A \times S$.
3) NOMA-CAS: For NOMA, the matching technique is divided into two stages: assignment of subbands followed by user pairing on the subbands assigned to RT users. In the first stage, since we have S subbands in the system, a user (or user pair) can propose to at most S subbands, meaning that the maximum number of iterations before reaching convergence is also S. For the second stage, the maximum number of iterations is given by the number of subbands assigned to RT users, which is upper bounded by S also. Hence, the maximum number of iterations before the CAS version of Algorithm 3 converges is $2 \times S$.
4) NOMA-DAS: The maximum number of iterations in this case is an extension of the OMA-DAS and NOMA-CAS ones. In the first step of the allocation technique, $A \times S$ is the maximum number of iterations followed by a maximum of S iterations for the second part. Hence, an upper bound for the maximum number of iterations is $S \times (A + 1)$.

APPENDIX C
PROOF OF THEOREM 3

The complexity of the proposed matching technique in all system settings is dominated by two steps: 1. sorting the subbands to form the preference lists, 2. the matching step which involves making proposals and decisions. In Table III, the complexity of each system setting is evaluated. For comparison, the complexity of the optimal method based on exhaustive search is also given.

TABLE III: Complexity Analysis

	OMA-CAS	OMA-DAS	NOMA-CAS	NOMA-DAS
Sorting Complexity	$\mathcal{O}(KS^2)$	$\mathcal{O}(KAS^2)$	$\mathcal{O}((K_{RT}+\lvert\mathcal{US}\rvert)S^2)$	$\mathcal{O}((K_{RT}+\lvert\mathcal{US}\rvert)AS^2)$
Matching Complexity	$\mathcal{O}(KS)$	$\mathcal{O}(KAS)$	$\mathcal{O}((K_{RT}+\lvert\mathcal{US}\rvert)S)$	$\mathcal{O}((K_{RT}+\lvert\mathcal{US}\rvert)AS)$
Overall Complexity	$\mathcal{O}(KS^2)$	$\mathcal{O}(KAS^2)$	$\mathcal{O}((K_{RT}+\lvert\mathcal{US}\rvert)S^2)$	$\mathcal{O}((K_{RT}+\lvert\mathcal{US}\rvert)AS^2)$
Exhaustive Search Complexity	$\mathcal{O}(K^S)$	$\mathcal{O}((K\times A)^S)$	$\mathcal{O}\left((K+P(K,2))^S\right)$	$\mathcal{O}\left(((K+P(K,2))\times A)^S\right)$

APPENDIX D
PROOF OF THEOREM 4

The number of subbands and antennas of the considered system is limited. Hence, the number of potential assignments of subbands to antennas is finite. Furthermore, system performance is evaluated in terms of RT users satisfaction and BE users sum rate. It can be shown that system performance is enhanced after each iteration. Since system performance has an upper bound because of the limited spectral resources, the iterative approach terminates when this upper bound is reached. Therefore, the number of subbands per antenna can be found in a limited number of iterations.

REFERENCES

[1] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Seln, and J. Skld, "5G wireless access: requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42–47, Dec. 2014.

[2] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 66–71, Jan. 2009.

[3] W. Chung, C. Chang, and L. Wang, "An Intelligent Priority Resource Allocation Scheme for LTE-A Downlink Systems," *IEEE Wireless Commun. Lett.*, vol. 1, no. 3, pp. 241–244, June 2012.

[4] Y. Chung and C. Chang, "A balanced resource scheduling scheme with adaptive priority thresholds for OFDMA downlink systems," *IEEE Trans. Veh. Technol.*, vol. 61, no. 3, pp. 1276–1286, Mar. 2012.

[5] R. Balakrishnan and B. Canberk, "Traffic-aware QoS provisioning and admission control in OFDMA hybrid small cells," *IEEE Trans. Veh. Technol.*, vol. 63, no. 2, pp. 802–810, Feb. 2014.

[6] M. Pischella and J. Belfiore, "Resource allocation for QoS-aware OFDMA using distributed network coordination," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1766–1775, May 2009.

[7] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–7.

[8] M. J. Youssef, J. Farah, C. A. Nour, and C. Douillard, "Resource allocation for mixed traffic types in distributed antenna systems using NOMA," *Proc. IEEE Veh. Techn. Conf. Fall (VTC)*, Aug. 2018.

[9] A. Brighente and S. Tomasin, "Power allocation for non-orthogonal millimeter wave systems with mixed traffic," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 432–443, Jan. 2019.

[10] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Netw.*, vol. 31, no. 4, pp. 8–14, July 2017.

[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[12] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Annual Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Sept. 2013, pp. 611–615.

[13] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proc. IEEE Global Commun. Conf. Workshops*, Dec. 2013, pp. 66–70.

[14] M. Youssef, J. Farah, C. A. Nour, and C. Douillard, "Waterfilling-based resource allocation techniques in downlink non-orthogonal multiple access (NOMA) with single-user MIMO," in *IEEE Symp. on Comput. and Commun. (ISCC)*, July 2017, pp. 499–506.

[15] 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.

[16] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[17] L. Lei, D. Yuan, and P. Vrbrand, "On power minimization for non-orthogonal multiple access (NOMA)," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2458–2461, Dec. 2016.

[18] J. Farah, E. Sfeir, C. A. Nour, and C. Douillard, "New resource allocation techniques for base station power reduction in orthogonal and non-orthogonal multiplexing systems," in *Proc. Int. Conf. on Commun. Workshops*, May 2017, pp. 618–624.

[19] Q. Vien, N. Ogbonna, H. X. Nguyen, R. Trestian, and P. Shah, "Non-orthogonal multiple access for wireless downlink in cloud radio access networks," in *21th European Wireless Conf.*, May 2015, pp. 1–6.

[20] Q. Vien, T. A. Le, C. V. Phan, and M. O. Agyeman, "An energy-efficient NOMA for small cells in heterogeneous CRAN under QoS constraints," in *23th European Wireless Conf.*, May 2017, pp. 1–6.

[21] K. N. Pappi, P. D. Diamantoulakis, and G. K. Karagiannidis, "Distributed uplink-NOMA for cloud radio access networks," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2274–2277, Oct. 2017.

[22] X. Gu, X. Ji, Z. Ding, W. Wu, and M. Peng, "Outage probability analysis of non-orthogonal multiple access in cloud radio access networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 149–152, Jan. 2018.

[23] J. Farah, A. Kilzi, C. A. Nour, and C. Douillard, "Power minimization in distributed antenna systems using non-orthogonal multiple access and mutual successive interference cancellation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11 873–11 885, Dec. 2018.

[24] B. Zhang, X. Mao, J. Yu, and Z. Han, "Resource allocation for 5G heterogeneous cloud radio access networks with D2D communication: A matching and coalition approach," *IEEE Trans. Veh. Commun.*, vol. 67, no. 7, pp. 5883–5894, July 2018.

[25] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319–5332, Dec. 2017.

[26] B. Di, S. Bayat, L. Song, and Y. Li, "Radio resource allocation for downlink non-orthogonal multiple access (NOMA) networks using matching theory," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[27] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, "Joint subchannel and power allocation for NOMA enhanced D2D communications," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 5081–5094, Nov. 2017.

[28] Y. Li, M. Sheng, X. Wang, Y. Shi, and Y. Zhang, "Globally optimal antenna selection and power allocation for energy efficiency maximization in downlink distributed antenna systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 3856–3861.

[29] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[30] C. He, G. Y. Li, F. Zheng, and X. You, "Energy-efficient resource allocation in OFDM systems with distributed antennas," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1223–1231, Mar. 2014.

[31] Y. L. Lee, L. Wang, T. C. Chuah, and J. Loo, "Joint resource allocation and user association for heterogeneous cloud radio access networks," in *28th Int. Teletraffic Congress (ITC 28)*, vol. 01, Sept. 2016, pp. 87–93.

[32] J. G. Andrews, "Interference cancellation for cellular systems: a contemporary overview," *IEEE Wireless Commun.*, vol. 12, no. 2, pp. 19–29, Apr. 2005.

[33] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[34] 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.

[35] A. E. Roth and M. Sotomayor, "Two-sided matching, a study in game-theoretic modeling and analysis," *Econometric Society Monographs. Cambridge University Press*, 1990.

[36] T. Quint, "The core of an *m*-sided assignment game," *Games and Economic Behavior*, vol. 3, 1990.

[37] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[38] M. Hojeij, J. Farah, C. A. Nour, and C. Douillard, "Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access," in *Proc. IEEE Veh. Techn. Conf. Spring (VTC)*, May 2015, pp. 1–6.

[39] 3GPP, "TR25-814 (V7.1.0), Physical Layer Aspects for Evolved Universal Terrestrial Radio Access (UTRA)," 2006.

[40] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *DEC Technical Report 301*, Sept. 1984.