A New Contention-Based PUSCH Resource Allocation in 5G NR for mMTC Scenarios

Luciano Miuccio, *Student Member, IEEE*, Daniela Panno, *Member, IEEE*, Salvatore Riolo, *Student Member, IEEE*

***Abstract*—This letter addresses the uplink radio resource allocation problem in a 5G massive Machine Type Communication (mMTC) scenario, where a large number of MTC devices generate small-sized packets. The current cellular networks are unsuitable for this scenario, due to the limited resources allocated to the Physical Random Access Channel (PRACH) and to the Physical Uplink Shared Channel (PUSCH). To overcome this issue, instead of improving the succeeded access attempts in the PRACH, as widely addressed in literature, we focus on the PUSCH. We present the innovative idea to exploit the unused PUSCH resources to serve also the MTC devices that have failed their access attempt, by assigning them a pool of resources, in a contention-based mode. The simulation results show that the proposed PUSCH Resource Reallocation Algorithm improves significantly any static and dynamic system performance in terms of number of succeeded communications, at the cost a negligible increment of energy consumption.**

***Index Terms*—mIoT, SCMA, NOMA, random access**

I. Introduction

The massive Machine Type Communication (mMTC) usage scenario [1] is characterized by a large number of devices, transmitting small amounts of data, and having high requirements on increased battery lifetime. In this scenario, thanks to the relaxed delay requirements, a contention-based Random Access (RA) procedure can be adopted. Currently, in the LTE, NarrowBand-IoT (NB-IoT) and enhanced MTC (eMTC) systems, the MTC device performs a 4-step contention-based RA procedure by firstly transmitting a preamble sequence in the Physical Random Access Channel (PRACH). Then, if the access attempt has been successful and there are available radio resources into the Physical Uplink Shared Channel (PUSCH), the MTC device transmits its data in it. Otherwise, the device re-attempts in successive RA cycles for a maximum number of times, M_A. If several hundred devices simultaneously request to send data, the preamble collision probability increases, thus the number of successful accesses decreases. This problem could be alleviated by increasing the resources allocated to the PRACH. However, due to the limited uplink resources, the higher the RA resources, the lower the amount of resources available for the PUSCH.

In literature, several works (e.g., [2]) take into account only the limit of PRACH resources and define new RA procedures to decrease the preamble collision probability. These works are based on Access Class Barring (ACB) schemes [3], i.e., congestion control schemes designed for limiting the number of simultaneous access attempts. Also, the number of successful communications can be further improved if an optimal division of uplink resources is made. In [4] we proposed a Dynamic Uplink Resource Dimensioning (DURD), which optimizes the resource allocation between PRACH and PUSCH according to the traffic load in every RA cycle. The dynamic dimensioning allows to achieve significant improvements in terms of both system throughput and energy consumption compared to any static dimensioning. In the view of this, in [5], [6] an innovative ACB scheme well integrated with that dynamic dimensioning control is proposed.

In this letter, we aim to improve the RA performance with the innovative idea to exploit the unused PUSCH resources, if any, to serve also some MTC devices that have failed their access attempt. Given the RA cycle i, instead of limiting the number of attempting MTC devices, our idea is to serve a larger number of attempting MTC devices in i in order to reduce the amount of re-attempting devices in the successive RA cycles, thus mitigating the congestion. In particular, for a subset of collided preambles, the scheduler allocates a pool of resources to each of them in a contention-based mode. This resource allocation in the PUSCH does not replace the conventional RA procedure in the PRACH, but it is an additional operation for increasing the number of succeeded communications, if possible. The main drawback of this approach is that contention-based transmission in the PUSCH may collide causing vain additional energy consumption. To attenuate this issue, we need to determine the optimal size of the subset of collided preambles that maximizes the number of successes in the PUSCH, thus reducing both the number of collisions in it and the additional energy consumption. At this aim, we present an analytic problem formulation and, accordingly, we propose an iterative algorithm, called PUSCH Resource Reallocation Algorithm (PRRA), to implement our strategy. The proposed PRRA is applied assuming both a static and dynamic uplink PRACH/PUSCH dimensioning. The simulation results show that the proposed PRRA improves any static and dynamic system performance in terms of succeeded communications, at the cost of a similar or slightly higher energy consumption.

II. System Model and Background

In this letter, we consider a 5G network scenario with one Next Generation Node B (gNB) that supports N_{MTC} MTC devices using a licensed band of $B = 1.08$ MHz exclusively dedicated to the mMTC scenario. Each device

Manuscript received ...; revised ...

This work has been partially funded by the University of Catania under project "MANGO - Piaceri 2020-2022 - Linea 2", and by the Ministero dello Sviluppo Economico (MISE) under Project M9 (CDS 000448) - CUP C32F18000100008, "UE-PON Imprese e Competitività 2014–2020."

The authors are with the DIEEI, University of Catania, Italy and also with the CNIT, Italy (e-mail: luciano.miuccio@phd.unict.it; daniela.panno@unict.it; salvatore.riolo@unict.it).

generates a single data at time $t \in [0, T_{arrival}]$ according to the Beta distribution, as recommended by the 3GPP [7]. As regards the uplink radio interface, we adopt the smallest 5G NR numerology, i.e., subcarrier spacing of 15 kHz, time slot duration (T_s) of 1 ms, containing 14 OFDM symbols. This choice has been made because delay tolerant services, such as MTC services, can benefit from small subcarrier spacing to reduce bandwidth consumption [8]. A typical RA cycle lasts $T_{ra} = 5$ time slots and the uplink radio resources are divided into a PRACH subset and a PUSCH subset. For the preamble sequences in the PRACH we adopt the NR Preamble Format 0, i.e., the PRACH access resources consist of 64 orthogonal preamble sequences that are mapped to 839 subcarriers of 1.25 kHz, and each preamble lasts 1 ms [9]. This choice has been made because the 1.25 kHz numerology is the only option available to support the bandwidth considered, and this Preamble Format maximizes the number of orthogonal preambles available per time slot. The total 64 preambles are divided into two groups: the first, consisting of L_0 preambles, is dedicated for the contention based procedure, while the second one is reserved for the contention free procedure. In the time domain PRACH lasts $T_{pr} \in \{1, 2, \ldots, T_{ra}-1\}$ time slots and PUSCH the remaining T_{pu} slots. In summary, for each RA cycle we have $T_{ra} = T_{pr} + T_{pu}$. Since each preamble lasts 1 T_s, the total number of preamble sequences available for the contention based procedure in each RA cycle is $L = L_0 T_{pr}$.

In order to reduce the MTC device energy consumption, we consider the 2-step connectionless RA procedure proposed in [6]. In summary, during Step 1, each MTC device randomly selects one preamble out of the L available and transmits a tagged preamble sequence on the PRACH. Obviously, there is a non-zero collision probability, since the same preamble can be selected by more than one device. This tagged preamble sequence allows the gNB to detect whether a collision has been occurred immediately after the preamble reception. However, we underline that the gNB does not know how many MTC devices have selected each collided preamble. During Step 2, if the preamble is successfully transmitted, the gNB will send to the MTC a Random Access Response (RAR) message. More specifically, because for the MTC scenario a very small amount of data is expected, we set for each transmission request an upper bound value, θ_{max} bits, and we assume that the gNB assigns to each successful access attempt, the R_θ PUSCH resources, if any, enough to transmit θ_{max} bits. Consequently, the MTC device which has received the RAR message from the gNB, transmits its data packet in the PUSCH of the next RA cycle together with an UL context containing all necessary information related to the device identity and PDN-ID. Conversely, the device reattempts the RA procedure inside the back-off window (B_W). In addition, to further increase the transmission efficiency, we adopt the Sparce Code Multiple Access (SCMA) technique for PUSCH resources, that is a promising Non-Orthogonal Multiple Access (NOMA) technique to support massive MTC connectivity requirements with small-size data. By using SCMA scheme, multiple MTC devices can transmit on the same Resource Element (RE) with different sparse codebooks [10].

As regards the PRACH, we denote M as the number of MTC devices which are performing the contention-based RA procedure in the considered RA cycle, and P_S as the number of succeeded preambles in the PRACH, i.e., the number of successful access attempts. In particular, given M and L, P_S is a random variable whose expected value is

$$\bar{P}_S = M\left(1 - {}^{1}\!/_{L}\right)^{M-1} = M\left(1 - {}^{1}\!/_{L_0 T_{pr}}\right)^{M-1}. \quad (1)$$

For the same M value, the higher T_{pr}, the higher $\bar{P}_S$.

As regards the PUSCH resources, according to the SCMA encoder, one $SCMA_{block}$ is the minimum resource quantity which can be shared by different transmissions, called layers. The total number of data transmissions (DT_{max}) which can be satisfied in the PUSCH is:

$$DT_{max} = \left\lfloor \frac{\lfloor {}^{72}\!/_{Q} \cdot 14 \rfloor \, {}^{(QK_{max})}\!/_{S}}{\lceil {}^{\theta_{max}}\!/_{\log_2(I)} \rceil} \right\rfloor (T_{ra} - T_{pr}), \quad (2)$$

where Q is the the number of REs in one $SCMA_{block}$, S is the number of REs which a layer occupies respect to Q, K_{max} is the maximum number of overlapped layers in one RE, and $\log_2(I)$ is the number of bits per symbol [4]. It is evident that, the higher T_{pr}, the lower DT_{max}. Finally, the average number of MTC devices which successfully transmit inside a RA cycle, i.e., the number of succeeded communications, is

$$\bar{C}_S = \min(\bar{P}_S, DT_{max}). \quad (3)$$

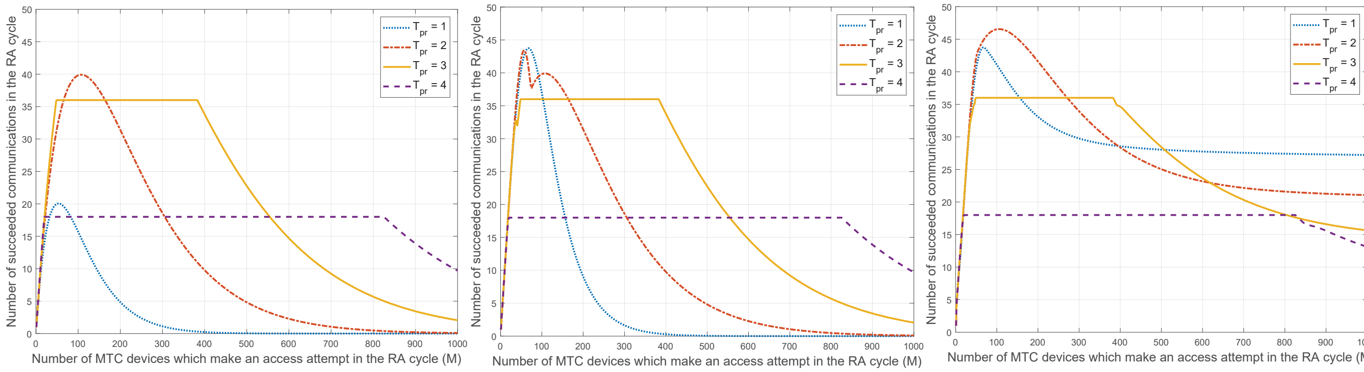
In Fig. 1a we plot $\bar{C}_S$ versus M[1] for any T_{pr} value. We observe that the straight lines are caused by the saturated PUSCH resources (e.g., with T_{pr} = 3, $\bar{C}_S = DT_{max} = 36$, for M ranging from 50 to 380).

III. Our proposal

Starting from (3), we note that when $P_S < DT_{max}$, i.e., the number succeeded access attempts is less than the amount of resources available for the data transmission, P_S out of DT_{max} data transmissions are scheduled in a collision free mode to the succeeded preambles, while $DT_U = DT_{max} - P_S$ available data transmissions remain unused in the PUSCH.

Our strategy aims to rise $\bar{C}_S$ trying to serve also the collided preambles by assigning to them the DT_U resources in a proper manner. In this case, the data transmission resources cannot be allocated in a collision free mode, since there is a 1 to many relationship between each collided preamble and the related MTC devices. So, the scheduler needs to allocate, in a contention based mode, a pool of R_θ resources consisting of DT_{P_C} data transmissions per collided preamble, with $DT_{P_C} > 1$. Specifically, the gNB reserves a single pool of DT_{P_C} resources to the MTC devices that have transmitted the same collided preamble. Then, each related MTC device draws independently a uniform random number $g \in \{1, \ldots, DT_{P_C}\}$ and sends its data in the gth data transmission resource. Similarly to the preamble transmission in the PRACH, there is a non-zero data transmission collision probability. We emphasize that, if the additional communication is successful, the gNB knows which device has transmitted data by inspecting the related UL context. Hence, the gNB is able to ACK the data to

[1] As shown in [4], the values assumed by M are consistent with the MTC arrival distribution suggested by the 3GPP in [7], with $N_{MTC} = 100000$.



(a) $\bar{C}_S$ (b) $\bar{C}'_S$ with the straightforward strategy (c) $\bar{C}'_S$ by using (6)

Fig. 1. Number of successful transmissions in an RA cycle vs the number of MTC devices (M) which carry out the RA procedure for different T_{pr} values.

the related device and to further process and forward the data to the expected network entity. Conversely, if the additional communication is collided, the MTC device waits in vain the ACK from the gNB within the related waiting window. Then, it re-attempts the RA procedure during a randomly selected PRACH inside the Backoff Window (B_W).

A. Analysis and Formulation of Proposed Strategy

The number of successful additional data transmission (A_S) is a random variable that depends on DT_{P_C} and on the number of MTC devices that have transmitted the same collided preamble. We denoted it as K_C and it is termed as "collision coefficient". The average number of successful additional data transmissions per collided preamble ($\bar{A}_{S,P_C}$) can be calculated as (1), with $M = K_C$ and $L = DT_{P_C}$, as follows:

$$\bar{A}_{S,P_C} = \begin{cases} K_C \left(1 - {}^{1}/_{DT_{P_C}}\right)^{K_C - 1} & \text{if } DT_{P_C} > 1 \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Since $\bar{A}_{S,P_C}$ and $\bar{P}_C$ are independent random variables, the average total number of successful additional data transmissions is $\bar{A}_S = \bar{A}_{S,P_C}\bar{P}_C$. Finally, we redefine the average number of succeeded communications ($\bar{C}'_S$) as:

$$\bar{C}'_S = \begin{cases} \bar{P}_S + \bar{A}_S & \text{if } \bar{P}_S < DT_{max} - 1 \\ \bar{C}_S & \text{otherwise.} \end{cases} \quad (5)$$

To calculate $\bar{C}'_S$, we need to derive K_C and DT_{P_C}. Let us start by calculating the total number of preambles out of L transmitted by M MTC devices. It is denoted as P_T. At this aim, we define Y_i, with $i \in \{1, \ldots, L\}$, as Boolean random variables, each one equal to 1 if and only if preamble i has been selected by at least one MTC device. Since Y_i are L random variables identically distributed, the mean value of P_T can be calculated as $\bar{P}_T = E\left\{\sum_{i=1}^{L} Y_i\right\} = LE\{Y_i\}$, where $E\{Y_i\} = \Pr(Y_i = 1) = 1 - \left(1 - {}^{1}/_{L}\right)^M$, for each i.

Now, the average number of collided preambles can be easily calculated as $\bar{P}_C = \bar{P}_T - \bar{P}_S$, and the average number of MTC devices that have selected a collided preamble can be approximated as $\bar{K}_C \simeq {}^{(M - \bar{P}_S)}/_{\bar{P}_C}$. Then, having $\overline{DT}_U$ unused data transmissions and $\bar{P}_C$ collided preambles, the straightforward strategy is to equally distribute the $\overline{DT}_U$ resources among the $\bar{P}_C$ preambles, i.e., $\overline{DT}_{P_C} = {}^{\overline{DT}_U}/_{\bar{P}_C}$.

Having derived $\bar{K}_C$ and $\overline{DT}_{P_C}$, we can calculate $\bar{A}_{S,P_C}$ by using (4), where we set $K_C = \bar{K}_C$ and $DT_{P_C} = \overline{DT}_{P_C}$. This treatment is valid when using the drift approximation [11]. Then, we calculate $\bar{C}'_S$ by using (5) and plot the variation of $\bar{C}'_S$ versus M for any T_{pr} value in Fig. 1b. As expected, the only work areas that permit to improve the number of successful transmissions are outside of straight lines, i.e., where the PUSCH resources are not saturated. More specifically, the best improvement in terms of $\bar{C}'_S$ is obtained by the $T_{pr} = 1$ dimensioning, thanks to the high availability of $\overline{DT}_U$ PUSCH resources. As regards the $T_{pr} = 2$ dimensioning, it exhibits a modest gain with respect to $\bar{C}_S$ only up to $M = 74$, since for $M \geq 75$, it results $\bar{A}_S = 0$, as can be verified. Instead, the remaining dimensionings show a very slight improvement only before the related straight lines, that can be neglected. Moreover, all the dimensionings do not show improvements for high M values because the higher M, the higher $\bar{P}_C$ and $\bar{K}_C$, the lower $\overline{DT}_{P_C}$. So, the straightforward strategy of serving all the collided preambles leads to $\bar{A}_S = 0$.

To overcome this issue, we propose an enhanced strategy to maximize the term $\bar{C}'_S$. We note that, given a fixed T_{pr} value, M, and θ_{max}, the values of $\bar{P}_S$ and DT_{max} are determined by (1) and (2), respectively. Then, in (5) the only term that can be maximized is $\bar{A}_S$. Unlike the straightforward strategy, we propose to divide, on average, the unused PUSCH resources among a proper subset of collided preambles (i.e., $\bar{P}_{C_S} \leq \bar{P}_C$). For this reason, we can formulate the goal as follows: "*Given M and T_{pr}, the aim is to maximize $\bar{A}_S$ by varying the number of collided preambles $\left(\bar{P}_{C_S}\right)$ to which the unused PUSCH resources should be allocated in the contention way*:

$$\max_{\bar{P}_{C_S} \in \left[0, \min\left(\overline{DT}_U/2, \bar{P}_C\right)\right]} \left\{\bar{A}_S\right\}." \quad (6)$$

We calculated $\bar{A}_S$ numerically and in Fig. 1c, we plot the variation of $\bar{C}'_S$ versus M for any T_{pr} value. The $T_{pr} = 1$ and $T_{pr} = 2$ dimensionings achieve a remarkable improvement for any value of M in comparison with Figs. 1a and 1b. As regards $T_{pr} = 3$, a light improvement occurs when $M \geq 500$. $T_{pr} = 4$ has no significant room for improvement. The lower the T_{pr} value, the larger $\overline{DT}_U$, the higher the improvement.

B. The PUSCH Resource Reallocation Algorithm (PRRA)

In light of the above theoretical average advantages, we implement the proposed strategy by presenting the PUSCH Resource Reallocation Algorithm (PRRA). It allows the scheduler to re-allocate the PUSCH resources efficiently, inside a given RA cycle, known M, P_S, P_C and T_{pr}. The purpose is to find the optimal $P^*_{C_S} \leq P_C$ value that maximizes the expected value of A_S inside that RA cycle, and the relative $DT^*_{P_C}$ value. Then, the $P^*_{C_S}$ collided preambles will be chosen randomly. The algorithm is described in pseudo-code 1, and its time complexity is $O(n)$, where $n = \min\{DT_U, P_C\}$.

Pseudo-code 1 PUSCH Resources Reallocation Algorithm

Inputs: M, Q, K_{max}, S, θ_{max}, T_{pr}, T_{ra}, P_S, and P_C

Iteration:

1: calculate DT_{max} by using (2), $K_C = {}^{(M-P_S)}/_{P_C}$, and $DT_U = DT_{max} - P_S$;
2: create an auxiliary vector $\mathbf{A} = [1, \ldots, \min\{DT_U, P_C\}]$;
3: calculate $\mathbf{B} = \lfloor DT_U \mathbf{A}^{\circ(-1)} \rfloor$, where $(\cdot)^\circ$ is the Hadamard power operator;
4: calculate $\mathbf{C} = K_C \left(\mathbf{J}_{1,|\mathbf{B}|} - \mathbf{B}^{\circ(-1)}\right)^{\circ(K_C-1)}$ where $\mathbf{J}_{1,|B|}$ is the all-ones matrix of size $1 \times |\mathbf{B}|$;
5: calculate $\mathbf{D} = \mathbf{A} \circ \mathbf{C}$, where $\circ$ is the Hadamard product;
6: calculate $\bar{A}_S = \max\{\mathbf{D}\}$ and $P^*_{C_S} = \arg\max\{\mathbf{D}\}$;
7: calculate $DT^*_{P_C} = \mathbf{B}[P^*_{C_S}]$.

However, in order to apply the PRRA, the gNB should know for each RA cycle, inter alia, the total amount of MTC devices M attempting to access. Clearly, M is not known at the base station a priori, and therefore this parameter should be estimated. From an empirical study, we derived in [4] the estimated number of attempting devices ($\tilde{M}$) as function of P_S and P_C, which the gNB knows at Step 2 of the RA procedure:

$$\tilde{M} = P_S + 2 \cdot 1.97^{P_C/L} P_C. \quad (7)$$

This estimate exploits the fact that the process of total access attempts in two consecutive RA cycles is strongly correlated, as reported in [4] and proved in [6]. The accuracy of (7) is verified by means of the simulation results in Section IV.

C. A New Enhanced Dynamic Uplink Resource Dimensioning

Observing Fig. 1c, we note that there is not a unique value of T_{pr} which optimizes $\bar{C}'_S$ for any number of attempting devices M. In [4], [6] we showed, with a long time evaluation (i.e., considering thousands of successive RA cycles), that the optimal PRACH/PUSCH resource allocation can be achieved only if a dynamic load-aware control is applied. More specifically, we proposed to dynamically set the T_{pr} dimensioning for each RA cycle, on the basis of M. In [4] the proposed Dynamic Uplink Resource Dimensioning (DURD) was applied to the $\bar{C}_S$ values derived from (3) and reported in Fig. 1a. In this letter, we apply the same approach, considering the new curves derived from (5) and (6), and shown in Fig. 1c.
"Given M, the goal is to find the optimal T_{pr} value so that

$$T^*_{pr} = \underset{T_{pr} \in \{1,\ldots,T_{ra}-1\}}{\arg\max} \left\{\bar{C}'_S\right\}, \quad (8)$$

where the dependence on M is contained in $\bar{C}'_S$." Choosing $T_{pr} = T^*_{pr}$, we have the optimal resource dimensioning in the RA cycle. In this way, we obtain an Enhanced Dynamic Uplink Resource Dimensioning system, termed EDURD and, as shown in Fig. 1c, the optimal dimensioning results $T^*_{pr} = 2$ for $M \leq 271$, $T^*_{pr} = 3$ for $272 \leq M \leq 507$, and $T^*_{pr} = 1$ for $M \geq 508$. Compared to Fig. 1a, the high M values correspond to the lowest T^*_{pr} value because for $T_{pr} = 1$ a large portion of the preambles in the PRACH are collided and, consequently, the succeeded transmissions are related to the reallocation in the PUSCH, that works better with $T_{pr} = 1$. In summary, adopting the EDURD system, for each RA cycle, on the basis of M, the gNB sends in broadcast the optimal T^*_{pr} value for the next RA cycle. Then, on basis of P_S and P_C detected, the gNB scheduler assigns one R_θ resource to each of the P_S MTC devices in a collision free mode, and $DT^*_{P_C}$ resources to $P^*_{C_S} \leq P_C$ preambles in a contention-based mode, that were estimated as output of the PRRA.

TABLE I
SIMULATION PARAMETERS

Parameter	Symbol	Value
Preambles reserved for contention-based procedure	L_0	54
Number of total MTC devices	N_{MTC}	40000, 50000, 75000, 100000
Number of data transmission requests per MTC device		1
Maximum number of RA attempts	M_A	10
Data transmission size	θ_{max}	160 bits
Arrival distribution	Beta	α = 3, β = 4 $T_{Arrival}$ = 10s
RA cycle duration	T_{ra}	5 time slots
Simulation Length	T_{sim}	10s
SCMA parameters	Q, K_{max}, S	$4, 3, 2$
Backoff Window	B_W	20ms
Constellation Points	I	4

IV. Performance Evaluation

In this section, we consider the following uplink dimensioning systems, where the PUSCH resources are assigned on basis of the SCMA technique. We denote them as:

- S_i, with $i \in \{1, \ldots, 4\}$, as the system with static uplink dimensioning $T_{pr} = i$;
- $DURD$ [4] as the system with optimal dynamic uplink dimensioning based on Fig. 1a;
- $eDURD$ [4] as the $DURD$ system based on the predictive estimation of M by using (7).
- Ex, with $x = \{S_i, DURD, eDURD\}$, as the enhanced system x with the contention-based additional transmissions in the PUSCH, by using the PRRA.

We compare the performance of above systems by simulation in MATLAB environment. The simulation parameters are provided in Table I, and results are averaged over 50 independent simulations. We introduce two metrics.
1) *System Throughput Gain*. It is the percentage throughput gain compared to S_1, i.e., $\Delta Th_x = [{}^{(Th_x - Th_{S_1})}/_{Th_{S_1}}] \cdot 100$, where Th_x is the system throughput in terms of number of successful communications achieved by means of system x.
2) *Energy Consumption Indicator*. It is the average number of times in which an MTC device enters in the RX/TX active state

[5], denoted as $\bar{E}_I$. For each MTC device, E_I is calculated as $E_I = E_{RA} + E_T$, where $E_{RA} \in \{1, \ldots, M_A\}$ is the number of access attempts, and E_T counts the total number of contention-based and contention-free transmissions in the PUSCH. In particular, $E_T \in \{0, 1\}$ for the non-Enhanced systems, while $E_T \in \{0, \ldots, M_A\}$ for the other ones.

Figs. 2 and 3 show $\Delta Th_x(\%)$ and the Energy Consumption Indicator $\bar{E}_I$, respectively. Let us start considering a moderate load, i.e., $N_{MTC} = 40000$. Among all the static dimensioning systems, only ES_1 and ES_2 show a throughput gain with respect to the related standard systems. Instead, as expected, ES_3 does not show improvement with respect to S_3 because, in this light traffic conditions, $M \leq 500$ for each RA cycle. By comparing ES_2 with DURD, we underline that the PRRA improves the performance of a static system also compared to an optimized dynamic solution. As regards the EDURD, the results are the same of the ones obtained by ES_2, being the optimal dimensioning $T^*_{pr} = 2$ for any RA cycle. Finally, EeDURD obtains the same results as EDURD, confirming that the predictive estimation of M is efficient in the considered working zone. As regards a medium-high load ($N_{MTC} = 50000$), compared to the previous case, also ES_3 shows a gain with respect to S_3. As regards the dynamic systems, EDURD shows the similar performance of the ES_3, because the optimal dimensioning is equal to $T^*_{pr} = 3$ for almost the totality of the RA cycles. In addition, it achieves the best performance in terms of energy consumption. Also in this case, the predictive estimation of M is effective.In the high load scenarios ($N_{MTC} \in \{75000, 100000\}$), among the static systems the best performances in throughput are obtained with ES_1. The EDURD system has a good gain with respect to any static system. In particular, the highest gain is achieved for $N_{MTC} = 75000$ because the optimal dimensioning ranges from $T^*_{pr} = 3$ to $T^*_{pr} = 1$, while for $N_{MTC} = 100000$ the gain is lower, because the optimal dimensioning is $T^*_{pr} = 1$ for many RA cycles. The energy consumption is slightly higher with respect to the other dynamic systems, but in line with the best static energy consumption. As regards the estimated versions under very high load, for EeDURD the predictive estimation does not work so well, because the optimal dimensioning is $T^*_{pr} = 1$ and only $L = 54$ preambles are available. The L value is too much low with respect to the number of attempting MTC devices and the estimated value of K_C, as a function of P_C and L, is more approximated. In stead, the predictive estimation in the eDURD system shows good performance because the optimal dimensioning is $T^*_{pr} = 4$, see Fig. 1a.

V. Conclusion

In this letter, we proposed the innovative idea of exploiting unused PUSCH resources, i.e., not already allocated in contention-free mode. We presented an analytic problem formulation for determining the optimal number of collided preambles to which allocate, in contention-based mode, the unused PUSCH resources, and we proposed an iterative algorithm, termed PRRA, to implement our strategy. The proposed PRRA has been applied assuming both a static and

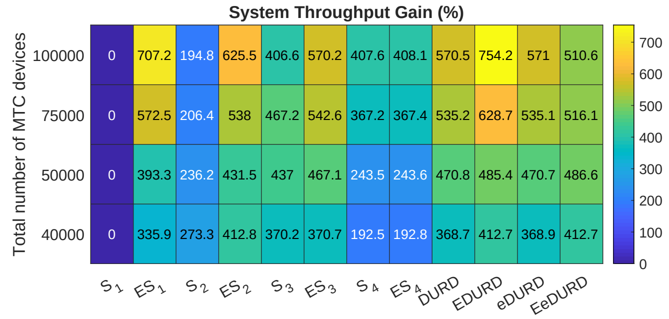


Fig. 2. Throughput Gain under different systems and M

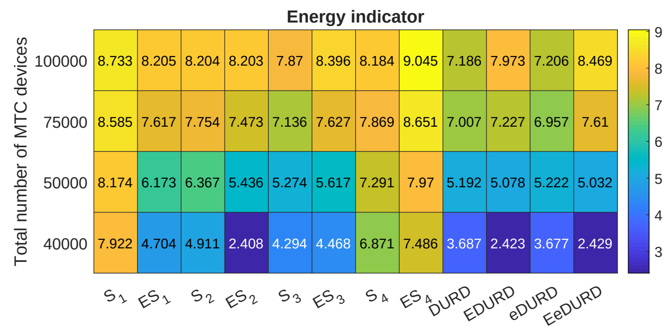


Fig. 3. Energy Consumption Indicator under different systems and M

dynamic uplink PRACH/PUSCH dimensioning. Simulation results showed that all the systems enhanced with PRRA exhibits better performance, in terms of throughput gain, with respect to the classic ones at the cost of negligible increment of the energy consumption in very high load condition.

References

[1] ITU-T, “IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond,” International Telecommunication Union, Geneva, Recommendation M.2083, 2015.

[2] J. Vidal, L. Tello-Oquendo, V. Pla, and L. Guijarro, “Performance study and enhancement of Access Barring for massive Machine-Type Communications,” *IEEE Access*, vol. 7, pp. 63 745–63 759, 2019.

[3] 3GPP, “Service accessibility,” 3rd Generation Partnership Project (3GPP), TS 22.011, 03 2020, version 17.1.0.

[4] L. Miuccio, D. Panno, and S. Riolo, “Dynamic uplink resource dimensioning for massive MTC in 5G networks based on SCMA,” in *European Wireless 2019; 25th European Wireless Conference*, 2019, pp. 1–6.

[5] ——, “Joint congestion control and resource allocation for massive MTC in 5G networks based on SCMA,” in *2019 15th Intern. Conf. on Telecomm. (ConTEL)*, 2019, pp. 1–8.

[6] ——, “Joint control of random access and dynamic uplink resource dimensioning for massive MTC in 5G NR based on SCMA,” *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5042–5063, 2020.

[7] 3GPP, “Study on RAN Improvements for Machine-type Communications,” 3rd Generation Partnership Project (3GPP), TS 37.868, 10 2011, version 11.1.0.

[8] A. B. Kihero, M. S. J. Solaija, and H. Arslan, “Inter-numerology interference for beyond 5G,” *IEEE Access*, vol. 7, pp. 146 512–146 523, 2019.

[9] 3GPP, “NR;Physical channels and modulation,” 3rd Generation Partnership Project (3GPP), TS 38.211, 04 2020, version 16.1.0.

[10] Y. Wu, C. Wang, Y. Chen, and A. Bayesteh, “Sparse code multiple access for 5G radio transmission,” in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Sep. 2017, pp. 1–6.

[11] C. Wei, G. Bianchi, and R. Cheng, “Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1940–1953, 2015.