Resource Allocation for Mixed Traffic Types in Distributed Antenna Systems Using NOMA

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

Resource Allocation for Mixed Traffic Types in Distributed Antenna Systems Using NOMA

Marie-Josepha Youssef[(1)], Joumana Farah[(2)], Charbel Abdel Nour[(1)], Catherine Douillard[(1)]
[(1)]IMT Atlantique, Department of Electronics, Lab-STICC - UMR 6285
Technopôle Brest Iroise, CS 83 818 - 29238 Brest Cedex, France
[(2)]Department of Electricity and Electronics, Faculty of Engineering,
Lebanese University, Roumieh, Lebanon

***Abstract*—This paper proposes a new traffic-aware resource allocation technique, employing non-orthogonal multiple access (NOMA) in a downlink distributed antenna system (DAS). The studied framework consists of users with mixed traffic types: real-time (RT) users having strict QoS requirements (in terms of amount of data and latency), and best-effort (BE) users aiming to maximize their throughput and fairness. After formulating the resource and power optimization problem, we propose a low complexity sub-optimal algorithm that aims at guaranteeing the requirements of RT users while maximizing the utility function of BE users. Simulation results show a remarkable performance enhancement of the proposed algorithm over baseline techniques in terms of RT users satisfaction. Also, the proposed technique achieves near optimal fairness for BE users while maximizing their average throughput.**

***Index Terms*—Mixed traffic types, latency, resource allocation, NOMA, DAS.**

I. Introduction

It is widely agreed that 5G systems are not going to be just a mere evolution of 4G networks by increasing the amount of available spectral bands and providing a higher spectral efficiency. In fact, 5G systems are expected to provide support to an increasingly growing number of diverse applications, while ensuring connectivity to a massive number of devices. In addition to traditional bandwidth-hungry applications (e.g. web browsing), some of the envisioned applications need relatively low throughput but demand strict latency and reliability (e.g. e-health, automated control, and autonomous vehicles), while others require both high throughput and low latency (e.g. video conferencing, augmented and virtual reality). As a result, mobile traffic is evolving into a more heterogeneous or mixed model. To cope with these requirements, 5G networks must benefit from new technologies and employ flexible and highly adaptable architectures.

Resource allocation for mixed traffic has been previously investigated in the literature. Using utility theory, a heuristic algorithm based on Lagrange multipliers is proposed in [1] in order to maximize the utility of a system consisting of RT and BE users. In [2], network coordination is employed to enhance the performance of RT users, and the amount of resources needed by RT users is minimized in order to increase the amount of resources available for BE users. This minimization is also the target of [3] where a scalable transmission time interval (TTI) is adapted to the needs of the users, expressed by a requested number of bits and a latency requirement.

All of the aforementioned works employ orthogonal multiple access (OMA) for resource allocation. However, allocating a resource to a RT user running an application with low throughput and stringent latency constraints leads into a sub-optimal bandwidth distribution among users, penalizing system spectral efficiency as also noted by [4].

Recently, NOMA emerged as a promising radio access technology. In contrast to OMA, it enables the cohabitation of multiple users on the same resource by multiplexing them in the power domain [5], [6]. At the receiver side, successive interference cancellation (SIC) is performed to retrieve superimposed signals. NOMA constitutes a more flexible radio access scheme than OMA; its appeal lies in its potential to serve a larger number of simultaneous users, increasing system throughput and fairness while reducing access latency.

To further improve system performance, NOMA can be used with the concept of DAS and their evolution to cloud radio access networks (C-RAN) [7], [8], recently introduced as promising network architectures. By using multiple remote radio heads (RRH) coordinated by a central controller, these technologies enable higher capacities and increased coverage. Some of the works in this context use NOMA in the transmission from the central controller to the RRHs as in [7]. Others use NOMA in the transmission from the RRHs to the users as in [8] which derives a closed-form expression for the outage probability of a two-user system.

In this paper, we consider the resource allocation problem in a system consisting of users with mixed traffic. Our goal is to come up with a solution that guarantees the requirements of RT users while maximizing both the data rates and fairness of BE users. To guarantee an efficient use of the spectrum, we propose to serve users by employing NOMA in the context of DAS. By doing so, whenever a RT user is allocated a resource exceeding its needs, the resource can be shared with another user (either RT or BE) to enhance system performance. Note that this work can be directly used in the context of C-RAN and heterogeneous networks. To the best of our knowledge, this is the first study that considers resource allocation for mixed traffic in a NOMA-DAS context.

II. System Model

Consider a downlink NOMA-DAS where R RRHs, uniformly deployed over the cell, serve K users. All transceivers are equipped with single antennas. The total bandwidth B is

partitioned into a set $\mathcal{S}$ of S subbands, leading to $B_c = B/S$ as the bandwidth per subband. Each subband s can be allocated to one transmit antenna within each timeslot to limit intra-cell interference. According to the NOMA principle [5], the messages of up to N_s users are superposed and transmitted over subband s during each transmission period. Consequently, the received signal by each user k suffers from the interference caused by the messages of users $k' \in \mathcal{S}_s \setminus \{k\}$, where $\mathcal{S}_s$ is the set of active users on s. Therefore, the receiver of user k generally applies SIC before demodulating its own signal, resulting in the following achieved rate:

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

In (1), H^t_{k,s,r_s} is the channel coefficient at timeslot t between user k and RRH r_s, the antenna to which s is assigned, applying both a small scale Rayleigh fading and a large scale fading (path-loss and log-normal shadowing). p^t_{k,s,r_s} is the power allocated to user k on subband s and N_0 is the noise power spectral density. The first term in the denominator reflects the interference experienced by user k from users in $\mathcal{I}_{s,k} = \{(k' \in \mathcal{S}_s \setminus \{k\}) \cap \left(H^t_{k',s,r_s} > H^t_{k,s,r_s}\right)\}$; i.e. users scheduled on subband s and having a higher channel gain than k on s.

Note that SIC results in a significant complexity increase at the receiver side; thus, the maximum number of multiplexed users on each subband is limited to 2, i.e. $N(s) = 2,\ \forall s \in \mathcal{S}$.

The studied system consists of heterogeneous users: among the K users, K_{BE} are classified as BE users and K_{RT} as RT users (having the highest priority). Each RT user k_{RT} has a strict latency requirement $L_{k_{RT}}$, defined as an integer number of timeslots, each with duration τ (ms). The measure of satisfaction for user k_{RT} depends upon receiving its requested amount of data bits $D^{\text{req}}_{k_{RT}}$ prior to the latency limit. For BE users, the goal is to maximize the received data rates while preserving system fairness.

III. Problem Formulation

Guaranteeing the satisfaction of RT users while maximizing the performance of BE users depends on an efficient resource allocation, in terms of subband and RRH assignment, user pairing, and power allocation (PA). In the following, we introduce the performance measures for RT and BE users before formulating the optimization problem.

A. Performance Measure of RT Users

We propose to take advantage from the fact that the satisfaction of RT users does not depend on the specific timeslots in which they receive the requested data as long as these slots precede $L_{k_{RT}}, \forall\ k_{RT} \in \mathcal{K}_{RT}$. Benefiting from this fact, each RT user will be given enough resources in order to reach $D^{\text{req}}_{k_{RT}}$ by the time $t = L_{k_{RT}}$. Therefore, from the start of the scheduling period till the end of timeslot t, k_{RT} needs to be allocated a number of bits $D^{\text{req},t}_{k_{RT}}$ given by:

$$D^{\text{req},t}_{k_{RT}} = t D^{\text{req}}_{k_{RT}} / L_{k_{RT}},\ \forall k_{RT} \in \mathcal{K}_{RT}. \tag{2}$$

Let $\mathbb{1}^t_{k_{RT}}(\boldsymbol{a}, \boldsymbol{p})$ be a measure of the satisfaction of RT user k_{RT} at each timeslot t, defined by:

$$\mathbb{1}^t_{k_{RT}}(\boldsymbol{a^t}, \boldsymbol{p^t}) = \begin{cases} 1 \text{ if } \sum\limits_{t'=1}^{t} \sum\limits_{r=1}^{R} \sum\limits_{s=1}^{S} R^{t'}_{k_{RT},s,r} a^{t'}_{k_{RT},s,r} \tau \geq D^{\text{req},t}_{k_{RT}}, \\ 0 \text{ otherwise.} \end{cases} \tag{3}$$

where $a^t_{k,s,r} = 1$ if k is scheduled on s, when the latter is assigned to r, and 0 otherwise. $\boldsymbol{p^t}$ is the PA vector. At each timeslot, we aim to maximize the following metric:

$$\sum_{k_{RT}=1}^{K_{RT}} \mathbb{1}^t_{k_{RT}}(\boldsymbol{a^t}, \boldsymbol{p^t}) \tag{4}$$

B. Performance Measure of BE Users

For BE users, the system utility, reflecting rate and fairness maximization, is defined by:

$$\sum_{k_{BE}=1}^{K_{BE}} \sum_{r=1}^{R} \sum_{s=1}^{S} R^t_{k_{BE},s,r} a^t_{k_{BE},s,r} f\left(T^t_{k_{BE}}, k_{BE} = 1, \ldots, K_{BE}\right), \tag{5}$$

where f is a measure of the fairness between BE users that depends on their average rates $T^t_{k_{BE}}$ until timeslot t. At the beginning of each timeslot, $T^t_{k_{BE}}$ is updated according to:

$$T^t_{k_{BE}} = (1 - \frac{1}{t_c}) T^{t-1}_{k_{BE}} + \frac{1}{t_c} R^{t-1}_{k_{BE}}. \tag{6}$$

In (6), t_c is the averaging window and $R^{t-1}_{k_{BE}}$ is the total rate of k_{BE} during timeslot $(t-1)$ given by:

$$R^{t-1}_{k_{BE}} = \sum_{r=1}^{R} \sum_{s=1}^{S} R^{t-1}_{k_{BE},s,r} a^{t-1}_{k_{BE},s,r}. \tag{7}$$

The PF scheduler [9] is known to achieve a tradeoff between total throughput and fairness maximization, by scheduling on each subband s the user (or user set in the case of NOMA) that maximizes the PF metric. However, the PF scheduler achieves fairness through the consideration of historical rates up to the last complete allocation slot; i.e. it does not take into consideration the rates achieved during the current slot. That is why we adopt the enhanced PF scheduler from [9] and schedule on each subband s the user set $\mathcal{U}^*$ that satisfies:

$$\mathcal{U}^* = \underset{\mathcal{U}}{\text{argmax}} \sum_{k_{BE} \in \mathcal{U}} \frac{R^t_{k_{BE},s,r^*_s}}{(1 - \frac{1}{t_c}) T^t_{k_{BE}} + \frac{1}{t_c} \sum_{i=1}^{s-1} R^t_{k_{BE},i,r^*_i}}. \tag{8}$$

In (8), RRH r^*_s is chosen so as to maximize the rate of users in $\mathcal{U}$, i.e. $r^*_s = \text{argmax}_{r \in \mathcal{R}} \sum_{k_{BE} \in \mathcal{U}} R^t_{k_{BE}s,r}$; whereas the second denominator term accounts for the rate achieved by users in $\mathcal{U}$ during the current timeslot (if any) before considering s for allocation, promoting better fairness. Note that (8) is computed after considering all $N_U = K_{BE} + P(K_{BE}, 2)$ possible user sets, where K_{BE} accounts for OMA signaling whereas $P(K_{BE}, 2)$ accounts for NOMA signaling.

For concision purposes, when there is no confusion, we will drop the time index t in the following.

C. Optimization Problem

The following optimization problem must be solved at each time slot:

$$\max_{\boldsymbol{a},\boldsymbol{p}} \quad (4),(5) \tag{9}$$

$$\text{such that} \sum_{r\in\mathcal{R}} a_{k,s,r} \leq 1,\ \forall (k,s) \in \mathcal{K}\times\mathcal{S} \tag{9a}$$

$$\sum_{k\in\mathcal{K}} a_{k,s,r} \leq 2,\ \forall (s,r) \in \mathcal{S}\times\mathcal{R} \tag{9b}$$

$$\sum_{k\in\mathcal{K}}\sum_{s\in\mathcal{S}} P_{s,r} a_{k,s,r} \leq P,\ \forall r \in \mathcal{R} \tag{9c}$$

$$p_{k_1,(s,r)} < p_{k_2,(s,r)}\ \forall (s,r), \in \mathcal{S}\times\mathcal{R} \tag{9d}$$

$$p_{k_1,(s,r)} + p_{k_2,(s,r)} = P_{s,r},\ \forall (s,r) \in \mathcal{S}\times\mathcal{R} \tag{9e}$$

$$a_{k,s,r} \in \{0,1\}. \tag{9f}$$

Constraint (9a) restricts each subband to be assigned to one antenna, while (9b) restricts each subband to be shared by at most 2 users. (9c) reflects the power constraint per antenna. Denoting by k_1 and k_2 the users multiplexed on (s,r) s.t. $H_{k_1,s,r} > H_{k_2,s,r}$, (9d) is necessary to guarantee SIC stability [10], while (9e) denotes the power sharing constraint on each subband. (9) is a mixed-integer multi-objective optimization problem; an optimal solution is computationally intractable. In the following, a sub-optimal algorithm is proposed to provide an alternative solution.

IV. Description of The Proposed Resource Allocation Technique

Since the system consists of users with heterogeneous traffic, the allocation technique must take into account the difference in priority between users. Therefore, whenever RT users are to be scheduled in the current slot, these users are assigned to subbands and antennas in an OMA manner. Once this OMA phase is done, we proceed with the user pairing phase. When no RT users are to be scheduled, BE users are scheduled following the PF principle. In the following, the different steps of the allocation procedure are described.

A. Phase 1: Assignment of Users and Subbands to Antennas

The goal of this first step in the allocation technique is to assign users and subbands to the antennas in such a way to guarantee the requirements of RT users while maximizing the utility of BE users.

Assigning subbands to RRHs is not a straightforward task. Indeed, basing the assignment solely on observed channel gain values may lead to an antenna r being assigned a large number of subbands. In that case, the power per subband on r decreases, while a better performance would have been achieved by assigning the subband to a less congested antenna. Therefore, we start by following the proposal in [11] that relies on large-scale fading $w_{k,r}$ between user k and antenna r to estimate N_r, the potential number of subbands assigned to r:

$$N_r = \left\lfloor \frac{S \times \sum_{k=1}^{K} w_{k,r}}{\sum_{r=1}^{R}\sum_{k=1}^{K} w_{k,r}} \right\rfloor, r \in \mathcal{R}. \tag{10}$$

The floor operation $\lfloor . \rfloor$ in (10) results in $(S - \sum_r N_r)$ unallocated subbands; this number will be provisionally given to the antennas having the smallest N_r. To get an initial approximation of the potential power per subband on r, we assume equal inter-subband power repartition leading to the following power per subband on r: $P_{s,r} = P_r = P/N_r$. However, contrary to [11], in our work, N_r is strictly used for initial power approximation and does not dictate the number of assigned subbands to each antenna r in each timeslot.

Algorithm 1 Assignment of subbands, single RT users and NOMA BE users to antennas

while $\mathcal{U}_{active} \neq \emptyset$ & $\mathcal{S} \neq \emptyset$ **do** // Assignment of RT users
  Update priorities for users in $\mathcal{U}_{active}$ and remove users having negative priorities from the active set
  $\mathcal{M} \leftarrow$ users having the highest priority
  while $\mathcal{M} \neq \emptyset$ **do**
    $(s_k, r_k) = \underset{s,r}{\text{argmax}} \left(P_r H^2_{\mathcal{M}(k),s,r}\right),\ \forall k \in |\mathcal{M}|$
    Schedule $k^* = \underset{k}{\text{argmin}} \left(P_{r_k} H^2_{\mathcal{M}(k),s_k,r_k}\right)$ on s_{k*}
    Update $d^t_{k^*}$. $\mathcal{S}_{r^*_k} = \mathcal{S}_{r^*_k} \cup \{s_{k^*}\}$ and $\mathcal{S} = \mathcal{S} \setminus \{s_{k^*}\}$ //$\mathcal{S}_{r^*_k}$ is the set of subbands assigned to r^*_k
    if $|\mathcal{S}_{r^*_k}| > N_{r_{k^*}}$ **then**
      Update $P_{r^*_k}$, rate, power and $d^t_{k_{RT}}$ for all users scheduled on r_{k^*}
    end if
    $\mathcal{M} = \mathcal{M} \setminus \{k^*\}$
  end while
end while
if $\mathcal{S} \neq \emptyset$ **then** //Assignment of BE users if all RT users are satisfied
  for $i = 1 : |\mathcal{S}|$ **do**
    $s = \mathcal{S}(i)$
    Remove from $\mathcal{R}$ the antennas which, if assigned another subband, would cause some RT users to become unsatisfied
    for $i = 1 : N_U$ **do**
      Divide $P_r,\ \forall r \in \mathcal{R}$ between users in $\mathcal{U}_i$ using FTPA
      Find $PF_{\mathcal{U}_i}$ on $r^*_{\mathcal{U}_i} = \underset{r}{\text{argmax}} \sum_{k_{BE}\in\mathcal{U}_i} R_{k_{BE},s,r}$ using (8)
    end for
    Schedule $\mathcal{U}^*_i = \text{argmax}_{\mathcal{U}_i} \text{PF}_{\mathcal{U}_i}$ on s
    if $|\mathcal{S}_{r_{\mathcal{U}^*_i}}| > N_{r_{\mathcal{U}^*_i}}$ **then**
      Update $P_{r_{\mathcal{U}^*_i}}$, rate, power and $d^t_{k_{RT}}$ for all users scheduled on $r_{\mathcal{U}^*_i}$
    end if
  end for
end if

With these estimations of power per subband on each antenna, we proceed with the resource allocation for RT users giving them the benefit of being allocated subbands with best perceived channel coefficient. Since RT users may have different data and latency requirements, we introduce a priority function in order to favor users with stringent requirements. Such a function should be increasing with $D^{\text{req},t}_{k_{RT}}$ and decreasing with $L_{k_{RT}}$ and with the number of bits so far transferred. Therefore, we propose the following priority measure:

$$\text{Priority}(k_{RT}, t) = \frac{\left(D^{\text{req},t}_{k_{RT}} - \left(D^{\text{ach},t-1}_{k_{RT}} + d^t_{k_{RT}}\right)\right)/D^{\text{req},t}_{k_{RT}}}{(L_{k_{RT}} - t)/L_{k_{RT}}}, \tag{11}$$

where $D^{\text{ach},t-1}_{k_{RT}}$ is the total number of bits received by k_{RT} until the end of timeslot $(t-1)$ and $d^t_{k_{RT}}$ is the number of

bits so far transferred during timeslot t. Note that the total number of transferred bits at the end of the second phase of the allocation technique (NOMA pairing), at timeslot t, is given by: $D^{\text{ach},t}_{k_{RT}} = D^{\text{ach},t-1}_{k_{RT}} + d^t_{k_{RT}}$.

Algorithm 1 describes the first phase of the allocation process applied while there are RT users in the system. After updating the priorities of all RT users in the active set $\mathcal{U}_{active}$, the best resource for users having the highest priority is found. This decision is based on the criterion $(P_r H^2_{k_{RT},s,r})$ rather than just $(H_{k_{RT},s,r})$ in order to account for different power levels on different antennas. When multiple users have the same highest priority (users in $\mathcal{M}$), the one having the worst best perceived channel coefficient on a particular subband/antenna couple is prioritized and scheduled on its preferred subband.

Steps 2 till 12 are repeated until $\mathcal{U}_{active} = \emptyset$ or no free subbands remain. If the former happens, BE users are scheduled on the remaining subbands, using NOMA, in a way that does not penalize the achieved level of satisfaction of RT users. BE users are scheduled according to the PF principle as shown between steps 14 and 27. In this work, fractional transmit power allocation (FTPA) [5] is used to partition the power on subbands assigned to BE users leading to the following power for user k_{BE} on s:

$$p_{k_{BE},s,r_s} = \frac{P_{r_s}\left(\frac{H^2_{k_{BE},s,r_s}}{N_0 B_c}\right)^{-\alpha_{FTPA}}}{\sum\limits_{k'\in\mathcal{K},a_{k',s,r_s}=1}\left(\frac{H^2_{k',s,r_s}}{N_0 B_c}\right)^{-\alpha_{FTPA}}}, \qquad (12)$$

where α_{FTPA} is a decay factor. (12) ensures a higher power for the user with a lower channel gain, guaranteeing SIC stability. After dividing the power between users in each candidate set $\mathcal{U}_i$, the antenna maximizing the sum rate of users in $\mathcal{U}_i$ is chosen to calculate the PF metric as in step 20.

Algorithm 2 NOMA pairing on subbands assigned to RT users

1: Solve (14) for satisfied RT users, add unsatisfied users to $\mathcal{U}_{active}$
2: **while** $\mathcal{U}_{active} \neq \emptyset$ & $\mathcal{S}_{RT} \neq \emptyset$ **do**
3: $R^{\text{lack},t}_k = R^{\text{req},t}_k - \sum_{s\in\mathcal{S}_k} R_{k,s,r_s},\ \forall k \in \mathcal{U}_{active}$
4: Compute priorities for users in $\mathcal{U}_{active}$ and remove those having negative priorities.
5: $\mathcal{M} \leftarrow$ Users with highest priorities
6: **while** $\mathcal{M} \neq \emptyset$ **do**
7: **for** $k = 1 : |\mathcal{M}|$ **do**
8: **for** $i = 1 : |\mathcal{S}_{RT}|$ **do**
9: Find k_s, the RT user scheduled on $s = \mathcal{S}_{RT}(i)$
10: **if** $H(k_s, s, r_s) > H(\mathcal{M}(k), s, r_s)$ **then**
11: Calculate $R^{\text{temp}}_{\mathcal{M}(k),s,r_s}$ with the power calculated using (18)
12: **else**
13: Calculate $R^{\text{temp}}_{\mathcal{M}(k),s,r_s}$ with the power calculated using (20)
14: **end if**
15: **end for**
16: $\text{dist}_{\mathcal{M}(k),s} = \left(R^{\text{temp}}_{\mathcal{M}(k),s,r_s} - R^{\text{lack},t}_{\mathcal{M}(k)}\right),\ s \in \mathcal{S}_{RT}$
17: **if** $\exists\ s \in \mathcal{S}_{RT}$ s.t. $\text{dist}_{\mathcal{M}(k),s} > 0$ **then**
18: $s^*_{\mathcal{M}(k)} = \underset{s\in\mathcal{S}_{RT}\ \text{s.t. dist}_{\mathcal{M}(k),s}>0}{\text{argmin}} \left(\text{dist}_{\mathcal{M}(k),s}\right)$
19: **else**
20: $s^*_{\mathcal{M}(k)} = \underset{s\in\mathcal{S}_{RT}}{\text{argmax}} \left(\text{dist}_{\mathcal{M}(k),s}\right)$
21: **end if**
22: **end for**
23: Schedule $\mathcal{M}(k^*) = \underset{\mathcal{M}(k)}{\text{argmin}} \left(\text{dist}_{\mathcal{M}(k),s^*_{\mathcal{M}(k)}}\right)$ on $s^*_{\mathcal{M}(k^*)}$
24: $\mathcal{S}_{RT} = \mathcal{S}_{RT} \setminus \left(s^*_{\mathcal{M}(k^*)}\right)$
25: **end while**
26: **end while**
27: **if** $\mathcal{S}_{RT} \neq \emptyset$ **then** //Assignment of BE users
28: **for** $i = 1 : |\mathcal{S}_{RT}|$ **do**
29: Find RT user scheduled on $s = \mathcal{S}_{RT}(i)$
30: Find the available power from (18) or (20) and calculate $\text{PF}_{k_{BE}}$ according to (8), with $\mathcal{U} = \{k_{BE}\}, \forall\ k_{BE} \in \mathcal{K}_{BE}$
31: Schedule $k^*_{BE} = \text{argmax}_{k_{BE}}\ \text{PF}_{k_{BE}}$ on s
32: **end for**
33: **end if**

B. Phase 2: NOMA pairing on subbands assigned to RT users

Assuming equal inter-subband power on each antenna may lead to some RT users having more power than needed for achieving their target throughput. Also, some RT users may not be allocated enough resources, either because the system is congested, or because of their bad channel conditions. That is why we perform a NOMA pairing phase in which we assign second users to subbands assigned to RT users, while keeping the subband-antenna assignment unvaried and guaranteeing the required rates to RT users already scheduled.

First, we need to determine unsatisfied RT users and find the amount of excess power allocated to satisfied RT users. For this purpose, we start by estimating the amount of required throughput in timeslot t, for each RT user, according to:

$$R^{\text{req},t}_{k_{RT}} = \frac{D^{\text{req},t}_{k_{RT}} - D^{\text{ach},t-1}_{k_{RT}}}{\tau}, \quad \forall k_{RT} \in \mathcal{K}_{RT}. \qquad (13)$$

For each user k_{RT} exceeding its required rate, we recalculate the amount of power needed to reach $R^{\text{req},t}_{k_{RT}}$ on its assigned set of subbands, $\mathcal{S}_{k_{RT}}$, by solving the following power minimization problem:

$$\min_{\boldsymbol{p}_{k_{RT}}} \sum_{s\in\mathcal{S}_{k_{RT}}} p_{k_{RT},s,r_s} \qquad (14)$$

$$\text{such that} \sum_{s\in\mathcal{S}_{k_{RT}}} R_{k_{RT},s,r_s} = R^{\text{req},t}_{k_{RT}} \qquad (14a)$$

$$0 \leq p_{k_{RT},s,r_s} \leq P_{r_s}. \qquad (14b)$$

(14b) is imposed to enforce that no subband is allocated a power larger than the one it was initially allocated.

Solving the above optimization problem leads to the well-known waterfilling solution where p_{k_{RT},s,r_s} is given by:

$$p_{k_{RT},s,r_s} = \left[\frac{\lambda_{k_{RT}} B_c}{\log(2)} - \frac{N_0 B_c}{H^2_{k_{RT},s,r_s}}\right]^{P_{r_s}}_0. \qquad (15)$$

In (15), $\lambda_{k_{RT}}$ is the Lagrange multiplier given by:

$$\lambda_{k_{RT}} = 2^{\frac{1}{S_{k_{RT}}}\left[\frac{R^{\text{req},t}_{k_{RT}}}{B_c} - \sum\limits_{s\in\mathcal{S}_{k_{RT}}} \log_2\left(\frac{H^2_{k_{RT},s,r_s}}{\log(2)N_0}\right)\right]}. \qquad (16)$$

The required rate on $s \in \mathcal{S}_{k_{RT}}, R^{\text{req,t}}_{k_{RT},s,r_s}$, can be found by replacing (15) into (1). This rate will be guaranteed for all satisfied RT users during the pairing step.

Since BE users are scheduled directly using NOMA in Algorithm 1 (if scheduled), only subbands assigned to RT users, $\mathcal{S}_{RT}$, are considered in the second phase of the allocation technique. For that purpose, we start by checking if the achieved rate of each scheduled RT user k_{RT} in Algorithm 1 over its assigned subband s exceeds the required one. If it doesn't, s is not available for NOMA pairing and k_{RT} will remain its sole occupier. In the opposite case, k_{RT} can share s and two scenarios are considered for each candidate user k' ($k' \in \mathcal{K}$) for pairing in order to guarantee $R^{\text{req,t}}_{k_{RT},s,r_s}$.

1) $H_{k_{RT},s,r_s} > H_{k',s,r_s}$: In this case, k_{RT} is paired as first user in NOMA on s and its required rate is given by:

$$R^{\text{req},t}_{k_{RT},s,r_s} = B_c \log_2 \left(1 + \frac{p^{1,\text{req}}_{k_{RT},s,r_s} H^2_{k_{RT},s,r_s}}{N_0 B_c}\right), \quad (17)$$

where the necessary power $p^{1,\text{req}}_{k_{RT},s,r_s}$ is given by (15). To guarantee SIC stability, $p^{1,\text{req}}_{k_{RT},s,r_s}$ must be less than the power allocated to the second user, k', multiplexed on s. This translates into considering s for NOMA pairing with k_{RT} as first user if and only if $p^{1,\text{req}}_{k_{RT},s,r_s} < P_{r_s}/2$. If this condition is verified, the power available for k' on s is given by:

$$p^{2,\text{av}}_{s,r_s} = P_{r_s} - p^{1,\text{req}}_{k_{RT},s,r_s}. \quad (18)$$

2) $H_{k_{RT},s,r_s} < H_{k',s,r_s}$: In this case, k_{RT} is paired as second user in NOMA with a required rate given by:

$$R^{\text{req},t}_{k_{RT},s,r_s} = B_c \log_2 \left(1 + \frac{p^{2,\text{req}}_{k_{RT},s,r_s} H^2_{k_{RT},s,r_s}}{p^{1,\text{av}}_{s,r_s} H^2_{k_{RT},s,r_s} + N_0 B_c}\right), \quad (19)$$

where $p^{1,\text{av}}_{s,r_s}$is the available power on s for a first user, k':

$$p^{1,\text{av}}_{s,r_s} = P_{r_s} - p^{2,\text{req}}_{k_{RT},s,r_s}. \quad (20)$$

After substituting (20) in (19), $p^{2,\text{req}}_{k_{RT},s,r_s}$ will be given by:

$$p^{2,\text{req}}_{k_{RT},s,r_s} = \frac{(a-1)\left(P_{r_s} H^2_{k_{RT},s,r_s} + N_0 B_c\right)}{a H^2_{k_{RT},s,r_s}}, \quad (21)$$

where $a = 2^{R^{\text{req},t}_{k_{RT},s}/B_c}$. SIC stability is guaranteed if $p^{2,\text{req}}_{k_{RT},s,r_s} > P_{r_s}/2$, which results in ensuring:

$$P_{r_s} H^2_{k_{RT},s,r_s} (0.5a - 1) + (a-1) N_0 B_c > 0. \quad (22)$$

(22) is guaranteed if $a \geq 2$. In other words, if being scheduled as a second user, the required rate of k_{RT} is chosen to be:

$$R^{\text{req},t}_{k_{RT},s,r_s} = \max\left(R^{\text{req},t}_{k_{RT},s,r_s}, B_c\right). \quad (23)$$

Note that (23) does not result in any degradation regarding the rate of k_{RT}.

In Algorithm 2, we describe the steps followed for NOMA pairing on subbands assigned to RT users. First, the amount of excess power on each subband and the set of unsatisfied RT users, $\mathcal{U}_{active}$, are found. If $\mathcal{U}_{active} \neq \emptyset$, we start by scheduling each user in $\mathcal{U}_{active}$ on the subbands that provide a total rate closest to its requested rate. To do that, we compute the rate each user $k \in \mathcal{U}_{active}$ lacks and the priorities of active users. For users having the same highest priority (users in $\mathcal{M}$), we compute for each user the achievable rate on available subbands $R^{\text{temp}}_{\mathcal{M}(k),s,r_s}, \forall s \in \mathcal{S}_{RT}$ as done between steps 8 and 15. Then, for each user in $\mathcal{M}$, we find the subband that minimizes the distance to $R^{\text{lack,t}}_{\mathcal{M}(k)}$ as shown between steps 16 and 21. We then prioritize the user having the smallest distance. When there is at least one user having a negative distance, such a user is prioritized and given its best subband. Also, when all users have positive distances, the metric allows us to satisfy the one for whom the distance is the smallest, i.e. it promotes the efficient use of the available spectrum.

Note that in step 30, we only need to calculate the PF metric with the considered BE user in the candidate set $\mathcal{U}$ because the rate of the RT user is constant; hence it can be dropped.

C. Global Resource Allocation Technique

Our allocation technique is summarized in Algorithm 3.

Algorithm 3 Proposed Allocation Technique

1: **if** there are RT users to be scheduled **then**
2: Find the assignment of subbands to the antennas as well as the assignment of single RT users and NOMA BE users using Algorithm 1
3: Perform NOMA pairing on subbands assigned to RT users using Algorithm 2
4: **else** // The system consists of BE users only
5: **for** $s = 1 : |\mathcal{S}|$ **do**
6: Schedule BE users using the PF scheduler as in steps 18 to 22 of Algorithm 1
7: **end for**
8: **end if**

As can be seen from Algorithm 3, when the system consists of BE users only (either because RT users have been satisfied in former timeslots, have reached their latency limit, or because the system does not support RT applications), these users are scheduled according to the PF algorithm as stated in step 6 of Algorithm 3.

V. Numerical Results

We consider a single cell having a radius of $R_d = 500$m with $R = 4$ RRHs. One antenna is located at the cell center while the others are equally distanced and positioned on a circle of radius $2R_d/3$ with an angular separation of $120°$. Each RRH has a power budget $P = 10$ W. The system bandwidth $B = 10$ MHz is divided into $S = 32$ subbands. Signals undergo frequency-selective Rayleigh fading with a root mean square delay spread of 500 ns and a distance-dependent path loss with a decay factor of 3.76. The noise power spectral density is $N_0 = 4.10^{-18}$ mW/Hz and the decay factor for FTPA is $\alpha = 0.5$. Unless otherwise stated, our system consists of $K_{RT} = 20$ RT users and $K_{BE} = 20$ BE users. In this study, we assume perfect channel estimation.

The performance of our allocation technique, denoted as Prop-NOMA-DAS, is compared with the enhanced PF scheduler [9] denoted by PF-DAS. The latter does not take into account the presence of RT users in the allocation process.

However, for fair comparison, when a RT user reaches its latency limit or its requested number of bits, it is removed from the set of users to be scheduled. Different versions of the proposed technique are also considered: an OMA version denoted as Prop-OMA-DAS, a CAS version denoted as Prop-OMA-CAS and a version that restricts the number of subbands per antenna as done in [11] denoted as Prop-NOMA-DAS-F. In the latter, the estimated number of subbands per antenna is not updated throughout the allocation process.

A. Evaluation of the Performance of RT Users

To reflect the requirements of different services, RT users are partitioned into 3 classes (C1, C2 and C3). While all users request 10^5 bits, a user in C1 has a latency limit of 6 ms, while a user in C2 (resp. C3) has a latency limit of 10 ms (resp. 15 ms). Three scenarios are simulated: the first one, S1, being the least strict, consists of 5 RT users belonging to C3; the second scenario, S2 consists of 2 RT users in C1, and 9 in each of C2 and C3, while in S3, we have 5 RT users in each of C1 and C2 and 10 users in C3. In all three scenarios, the number of BE users is 20.

In Fig. 1, we plot the percentage of satisfied RT users in the considered scenarios. As expected, S1 has the largest percentage of satisfied users. While our technique (in all its versions) guarantees the satisfaction of all RT users, PF-DAS can reach satisfaction in only 78% of the cases. As the simulated conditions become harder, the performance of PF-DAS rapidly degrades reaching less than 10% in both S2 and S3. Prop-NOMA-DAS has the best performance among the different versions of our proposed technique. For example, Prop-NOMA-DAS guarantees the satisfaction of RT users in 97% of the cases in S2, i.e. 45% more than Prop-NOMA-CAS and Prop-NOMA-DAS-F. In comparison with Prop-OMA-DAS, Prop-NOMA-DAS achieves 13% higher satisfaction in S2. In fact, in S2, some RT users achieve higher rates than needed in the OMA step and can be paired with other RT users in the NOMA step which increases the satisfaction percentage for NOMA. This is not the case in S3 where requirements are more strict and the number of RT users that cannot share their subbands is higher than in other scenarios, leading to a close performance for OMA and NOMA.

For unsatisfied users in S2 and S3, the measured average number of received bits for Prop-NOMA-DAS is 9.7×10^4. This number decreases to 5.75×10^4 for PF-DAS. Therefore, while Prop-NOMA-DAS cannot satisfy all users, it allows unsatisfied users to get closer to their requirement (10^5 bits).

Fig. 1 also shows that the performance gain obtained by varying the number of subbands per antenna throughout the allocation process is significant with respect to non-adaptive approach in [11].

B. Evaluation of the Performance of BE Users

For BE users, we consider two system-level performance indicators: the achieved system throughput and user fairness.

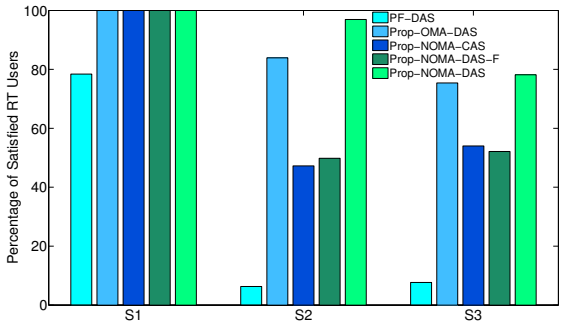

Fig. 1. Percentage of Satisfied RT Users

The latter is assessed through Jain's fairness index [12]:

$$J = \frac{\left(\sum_{k \in \mathcal{K}_{BE}} R_k\right)^2}{K_{BE} \sum_{k \in \mathcal{K}_{BE}} R_k^2}, \tag{24}$$

where R_k is the total achieved throughput by BE user k within a timeslot. Jain's fairness index ranges between 0 and 1 with the maximum achieved in the case of absolute fairness.

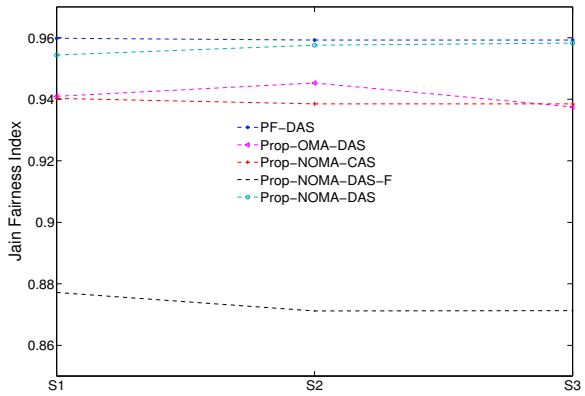

Fig. 2. Fairness Achieved in S1, S2 and S3 by the Different Techniques

Fig. 2 shows that PF-DAS and Prop-NOMA-DAS have a similar performance in terms of user fairness. This shows that the proposed allocation technique does not jeopardize the fairness between BE users. Moreover, Fig. 2 shows that Prop-NOMA-DAS outperforms Prop-NOMA-DAS-F in terms of fairness as the latter technique does not guarantee BE users to be served on their best perceived subbands.

Fig. 3 shows the evolution of the system throughput as time progresses, for S1 and S3, encompassing systems with relaxed and strict requirements. For both scenarios, Prop-NOMA-DAS outperforms Prop-OMA-DAS, Prop-NOMA-CAS and Prop-NOMA-DAS-F; hence, our choice to use NOMA and DAS for a mixed traffic system is justified.

Now moving to the comparison of Prop-NOMA-DAS with PF-DAS, for S1, we can see that Prop-NOMA-DAS outperforms PF-DAS in the first 12 timeslots, when most of the RT users are still in the active set of the allocation process. This is because Prop-NOMA-DAS gives the minimum necessary amount of resources to RT users in order to maximize the utility of BE users. Starting at timeslot 13, PF-DAS achieves a higher data rate than Prop-NOMA-DAS. This is due to the fact that the PF scheduler aims at maximizing system

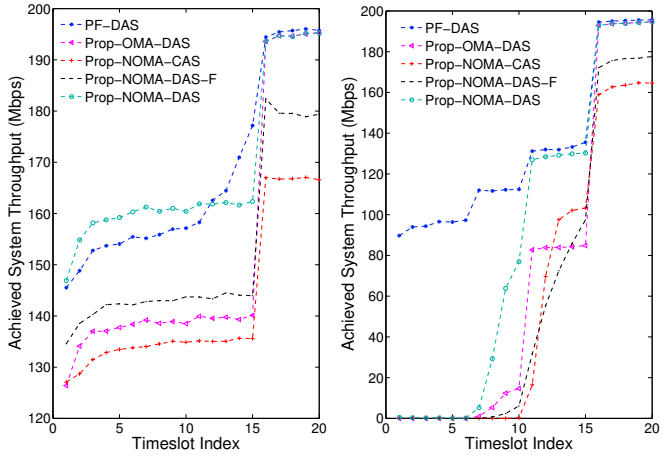

Fig. 3. Evolution of the Achieved System Throughput in Mbps per Timeslot a) for scenario 1. b) for scenario 3

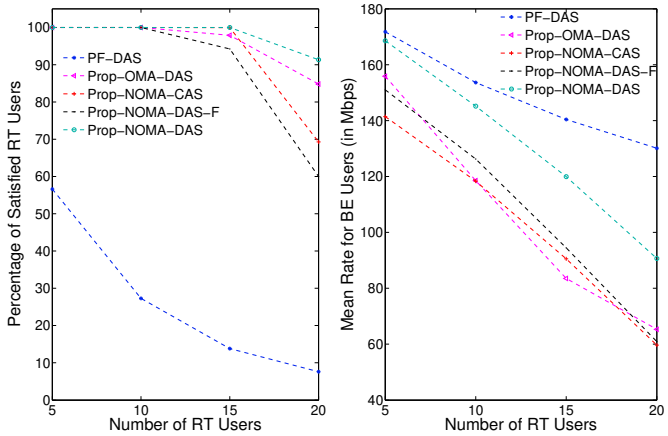

Fig. 4. Left: Percentage of Satisfaction for RT users in terms of their number. Right: Mean Rate for BE users in terms of the number of RT users

throughput. Therefore, some RT users might receive their required number of bits before reaching their latency limit and exit the system, leaving more resources for BE users. Once all RT users have exited the system (at the end of timeslot 15), Prop-NOMA-DAS achieves a performance very close to PF-DAS. However, when averaging system throughput over time, we find that PF-DAS achieves a mean BE throughput of 167 Mbps while our technique achieves 168 Mbps, hence has a superior performance. Also, recall that PF-DAS satisfies RT users in 78% of the cases in S1 compared to 100% satisfaction for Prop-NOMA-DAS as shown in subsection V-A.

Moving to S2, we notice that Prop-NOMA-DAS does not give any resource to BE users prior to timeslot 6. In fact, until timeslot 6, all 20 RT users are awaiting scheduling, and Prop-NOMA-DAS prioritizes them in the resource allocation, which is not the case for PF-DAS. Starting from timeslot 6, Prop-NOMA-DAS starts giving more and more resources to BE users as the system becomes less congested. After timeslot 15, all RT users exit the system, resulting in Prop-NOMA-DAS and PF-DAS achieving similar performance.

Finally, in Fig. 4, the performance of the different algorithms is shown for a varying number of RT users. The following cases are studied:

1) $K_{RT} = 5$ users (1 user in C1, 1 in C2 and 3 in C3)
2) $K_{RT} = 10$ users (2 users in C1, 2 in C2 and 6 in C3)
3) $K_{RT} = 15$ users (3 users in C1, 3 in C2 and 9 in C3)
4) $K_{RT} = 20$ users (4 users in C1, 4 in C2 and 12 in C3)

Fig. 4 shows the superiority of the proposed technique in terms of satisfying RT users. In fact, Prop-NOMA-DAS satisfies RT users in over 90% of the cases, regardless of their number. In comparison, PF-DAS satisfies RT users in 57% of the cases when $K_{RT} = 5$, and 9% of the cases when $K_{RT} = 20$. In contrast, PF-DAS outperforms the proposed technique in terms of maximizing the mean rate of BE users as the latter reserves more resources for RT users.

VI. Conclusion

In this paper, we proposed a new technique for allocating resources to users with mixed traffic in a NOMA-DAS setting. The proposed technique aims at satisfying all RT users, while maximizing the utility of BE users. Simulation results showed that our method outperforms the PF scheduler, especially in terms of the number of satisfied RT users. We also showed the performance improvement obtained by NOMA over OMA, and DAS over CAS, in the mixed traffic context.

Acknowledgment

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

References

[1] 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.

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

[3] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD systems," *IEEE Globecom Workshops*, pp. 1–7, December 2016.

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

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

[6] M. J. 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 Symposium on Computers and Communications (ISCC)*, July 2017, pp. 499–506.

[7] Q. T. 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.

[8] 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 Wireless Commun. Lett.*, vol. 22, no. 1, pp. 149–152, Jan 2018.

[9] E. Okamoto, "An improved proportional fair scheduling in downlink non-orthogonal multiple access system," *Proc. IEEE Veh. Techn. Conf. Fall*, pp. 1–5, Sept 2015.

[10] 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.

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

[12] 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.