Weighted Proportional Fair Scheduling for Downlink Non-Orthogonal Multiple Access

Marie Rita Hojeij, Charbel Abdel Nour, Joumana Farah, Catherine Douillard

Weighted Proportional Fair Scheduling for Downlink Non-Orthogonal Multiple Access

Marie-Rita Hojeij, Charbel Abdel Nour, Joumana Farah, and Catherine Douillard,

***Abstract*—In this paper, a weighted proportional fair (PF) scheduling method is proposed in the context of non-orthogonal multiple access (NOMA) with successive interference cancellation (SIC) at the receiver side. The new scheme introduces weights that adapt the classical PF metric to the NOMA scenario, improving performance indicators and enabling new services. The distinguishing value of the proposal resides in its ability to improve long term fairness and total system throughput while achieving a high level of fairness in every scheduling slot. Finally, it is shown that the additional complexity caused by the weight calculation has only a limited impact on the overall scheduler complexity while simulation results confirm the claimed improvements making the proposal an appealing alternative for resource allocation in a cellular downlink system.**

***Index Terms*—Non-orthogonal multiple access, scheduling, proportional fairness, resource allocation.**

I. Introduction

RADIO access technologies apply multiple access schemes to provide the means for multiple users to access and share resources at the same time. In the 3.9 and fourth generation of mobile communication systems, such as Long-Term Evolution (LTE) [1] and LTE-Advanced [2], [3], orthogonal multiple access (OMA) based on orthogonal frequency division multiplexing (OFDM) or single carrier frequency division multiple access (SC-FDMA) were adopted, respectively for downlink and uplink transmissions. Orthogonal multiple access techniques have gained their success from their ability to achieve good system-level throughput performance in packet-domain services, while requiring a reasonable complexity, especially due to the absence of multi-user detection.

However, with the proliferation of Internet applications, between the end of 2016 and 2022, total mobile traffic is expected to increase by 8 times [4]. At the same time, communications networks are required to further enhance system efficiency, latency, and user fairness. To this end, non-orthogonal multiple access (NOMA) has recently emerged as a promising candidate for future radio access. By exploiting an additional multiplexing domain, the power domain, NOMA allows the cohabitation of two or more users per subcarrier. User multiplexing is conducted at the transmitter side, on top of the OFDM layer, and multi-user signal separation takes place at the receiver side, using successive interference cancellation (SIC) [5]–[11].

The main appeal of NOMA is that it improves user fairness while maximizing the total user throughput. The majority of existing works dealing with scheduling in NOMA have investigated and proposed new techniques for improving the system-level performance in terms of system capacity and cell-edge user throughput.

In [12], throughput performance is assessed for an uplink non-orthogonal multiple access system where optimized scheduling techniques are proposed and evaluated. A cost function is assigned to each possible pair of users, in order to maximize either the sum-rate or the weighted sum rate. The user pairing problem is solved by the Hungarian method and significant improvements in sum rates and cell-edge rates are shown compared to OMA.

In [13], several new strategies for the allocation of radio resources (in terms of bandwidth and power) in a downlink NOMA system have been investigated and evaluated. The main objective of [13] is to minimize the number of allocated subbands, while guaranteeing a requested service data rate for each user. In this sense, several design issues have been explored: choice of user pairing, subband assignment, optimal and suboptimal power allocation, dynamic switching to OMA. Simulation results show that the proposed resource allocation techniques provide better performance when NOMA is used, compared to OMA.

In addition to proposing new scheduling techniques for a NOMA-based system, some papers have investigated the commonly used PF scheduler for the good tradeoff it provides between system capacity and user fairness. Several enhancements have been proposed to the PF scheduler in order to further improve the system-level performance of a NOMA system.

In [14], an improved downlink NOMA scheduling scheme based on the PF scheduler is proposed and evaluated. In this sense, modifications to the PF scheduling metric have been introduced in order to improve the fairness at every frame assignment. Results have shown improved performance compared to a NOMA-based system that considers the conventional PF scheduler.

In [10], a weighted PF-based multiuser scheduling scheme is

M. Hojeij, C. Abdel Nour and C. Douillard are with IMT-Alantique, CNRS UMR 6285 Lab-STICC, UBL, France (email: marie.hojeij@imt-atlantique.fr; charbel.abdelnour@imt-atlantique.fr; catherine.douillard@imt-atlantique.fr) and M. Hojeij is also with the Faculty of Engineering, Holy Spirit University of Kaslik, Lebanon.

J. Farah is with the Faculty of Engineering, Lebanese University, Lebanon email:joumana.farah@ul.edu.lb

Part of this work has been performed in the framework of the Horizon 2020 project FANTASTIC-5G (ICT-671660), which is partly funded by the European Union. This work has also been funded with support from the Lebanese University and the French-Lebanese CEDRE program.

The authors declare that there is no conflict of interest regarding the publication of this paper.

proposed in the context of a non-orthogonal access downlink system for the aim of further enhancing the gain of the cell-edge user. A frequency block access policy is proposed for cell-interior and cell-edge user groups using fractional frequency reuse (FFR), with significant improvements in the user fairness and system frequency efficiency.

Proposing enhancements to the PF scheduling metric has also been investigated in an OMA-based system. Similar to the work done in [10], several papers have proposed weighted versions of the PF scheduler, with the aim of improving user fairness in the OMA context.

In [15], fair weights have been implemented for opportunistic scheduling of heterogeneous traffic types for OMA networks. For designing fair weights, the proposed scheduler takes into account the average channel status as well as resource requirements in terms of traffic types. Simulation analysis demonstrates the efficiency of the proposed scheme in terms of resource utilization, and its flexibility with regards to network characteristics changes due to user mobility.

In [16], the problem of fairness deficiency encountered by the PF scheduler when the mobiles experience unequal pathloss is investigated. To mitigate this issue, a modified version of the PF scheduler introducing distance compensation factors has been proposed. This solution was shown to achieve both high capacity and high fairness.

In [17], a weighted PF algorithm is proposed in order to maximize best-effort service utility. The reason behind introducing weight factors into the PF metric is to exploit the inherent near-far diversity given by the pathloss. The proposed algorithm enhances both best-effort service utility and throughput performance, with a complexity similar to the complexity of the conventional PF scheduler.

Designing fair weights within the PF scheduling metric has shown remarkable improvements in the system's performance of OMA and NOMA-based systems, especially at the level of user fairness.

The above mentioned studies have shown the advantages of introducing fair weights within the PF scheduler, especially when combined with a NOMA system. However, increasing the achieved user rate at every slot has not been tackled in these studies. Indeed, such a feature can have a positive impact on the perceived quality of service especially for multimedia services on one side and can reduce the amount of required buffering and memory at the user terminal on the other side. Accordingly, we aim in this paper to combine the advantages of NOMA in terms of spectral efficiency with an implementation of fair weights at the scheduling level in order to improve both the achieved user rate and the user fairness at each time slot. We propose indeed a weighted PF metric where several designs of the introduced weights are evaluated. The proposed scheme aims at providing fairness among users for each channel realization. By doing so, not only short-term fairness is achieved but also user capacity and long-term fairness are enhanced accordingly. On the other hand, the proposed schemes mitigate the problem of zero-rate incidence, inherent to PF scheduling, by attempting to provide non-zero rate to each user in any time scale of interest. This will further enhance the quality of experience (QoE) of all users.

This paper is organized as follows: In Section II, we introduce the system model and give a general description of the NOMA-based PF scheduler. Section III details the proposed weighted schemes in the NOMA context. In Section IV, we apply the fair weights to a resource allocation system based on OMA. Simulation results are given and analyzed in Section V, while Section VI concludes the paper.

II. System Description

A. Basic NOMA System

In this section, we describe the basic concept of NOMA including user multiplexing at the transmitter of the base station (BS) and signal separation at the receiver of the user terminal.

In this paper, a downlink system with a single input single output (SISO) antenna configuration is considered. The system consists of K users per cell, with a total bandwidth B divided into S subbands.

Among the K users, a subset of users $U_s = \{k_1, k_2, ..., k_n, ..., k_{n(s)}\}$, is selected to be scheduled over each frequency subband s, ($1 \leq s \leq S$). The nth user ($1 \leq n \leq n(s)$) scheduled at subband s is denoted by k_n, and $n(s)$ indicates the number of users non-orthogonally scheduled at subband s. At the BS transmitter side, the information sequence of each scheduled user at subband s is independently coded and modulated resulting into symbol x_{s,k_n} for the nth scheduled user. Therefore, the signal transmitted by the BS on subband s, x_s, represents the sum of the coded and modulated symbols of the $n(s)$ scheduled users:

$$x_s = \sum_{n=1}^{n(s)} x_{s,k_n}, \ with \ \mathrm{E}\left[|x_{s,k_n}|^2\right] = P_{s,k_n} \tag{1}$$

where P_{s,k_n} is the power allocated to user k_n at subband s. The received signal vector of user k_n at subband s, y_{s,k_n}, is represented by:

$$y_{s,k_n} = h_{s,k_n} x_{s,k_n} + w_{s,k_n} \tag{2}$$

where h_{s,k_n} is the channel coefficient between user k_n and the BS, at subband s. w_{s,k_n} represents the received Gaussian noise plus inter-cell interference experienced by user k_n at subband s. Let P_{max} be the maximum allowable power transmitted by the BS. Hence, the sum power constraint is formulated as follows:

$$\sum_{s=1}^{S} \sum_{n=1}^{n(s)} P_{s,k_n} = P_{\max} \tag{3}$$

The SIC process [18] is conducted at the receiver side, and the optimal order for user decoding is in the increasing order of the channel gains observed by users, normalized by the noise and inter-cell interference $h^2_{s,k_n}/n_{s,k_n}$, where n_{s,k_n} is the average power of w_{s,k_n}. Therefore, any user can correctly decode the signals of other users whose decoding order comes before that user. In other words, user k_n at subband s can remove the inter-user interference from the jth user, k_j, at

subband s, provided $h^2_{s,k_j}/n_{s,k_j}$ is lower than $h^2_{s,k_n}/n_{s,k_n}$, and it treats the received signals from other users with higher $h^2_{s,k_j}/n_{s,k_j}$ as noise [6], [19].
Assuming successful decoding and no error propagation, and supposing that inter-cell interference is randomized such that it can be considered as white noise [8], [12], the throughput of user k_n, at subband s, R_{s,k_n}, is given by:

$$R_{s,k_n} = \frac{B}{S}\log_2\left(1 + \frac{h^2_{s,k_n}P_{s,k_n}}{\sum\limits_{j \in \mathrm{N}_s,\ \frac{h^2_{s,k_n}}{n_{s,k_n}} < \frac{h^2_{s,k_j}}{n_{s,k_j}}} h^2_{s,k_n}P_{s,k_j} + n_{s,k_n}}\right) \tag{4}$$

It should be noted that most of the papers dealing with resource allocation in downlink NOMA [9], [19]–[21], consider a maximum number of users per subband of two, in order to limit the SIC complexity in the mobile receiver, except for [8] and [22] where this number respectively reaches 3 and 4. However, in the last two cases, static power allocation is assumed, which simplifies the power allocation step but degrades throughput performance. It has also been stated that the performance gain obtained with 3 or 4 users per subband is minor in comparison to the case with 2 users.

B. Conventional PF Scheduling Scheme

The PF scheduling algorithm has been proposed to ensure balance between cell throughput and user fairness. Kelly et al. [23] have defined the proportional fair allocation of rates, and used a utility function to represent the degree of satisfaction of allocated users. In [24], the practical implementation of the PF scheduler is detailed: at the beginning of every scheduling slot, each user provides the base station with its channel state (or equivalently its feasible rate). The scheduling algorithm keeps track of the average throughput $T_k(t)$ of each user in a past window of length t_c. In the scheduling slot t, user k^* is selected to be served based on [24]:

$$k^* = \arg\max_k \frac{R_k(t)}{T_k(t)} \tag{5}$$

where $R_k(t)$ is the feasible rate of user k for scheduling slot t.

In [25], an approximated version of the PF scheduler for multiple users transmission is presented. This version has been adopted in the majority of the works dealing with NOMA [19], [20], [22] in order to select users to be non-orthogonally scheduled on available resources.
For a subband s under consideration, the PF metric is estimated for each possible combination of users U, and the combination that maximizes the PF metric is denoted by U_s:

$$U_s = \arg\max_U \sum_{k \in U} \frac{R_{s,k}(t)}{T_k(t)} \tag{6}$$

$R_{s,k}(t)$ denotes the instantaneous achievable throughput of user k at subband s and scheduling time slot t.

Note that the total number of combinations tested for each considered subband is:

$$N_U = \binom{1}{K} + \binom{2}{K} + ... + \binom{N(s)}{K} \tag{7}$$

$R_{s,k}(t)$ is calculated based on Eq. 4, whereas $T_k(t)$ is recursively updated as follows [25] :

$$T_k(t+1) = \left(1 - \frac{1}{t_c}\right)T_k(t) + \frac{1}{t_c}\sum_{s=1}^{S} R_{s,k}(t) \tag{8}$$

Parameter t_c defines the throughput averaging time window. In other words, this is the time horizon in which we want to achieve fairness. t_c is chosen to guarantee a good tradeoff between system performance (in terms of fairness) and system capacity. We assume in the following a t_c window of 100 time slots. With a time slot duration equal to 1 ms, a 100 ms average user throughput $T_k(t)$ is therefore considered.

III. Proposed Weighted NOMA-Based Proportional Fairness (WNOPF) Scheduler

The PF scheduler both aims at achieving high data rates and at ensuring fairness among users, but it only considers long-term fairness. In other words, a duration of t_c time slots is needed to achieve fairness among users. However, short-term fairness and fast convergence towards required performance is an important issue to be addressed in upcoming mobile standards [4].
Since all possible combinations of candidate users are tested for each subband, a user might be selected more than once and attributed multiple subbands during the same time slot. On the other hand, it can also happen that a user will not be allocated any subband whenever its historical rate is high. Then, the user will not be assigned any transmission rate for multiple scheduling slots. This behavior can be very problematic in some applications, especially those requiring a quasi-constant QoE such as multimedia transmissions. In such cases, buffering may be needed. However, such a scenario may not be compatible with applications requiring low latency transmission.
Therefore, we propose several weighted PF metrics that aim at:

- enhancing the user capacity, thus increasing the total achieved user throughput;
- reducing the convergence time towards required fairness performance;
- enhancing fairness among users (both long-term and short-term fairness);
- limiting the fluctuations of user data rates;
- incorporating the delivery of different levels of quality of service (QoS).

The proposed scheduler consists of introducing fair weights into the conventional PF scheduling metric. The main goal of the weighted metrics is to ensure fairness among users in every scheduling slot.
To do so, we start by modifying the PF metric expression so as to take into account the status of the current assignment

in time slot t. Therefore, the scheduling priority given for each user is not only based on its historical rate but also on its current total achieved rate (throughput achieved during the current scheduling slot t), as proposed in [14].
Scheduling is performed subband by subband and on a time slot basis. For each subband s, the conventional PF metric PF_s^{NOMA} and a weight factor $W(U)$ are both calculated for each candidate user set U. Then, the scheduler selects the set of scheduled users U_s that maximizes the weighted metric $PF_s^{NOMA}(U) \times W(U)$. The corresponding scheduling method is referred to as Weighted NOMA PF scheduler, denoted by WPF^{NOMA}. The resource allocation metric can be formulated as follows:

$$\begin{aligned} WPF_s^{NOMA}(U) &= PF_s^{NOMA}(U) \times W(U) \\ U_s &= \arg\max_U WPF_s^{NOMA}(U) \end{aligned} \tag{9}$$

The proposed weight calculation for each candidate user set U relies on the sum of the weights of the multiplexed users.

$$W(U) = \sum_{k \in U} W_k(t) \tag{10}$$

with

$$W_k(t) = R_{avg}^e(t) - R_k(t), \ k \in U \tag{11}$$

$R_{avg}^e(t)$ is the expected achievable bound for the average user data rate in the current scheduling slot t. It is calculated as follows:

$$R_{avg}^e(t) = b.R_{avg}(t-1) \tag{12}$$

Since we tend to enhance the achieved user rate in every slot, each user must target a higher rate compared to the rate previously achieved. Therefore, parameter b is chosen to be greater than 1.
The average user data rate, $R_{avg}(t)$, used in (12), is updated at the end of each scheduling slot based on the following:

$$R_{avg}(t) = \frac{1}{K} \sum_{k=1}^{K} \sum_{s=1}^{S} R_{s,k}(t) \tag{13}$$

where $R_{s,k}(t)$ is the data rate achieved by user k on subband s.

On the other hand, $R_k(t)$, the actual achieved data rate by user k during scheduling slot t, is calculated as:

$$R_k(t) = \sum_{s \in S_k} R_{s,k}(t), \ k \in U \tag{14}$$

with S_k the set of subbands allocated to user k during time slot t. At the beginning of every scheduling slot, S_k is emptied; each time user k is being allocated a new subband, S_k and $R_k(t)$ are both updated.

The main idea behind introducing weights is to minimize the rate gap among scheduled users in every scheduling slot, thus maximizing fairness among them. A user set U is provided with a high priority among candidate user sets if it contains non-orthogonally multiplexed users experiencing a good channel quality on subband s, having low or moderate historical rates, or/and having large rate distances between their actual achieved rates and their expected achievable average user throughput. The highest level of fairness is achieved when all users reach the expected user average rate $R_{avg}^e(t)$. By applying the proposed scheduling procedure, we aim to enhance long-term and short-term fairness at the same time.

It was shown in [25] that the scheduling metric PF^{NOMA}, defined in (6), strikes a good trade-off between throughput and fairness, since it maximizes the sum of users service utility which can be formally written as:

$$PF^{NOMA} = \max_{scheduler} \sum_{k=1}^{K} \log T_k \tag{15}$$

Therefore, any enhanced scheduling metric like WNOPF can strike a better throughput-fairness balance (by achieving a higher service utility), compared to the conventional PF scheduler, provided that:

$$\sum_{k=1}^{K} \log T_k \geq \sum_{k=1}^{K} \log T_k^{'} \tag{16}$$

where the historical rates T_k and $T_k^{'}$ correspond to the schedulers using the WNOPF metric and the conventional PF metric, respectively.

Inspired by the work in [17], in the sequel, we show how this goal can be achieved by an appropriate design of the weights that verifies the constraints provided in Proposition 1.

Proposition 1: To make (16) valid, for a NOMA-based system, the following inequality should be verified:

$$\prod_{k=1}^{K} W(U_k) \prod_{k=1}^{K} E[R_{s,k}] \geq \prod_{k=1}^{K} E\left[R_{s,k}^{'}\right] \tag{17}$$

$E[R_{s,k}]$ and $E\left[R_{s,k}^{'}\right]$ are the statistical average of the instantaneous transmittable rate of user k on a subband s, when WNOPF and the conventional PF scheduler are applied respectively. U_k denotes a scheduled user set containing user k, U is a possible candidate user set, and $W(U_k)$ is the weight of the set U_k.

Proof. Equation (16) can be written as:

$$\prod_{k=1}^{K} T_k \geq \prod_{k=1}^{K} T_k^{'} \tag{18}$$

If we consider that $T_k = I_{k,tot} / (t_c \Delta T)$, where $I_{k,tot}$ is the total amount of information that can be received by user k, for a total observation time $t_c \Delta T$, and ΔT is the scheduling time slot length, we obtain:

$$\prod_{k=1}^{K} \frac{I_{k,tot}}{t_c \Delta T} \geq \prod_{k=1}^{K} \frac{I_{k,tot}^{'}}{t_c \Delta T} \tag{19}$$

If we denote by N_k the number of allocated time slots for user k within t_c, and n_k the statistical average of the number

of allocated subbands to user k per time slot, (19) can be re-written as:

$$\prod_{k=1}^{K} \frac{N_k n_k E\left[R_{s,k}\right] \Delta T}{t_c \Delta T} \geq \prod_{k=1}^{K} \frac{N_k^{'} n_k^{'} E\left[R_{s,k}^{'}\right] \Delta T}{t_c \Delta T} \tag{20}$$

Using a simple rearrangement, we get:

$$\frac{\prod_{k=1}^{K} \left(N_k/t_c\right) S\left(n_k/S\right)}{\prod_{k=1}^{K} \left(N_k^{'}/t_c\right) S\left(n_k^{'}/S\right)} \geq \frac{\prod_{k=1}^{K} E\left[R_{s,k}^{'}\right]}{\prod_{k=1}^{K} E\left[R_{s,k}\right]} \tag{21}$$

If Pr_k $(= N_k/t_c)$ denotes the probability of user k being scheduled per time slot and pr_k $(= n_k/S)$ the probability of user k being scheduled per subband, (21) can be reformulated as:

$$\frac{\prod_{k=1}^{K} Pr_k pr_k}{\prod_{k=1}^{K} Pr_k^{'} pr_k^{'}} \geq \frac{\prod_{k=1}^{K} E\left[R_{s,k}^{'}\right]}{\prod_{k=1}^{K} E\left[R_{s,k}\right]} \tag{22}$$

pr_k can be regarded as the probability of a set U_k (i.e. $\Pr\left(U_k\right)$, being chosen among all possible candidate sets U to be scheduled per subband.

Let's consider two sets of users U_1 and U_2. If the probability of the user set U_1 is greater than that of U_2, U_1 will be chosen to be scheduled. Equivalently, the corresponding scheduling metric for user set U_1 will be in this case larger than that of U_2. Therefore, In fact, a user set is chosen if the corresponding PF metric is the largest.

Thus, Eq.22 can be equivalent to the following equation:

$$\frac{\prod_{k=1}^{K} Pr_k PF^{NOMA}\left(U_k\right) W\left(U_k\right)}{\prod_{k=1}^{K} Pr_k^{'} PF^{NOMA}\left(U_k\right)} \geq \frac{\prod_{k=1}^{K} E\left[R_{s,k}^{'}\right]}{\prod_{k=1}^{K} E\left[R_{s,k}\right]} \tag{23}$$

Note that, in a NOMA-based system, the probability of a user being scheduled per time slot remains the same when using the proposed weighted metric or the conventional PF metric, since users are distributed with uniform and random probability over the entire network in each time slot. Thus, we adopt the following approximation:

$$Pr_k \simeq Pr_k^{'} \tag{24}$$

Additional observations and verifications related to this approximation are given in VII. Therefore, (23) and (24) can also be formulated as (17).

□

Other configurations of rate-distance weights can also be introduced. A promising one is obtained by substituting (25) for (9) and (10):

$$U_s = \arg\max_{U} \sum_{k \in U} \frac{R_{s,k}(t)}{T_k(t)} W_k(t), \;\; k \in U \tag{25}$$

Here, the conventional NOMA-based PF metric and the weights are jointly calculated for each user k in candidate user set U. By doing so, we assign to each user its weight while ignoring the cross effect $\frac{R_{s,k|U}(t)}{T_{k|U}(t)} W_{k'|U}(t)$ produced by (9), where k and k' are non-orthogonally multiplexed users in the same U. This joint-based incorporation of weights is denoted by J-WNOPF in the following evaluations.

IV. Proposed Weighted OMA-Based PF Scheduler (WOPF)

In the majority of existing works dealing with fair scheduling, OMA-based systems are considered. For this reason, we propose to apply the weighted proportional fair scheduling metric introduced in this paper to an OMA-based system as well. This allows the contribution of NOMA within our framework to be evaluated. In the OMA case, non-orthogonal cohabitation is not allowed. Instead, a subband s is allocated to only one user, based on the following metric:

$$k^* = \arg\max_{k} \frac{R_{s,k}(t)}{T_k(t)} W_k(t) \tag{26}$$

where $W_k(t)$ is the weight assigned to user k, calculated similarly to the weights in WNOPF. The conventional OMA-based PF scheduling metric is denoted by PF^{OMA}, whereas the resulting scheduling algorithm combining OMA with the proposed weighted PF is denoted by WOPF.

OMA can be regarded as a special case of NOMA where only one user is allowed to be scheduled per subband. Therefore, in order to achieve a higher user service utility with WOPF than with the conventional PF scheduler in OMA, Proposition 1, detailed and proven in Section III, should also be verified for an OMA-based system. For this purpose, (17) is modified as follows:

$$\prod_{k=1}^{K} W_k \prod_{k=1}^{K} E\left[R_{s,k}\right] \geq \prod_{k=1}^{K} E\left[R_{s,k}^{'}\right] \tag{27}$$

where W_k is the weight assigned to user k.

Note that, as in the NOMA case, we assume that the probability of a user being scheduled per time slot remains the same when using the proposed weighted metric or the conventional PF metric.

V. Proposed Scheduling Metric for the First Scheduling Slot

In the first scheduling slot, the historical rates and the expected user average data rate are all set to zero. Hence, the selection of users by the scheduler is only based on the instantaneous achievable throughputs. Therefore, fairness is not achieved in the first scheduling slot, and the following slots are penalized accordingly. To counteract this effect, we propose to treat the first scheduling slot differently, for all the proposed weighted metrics.

For each subband s, the proposed scheduling process selects

U_s among the candidate user sets based on the following criterion:

$$U_s = \arg\max_{U} \sum_{k \in U} \frac{R_{s,k}(t=1)}{R_k(t=1)} \quad (28)$$

Note that when WOPF is considered, the maximum number of users per set U is limited to 1.
$R_k(t=1)$, the actual achieved throughput, is updated each time a subband is allocated to user k during the first scheduling slot. By doing so, we give priority to the user experiencing a good channel quality with regard to its actual total achieved data rate, thus enhancing fairness in the first slot.

VI. Incorporation of Premium Services

In this section, we propose some changes to the proposed weighted metrics in order to give the possibility of delivering different levels of suality of service. In other words, the proposed metrics should have the ability to provide different priorities to different users or to guarantee a certain level of performance to a data flow. To do so, (11) is modified as follows:

$$W_k(t) = R_{service} - R_k(t), \ k \in U \quad (29)$$

where $R_{service}$ is the data rate requested by a certain group of users, corresponding to a certain level of performance. As an example, we detail an example of 3 services, although the proposed modifications can be applied to an arbitrary number of services. $R_{service}$ is then defined as follows:

$$R_{service} = \begin{cases} R_{basic}, \text{ if } k \text{ requests a basic service} \\ R_{silver}, \text{ if } k \text{ requests a silver service} \\ R_{gold}, \text{ if } k \text{ requests a gold service} \end{cases}, k \in U \quad (30)$$

This modification aims to guarantee a minimum requested service data rate for each user and also tends to enhance the overall achieved fairness between users belonging to the same group, i.e. asking for the same service.

VII. Numerical Results

A. System Model Parameters and Performance Evaluation

This subsection presents the system level simulation parameters used to evaluate the proposed scheduling techniques. The parameters considered in this work are based on existing LTE/LTE-Advanced specifications [26]. We consider a baseline SISO antenna configuration. The maximum transmission power of the base station is 46 dBm. The system bandwidth is 10 MHz and is divided into 128 subbands when not further specified. The noise power spectral density is 4.10^{-18} mW/Hz. Users are deployed randomly in the cell and the cell radius is set to 500 m. Distance-dependent path loss is considered with a decay factor of 3.76. Extended typical urban (ETU) channel model is assumed, with time-selectivity corresponding to a mobile velocity of 50 km/h, at the carrier frequency of 2 GHz. In both OMA and NOMA scenarios, equal repartition of power is considered among subbands, as considered in [8], [20], [21]. In the case of NOMA, fractional transmit power allocation (FTPA) [27] is used to allocate power among scheduled users within a subband. Without loss of generality, NOMA results are shown for the case where the maximum number of scheduled users per subband is set to 2 ($n(s) = 2$).

As for parameter b in (12), after several testings, the best performance was observed for b equal to 1.5. In fact, the system has a rate saturation bound with respect to parameter b, since when we further increase b, similar performance is maintained.

B. Performance Evaluation

In this part, we mainly consider four system-level performance indicators: achieved system capacity, long-term fairness, short-term fairness, and cell-edge user throughput.
Several techniques are evaluated and compared. The following acronyms are used to refer to the main studied methods:

- PF^{NOMA}: conventional PF scheduling metric in a NOMA-based system;
- $WNOPF$: proposed weighted PF scheduling metric in a NOMA-based system;
- $J-WNOPF$: proposed Weighted PF scheduling metric with a joint incorporation of weights in a NOMA-based system;
- $PF^{NOMA}_{modified}$: a modified version of the PF scheduling metric proposed in [14], where the actual assignment of each frame is added to the historical rate;
- PF^{OMA}: conventional PF scheduling metric in an OMA-based system;
- $WOPF$: proposed weighted PF scheduling metric in an OMA-based system.

In order to assess the fairness performance achieved by the different techniques, a fairness metric needs to be defined first. Gini fairness index [28] measures the degree of fairness that a resource allocation scheme can achieve. It is defined as:

$$G = \frac{1}{2K^2\overline{r}} \sum_{x=1}^{K} \sum_{y=1}^{K} |r_x - r_y| \quad (31)$$

with

$$\overline{r} = \frac{\sum_{k=1}^{K} r_k}{K} \quad (32)$$

r_k is the throughput achieved by user k. When long-term fairness is evaluated, r_k is considered as the total throughput achieved by user k averaged over a time-window length t_c:

$$r_k = \frac{1}{t_c} \sum_{t=1}^{t_c} R_k(t) \quad (33)$$

Otherwise, when fairness among users is to be evaluated within each scheduling slot, short-term fairness is considered and r_k is taken equal to $R_k(t)$, the actual throughput achieved by user k during scheduling slot t.
Gini fairness index takes values between 0 and 1, where $G = 0$ corresponds to the maximum level of fairness among users, while a value of G close to 1 indicates that the resource

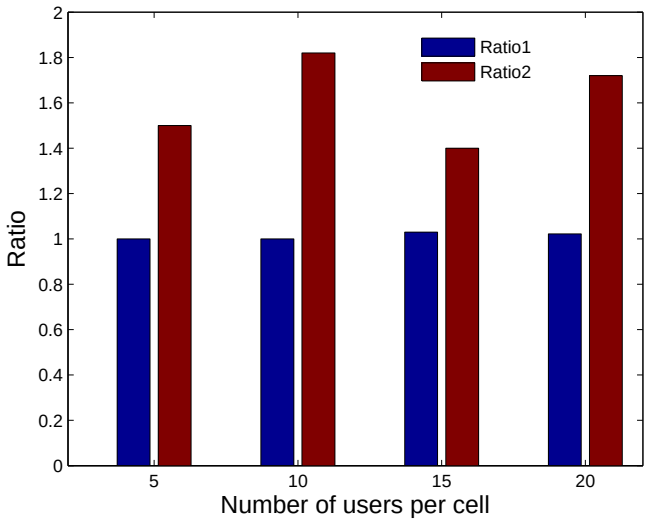

Fig. 1. Observed ratios related to (17) and (26) vs. number of users per cell, NOMA-based system.

allocation scenario is highly unfair.

First, we check the validity of Proposition 1 detailed in Section III and IV, and of the assumption done in (24). Fig. 1 shows the observed ratio between Pr_k and $Pr_k\prime$, denoted by Ratio1, for different values of the number of users per cell. Fig. 1 also shows the ratio between the left hand and the right hand expressions of (17), denoted by Ratio2. Results show that Ratio1 is very close to 1, which means that the probability of a user being scheduled per time slot remains the same, under the proposed weighted metric or under the conventional PF metric. In addition, Ratio2 is shown to be greater than 1 regardless of the number of users per cell, which verifies Proposition 1, defined in (17). The results of a similar verification for an OMA system are observed in Fig. 2 .

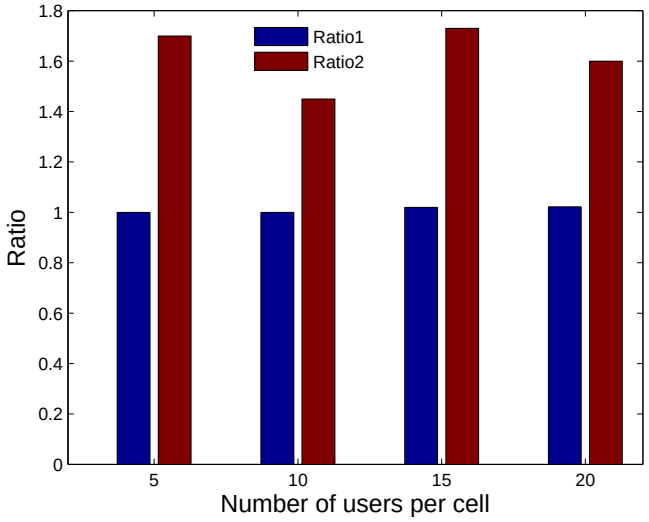

Fig. 2. Observed ratio related to (26) and (29) vs. number of users per cell, OMA-based system.

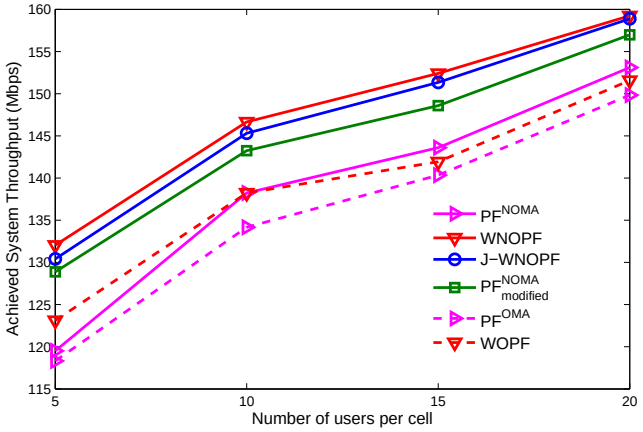

Fig. 3. System throughput achieved with the proposed scheduling schemes vs. number of users per cell.

Fig. 3 shows the system capacity achieved with each of the simulated methods for different numbers of users per cell. Curves in solid lines represent the NOMA case, whereas curves with dotted lines refer to OMA.

We can observe that the throughput achieved with all the simulated methods increases as the number of users per cell is increased, even though the total number of used subbands is constant. This is due to the fact that the higher the number of users per cell, the better the multi-user diversity is exploited by the scheduling scheme, as also observed in [15].

The gain achieved by WNOPF, when compared to the other proposed weighted metric J-WNOPF, is mainly due to the fact that the joint incorporation of weights does not take into consideration the cross effect produced by non-orthogonally multiplexed users.

The gain in performance obtained by the introduction of weights in the scheduling metric, compared to the conventional PF^{NOMA} metric, stems from the fact that for every channel realization, the weighted metrics try to ensure similar rates to all users, even those experiencing bad channel conditions. With PF^{NOMA}, such users would not be chosen frequently, whereas appropriate weights give them a higher chance to be scheduled more often.

Fig. 3 also shows an improved performance of the proposed metrics when compared to the modified PF scheduling metric $PF^{NOMA}_{modified}$ described in [14]. Although they both consider the current assignment in their metric calculation, they still differ by the fact that the proposed weighted metrics target a higher rate compared to the rate previously achieved, therefore tending to increase the achieved user rate in every slot.

When the proposed scheduling metrics are applied in an OMA context, WOPF provides higher throughputs than PF^{OMA}, due to the same reason why WNOPF outperforms PF^{NOMA}. Fig. 3 also shows a significant performance gain achieved by NOMA over OMA. All weighted scheduling metrics applying NOMA outperform the simulated metrics based on OMA, including WOPF. This gain is due to the efficient non-orthogonal multiplexing of users. It should also

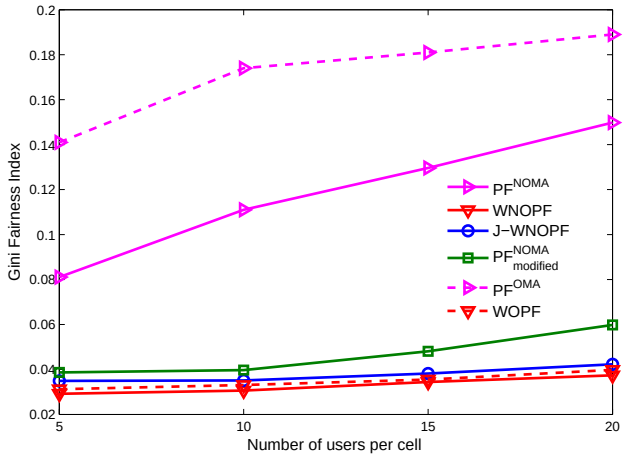

Fig. 4. Gini fairness index of the proposed scheduling schemes vs. number of users per cell.

be noted that the gain achieved by WNOPF over PF^{NOMA} is greater than the one achieved by WOPF over OPF: combining fair weights with NOMA definitely yields the best performance.

Long-term fairness is an important performance indicator for the allocation process. Fig. 4 shows this metric as a function of the number of users per cell. Long-term fairness is improved when fair weights are introduced, independently of the access technique (OMA or NOMA). The reason is that, when aiming to enhance fairness in every scheduling slot, long-term fairness is enhanced accordingly. Again, in terms of fairness, the proposed weighted metrics outperform the modified PF metric [14], $PF^{NOMA}_{modified}$. This is due to the fact that WNOPF and J-WNOPF do not only consider the current rate assignment, but also tend to minimize the rate gap among scheduled users in every channel realization, thus maximizing fairness among them.

Fig. 5 shows the achieved system throughput as a function of the number of subbands S, for 15 users per cell. We can see that the proposed weighted metrics outperform the conventional PF scheduling scheme, for both access techniques OMA and NOMA, even when the number of subbands is limited.

Since WNOPF proves to give better performance than J-WNOPF, in terms of system capacity and fairness, J-WNOPF won't be considered in the subsequent results.
Since one of the main focuses of this study is to achieve short-term fairness, the proposed techniques should be compared based on the time required to achieve the final fairness level. Fig. 6 shows the Gini fairness index versus the scheduling time index t. The proposed weighted metric WNOPF achieves a high fairness from the beginning of the allocation process, and converges to the highest level of fairness (lowest value of index $G = 0.0013$) in a limited number of allocation steps or time slots. On the contrary,

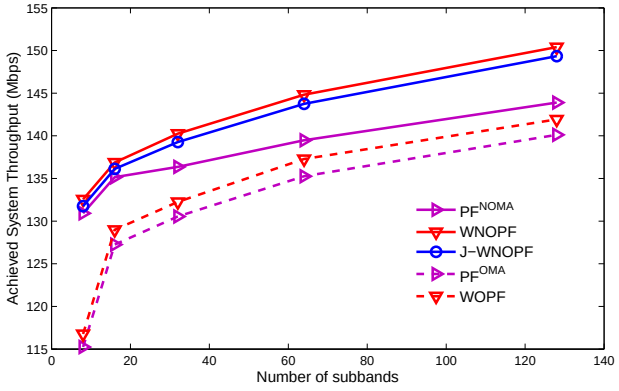

Fig. 5. Achieved system throughput vs. S, for $K = 15$.

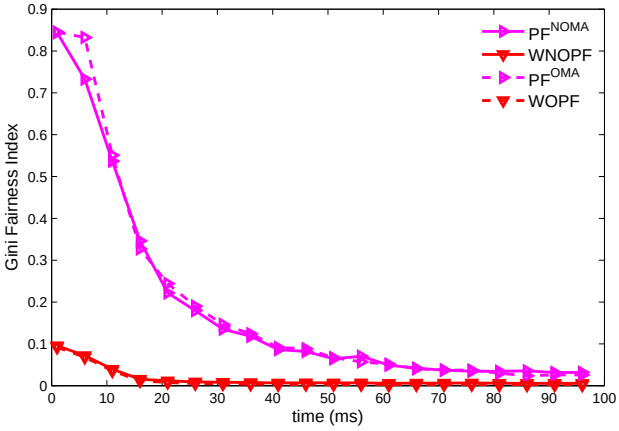

Fig. 6. Gini fairness index vs. scheduling time index t.

PF^{NOMA} shows unfairness among users for a much longer time. Weighted metrics not only show faster convergence to a high fairness level, but also give a lower Gini indicator at the end of the window length, when compared to conventional PF^{NOMA}.

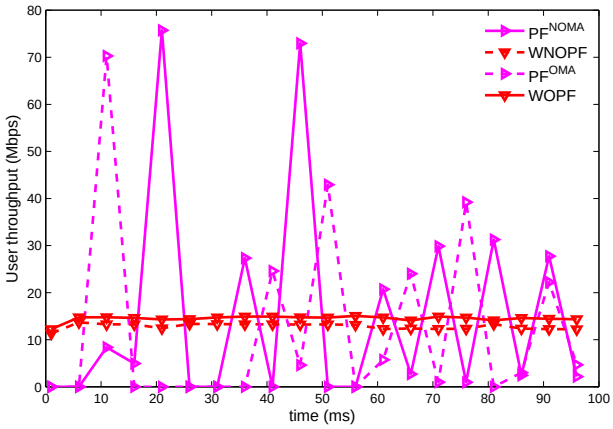

Fig. 7. User throughput vs. time for NOMA-based scheduling schemes.

In order to assess the QoE achieved by the proposed scheduling schemes, we evaluate the time required for each user to be served for the first time, referred to as the rate latency, as well the variations of its achieved rate over time. For this purpose, Fig. 7 shows the achieved rate versus time for the user experiencing the largest rate latency, for the different scheduling schemes.

When the conventional PF^{NOMA} is used, no rate is provided for this user, for the first five scheduling slots. In addition, large rate fluctuations are observed through time. In contrast, when weighted metrics and a special treatment of the first time slot are considered, a non-zero rate is assigned for the least privileged users from the first scheduling slot, and remains stable for all the following slots. This behavior results from the fact that, at the beginning of the scheduling process (first scheduling slot), historical rates are set to zero, and PF^{NOMA} uses only instantaneous achievable throughputs to choose the best candidate user set. Therefore, users experiencing bad channel conditions have a low chance to be chosen. The corresponding achieved data rates are then equal to zero. On the other side, using the proposed scheduling, the treatment of the first scheduling slot is conducted differently and users are chosen depending on their actual rates (measured during the actual scheduling period). In this case, zero rates are eliminated. Hence, latency is greatly reduced.

For the next scheduling slots, historical rates are taken into account. For PF^{NOMA}, users experiencing a large $T_k(t)$ have less chance to be chosen, and may not be chosen at all. In this case, the use of buffering becomes mandatory and the size of the buffer should be chosen adequately to prevent overflow when peak rates occur, as a result of a high achieved throughput (high $R_{s,k}(t)$). Based on calculation, the average size of the buffer should be around 110 Mbit, for the simulation case at hand. However, in the case of the weighted proposed metrics, buffering is not needed, since only small variations between user data rates are observed, and a better QoE is achieved. Similar performance improvement is obtained for the orthogonal case in the same aforementioned conditions.

Finally, we have analyzed the effect of the proposed scheduling scheme on the cell-edge user throughput in Fig. 8. Again, the proposed weighted metrics outperform the conventional PF scheduling scheme for both access techniques, OMA and NOMA. In addition, WNOPF shows the best performance. Therefore, we can state that the incorporation of fair weights with a NOMA-based system proves to be the best combination.

In order to evaluate the performance of the proposed weighted metrics when premium services are considered, Tables 1 and 2 show the Gini fairness index values for two different scenarios, where three levels of services are requested: basic, silver, and gold. The number of users per group is set to 5.

Scenario 1:

The corresponding data rates of the three levels are set to 5 Mbps, 10 Mbps, and 15 Mbps respectively.

Scenario 2:

The corresponding data rates of the three levels are set to 10 Mbps, 20 Mbps, and 30 Mbps respectively.

In scenario 1, all users succeed in reaching their requested service data rates, and results of Table 1 show a high level of fairness achieved among users requesting the same service. However, when scenario 2 is applied, no success could be obtained but fairness is still maintained among users.

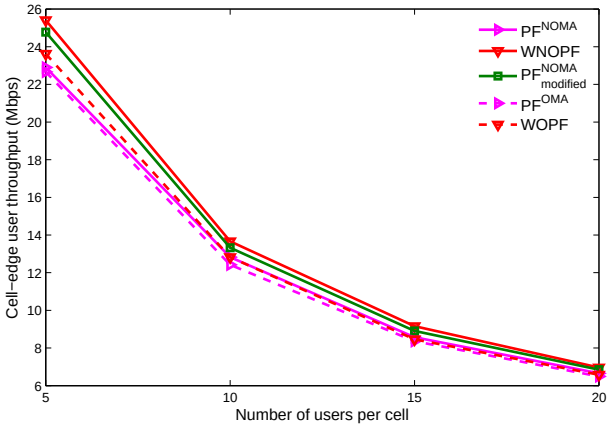

Fig. 8. Cell-edge user throughput vs. number of users per cell.

TABLE I
GINI FAIRNESS INDEX AND DATA RATE ACHIEVED PER GROUP FOR SCENARIO 1 (100% SUCCESS)

Service	Gini fairness index	Achieved data rate per group (Mbps)
Basic	0.0491	25.7
Silver	0.0724	51
Gold	0.0042	76.3

TABLE II
GINI FAIRNESS INDEX AND DATA RATE ACHIEVED PER GROUP FOR SCENARIO 2 (NO SUCCESS)

Service	Gini fairness index	Achieved data rate per group (Mbps)
Basic	0.0522	30.2
Silver	0.0613	49.6
Gold	0.0049	75.2

C. Computational Complexity

With the aim of assessing the implementation feasibility of the different proposed schedulers, we measured the computational load of the main allocation techniques to be integrated at the BS.

From a complexity point of view, the proposed scheduling metric WNOPF differs from the conventional PF metric in the weight calculation. For a number of users per subband limited to 2 in NOMA, the number of candidates per subband is $\binom{1}{K} + \binom{2}{K}$. When listing the operations of the proposed allocation technique, we obtain that the proposed metric

WNOPF increases the PF computational load by $\frac{26}{3}KS+S$ ($\simeq O(KS)$) multiplications and $-K^3S+\frac{3}{2}K^2S^2-\frac{4}{6}K^2S-\frac{3}{6}KS$ ($\simeq O(\frac{3}{2}K^2S^2-K^3S)$) additions.

In order to compute the PF metric for a candidate user set containing only 1 user, $4+S$ multiplications and $1+\frac{3}{2}S$ additions are needed. For each candidate user set containing 2 multiplexed users, $13+2S$ multiplications and $6+3S$ additions are required.

By taking account of the calculations of the terms $h^{-2\alpha}$, h^2, and $h^2/(N_0B/S)$ performed at the beginning of the allocation process, the classical NOMA PF requires a total of $3KS+C_K^1S(4+S)+C_K^2S(13+2S)$ multiplications which is equal to $K^2S^2+\frac{1}{2}KS+\frac{13}{2}K^2S$ ($\simeq O(K^2S^2)$) and $C_K^1S(1+3S/2)+C_K^2S(6+3S)$ additions which is equal to $\frac{3}{2}K^2S^2+\frac{1}{2}KS+\frac{13}{2}K^2S$ ($\simeq O(K^2S^2)$). Therefore, we can see that the increase in the number of multiplications in minor in comparison with that of the conventional PF, while the number of additions is almost doubled.

VIII. Conclusion

In this paper, we have proposed new weighted scheduling schemes for both NOMA and OMA multiplexing techniques. They target maximizing fairness among users, while improving the achieved capacity. Several fair weights designs have been investigated. Simulation results show that the proposed schemes allow a significant increase in the total user throughput and the long-term fairness, when compared to OMA and classic NOMA-based PF scheduler. Combining NOMA with fair weights shows the best performance. Furthermore, the proposed weighted techniques achieve a high level of fairness within each scheduling slot, which improves the QoE of each user. In addition, the proposed weighted metrics give the possibility of delivering different levels of QoS which can be very useful for certain applications. The study conducted here with two scheduled users per subband can be easily adapted to a larger number of users. Initially developed for single-antenna systems, the proposed scheduling technique can be extended to support multi-antenna systems. Such extension could be performed using our previous work in [29]. We are currently undergoing further research to reduce the complexity of the PF scheduler by introducing an iterative allocation scheme.

References

[1] 3GPP TS36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description".

[2] 3GPP TR36.913 (V8.0.0), 3GPP; TSG RAN; "Requirements for further advancements for E-UTRA (LTE-Advanced)", 2008.

[3] 3GPP TR36.814 (V9.0.0), "Further advancements for E-UTRA physical layer aspects", 2010.

[4] "Ericsson Mobility Report, on the pulse of the networked society", June 2017, available at: https://https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-june-2017.pdf

[5] G. Caire, and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel", *IEEE Trans. Inf. Theory*, 2003, 49(7), pp.1692-1706.

[6] T. Takeda, and K. Higuchi, "Enhanced user fairness using non-orthogonal access with SIC in cellular uplink", in *Proc. IEEE Vehicular Technology Conference (VTC) Fall*, 2011.

[7] K. Higuchi, and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink", in *Proc. IEEE Vehicular Technology Conference (VTC) Fall*, 2013.

[8] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation", in *Proc. Int. Symp. on Wireless Commun. Syst.*, 2012, pp. 476-480.

[9] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for Cellular future radio access", in *Proc. IEEE Vehicular Technology Conference (VTC) Spring*, 2013.

[10] J. Umehara, Y. Kishiyama, and K. Higuchi, "Enhancing user fairness in non-orthogonal access with successive interference cancellation for cellular downlink", in *Proc. Int. Conf. on Commun. Syst.*, 2012.

[11] Sharp corporation, "Evolving RAN towards Rel-12 and beyond, RWS-120039", 3GPP Workshop on Release 12 Onward, Ljubljana, Slovenia, 2012.

[12] J. Schaepperle, "Throughput of a wireless cell using superposition based multiple-access with optimized scheduling", in *Proc. IEEE Personal Indoor and Mobile Radio Communications (PIMRC)*, 2010, pp. 212-217.

[13] M.R. Hojeij, J. Farah, C. Abdel Nour, and C. Douillard, "New Optimal and Suboptimal Resource Allocation Techniques for Downlink Non-orthogonal Multiple Access", *Wireless Pers. Commun.*, 2015, pp. 1-31.

[14] E. Okamoto, "An improved proportional fair scheduling in downlink non-orthogonal multiple access system", in *Proc. IEEE VTC Fall*, pp. 1-5, 2015.

[15] M. Mehrjoo, M.K. Awad, M. Dianati, and S. Xuemin, "Design of Fair Weights for Heterogeneous Traffic Scheduling in Multichannel Wireless Networks", *IEEE Trans. Commun.*, 2010, 58(10).

[16] C. Gueguen, and S. Baey, "Compensated Proportional Fair Scheduling in Multiuser OFDM Wireless Networks", in *Proc. IEEE Int. conf. on Wireless & Mobile Comp., Netw. & Commun.*, 2008.

[17] C. Yang, W. Wang, Y. Qian, and X. Zhang, "A Weighted Proportional Fair Scheduling to Maximize Best-Effort Service Utility in Multicell Network", in *Proc. IEEE PIMRC*, 2008.

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

[19] S. Tomida, and K. Higuchi, "Non-orthogonal Access with SIC in Cellular Downlink for User Fairness Enhancement", in *Proc. Int. Symp. on Intelligent Signal Processing and Comm. Systems (ISPACS)*,2011, pp. 1-6.

[20] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access", in *Proc. Int. Symp. on Intelligent Signal Processing and Commun. Systems (ISPACS)*, 2013.

[21] B. Kim, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal Multiple Access in a Downlink Multiuser Beamforming System", in *Proc. Military Commun. Conf.*, 2013.

[22] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)", in *Proc. IEEE Personal Indoor and Mobile Radio Communications (PIMRC)*, September 2013.

[23] F.P. Kelly, A.K. Maulloo, Tan, and D.K.H., "Rate control for communication networks: shadow prices, proportional fairness and stability", *Journal of the Operational Research Society*, 1998, 49(3).

[24] P. Viswanath, D. Tse, and R. LAroia, "Opportunistic Beamforming using Dumb Antennas", *IEEE Trans. Inf. Theory*, 2002, vol 48(6), pp. 1277-1294.

[25] M. Kountouris, and D. Gesbert, "Memory-based Opportunistic Multi-user Beamforming", in *Proc. IEEE Int. Symp. Inf. Theory*, 2005.

[26] 3GPP, TR25-814 (V7.1.0), "Physical Layer Aspects for Evolved UTRA", 2006.

[27] 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 Globecom*, 2013.

[28] M. Dianati, X. Shen, and S. Naik, "A new fairness index for radio resource allocation in wireless networks", in *Proc. IEEE Wireless Commun. and Netw. Conf.*, 2005.

[29] M.-J. Youssef, J. Farah, C. Abdel Nour, C. Douillard, "Waterfilling-based Resource Allocation Techniques in Downlink Non-Orthogonal Multiple Access (NOMA) with Single-User MIMO", the 22nd IEEE Symposium on Computers and Communications (ISCC2017), July 2017.