Rate-Latency Optimization for NB-IoT with Adaptive Resource Unit Configuration in Uplink Transmission

Osama Elgarhy, Luca Reggiani, Hassan Malik, Muhammad Mahtab Alam, Muhammad Ali Imran, *Senior Member, IEEE*

***Abstract*—Narrowband Internet of Things (NB-IoT) is a cellular IoT communication technology standardized by $3rd$ Generation Partnership Project for supporting massive machine type communication and its deployment can be realized by a simple firmware upgrade on existing LTE networks. The NB-IoT requirements in terms of energy efficiency, achievable rates, latency, extended coverage, make the resource allocation, in a limited bandwidth, even a more challenging problem w.r.t. to legacy LTE. The allocation, done with sub-carrier granularity in NB-IoT, should maintain adequate performance for the devices while keeping the power consumption as low as possible. Nevertheless, the optimal solution of the resource allocation problem is typically unfeasible since non-convex, NP-hard and combinatorial because of the use of binary variables. In this paper, after the formulation of the optimization problem, we study the resource allocation approach for NB-IoT networks aiming to analyze the trade-off between rate and latency. The proposed sub-optimal algorithm allocates radio resource (i.e. sub-carriers) and transmission power to the NB-IoT devices for the uplink transmission and the performance is compared in terms of latency, rate, and power. By comparing the proposed allocation to a conventional Round Robin (RR) and to a brute-force approach, we can observe the advantages of the formulated allocation problem and the limited loss of the sub-optimal solution. The proposed algorithm outperforms the RR by a factor 2 in terms of spectral efficiency and, moreover, the study includes techniques that reduce the dropped packets from 29% to 1.6%.**

***Index Terms*—Maximum Throughput, Resource allocation, Power allocation, NB-IoT, Uplink scheduling.**

I. Introduction

Internet of Things (IoT) is becoming a fundamental part of our society as it brings the digitalization in every sector of life; smart homes, smart cities, smart health care, smart agriculture are just some of the examples in which IoT is playing a crucial role [1], [2]. However, in such digitalized ecosystems, wireless IoT devices are expected to grow exponentially and this continuous process requires more efficient ways for using the available spectrum. Therefore, the use of the spectrum should be optimized for the long-term sustainability of the future digitalized ecosystems. Therefore, the key question is *how to use efficiently the scarce spectrum resources for massive IoT devices connectivity?*

O. Elgarhy and L. Reggiani are with Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

H. Malik and M. Alam are with Thomas Johann Seebeck Department of Electronics, Tallinn University of Technology, Estonia

M. Imran is with James Watt School of Engineering, University of Glasgow, UK

In order to address this issue, 3^{rd} Generation Partnership Project (3GPP) introduced a cellular based technology named Narrowband Internet of Things (NB-IoT). NB-IoT inherits most of the design from the Long Term Evolution (LTE) system with a bandwidth requirement of 180 kHz. Due to this limited bandwidth availability, it can be deployed in three different modes, i.e. in-band (within the LTE band), guard-band (within the guard band of LTE) and standalone (exploiting the GSM band) [3]–[7].

NB-IoT aims to provide low power, long range communication for massive Machine Type Communication (mMTC). In order to enable such feature, the main design change in NB-IoT is the allocation of resources to the devices based on a sub-carrier level in uplink rather than the whole physical Resource Block (RB), typical of LTE allocation. In this regard, 3GPP has introduced the concept of Resource Unit Configuration (RUC), which is the combination of a number of Sub-Carriers (SCs), also refereed as tones in the specifications documents, on a TS duration. Like LTE, NB-IoT uses orthogonal frequency-division multiple access (OFDMA) in downlink and single carrier frequency-division multiple access (SC-FDMA) in uplink with sub-carrier spacing of 15 kHz, comprising 12 sub-carrier and 14 symbols in each TS of 1 ms; there is also the possibility of using a 3.75 kHz sub-carrier spacing in uplink, consisting of 48 sub-carriers and 14 symbols for a TS equal to 2 ms. In uplink, NB-IoT supports both single tone transmission or multi-tone transmission and the 3GPP specifications recommend tone allocations made of 1 tone (i.e. 1 sub-carrier for 8 ms), 3 tones (3 sub-carriers for 4 ms), 6 tones (6 sub-carriers for 2 ms) and 12 tones (i.e., 12 sub-carriers for 1 ms) with 15 kHz spacing. On the other hand, with 3.75 kHz spacing, only 1 tone transmission is recommended.

Nevertheless, such granularity in the resource allocation raises serious challenges for the radio resource scheduling, such as how to find the optimal tone configuration for each device and at the same time maximize the overall system performance in terms of spectral efficiency and quality of service. In order to address this challenge, the resource allocation algorithm needs a joint, multi-variable optimization strategy with multi-dimensional performance targets.

Furthermore, there are a lot of challenges from the resource allocation point of view, as the NB-IoT should be able to serve a particularly large number of devices. Furthermore, there are coverage, rate and latency constraints and, also, the necessity to address adaptively the quality of service of a

variety of applications and use cases, all in just a single, shared resource block. Of course, the other fundamental parameter to be optimized is the power consumption, typically associated to the battery life in stand alone NB-IoT devices. Finally, all these challenges are more significant in the uplink of the NB-IoT networks, since most of the traffic is typically in the uplink and, moreover, the NB-IoT devices are clearly much more limited in their energy and performance than the eNodeBs.

According to the above discussion, the problem statement faced here can be summarized as follows. Performance metrics such as throughput and latency are important for almost all the categories of NB-IoT systems: even if NB-IoT devices do not typically require high data rates, the throughput maximization is clearly crucial for serving them faster and increasing the potential of very high number of devices in urban areas in the available limited bandwidth resources. At the same time, even if latency is not typically considered as a key performance metric for most of the NB-IoT devices, the transmission repetition mechanism prescribed in the standard makes latency a factor that might significantly affect the number of connectable devices and the overall allocation efficiency. In addition, in NB-IoT there are some different and peculiar design constraint w.r.t. the standard terminals in the mobile system, in particular the resource unit configurations constraints and the really limited available resources. The idea behind this paper is to show that rate and latency optimization are two faces of the same problem and they can concur a remarkable increase of the overall system performance with an appropriate formulation of the allocation problem and of the related constraints. Therefore, the strategy followed in the sequel is composed by the following steps:

- The sum throughput maximization problem is formulated, with the addition of a minimum rate constraint for each device, the maximum power constraint and the resource unit configurations constraint, which is one of the challenging aspects of NB-IoT resource allocation and to try utilize the resource grid efficiently by using allocation shape constraint.
- The optimization problem is extended in order to include latency; however, the main component of the latency in NB-IoT, i.e. repetitions, turns out to be coherent with the throughput maximization approach, allowing the simplification of the problem.
- As the optimal solution for the proposed problem is unfeasible, some simplifications are justified and introduced in order to provide a sub-optimal solution for the optimization problem.
- The different variables of the allocation problem, including those related to a crucial transmission mechanism in the NB-IoT standard (repetition and its triggering parameters), are analyzed in order to appreciate their impact on the performance targets, throughput and latency.

The final sub-optimal allocation algorithm is presented in two versions. The former is more sophisticated and challenging w.r.t. a real implementation but it is used here as a benchmark for demonstrating the advantages of the proposed optimization approach, considering that an optimal solution is not feasible. The latter removes the part that poses some challenges for the real implementation. In Sect. IV-E, the aspects related to practical implementation and signaling are discussed.

The rest of the paper is organized as follows: after a review of the state of the art and a detailed description of the novel aspects of the proposed approach in Sect. II. A description of the system model and the optimization problem formulation are given in Sect. III. Sect. IV discusses the proposed sub-optimal solution. Then Sect. V contains the simulation results with their analysis and discussion; finally, the conclusions are given in Sect. VI.

II. Related works and the novel contributions

The related works include contributions in the NB-IoT resource allocation, throughput optimization, and latency optimization.

Several papers have studied resource allocation and scheduling for NB-IoT. Some of these papers did not develop optimization functions and others tested known schedulers according to given criteria without introducing strategies specifically designed for NB-IoT, exploiting adaptively, for example, the multiple RUCs in NB-IoT. In [8], the focus is on the massive number of devices and the effect of the control plane optimization on the data scheduling has been investigated with some sort of a greedy algorithm; however, there is no optimization problem for either throughput or latency as a function of the transmission powers. In [9], the authors have reached the conclusion that channels, such as the random access channel (NPRACH), uplink shared channel (NPUSCH), downlink shared channel (NPDSCH), downlink control channel (NPDCCH), should not be scheduled separately. Furthermore, they proposed a tractable queuing model in order to study the effect of scheduling on latency and battery lifetime; however, the multi-tone allocation and the throughput optimization are not considered. Then, the performance of each individual RUC has been studied in [10] for different types of traffic and using three schedulers: round robin, proportional fair and maximum throughput; the scheduler was designed for each resource unit configuration separately, in order to measure the difference in performance between all of them but not for selecting the best configuration with adaptive allocation, according to an optimization goal. In [7], it has been proposed a two level link adaptation, composed by an inner loop for adjusting periodically the repetition number and an outer loop for selecting the MCS and deciding the repetition number. The link adaptation uses a long open loop power control, analyzed with only single-tone uplink scheduler; in this paper the resource unit is already allocated and there is no adaptive allocation or throughput optimization problem for resource unit allocation. In a single cell, a First In First Out (FIFO) scheduler with one RUC has been used in [11] and the analysis of resource utilization and average delay takes into account the scheduling delay; however, the uplink scheduling was designed for a single tone resource unit without adaptive allocation and for a single cell, without intercell interference. The authors of [12] have tested the different RUCs separately and developed a scheduling technique based on maximum

allowed latency (maximum delay tolerance), comparing it to FIFO and Minimum Transmission Time (MTT) with the number of served devices as the performance measure. For a single cell and non-adaptive RUC allocation. In [13] for single cell, as well, the authors have developed an adaptive allocator that minimizes the consumed resources. Finally, in [14], an adaptive RUC allocation has been used and, for addressing the energy efficiency optimization, the authors have proposed a heuristic algorithm composed of two steps: first, finding the configuration parameters that minimize energy consumption and satisfy QoS requirements and, second, optimizing these parameters in order to respect latency constraints.

A. Novel contributions

In this paper, we formulate a resource allocation optimization problem whose objective is to maximize the uplink achievable rate maintaining the lowest possible latency and taking into account, obviously, the transmitted power constraints of the devices. This optimization is based on the weighted sum throughput maximization problem and it turns out to be non-convex, NP-hard and combinatorial because of the use of binary variables. Hence, we propose a sub-optimal algorithm for its practical solution: the algorithm splits the problem into two problems, the uplink scheduling and the power allocation. For the power allocation, we exploit a modification of a global optimization technique, known as MAPEL [15] for distributing the interfering powers and the water filling principle for distributing the power of each device among its assigned SCs.

Moreover, we study the performance of the different RUCs and we use an adaptive RUC allocator, i.e. able to select the best RUC instead of using just a predefined one; this issue is also related to the effect of the shape of the allocation grid on the overall performance and we compare the performance of the proposed algorithm to some basic strategies, such as the round robin and the maximum power allocation. Since the 3GPP specifications introduce the repetition mechanism as one of the key-enablers for coverage enhancement in NB-IoT, its impact on performance and latency will be measured as well. As the granularity level in the NB-IoT is deeper than that in the conventional LTE, the allocation is done on the SC level instead of the RB. It is very natural to have a completely different interference on SCs belonging to the same device and most of the related works consider a single cell scenario, without the effect of interference on the performance. Moreover, a new factor of optimization has been considered: since we have multi-tone allocation and a maximum allowed power per device, an optimal power distribution operated among the SCs allocated to each single device is proposed and evaluated. Therefore, the novel contributions can be summarized as follows:

- The study of the techniques that can be used to distribute the transmitted power among the sub-carriers of each device, with their fundamental impact on their power consumption. In this context, we have used and compared the water-filling principle, for optimizing Power Distribution among the SCs of the same Device (PDSD), and the global optimization technique MAPEL, for optimizing power allocation among interfering devices. To the best of our knowledge there is no study in the literature on the impact of the PDSD on the performance. However, it will be shown that it affects greatly the main performance indicators, such as latency, throughput and packet drop rate.
- The investigation of the impact of the repetition triggering techniques, i.e. how to the decide the multiple transmissions of a RU. We will show that this part is essential for applications sensitive to latency, even though there are not specific studies about this issue.
- The formulation of the sum throughput optimization problem for NB-IoT, with the explicit inclusion of the RUC constraints.
- The study of the relation between throughput optimization and latency; to the best of our knowledge, it is the first time that a study on the joint optimization of these two performance indicators is presented.
- The proposal of an allocation approach, for the sum throughput optimization problem for NB-IoT. This novel adaptive allocator for NB-IoT is used as a sub-optimal solution for the aforementioned optimization problem.
- The application and comparison among different allocation and power distribution techniques, developed or selected from the literature for the proposed approaches and for having some performance benchmarks. Table I reports the list of the techniques considered in the paper and described in the dedicated Sections.

TABLE I
THE ALLOCATION TECHNIQUES ANALYZED IN THE PAPER

1. DAL (aDaptive ALlocator)	Adaptive resource allocator, based on best RUC selection and user diversity exploitation.
2. MAPEL [15]	The DAL is combined with a modified version of the MAPEL [16] for managing power allocation between the interfering users.
3. PDSD	The proposed approaches for distributing the power of each device: equal power distribution, water filling and max-min SINR.
4. ASC (Allocation Shape Constraint)	This proposed technique reduces the holes in the allocation grid due to RUCs with different numbers of SCs and sub-frames.
5. Repetition triggering	Two triggering mechanisms: (i) minimum SINR and (ii) maximum SINR within the RU.
6. RR (Round Robin)	Baseline scheduler for Throughput testing.

III. Mathematical Model and Problem Formulation

The methodology followed in the problem formulation is described by the following two steps:

- in Sect. III-A, the sum throughput maximization problem for the uplink of a set of cells is expressed taking into account all the interference relations (channel gains) among the devices and the eNBs. This assumption configures

TABLE II
MAIN ABBREVIATIONS

Abbreviation	Definition
DAL	aDaptive ALlocator
RR	Round Robin
RRc	Round Robin with configuration diversity
ASC	Allocation Shape Constraint
MAPEL	MLFP-bAsed PowEr aLlocation
PDSD	Power Distribution among the SCs of the same Device
MAMI	MAx-MIn SINR algorithm
BF	Brute Force search
TS	Time Slot

TABLE III
SUMMARY OF THE MAIN NOTATIONS

Notation	Definition
C	The set of the cells in the system, i.e. the eNBs
c	Index of a cell belonging to C
U_c	The set of devices within cell c
u	Index of a device belonging to a generic U_c
T	The set of available sub-frames
t	Index of a given sub-frame belonging to T
S	The set of the available sub-carriers
s	Index of a given sub-carrier belonging to S
$x^u_{c,t,s}$	Binary allocation variable for device u in cell c at slot t, sub-carrier s
$R^u_{c,t,s}$	The rate of device u
$G^u_{c,c,t,s}$	The channel gain of device u w.r.t. to its eNB c
$G^u_{a,c,t,s}$	The channel gain of device u, affiliated to eNB c, w.r.t. to another eNB $a \in C$
$P^u_{c,t,s}$	The transmission power of device u

the optimal, centralized control of the resource allocation with the maximum complexity. The optimization problem is completed by the set of specific constraints for NB-IoT: non-overlapping devices in the same cell, power, allocation and shape of the resource units.

- In Sect. III-B, the latency is divided into 3 components and, for the first 2 the minimization problem is formulated, giving raise to a multi-objective optimization function. It is anticipated also that the third component turns out to be dominant and it is minimized when the Signal-to-Interference and Noise Ratio (SINR) is maximized. Therefore, the dominant component of the NB-IoT latency is minimized when the rate is maximized, given that the rate is a logarithmic function of the SINR; obviously this constitutes one of the basis for the simplifications introduced in the next Sect. IV for achieving some numerical results.

Table III lists the main variables and notations.

A. Rate maximization

Here we formulate the problem for finding the optimal uplink power and scheduling according to the maximization of the sum throughput. Let us consider the following optimization function:

$$\max f_R(x^u_{c,t,s}, P^u_{c,t,s}) = \sum_{c\in C}\sum_{t\in T}\sum_{u\in U_c}\sum_{s\in S} x^u_{c,t,s} R^u_{c,t,s}$$

subject to

$$\sum_{t\in T}\sum_{s\in S} x^u_{c,t,s} R^u_{c,t,s} \geq r^u_c,\ \forall u \in U_c, \forall c \in C \quad (1)$$

In (1), the rate [bit/SC/symbol] of device u in cell c at time slot t and SC s is computed by

$$R^u_{c,t,s} = log_2\left(1 + \frac{P^u_{c,t,s} G^u_{c,c,t,s}}{\sum_{a\neq c, a\in C}\sum_{j\in U_a} P^j_{a,t,s} G^j_{a,c,t,s} x^j_{a,t,s} + P_n}\right) \quad (2)$$

where $x^u_{c,t,s}$ is a binary allocation variable within the allocation matrix X, which indicates that a device u, belonging to the set of devices U_c in a cell c of the set C, is allocated a time slot t within the scheduling interval T, in a sub-carrier s from the available sub-carriers S. The device transmitted power is $P^u_{c,t,s}$ and $P^j_{a,t,s}$ is the transmitted power of the interfering device j belonging to cell a. $G^u_{c,c,t,s}$ is the channel gain between device u belonging to cell c and its eNB at the given resource element (t,s). $G^j_{a,c,t,s}$ is the channel gain between device j belonging to cell a and eNB c, at the given resource element (t,s). The term P_n denotes the noise power at any eNB. In the optimization function (1) and the related (2), it is clear that the problem has two fundamental parts, the scheduling one (the allocation of the resources in the grid, or $x^u_{c,t,s}$) and the power allocation one (the transmit powers $P^u_{c,t,s}$). This twofold perspective will be the basis of the sub-optimal solution proposed in the next Sect. IV.

In addition, we formulate the following constraints.

1) Non-overlapping devices in the same cell:
$$\sum_{u\in U_c} x^u_{c,t,s} \leq 1,\ \forall s \in S, \forall t \in T, \forall c \in C, \quad (3)$$
which guarantees that the resource elements are used just by one device within the same cell or there is no intra-cell interference.
2) Power constraint:
$$\sum_{s\in S} x^u_{c,t,s} p^u_{c,t,s} \leq P_{max},\ \forall u \in U_c, \forall t \in T, \forall c \in C, \quad (4)$$
which means that the sum of power at a given time t for the same device should not exceed P_{max}, where P_{max} is the maximum allowed power per device in the uplink.
3) Constraints on the allocation and shape of the resource units, which force the optimizer to choose one of the 4 RUCs or none. In the time dimension, for device u at a given time slot t, the term
$$\sum_{s\in S} x^u_{c,t,s} = N^u_{c,t},\ \forall t \in T, \forall u \in U_c, \forall c \in C \quad (5)$$
is the number $N^u_{c,t}$ of SCs allocated to the device and
$$\sum_{s=1}^{end-1} |(x^u_{c,t,s} - x^u_{c,t,s+1})| \leq 2,\ \forall t \in T, \forall c \in C. \quad (6)$$
In the frequency dimension, for device u at a given sub-carrier s, the term
$$\sum_{t\in T} x^u_{c,t,s} = M^u_{c,s},\ \forall s \in S, \forall c \in C \quad (7)$$
is the number $M^u_{c,s}$ of time slots allocated to the device and
$$\sum_{t=1}^{end-1} |(x^u_{c,t,s} - x^u_{c,t+1,s})| \leq 2,\ \forall s \in S, \forall c \in C. \quad (8)$$
The relations in (6) and (8) impose that the allocated time slots and sub-carriers should be consecutive. In addition, introducing the binary variable $v^u_{c,q}$ (q =

$\{1, 2, 3, 4\}$ as the index for the 4 RUCs, we can state that

$$\sum_{q} v^u_{c,q} = 1, \tag{9}$$

since each device has just one assigned RUC. In addition, the values $N^u_{c,t}$ and $M^u_{c,s}$ can assume, at each slot t or SC s, only two values: 0 if the slot or the SC are not allocated to that user or the values z^u_c, q^u_c, which are the possible sides of the RUCs in the time and frequency domain respectively. Now we force the number of allocated SCs per time slot to be equal to a value z^u_c per user, i.e.

$$(N^u_{c,t})(N^u_{c,t} - z^u_c) = 0, \ \forall t \in T, \forall c \in C \tag{10}$$

The above equation make sure all $N^u_{c,t}$ per user are equal to z^u_c. Then, to force the number of allocated time slots per SC to be equal to a value q^u_c per user:

$$(M^u_{c,s})(M^u_{c,s} - q^u_c) = 0, \ \forall s \in S, \forall c \tag{11}$$

Therefore, considering just 1 RUC per user, we have that

$$z^u_c = 1v^u_{c,1} + 3v^u_{c,2} + 6v^u_{c,3} + 12v^u_{c,4}, \tag{12}$$

$$q^u_c = 8v^u_{c,1} + 4v^u_{c,2} + 2v^u_{c,3} + 1v^u_{c,4}. \tag{13}$$

Equ. (9)-(13) characterize the constraints on the 4 RUCs: (9) guarantees that only one value of $v^u_{c,q}$ is 1, while the values from (12) and (13) force the total numbers of SCs and time-slots per user to be equal to one of the predefined RUC shapes.

The following issues make this problem not solvable: (i) due to (3), i.e. sharing the same resource within the same cell is prohibited, the problem is NP-hard [17], (ii) the objective function is the well known sum throughput maximization, which is non-convex, and (iii) the use of the binary variables $x^u_{c,t,s}$, $v^u_{c,q}$ make the problem combinatorial (this last difficulty can be faced by using time sharing property as in [18], [19], [20]). Therefore, the optimal solution cannot be computed and this has motivated the search of some simplifications and heuristic methods, in order to propose a sub-optimal, feasible solution, as shown in Sect. IV.

B. Latency minimization

In order to include also the latency into the optimization problem, we have considered the different components of the latency:

1) Scheduling waiting time.
2) Transmission time of the assigned RUC.
3) Time due to repetitions of the transmission used for the coverage enhancement.

The first and second components can be included in the optimization problem, by considering the function, to be minimized,

$$\min f_L(x^u_{c,t,s}) = -\sum_{c\in C}\sum_{t\in T}\sum_{u\in U_c} t \cdot \frac{\sum_{s\in S}(x^u_{c,t,s} - x^u_{c,t-1,s} - |x^u_{c,t,s} - x^u_{c,t-1,s}|)}{2N^u_{c,t-1} + \epsilon}, \tag{14}$$

which is the minimization of the sum of the latencies due to the scheduling process; in fact, when the binary allocation variable is different from 0 at any time slot, the term in the internal sum of (14) intercepts the transitions of the variable $x^u_{c,t,s}$ from the value 1 to 0 for any $\{u, c, t\}$ taking into account the rectangular shape of each RUC (when different from 0, each RUC starts and finishes with the same number of ones in the SC dimension s and ϵ is a very small constant for avoiding the division by zero). The new formulation gives raise to a multi-objective optimization, which can be expressed by the minimization of the following function

$$\min f_{RL}(x^u_{c,t,s}, P^u_{c,t,s}) = - K_R \cdot [f_R(x^u_{c,t,s}, P^u_{c,t,s})]^2 + K_L \cdot [f_L(x^u_{c,t,s})]^2, \tag{15}$$

where K_R and K_T are the weights for the two terms, equal both to 1 if rate and latency are considered equally important.

Finally, the third component of the latency is impossible to be expressed in a cost function since it is function of the repetition mechanisms of the NB-IoT network. However this component is also coherent with the rate and, consequently, the SINR maximization principle, since the probability of retransmission or the number of prescribed repetitions are minimized when the SINR is maximized. Therefore, the problem formulation in (1) and (2) is also addressing the minimization of the third component of the latency; in Sect. IV-D and V-C we will justify how this third component can be considered the dominant one in this study and we will present a sub-optimal strategy specifically devoted to its minimization.

IV. The Sub-optimal Solution

Here we develop a sub-optimal solution for the problem in Sect. III; the motivation is to achieve a reasonable implementation of the optimization problem in order to provide the numerical results necessary for appreciating the potential advantages of resource allocation strategies for NB-IoT and the way for its practical implementation. The adopted techniques, anticipated in Table I, respond to the problem formulation and the related constraints according to the following rationale.

- Rate maximization w.r.t. SINR measured in the different channels and slots w.r.t. (1)-(3) by using DAL and possible power allocation among the resources assigned to each single device (Sect. IV-A).
- Rate maximization w.r.t. power allocation among mutually interfering devices by using MAPEL (Sect. IV-C).
- The constraint on the power distribution for the same device in (4) respected by means of the adopted PDSD techniques; in particular, the rate optimization is pursued by using the water filling principle among the PDSD techniques (Sect. IV-A).
- The shape constraints in (5)-(13) respected by the allocator rules; a strategy for optimizing their allocation taking into account their different shapes is proposed by means of the ASC (Sect. IV-B).
- The latency optimization by neglecting the scheduling and transmission time components and including the repetition mechanism, recognized as the dominant one. Therefore, the inclusion of (14)-(15) in the sub-optimal solution is not necessary.

TABLE IV
PROCEDURE FOR SUB-OPTIMAL NB-IOT RESOURCE ALLOCATION

Resource allocation procedure
A. Scheduling and device power distribution (DAL)
1) Start from the first available resource unit (RU) and first device.
2) Test the 4 RUCs and select the one with the highest rate. The test computes the SINRs using the PDSD techniques (in particular WF) to distribute the power in case of multi-tone RU.
3) Repeat (2) for all the devices.
4) Schedule the RU to the device with the highest rate, taking into account the ASC.
5) Repeat from (1) until all the devices are scheduled.
B. After the uplink scheduling, the modified MAPEL finds the optimal power allocation among interfering devices.

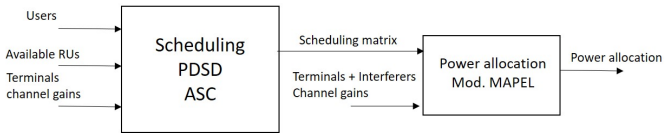

Fig. 1. Simple flowchart of the procedure for NB-IoT resource allocation.

Therefore, the proposed sub-optimal solution is done by splitting the optimization problem into two sub-problems, a scheduling and a power allocation problem. About the scheduling, we use a heuristic search algorithm combined with the water filling principle for distributing the power of each device among its own SCs. For the power optimization problem, a optimal global optimization technique called MAPEL [15] is used and, in particular, a modified version of it, developed in [16] for decreasing the computational load.

The whole algorithm is summarized in Table IV and a simple illustration highlighting the necessary inputs and outputs is in Fig. 1. It is possible to notice immediately the two fundamental blocks, the former devoted to a scheduler rate maximization based on the SINRs and rate requirements and the former devoted to an optimization of the transmit powers that takes into account the mutual interference impact in the uplink. In the sequel, we provide a detailed description of each part of the overall process. Finally, in Sect. IV-E, it is discussed the aspects related to the information necessary for this resource allocation procedure and its practical implementation.

A. Adaptive allocation and power distribution

The first part of the proposed solution has been called aDaptive ALlocator (DAL), for performing the uplink scheduling. Its basic operations are:

- Since there are 4 possible RUCs, starting from the first available point in the resource grid, these RUCs are tested for each device in terms of the achievable rate in order to find the best RUC. This operation is repeated for all the devices at each point in the resource grid and the best RUC among all the devices is selected.
- In case of multi-tone transmission, each device has multiple SCs per TS and the maximum transmit power per device needs to be distributed among these SCs; the water filling (WF) is the default operation associated to the DAL in order to have a power distribution that achieves maximum sum throughput. It is worth noting that we have tested other techniques for the device power distribution, such as equal power distribution and power distribution that maximize the minimum SINR. The results will clarify the advantage of WF.
- After the allocation for the best device, the next free point in the resource grid is selected and the same procedure is repeated until all the devices are allocated.

It is worth noting that the allocator is a form of coordinated, centralized strategy since it knows the SINRs of all the devices in the given resource grid or, equivalently, the channel gains of the devices in a group of cells, as shown in Fig. 1 for performing the optimization.

B. Allocation Shape Constraint

Since the algorithm can select different RUCs for each device and each RUC has a different shape, the final occupation of the resource allocation grid could not be optimal since there will be some unassigned SCs or holes in the grid. This particular operation takes into account this shape issue, forcing (or not) the allocator to fill these holes: in practice, when a device is selected with the best achievable rate, the best RUC is considered the one that fits better the space left free from the previous allocations (as the RUCs for the same device occupy adjacent SCs and slots, their SINRs will be similar and a priority will be given to the best shape fit). We refer to this operation step as *Allocation Shape Constraint* (ASC).

C. MAPEL: Global optimization for power control problem

One of the known methods for finding the global optimum solution for the weighted throughput sum rate maximization problem is the MLFP-bAsed PowEr aLlocation (MAPEL), where MLFP stands for multiplicative linear fractional programming. MAPEL is based on polyblock outer approximation and the problem is transformed from a weighted throughput sum rate maximization to a product of linear fractional functions. The MAPEL is used to find the optimal power allocation for a group of interfering devices. For a system with a given number of links, it is necessary to know all the channel gains, including those of the interfering signals, and the receiver noise powers; in our case, these gains are clearly the uplink channels between each device and the eNBs of the multi-cell system and they realize, as already observed, a form of coordinated, centralized allocation strategy.

Therefore, the MAPEL algorithm role is to find the optimal power allocation for the interfering devices in the multi-cell system and it gives the optimal power allocation for the devices interfering on the same shared channel, e.g shared SCs or RBs in different cells. More details about the algorithm and its modifications for increasing the computational speed can be found in [15], [16].

D. Latency

In order to include also the latency into the allocation problem, in Sect. III-B we have introduced its three components, i.e. the scheduling waiting time, the transmission

time and the number of repetitions. These sources have a different origin and weight in the determination of the final latency. The first component depends on the traffic and a scheduler that allocates serially the devices (excluding those already served, so with 1 RUC for each device as forced by the constraint (9)) makes the minimization of this term not relevant at least approximately. In fact, the difficulty of (14) and (15) can be skipped by considering that, under medium - high SINR conditions (coherent with the overall SINR maximization objective), the scheduling waiting time for each device will be approximately given by the total number of available RUs divided by the number of devices in the queue, namely a term approximately constant and consequently not subject to a real minimization.

The second component, the pure transmission time, depends on the particular RUC that is assigned to a device, from the minimum of 1 time slot (TS) to the maximum, equal to 8 TSs. On one hand, it is clear that an a-priori use of the RUCs with 8 SCs and lasting 1 TS would minimize the average latency, at the expense of the scheduler adaptivity, which is relevant for the rate maximization as it will be shown in the results. Therefore, in order to not affect the RUC adaptivity, we have assumed that this term is not specifically minimized in the general problem faced here and all the devices will be subject, on the average, to the same transmission time, given by the distribution of the allocated RUCs. On the other hand, in a real implementation, the presence of devices with stringent latency requirements can be easily solved associating them just to the RUC with the minimum transmission time.

Finally, the third component acquires here the key role for the latency minimization since the transmission time due to the repetitions spans multiple TSs and the number of repetitions can reach up to 128 RUs, according to the standard. Therefore, this component turns out to be the dominant one and, even if excluded from the theoretical formulation, it finds in the sub-optimal implementation, a feasible way for its practical solution: as the SINR per SC determines the number of repetitions, the maximization of the SINR per SC is coherent either with the rate maximization or with the minimization of the number of repetitions and, consequently, of this latency component. One of the crucial aspects of the repetition process is to decide when to trigger this process into the transmission: minimizing the activation or triggering of the process minimizes the latency. We have considered two options looking at the SINRs obtained at the SCs assigned to each user; in fact, due to channel and interference frequency selectivity, each SC has generally a different SINR.

1) Option *min-sinr*: to trigger the repetition when the SC with the minimum SINR is below a threshold.
2) Option *max-sinr*: to trigger the repetition when the SC with the maximum SINR is below a threshold.

The threshold is set to the lowest SINR acceptable in the system ($SINR_{min}$), below which the allocator cannot assign any type of RUCs. Therefore, the latency is increased when the allocator is not able to guarantee this minimum SINR level.

The *min-sinr* strategy is meant to be more prudent: the simulation results will show that the best performance is achieved by maximizing the minimum SINR when we distribute the transmission power of the UE among its SCs. On the other hand, in the case of *max-sinr*, the WF is the best technique to be used, since it distributes the power in a way that maximizes the sum throughput, so enhancing also latency. Therefore, as a matter of fact, we can observe that latency is primarily affected by the PDSD part of the scheduling process since it has immediate impact on the repetition triggering mechanisms.

E. System implementation and signaling

The sub-optimal algorithm, sketched in Fig. 1, is formed by two fundamental blocks: the former, composed by the scheduling (DAL), ASC and PDSD functions, is directly implementable in a real system since the complexity is limited and it exploits information that are available in real systems at the eNBs, i.e. the uplink channel estimations and the related SINRs for the different affiliated devices. On the contrary, the latter, which performs the modified MAPEL algorithm, poses some challenges for a real implementation, mainly due to the type of information needed for running the algorithm: in fact, it is necessary the knowledge of all the channel gains, including those between the interfering devices and the eNBs. This knowledge is not currently supported by the standard through specific control channels and it is generally more challenging to obtain. However, the implementation of this part of the sub-optimal solution is not impossible: it would require a central unit sharing the channel gains estimations, at least of the strongest signals, among groups of adjacent base stations and performing the resource allocation for the overall group of cells in a cooperative and centralized approach. This type of coordination is expected to acquire more relevance for the efficiency improvement of 5G networks; in addition, the required channel gains, i.e. the values $G^u_{a,c,t,s}$ between each device and the eNBs belonging to the considered set of cells C (Table III), can be estimated just allowing some form of coordination among the adjacent eNBs, e.g. transmitting the reference symbols of the devices in different resource slots. As mentioned in Sect. V-E, these aspects will be subject of future investigation.

Therefore, with the aim of a practical system implementation, we currently distinguish between the following algorithm levels:

- Sub-optimal algorithm with integrated MAPEL, not immediately applicable for implementation in all current networks but used here for obtaining a numerically tractable procedure being as close as possible to the optimal problem formulation. This algorithm is used as a benchmark and it is denoted as DAL-MAPEL.
- Sub-optimal algorithm w/o integrated MAPEL, suitable for a direct system implementation in current networks, simply denoted as DAL with the possible integration of ASC and PDSD functions.

In Table V we summarize the two algorithm levels with the implications on the network signaling. After the allocation processing, the eNBs should forward to each NB-IoT device the allocation information for the uplink transmissions according to the computed resource allocation frequency/time grid.

TABLE V
THE ALGORITHM IMPLEMENTATION LEVELS AND RELATED SIGNALING.

Level implementation	Use	Signaling
1. DAL + MAPEL	Benchmark. Future applications with eNBs coordination.	Channel gains $G^{u}_{a,c,t,s}$ between each device and the eNBs belonging to a set of cells C with centralized processing.
2. DAL	Real network	Channel gains between each device and its eNB with decentralized implementation of the allocation at each eNB.

TABLE VI
SIMULATION PARAMETERS

Channel model	
Attenuation (d = distance [m], f = frequency [GHz])	$(44.9 - 6.5 \cdot log(32)) \cdot log(d) + 34.46 + 5.83 \cdot log(32) + 23 \cdot log(f/5)$
Shadowing	Standard deviation $\sigma = 8$ dB
Multipath	WINNER phase 2 model, Urban macro
Radio network	
Cell layout	7 hexagonal cells (in the set C)
Cell radius	250 m (Inter eNB distance 500 m)
Bandwidth	one RB = 180 KHz, number of elements in S is 12 SCs
Max UE transmit power	24 dBm
Min. SINR for RUC assignment	$SINR_{min} = 1.74$ dB
Scheduling	RR (+PDSD), DAL (+PDSD, ASC)
eNodeB	Single antenna
Number of users $\times$ cell	from 10 to 55

V. PERFORMANCE EVALUATION AND SIMULATION RESULTS

The algorithm is simulated under the conditions and parameters assumptions listed in Table VI. All the cells share obviously the same values and they operate with a full traffic queue. For testing the system under challenging interference conditions, we consider a dense urban macro environment with 7 adjacent cells, each with 3 sectors and radius equal to 250 m; devices are uniformly distributed in each cell. The 95% confidence interval of the presented numerical results turns out to be less than 2% of the mean values, with less precise results for the MAPEL, which is much heavier from a computational point of view and difficult to simulate for a large number of channel realizations. Unless specified differently, the number of channel realizations for each simulation is equal to 100 and 25 for the MAPEL. The spectral efficiency is computed selecting adaptively the best modulation and code in each RUC, according to the available levels in NB-IoT.

The following results represent the validation of the proposed techniques and the basis for a discussion about the impact of the different components of the scheduling and power allocation process.

1) Optimal formulation: a brute force (BF) approach on the original problem formulation (Sect. III) is adopted for achieving a reference and evaluating the sub-optimal solution.
2) Sub-optimal solution: impact of resource and power allocation on the *throughput* (Sect. V-A, without repetition triggering):
   - The DAL is compared to a standard RR (operating with a fixed RUC).
   - The DAL is compared to an adaptive version of RR (able to select the best RUC for the current allocated device), and DAL + MAPEL.
   - The ASC is evaluated separately.
3) Sub-optimal solution: impact of PDSD on the *throughput* and *latency* (Sect. V-C, with repetition triggering, responsible of the dominant latency component):
   - The DAL is compared to the adaptive RR with different types of PDSD in case of min-sinr repetition triggering.
   - The DAL is compared to the adaptive RR with different types of PDSD in case of max-sinr repetition triggering.

W.r.t. the system implementations, as discussed in Sec. IV-E (see also Table V):

- The BF curves have to be considered benchmarks for performance validation and comparison.
- The DAL+MAPEL curves have to be considered potential performance improvements for future system implementations.
- All the other DAL and RR curves with their several options and integrations (PDSD, ASC, triggering) can be considered immediately suitable for a real implementation.

Before the presentation of the numerical results, it is useful to summarize the 4 key points of the scheduling and allocation process in Table IV, responsible of the main impact on the measured throughput and latency. Each of these four key steps has a different impact on performance and some of them can compensate the gain or loss of the others:

1) From step 2 of the algorithm: the power is distributed among the SCs of the same device according to the PDSD techniques. This operation will distribute power among the SCs in order to achieve a rate maximization and, at the same time, a latency minimization: we will study WF, equal power distribution, and max-min SINR.
2) From step 2, we select the best configuration among the 4 RUCs, observing that each configuration occupies a different part of the allocation matrix, with a different diversity effect for each device. We refer to this type of diversity effect as *configuration diversity*.
3) At step 4, the RU is given to the device that achieves the highest performance target, i.e. rate in the DAL. This realizes a form of the well known *device or user diversity*.
4) The effect of the RUC shapes on the allocation grid, i.e. the number of unused RUs. ASC is responsible for the mitigation of this effect.

A. Impact of allocation on throughput

Here we focus on the impact of the scheduling and allocation process without considering the repetition triggering mechanism; this choice allows to compare the strategies impact on the throughput leaving, as discussed in Sect. IV-D,

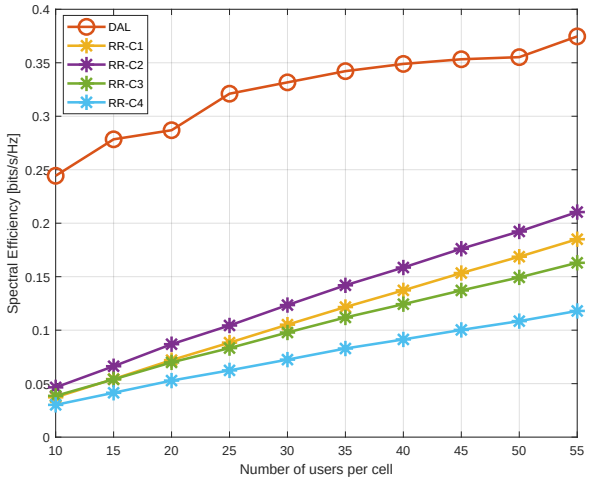

Fig. 2. Spectral efficiency per cell: comparison between DAL and RR for different numbers of devices per cell. *RR-Cq* is a standard RR operating with a fixed RUC q ($q = \{1, 2, 3, 4\}$).

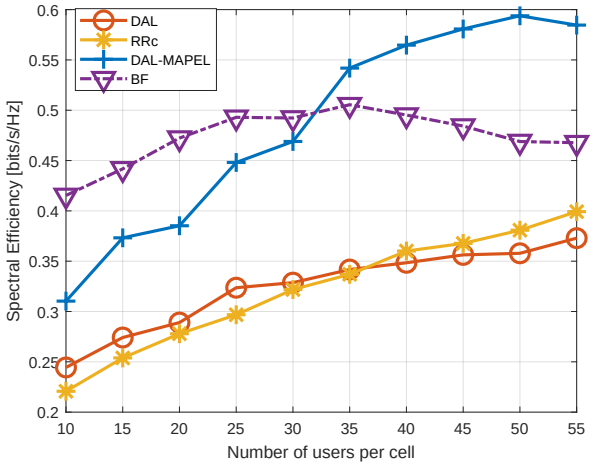

Fig. 3. The spectral efficiency per cell with BF, DAL integrated with MAPEL, DAL without MAPEL and RRc.

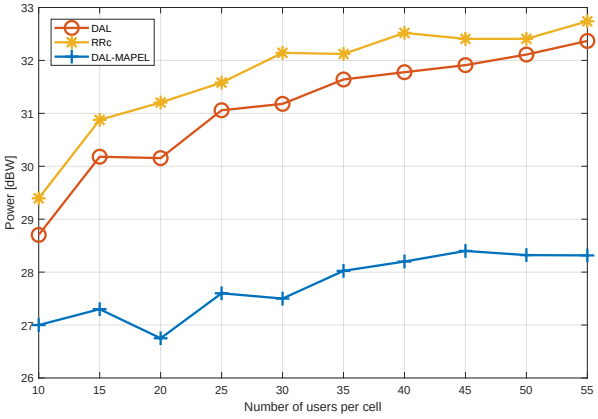

Fig. 4. The total consumed power (for all the 7 cells) with DAL integrated with MAPEL, DAL without MAPEL and RRc.

the impact on latency to the case of repetitions.The DAL is studied with and without the MAPEL and compared with a RR allocator to highlight the potential gain. The spectral efficiency is simulated for a simple RR that uses fixed allocation, i.e. one of the possible 4 RUCs. The comparison is done for all the 4 possible configurations.

In Fig. 2 the spectral efficiency is simulated for different numbers of devices per cell. It can be noticed that DAL outperforms the RR; the acronym *RR-Cq* indicates a standard RR operating with a fixed RUC q ($q = \{1, 2, 3, 4\}$, see (9)-(13). Furthermore, the DAL provides adaptively the flexibility for selecting the configuration that fits better the different design criteria, i.e. latency and throughput.

In order to understand the impact of MAPEL in the DAL performance, we have to consider the configuration diversity effect also in the RR, using an adaptive RR (RRc), able to select the best RUC for each served device. It turns out that the RRc performance is very similar to the DAL w/o MAPEL just because of the configuration diversity that appears the dominant gain factor in the scheduling. However, once we integrate the MAPEL into the DAL (DAL-MAPEL), the performance gain is clear. From Figs. 3 and 4, we can notice an increase in the spectral efficiency and a corresponding decrease in the power consumption as a result of the MAPEL adoption. At the same time, from Fig. 3 we can observe that the DAL-MAPEL solution performance is comparable to the greedy approach (BF) and this means that the DAL-MAPEL approaches the optimal performance regions for the system. More results for the BF benchmark are provided in Sect. V-B.

Finally, in order to appreciate the impact of ASC, we considered 1000 simulation runs, each with a different channel realization, for the case without and with ASC: from the distribution of the throughput ratio between DAL and RRc in case of ASC and no ASC, we have obtained that ASC has a positive impact on the DAL throughput since the DAL throughput is higher than the RRc one for 61.0% of the cases without ASC and 88.1% with ASC; the corresponding rate improvement with ASC w.r.t. RRc is in the range $0 - 25\%$.

B. Comparison with a brute force search algorithm

In order to assess the performance of the sub-optimal solution, considering the high number of involved parameters (4 RUCs for each user, 12 sub-carriers, etc.), we have implemented the BF (or greedy) algorithm based on a large number of random allocations and a final selection of the best one; in each random allocation, the terminals are allocated according to a random order and selecting first a random RUC for each one and then a random location in the available resource grid.

In Fig. 5 the DAL and the RR are compared against the BF in terms of spectral efficiency, under the same conditions of Fig.2, within a resource grid of 12 sub-carriers and 150 TSs. It can be seen that the DAL performance is not so far from the BF one. At the same time, we can observe that this performance difference depends strongly on the size of the resource grid. When we compare resource grids with

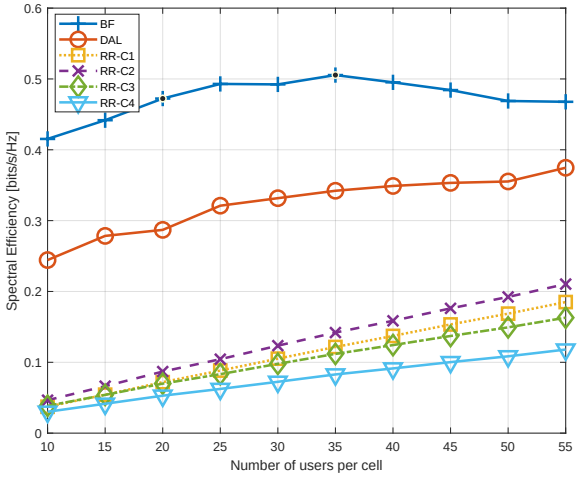

Fig. 5. Spectral efficiency per cell comparison between BF, DAL and RR for different numbers of devices per cell for 150 TSs allocation grid.

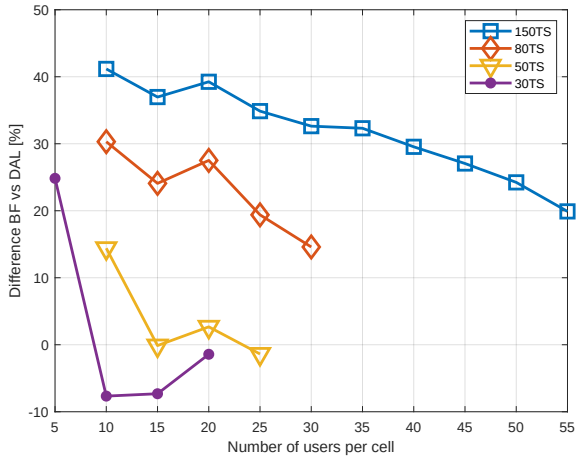

Fig. 6. Difference between the DAL and BF spectral efficiencies for different numbers of TSs in the allocation resource grid.

different number of TSs, we see that DAL and BF performance become closer when (i) the number of TSs decreases, since the random space of the BF search is reduced, or (ii) the number of devices increases, since the random search becomes less efficient keeping the same number of random extractions. Fig. 6 shows the difference between BF and DAL for 150, 80, 50, and 30 TSs, where each point represents the percentage of the spectral efficiency difference between BF and DAL w.r.t. BF result.

C. Impact of PDSD on throughput and latency

Table VII and Table VIII summarize the results for the final part of our analysis, which concerns with power distribution, latency and the number of dropped packet. We have already observed that, in this case, the repetition triggering mechanism is taken into account. In these simulations, the DAL is compared to the RRc. Different techniques of power distribution (PDSD) have been used with DAL and the total throughput and latency for delivering correctly the assigned RUs are simulated for each device. Furthermore, each device has a maximum allowed latency and, since this threshold is application specific, it is set here to the value of 50 ms, small enough for being challenging for the schedulers: when this maximum latency is exceeded, the packet is dropped and the results report also the average number of dropped packets.

The simulation has been performed with 1000 runs for each technique, the number of sub-frames varies, and the tested number of users is 15 per cell. The performance is simulated for the following power distribution techniques: water filling (WF), equal distribution (ED) and maximum-minimum SINR power distribution (MAMI). In addition, we have considered the two repetition triggering strategies, i.e. the *min-sinr* and *max-sinr*. The MAPEL has not been applied in this case since the PDSD techniques have shown the main impact on the number of repetitions and hence on the latency. In Tables VII and VIII, each row contains the comparison results of two cases (A vs B). Each case represents an allocation technique, e.g. DAL or RRC, with a given PDSD technique, e.g. DAL(WF). There are 3 performance indicators: throughput, latency and number of dropped packets. Each performance indicator shows 3 values, (i) the average value achieved by A, (ii) the average value achieved by B and the probability (percentage of the cases) that A performance value is greater than B one. Please notice that having a greater throughput is positive but having a greater latency is not.

From Table VII, the following results can be observed:

1) The WF improves the throughput performance of DAL, as expected.
2) When ED is used, the DAL and RRc performance becomes really similar. However, there is a small advantage for the DAL in terms of throughput because of the device diversity.
3) MAMI shows a remarkable advantage in terms of latency.

On the other hand, from Table VIII, we can see that:

1) The WF achieves the best performance in all the performance indicators, including latency and number of dropped packets.
2) MAMI performance is not totally satisfactory, especially in terms of latency.
3) There is a great advantage for the max-sinr triggering strategy w.r.t. the min-sinr one in all the performance indicators, especially latency and dropped packets.

D. Discussion of the results

We have presented the resource allocation optimization problem for the weighted sum throughput taking into account the different constraints, specifically introduced for the NB-IoT systems. The numerical results are simulated by means of a sub-optimal solution for the problem formulation. Here we summarize the main observations that can be derived from the simulation results.

- The brute force approach and the best performance obtained from the sub-optimal solution show that the choice of the DAL is a reasonable choice.

TABLE VII
AVERAGE VALUE OF CASE (A), CASE (B), AND PERCENTAGE THAT A IS GREATER THAN B, FOR THE GIVEN METRICS, FOR 1000 DIFFERENT CHANNEL REALIZATION AND THE *min-sinr* OPTION USED TO TRIGGER THE REPETITION PROCESS.

Comparison (A vs B)		Throughput [bps] 10^5			Percentage of dropped packets			Latency [ms]		
A	B	Average(A)	Average(B)	A>B	Average(A)	Average(B)	A>B	Average(A)	Average(B)	A>B
DAL(WF)	DAL(ED)	4.2	4.0	74.7%	29.1%	23.4%	69.0%	35.6	30.8	97.9%
DAL(ED)	RRc(ED)	4.0	3.8	68.6%	24.5%	18.5%	67.2%	31.0	30.8	53.5%
DAL(WF)	RRc(ED)	4.2	3.8	88.5%	29.9%	18.8%	73.2%	35.9	30.8	96.4%
DAL(MAMI)	RRc(ED)	3.9	3.8	55.1%	13.7%	19.3%	15.6%	23.9	30.9	0.3%

TABLE VIII
AVERAGE VALUE OF CASE (A), CASE (B), AND PERCENTAGE THAT A IS GREATER THAN B, FOR THE GIVEN METRICS, FOR 1000 DIFFERENT CHANNEL REALIZATIONS AND THE OPTION *max-sinr* USED TO TRIGGER THE REPETITION PROCESS.

Comparison (A vs B)		Throughput [bps] 10^5			Percentage of dropped packets			Latency [ms]		
A	B	Average(A)	Average(B)	A>B	Average(A)	Average(B)	A>B	Average(A)	Average(B)	A>B
DAL(WF-max)	RRc(ED-min)	4.8	3.8	80.2%	1.6%	19%	4.6%	11.9	30.9	0%
DAL(WF-max)	RRc(ED-max)	4.8	4.0	84.3%	1.5%	2.7%	6.9%	11.9	12.3	33.7%
DAL(MAMI-max)	RRc(ED-max)	3.7	4.0	23.2%	3.5%	2.8%	54.2%	13.3	12.4	80.8%

- Both device diversity and configuration diversity give the highest contribution to the performance and gain of the DAL in terms of throughput. However, when one of them is used, adding the other provides just a slight increase in the performance.
- Power distribution and repetition triggering criteria: for the *min-sinr* case, when the WF is used, it optimizes the power distribution w.r.t. the sum-rate of the device SCs and the rate improves. However, since the max sum-rate of the SCs can give very low power to some of the SCs and consequently zero or low SINR, the trigger repetition can be activated, increasing latency; the severity of this problem can be reduced by using ED or MAMI.
- Power distribution and repetition triggering criteria: for the *max-sinr* case, when the WF is used, it maximises the sum-rate, regardless of the individual rate of each SC and, as a result, devices with extreme channel conditions will be given very little or no power and devices with good channel conditions will get most of the power achieving higher SINRs. Therefore, the possibility of repetition becomes lower, improving also latency and dropped packets percentage.
- Power distribution and repetition triggering criteria: a totally different case is the MAMI, since it achieves more fairness between the SCs of the device, increasing the SINR of each of them. This way of distributing the power is not advantageous especially w.r.t. its impact on latency.
- The shape of the allocation grid, ASC, improves allocation performance.

Parts of the sub-optimal, heuristic algorithm have some shortcomings and possible improvements to be addressed, partially anticipated in Sect. IV-E:

- Interference: the interference of the previous TS is taken as valid also for the current TS, since we cannot know the current allocation and interference from all the cells. This approach is feasible also in a practical implementation; for improving the numerical benchmark, a possible method is to run the algorithm (DAL with or w/o MAPEL) iteratively until the achievement of a final convergence of the interference measured in each subcarrier.
- ASC: we have implemented the ASC by adding a constraint that prioritizes the filling of empty slot/subcarriers within each frame. However, this strategy could be improved for finding an optimal occupation of the resource grid. The really challenging aspect of this part is the combination between the best SINR and the best shape searches.
- Power allocation: the MAPEL needs an intense centralized processing and a lot of information from all the involved cells and this part should be subject to further simplifications.
- Repetition triggering: to the best of our knowledge, there is no research done about the repetition triggering criteria in NB-IoT. However, we have observed its importance for the impact on latency and the two criteria for triggering repetitions could be extended to other options.

E. Future work

Based on the previous remarks and the current state of the art, the following points could be of interest for future work:

- Inclusion of energy efficiency in the problem trade-offs: for NB-IoT devices, the energy efficiency is clearly vital and, even if a SINR or rate maximization is generally valid also from an energy point, a deeper investigation on the aspects related to energy efficiency could provide further insight into the specific trade-offs for NB-IoT devices.
- A study on cooperative approaches among adjacent eNBs for realizing the MAPEL integration into the algorithm with a realistic protocol, compliant with the current and future 5G releases.
- A simplification of some of the assumptions and inputs necessary for the MAPEL integration into the algorithm will lead to solutions suited to an easier implementation. The other parts of the sub-optimal algorithm can be already implemented in current systems.
- Repetition triggering: in addition to the two criteria used here, other mechanisms could be investigated.

VI. Conclusions

In this paper, we have presented a study on resource allocation for NB-IoT. After writing the optimization problem for maximizing the sum throughput, we have modified this problem in order to achieve satisfactory throughput and latency. Since the optimization problem is non-convex, NP-hard and with binary variables, we have developed a sub-optimal algorithm for finding the correct power allocation and uplink scheduling. The solution is based on dividing the problem into two parts, i.e. the uplink scheduling problem and the power allocation.

The proposed algorithm comparable performance w.r.t. a brute force approach and better performance w.r.t. to a basic standard RR with a fixed RU allocation. Furthermore, by using the MAPEL algorithm, we have shown that power consumption can be decreased by 3 dB, and doubled the throughput in case of high number of devices.

The results show the importance of using an adaptive allocation of RUCs, which creates a configuration diversity. When this configuration diversity is used, the throughput doubles approximately either with DAL or even with the simple RR.

The study of the power distribution among the SCs of the device leads to understanding also the main impact on the latency. By selecting the proper power distribution algorithm, in case of *min-sinr* repetition triggering, we have improved drastically the latency. In fact, the results show the importance of the triggering mechanism and using the *max-sinr* option reduces further the latency and the number of reduced packets (for example the percentage of the number of dropped packets is decreased from 29% to 1.6%).

Finally, we have seen that the imperfect alignment of the allocated resources, due to the adaptive RUC allocation and to the different shapes of each RUC, affects performance of the allocators and the final throughput. By realizing a proper management of the shape constraints (ASC), the performance increases.

References

[1] H. Malik, M. M. Alam, H. Pervaiz, Y. L. Moullec, A. Al-Dulaimi, S. Parand, and L. Reggiani, "Radio Resource Management in NB-IoT Systems: Empowered by Interference Prediction and Flexible Duplexing," *IEEE Network*, pp. 1–8, 2019.

[2] H. Malik, N. Kandler, M. M. Alam, I. Annus, Y. Le Moullec, and A. Kuusik, "Evaluation of low power wide area network technologies for smart urban drainage systems," in *IEEE International Conference on Environmental Engineering (EE)*, 2018, pp. 1–5.

[3] M. Chen, Y. Miao, Y. Hao, and K. Hwang, "Narrow band internet of things," *IEEE access*, vol. 5, pp. 20 557–20 577, 2017.

[4] R. Ratasuk, N. Mangalvedhe, Y. Zhang, M. Robert, and J.-P. Koskinen, "Overview of narrowband IoT in lte rel-13," in *2016 IEEE conference on standards for communications and networking (CSCN)*. IEEE, 2016, pp. 1–7.

[5] Y.-P. E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A primer on 3GPP narrowband Internet of Things," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 117–123, 2017.

[6] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Narrowband IoT data transmission procedures for massive machine-type communications," *IEEE Network*, vol. 31, no. 6, pp. 8–15, 2017.

[7] C. Yu, L. Yu, Y. Wu, Y. He, and Q. Lu, "Uplink scheduling and link adaptation for narrowband internet of things systems," *IEEE Access*, vol. 5, pp. 1724–1734, 2017.

[8] C.-W. Huang, S.-C. Tseng, P. Lin, and Y. Kawamoto, "Radio resource scheduling for narrowband internet of things systems: A performance study," *IEEE Network*, vol. 33, no. 3, pp. 108–115, 2019.

[9] A. Azari, G. Miao, C. Stefanovic, and P. Popovski, "Latency-energy tradeoff based on channel scheduling and repetitions in NB-IoT systems," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–7.

[10] R. C. J. Neto, E. B. Rodrigues, and C. T. de Oliveira, "Performance analysis of resource unit configurations for m2m traffic in the narrowband-IoT system," in *Proc. 35th Brazilian Commun. Signal Process. Symp.*, 2017, pp. 816–820.

[11] B.-Z. Hsieh, Y.-H. Chao, R.-G. Cheng, and N. Nikaein, "Design of a ue-specific uplink scheduler for narrowband internet-of-things (NB-IoT) systems," in *2018 3rd International Conference on Intelligent Green Building and Smart Grid (IGBSG)*. IEEE, 2018, pp. 1–5.

[12] S.-F. Cheng, H.-A. Hou, L.-C. Wang, K.-T. Feng, and J.-Y. Hsu, "Multi-user scheduling for narrow band internet of things," *SPACOMM 2018 : The Tenth International Conference on Advances in Satellite and Space Communications*, 2018.

[13] Y.-J. Yu and J.-K. Wang, "Uplink resource allocation for narrowband internet of things (NB-IoT) cellular networks," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 466–471.

[14] J.-M. Liang, K.-R. Wu, J.-J. Chen, P.-Y. Liu, and Y.-C. Tseng, "Energy-efficient uplink resource units scheduling for ultra-reliable communications in NB-IoT networks," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.

[15] L. P. Qian, Y. J. Zhang, and J. Huang, "Mapel: Achieving global optimality for a non-convex wireless power control problem," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1553–1563, 2009.

[16] O. Elgarhy and L. Reggiani, "Application of the water filling algorithm to the sum rate problem with minimum rate and power constraint," in *2018 Advances in Wireless and Optical Communications (RTUWO)*. IEEE, 2018, pp. 12–16.

[17] S. Hayashi and Z.-Q. Luo, "Spectrum management for interference-limited multiuser communication systems," *IEEE Transactions on Information Theory*, vol. 55, no. 3, pp. 1153–1175, 2009.

[18] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Low complexity subcarrier and power allocation algorithm for uplink ofdma systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, p. 98, 2013.

[19] K. Kim, Y. Han, and S.-L. Kim, "Joint subcarrier and power allocation in uplink ofdma systems," *IEEE Communications Letters*, vol. 9, no. 6, pp. 526–528, 2005.

[20] H. Malik, H. Pervaiz, M. M. Alam, Y. Le Moullec, A. Kuusik, and M. A. Imran, "Radio resource management scheme in NB-IoT systems," *IEEE Access*, vol. 6, pp. 15 051–15 064, 2018.