Connection Throughput Maximization for Grant-Based NOMA Massive IoT with Graph Matching

Shashwat Mishra, Lou Salaün, Jean-Marie Gorce, Chung Shue Chen

Connection Throughput Maximization for Grant-Based NOMA Massive IoT with Graph Matching

Shashwat Mishra*†, Lou Salaün*, Jean-Marie Gorce†, Chung Shue Chen*
*Nokia Bell Labs, 12 Rue Jean Bart, 91300 Massy, France
†Institut National des Sciences Appliquées (INSA) de Lyon, 6 Av. des Arts, 69621 Villeurbanne, France
Email: shashwat.mishra@nokia.com, lou.salaun@nokia-bell-labs.com, jean-marie.gorce@insa-lyon.fr,
chung_shue.chen@nokia-bell-labs.com

***Abstract*—We propose a framework for maximizing the number of machine-type devices connected in the uplink of a Narrow-band Internet of Things (NB-IoT) network using non-orthogonal multiple access (NOMA). The system is based on the fast-uplink grant (FUG), where the base station (BS) schedules the access for active devices requesting connection. This problem is a mixed-integer non-convex problem and real-time solutions using general solvers are computationally prohibitive. The proposed scheduling solution comprises efficient device clustering and optimum power allocation using a bipartite graph matching approach, termed connection throughput maximizing full matching with pruning (CTMBM). Different from the other solutions of state-of-the-art, our proposed scheme considers scheduling over multiple transmission time intervals while considering the transmission deadlines and quality of service (QoS) for the devices. Additionally, we provide a method for priority scheduling of a subset of devices. We compare our solution to the state-of-the-art schemes and analyze the achieved gains through Monte-Carlo computer simulations.**

***Index Terms*—Connection Throughput, mMTC, Fast-Uplink Grant, NOMA, Graph Matching**

I. Introduction

The trends in recent network growth suggest a paradigm shift in the traffic demands, as they move dominantly towards machine-to-machine (M2M) communication supported over legacy LTE technology [1]. The distinguishing feature of M2M devices is minimal human interaction and sporadic short packet transmission. Massive machine type communication (mMTC), coined by 3GPP (3rd Generation Partnership Program), is expected to be adopted for industrial IoT (IIoT) use cases and beyond-5G (B5G) systems [2] which will be driven by the proliferation of M2M deployments. Consequently, there is a compelling need to develop systems that can support a massive number of connected devices using limited radio spectrum and computational capabilities. A common feature among all the deployments is predominantly uplink transmission where low-power transmission is desirable. As the competition for radio resources between coexisting technologies increases, the importance of efficient power control and device scheduling becomes increasingly critical.

Non-orthogonal multiple access (NOMA) is a promising solution to alleviate the need for additional spectrum for supporting massive access [3]. NOMA operates through the superposition of signals from multiple transmitters on a given time-frequency resource by using successive interference cancellation at the base station (BS). Several studies propose grant-free solutions for NOMA-based mMTC systems, where the detection at the BS is carried out using compressive sensing-based solutions founded on the sparse activity assumption. Typical approaches in these kinds of solutions include approximate message passing-based user-activity detection and channel estimation [4]. However, these solutions are usually computationally complex, sometimes requiring customized machine-learning models for implementation. Additionally, as the density of devices increases, the sparsity assumption may not hold which is crucial for compressive sensing-based approaches. It is important to note that uncoordinated random access leads to excessive collisions among device packets and subsequently increases the latency. As such, it is difficult to guarantee the quality of service (QoS) to devices. On the other hand, fully coordinated schemes will result in intensive signaling overheads and may not be useful for short packets. Therefore, some solutions like [5] become less effective due to excessive scheduling overheads.

To support an expanding category of low mobility devices such as smart grids, smart homes, and environment monitoring, 3GPP has proposed fast-uplink grant [6] as a practical compromise between signaling overheads and access latency. In a fast-uplink grant scheme, devices do not send random access (RA) scheduling requests. Instead, the BS will actively allocate uplink resources to those devices. Moreover, in contrast to uncoordinated transmission, devices are scheduled by the BS, and hence collisions can be avoided. Recently, there has been a growing interest in leveraging fast-uplink grants to support mMTC. The authors in [7] propose a fast-uplink grant method that additionally allows for NOMA-based user-pairing, outperforming the scheme in [8]. However, this scheme relies on a source traffic predictor along with the use of a probabilistic sleeping multi-armed bandit, which is computationally prohibitive for high device density scenarios.

Motivated by the current state-of-the-art solutions, we propose a bi-partite graph-matching framework that performs joint user-pairing and power allocation to users, utilizing NOMA to enhance the uplink connection throughput, which is the total number of devices in the system that achieve their target

data rate under the power budget, transmission deadline and successive interference cancellation (SIC) constraint. In our previous work [9], we addressed the problem of connectivity maximization without any deadlines in the downlink for NB-IoT networks using a similar graph-matching approach with the key distinction of per-PRB power budget instead of the per-device power budget which is important in the uplink. We believe that the metric of user connection throughput is much more relevant for the presented system model as opposed to other metrics like sum-data rate and user access delay as studied in the literature. Note that our model already takes into account the user deadlines and scheduling priority. The proposed framework enhances connectivity, executed at the BS to schedule devices, and allocates them power using the fast-uplink grant mechanism. The key contributions of this work are as follows:

- We propose a graph matching-based technique called CTMBM, for user association and power allocation. Unlike the work in [5], our framework is not limited to 2 devices per time-frequency resource and allows the scheduling of devices with constraints of deadlines.
- Furthermore, we compare CTMBM with the single-tone sub-carrier and power assignment Algorithm 1 in [5]. Since the latter algorithm is originally limited to a single time interval, we propose a greedy strategy that accommodates multiple time intervals and device deadlines. We analyze the complexity of this heuristic against the CTMBM algorithm.
- Additionally, we evaluate the performance of CTMBM considering different service classes for devices, each with a different priority of transmission, and analyze the trade-off between priority scheduling and connection throughput.

II. System Description

Consider a set of single antenna machine-type devices $\mathcal{D} = \{1, \cdots, D\}$, each with the same target data rate of R kbps sending data using the fast-uplink grant-based access protocol [6]. We consider a single-cell system with a single antenna BS located at the center. We assume that the channel gain for device d, consisting of the path-loss and the Rayleigh fast fading, is available at the BS without error for the scheduling operation. The channel gain for device d is $g_d \triangleq |h_d|^2 PL_d$. where $h_d \sim \mathcal{CN}(0,1)$ represents the Rayleigh fast fading and PL_d is the distance-dependent path-loss as specified in the cellular IoT specification [10]. We chose this specific model to cover a broad range of deployment scenarios. We assume that the devices have a quasi-static channel gain for the duration of one allocation round, i.e. starting from the initiation of the access request to the transmission of the packet, which is a realistic assumption for narrow-band channel under low mobility conditions [11].

A sub-carrier is denoted as f and has a bandwidth of 3.75 kHz for the single-tone uplink operation. This gives us the set $\mathcal{F}$ of 48 sub-carriers in a system with 180 kHz bandwidth. The resource allocation grid in the present work consists of these 48 sub-carriers and extends across 10 frames, the set of which is denoted as $\mathcal{T}$. Each frame t has a duration of 10 ms in the time domain, following LTE frame definition [12]. We consider single-tone allocation, meaning each user equipment (UE) gets assigned only one frequency sub-carrier f from the set of sub-carriers $\mathcal{F}$. We consider that each user is assigned a resource grant (RG), characterized by the tuple (t, f), consisting of one sub-carrier and extending for one frame.

A. Signaling Characterization for NOMA

We employ power domain NOMA for user multiplexing in the uplink. The index of the device receiving the m-th packet on sub-carrier f at time t is denoted by $x_{t,f}(m)$. In other words, the first decoded packet is $x_{t,f}(1)$, followed by $x_{t,f}(2)$, etc. We define $\boldsymbol{X}_{t,f} \triangleq \{x_{t,f}(1), \ldots, x_{t,f}(m)\}$ as the set of devices allocated to the RG corresponding to sub-carrier f at time t. The cardinality of this set is denoted by $|\boldsymbol{X}_{t,f}|$ and must satisfy $|\boldsymbol{X}_{t,f}| \leq M$, due to the system SIC constraint, i.e., one RG cannot support more than M superposed devices. In a practical implementation, $\boldsymbol{X}_{t,f}$ is represented by a list sorted in the SIC decoding order so that accessing any element $x_{t,f}(m)$ from its decoding order m can be done in constant time.

We choose the SIC decoding order on RG (t, f) in the decreasing order of the users' received power [13], i.e. the strongest received signal is decoded first, as this guarantees minimum power consumption for the users. Therefore we have the following relation among users superimposed on RG (t, f):

$$g_{x_{t,f}(1)} p_{x_{t,f}(1)} \geq \cdots \geq g_{x_{t,f}(m)} p_{x_{t,f}(m)}.$$

The signal-to-interference-plus-noise ratio (SINR) for the m-th decoded device on RG (t, f) can be expressed as:

$$\gamma_{x_{t,f}(m)} = \frac{g_{x_{t,f}(m)} p_{x_{t,f}(m)}}{I_{t,f,m} + N}, \tag{1}$$

where $p_{x_{t,f}(m)}$ is the transmit power for device $x_{t,f}(m)$, and $I_{t,f,m}$ is the interference caused by other devices on the same sub-carrier given by:

$$I_{t,f,m} \triangleq \sum_{i=m+1}^{|\boldsymbol{X}_{t,f}|} g_{x_{t,f}(i)} p_{x_{t,f}(i)}. \tag{2}$$

$N = N_0 BF$ is the additive white Gaussian noise for the devices where N_0 is the noise spectral density, B is the sub-carrier bandwidth and F is the noise figure.

B. System Constraints

We recall again the system SIC constraint specified in the description of the NOMA signaling in II-A such that each RG can support a superposition for at most M devices. Additionally, each device in the system must obey the following constraints. Firstly, the uplink transmit power of each scheduled device must be under the device's power budget:

$$p_{x_{t,f}(m)} \leq P, \ \forall m \leq M, t \in \mathcal{T}, f \in \mathcal{F}. \tag{3}$$

Here, P is the power budget of the device, as specified by the 3GPP cellular IoT standard [10] and subsequently elaborated in Section V.

The data rate for device $x_{t,f}(m)$ is expressed as:

$$r_{x_{t,f}(m)} = B \log_2 \left(1 + \gamma_{x_{t,f}(m)}\right) \tag{4}$$

Each device must achieve a minimum data rate of R kbps, therefore the device $x_{t,f}(m)$ is considered to be connected if:

$$r_{x_{t,f}(m)} \geq R. \tag{5}$$

Note that R is the instantaneous rate achieved by devices by transmitting their packets on the assigned resource grant. We maintain this definition for the present work to make a consistent comparison with the state-of-the-art techniques [5]. However, in practice, we are concerned with the total number of packets transmitted by the device over the scheduled resource grant rather than the instantaneous rate. Therefore, we assume that each device has a fixed number of bits to transmit during our system simulations.

The power required to achieve this rate R can be obtained from (4) and (2) as:

$$p_{x_{t,f}(m)} = \frac{\Gamma}{g_{x_{t,f}(m)}}(I_{t,f,m} + N), \tag{6}$$

where $\Gamma = 2^{\frac{R}{B}} - 1$ is the target SINR to achieve the target rate of R kbps. Additionally, the grant-based devices must respect the delay requirements for data transmission. As a result of this, the packet of device d must be received before its deadline t_d. It is expected that the deadlines for devices in B5G networks may be a few tens of milliseconds even though the devices have limited mobility and relatively static fading. Therefore even within the current resource allocation grid, all RGs may not be usable by all devices.

III. Problem Formulation

We define the following function $Z(\cdot,\cdot)$ that represents the system's connection throughput, which is the number of devices successfully connected under their QoS and delay requirements:

$$Z(\boldsymbol{p}_{x_{t,f}}, \boldsymbol{X}_{t,f}) \triangleq \sum_{m=1}^{|\boldsymbol{X}_{t,f}|} \mathbb{1}\left(r_{x_{t,f}(m)} \geq R \wedge t \leq t_{x_{t,f}(m)}\right).$$

Here, $\mathbb{1}$ is the indicator function that takes value 1 if the rate for device $x_{t,f}(m)$ is greater than or equal to the target service rate R, and the device follows its delay constraints. Thus, $Z(\boldsymbol{p}_{x_{t,f}}, \boldsymbol{X}_{t,f})$ embeds both the QoS constraint (5) and the delay constraint. Thus, the uplink connection throughput maximization problem can now be formulated as follows:

$$\begin{aligned}
\text{maximize} \quad & \sum_{t=1}^{|\mathcal{T}|}\sum_{f=1}^{|\mathcal{F}|} Z(\boldsymbol{p}_{x_{t,f}}, \boldsymbol{X}_{t,f}) && (\mathcal{P}) \\
\text{subject to} \quad & C1: \ p_{x_{t,f}(m)} \leq P, \forall\, m, x \in \mathcal{D}, t \in \mathcal{T}, f \in \mathcal{F}, \\
& C2: \ |\boldsymbol{X}_{t,f}| \leq M, \forall\, t \in \mathcal{T}, f \in \mathcal{F}, \\
& C3: \ \sum_{f \in \mathcal{F}} |\{d\} \cap \boldsymbol{X}_{t,f}| \leq 1, \forall\, d \in \mathcal{D}.
\end{aligned}$$

The objective function in (**??**) maximizes the total number of devices in the system that satisfies their QoS and deadline requirement. Constraint C1 signifies the maximum transmit power for each device as defined in (3). Constraint C2 stands for the system limit that supports at most M devices superimposed per RG. Constraint C3 enforces that each device is allocated at most one sub-carrier, which is aligned to the single-tone uplink operation [10] for supporting massive connectivity. We can readily verify that this problem is a mixed integer problem due to the binary nature of $Z(,.,)$ and non-convex due to the rate constraint (5). The problem is known to be NP hard [14] for the general case but can be solved under practical system considerations.

IV. Proposed Framework

The following construction is based on the insight that the uplink transmit power for a device essentially depends on the SIC level m at which it wants to transmit and its channel gain. The transmit power for device $x_{t,f}(m)$ in order to achieve the target data rate R can be written as:

$$p_{x_{t,f}(m)} = \frac{\Gamma N (1+\Gamma)^{M-m}}{g_{x_{t,f}(m)}}, \tag{7}$$

which is obtained by iteratively evaluating (6) using (2) [15].

We shall now elaborate on the construction of bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ as shown in Figure 1, where $\mathcal{V}$ is the set of vertices, divided into two parts. The first part, shown on the left side of Figure 1, contains the set of devices $\mathcal{D}$ requesting a connection, and the second part, on the right side in Figure 1, corresponds to the resource vertices made of the set of RGs (t, f) for all $t \in \mathcal{T}, f \in \mathcal{F}$, where each RG has up to M devices superimposed on it, one on each of the SIC levels $m \in \{1, \ldots, M\}$. A slot is uniquely identified by the triplet (t, f, m). The set of edges is denoted by $\mathcal{E}$. We put an edge between device d and slot (t, f, m) if and only if $t \leq t_d$ and its power calculated using (7) is less than or equal to P. Due to the deadline requirement for devices, there does not necessarily exist an edge between all devices and all resource vertices, making the bipartite graph $\mathcal{G}$ incomplete. The set of weights is denoted by $\mathcal{W}$. The weight of an edge, $w_{d,t,f,m}$, is the transmit power of device d when connected to slot (t, f, m), obtained using (7) as:

$$w_{d,t,f,m} = \frac{\Gamma N (1+\Gamma)^{(M-m)}}{g_d}, \tag{8}$$

We now elaborate Algorithm 1 for maximizing user connection throughput in a frame. On line 1, we construct the bipartite graph $\mathcal{G}$ at the BS after obtaining the CSI from the contending devices, as described in the last paragraph. Then we obtain the minimum weight full matching $\mathcal{E}^*$ of $\mathcal{G}$ using an optimal linear sum assignment algorithm according to [16] as shown on line 2. More precisely, $\mathcal{E}^*$ is a subset of $\mathcal{E}$ such that the sum of all edge weights is the least among all full matchings, where each device gets connected to one RG. The proposed formulation effectively considers one transmit packet per device per RB, however, it would be possible to offer differentiated services

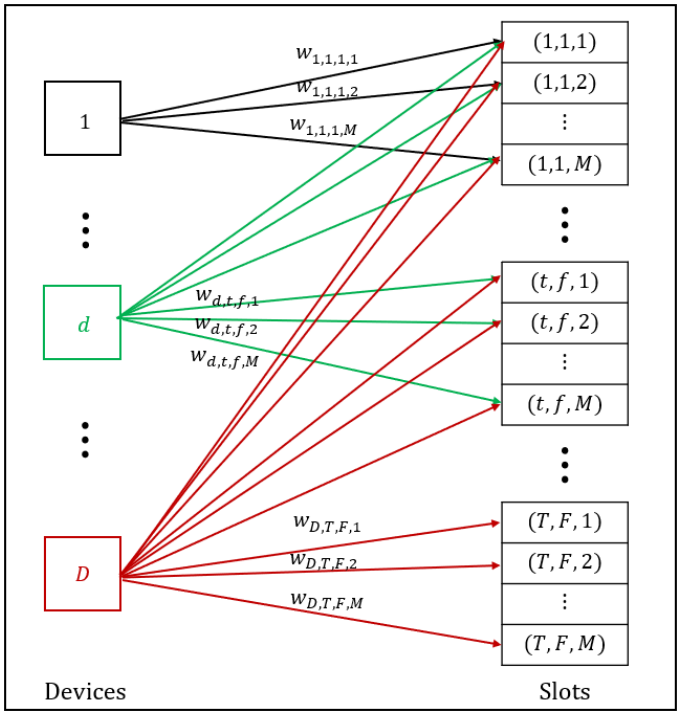

Fig. 1. The construction of bipartite graph for device-RG matching

to devices by allocating multiple RGs to a single UE. In this case, the connection throughput of the UE will be evaluated over the aggregated packets, but for the sake of brevity, we have limited the discussion here to one RG per device.

Additionally, we demonstrate how this framework can be efficiently adapted to serve devices distributed into different service classes. It is common for mMTC deployments to consist of devices that require different levels of service priority based on constraints such as delay sensitivity and fluctuating traffic patterns. Assume that the devices can be classified in the following classes $\mathcal{C}_1, \ldots, \mathcal{C}_n$, where the devices in class $\mathcal{C}_1$ have the highest priority and devices in class $\mathcal{C}_n$ have the lowest priority. The classes can be thought of as a partition of the device set $\mathcal{D}$ into n disjoint sets such that each device belongs to a unique class. CTMBM can address the priority of scheduling through the inclusion of a factor α_i in (8) in the edge weight of device d belonging to class $\mathcal{C}_i$ as follows:

$$w_{d,t,f,m} = \frac{\Gamma N (1+\Gamma)^{(M-m)}}{g_d \alpha_i}. \tag{9}$$

Effectively, this formulation decreases the edge weight for high-priority devices, thus incentivizing their inclusion into the full matching despite having a lower channel gain. It is evident from (9) that this formulation does not affect the weights of other devices on the same RG. However, we must note that the actual transmit power assignment for the device must still be in accordance to (7) and must be upper-capped at the power budget P due to (3). We set $\alpha_n = 1$. For all $i \in \{1, \ldots, n-1\}$, the value of α_i is chosen as follows:

$$\alpha_i > \frac{\max\limits_{(d,t,f,m) \in \mathcal{E} \wedge d \in \mathcal{C}_i} \Gamma N (1+\Gamma)^{(M-m)}/g_d}{\min\limits_{(d',t',f',m') \in \mathcal{E} \wedge d \in \mathcal{C}_{i+1} \cup \cdots \cup \mathcal{C}_n} w_{d',t',f',m'}}. \tag{10}$$

Equation (10) guarantees that $w_{d,t,f,m} < w_{d',t',f',m'}$, for any edges $(d,t,f,m), (d',t',f',m') \in \mathcal{E}$, where d belongs to class $\mathcal{C}_i$, and d' belongs to a lower priority class $\mathcal{C}_j$, $j > i$. In other words, device scheduling priority is guaranteed.

Algorithm 1 Connection Throughput Maximizing Bipartite Matching (CTMBM)

Input: $\mathcal{D}$, $\mathcal{F}$, $\mathcal{T}$, B, M, $P_{\max}$, and g_d for all $d \in \mathcal{D}$

1: **Initialization:** Form bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ as described in Section IV
2: Compute a minimum weight full matching $\mathcal{E}^*$ by running the matching algorithm in [16] on $\mathcal{G}$
3: **for all** $(d,t,f,m) \in \mathcal{E}^*$ **do**
4: $\quad x_{t,f}(m) \leftarrow d$
5: **end for**
6: Derive the power vector $\boldsymbol{p}$ using Eqn. (2) and (6)

Output: $\boldsymbol{p}$ and $\boldsymbol{X}_{t,f}$ for all $t \in \mathcal{T}, f \in \mathcal{F}$

V. Simulation Results

We now present the performance of the proposed algorithms, evaluated through system-level simulations. The key simulation parameters are given in Table I, which are taken from the experimental studies and standards [10], [17], [18]. These parameters are suitable for the industrial wireless IoT use case. The fast fading in the system is frequency flat Rayleigh fading. We assume that perfect SIC for NOMA can be carried out by the receiver unless otherwise stated. We consider that all UEs have a single transmit antenna and that the BS uses one receive antenna. Devices are deployed randomly following a uniform distribution in a square cell of side 1000 m unless specified otherwise. All devices have the same target data rate of R kbps. The presented simulation results are averaged over 1000 independent trials. Each RB of 180 kHz bandwidth has 48 sub-carriers and spans 10 frames each of duration 10 ms. We analyze the performance in terms of the system connection throughput in a critically loaded system with 480 RGs and {480, 960, 1440} devices requesting a connection with $M \in \{1, 2, 3\}$ devices superimposed on each sub-carrier respectively.

In Figure 2, we compare our proposed algorithm with the state-of-the-art algorithms. Here, we set $|\mathcal{T}| = 1$, i.e., only one-time frame and we remove the deadline constraint $t \leq t_{x_{t,f}(m)}$ from the counting function Z so that there are not any deadlines for the devices anymore. We set $M = 2$ for all the algorithms so that there are at most two devices on each sub-carrier. `ALG 1` represents Algorithm 1 in [5], which is a heuristic based on binary integer programming. `Optimal` is the mixed-integer linear programming formulation in [5], which is solved by using CVX [19] to obtain the optimal solution with the branch and bound method. We see that CTMBM achieves similar connection throughput as the `Optimal` scheme. However, `Optimal` and `ALG 1` have a computational

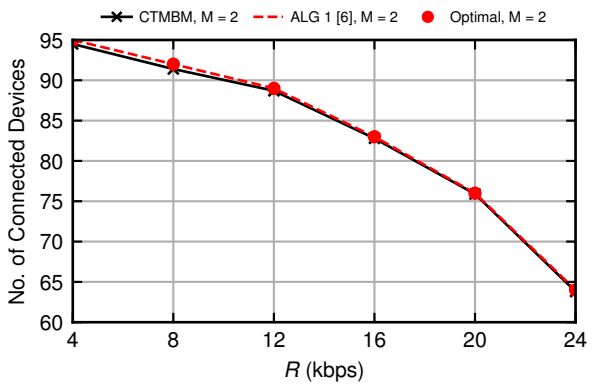

Fig. 2. Comparison of CTMBM with Algorithm 1 in [5], `Optimal` represents the solution obtained using CVX solver

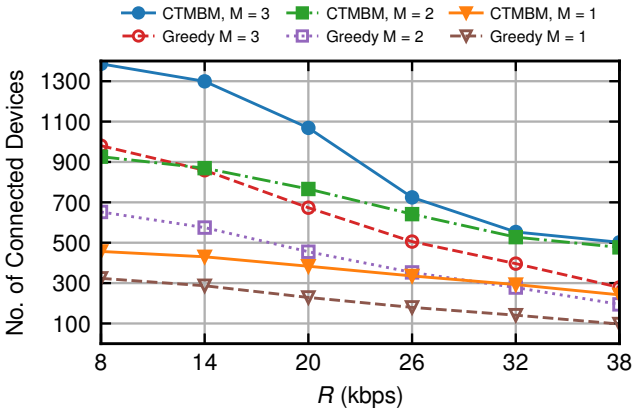

Fig. 3. Performance with varying data rate requirement

complexity of $O\left(2^D D^3\right)$ and $O\left(2^D D^2\right)$ respectively, whereas CTMBM has a significantly lower computational complexity of $O\left(D^2\sqrt{D}\right)$ [16]. Furthermore, CTMBM generalizes the connection throughput optimization problem in [5] to multiple time-frames, while also allowing for transmission deadlines and priority scheduling as shown in the results below.

Figure 3 shows the connection throughput with the variation of data rate. Effectively, we consider that the device has x bits to transmit in the 10 ms time-frame, giving it a data rate of $x/10$ kbps. Starting with 80 bits, we go up to 390 bits per frame, which is typical for status update packets [11], [20]. We introduce a baseline solution, labeled in the figure as `Greedy`, by augmenting Algorithm 1 from [5] which uses binary integer programming for joint sub-carrier and power allocation, The solution is near-optimal for a single transmit time interval, but the original formulation does not take into account the users' transmission deadlines. The proposed extension is the greedy strategy `Greedy`, which runs sequential allocation per time interval over the subset of devices that are yet unassigned and have deadlines falling within the current time interval. We observe that CTMBM outperforms `Greedy` due to its flexibility in assigning the devices to the best possible RG over all the time intervals in the RB. The gain in connection throughput is 52% for $M = 3$ and 79% for $M = 2$ respectively over the `Greedy` strategy. CTMBM provides a gain of 149% at $M = 3$ and 95% at $M = 2$ respectively over CTMBM OMA with $M = 1$, where we assign a single device to each RG.

In Figure 4, we show the performance of CTMBM under different deployment sizes at $R = 26$ kbps. We see that CTMBM achieves a superior connection throughput to OMA over a broad range of cell sizes. The connection throughput of all the algorithms naturally declines as the cell size becomes bigger due to the increase in the pathloss for distant users. However, CTMBM consistently outperforms the OMA solution of CTMBM with $M = 1$, connecting 125% more devices with $M = 3$ and 95% more devices with $M = 2$. Furthermore, CTMBM with $M = 3$ connects 59% more devices than the `Greedy` solution with $M = 3$.

We highlight the performance of CTMBM when considering 2 different service classes with $M = 2$ in Figure 5. Here, the plots in red and blue are priority-based CTMBM (P-CTMBM), where edge weights are assigned using (9). The target data rate is fixed as $R = 26$ kbps. We consider that the devices are divided into two classes, with the high-priority class labeled Class 1 and the rest of the devices labeled Class 2. Such a system may be used to support heterogeneous traffic, for example, to prioritize alarms over sensor observation. Further, we assume an overloaded scenario where there are 25% more devices than the available RGs, i.e. 1200 devices and 480 RGs. This assumption is made since assigning priorities is meaningful only when we have sufficient competing devices otherwise the total connectivity will be affected adversely by a marginal increase in the performance of high-priority devices, since the high-priority devices may have a worse channel condition than other candidate devices. As a result, the opportunity for better devices to be connected in the same RG is reduced, hence decreasing the connection throughput. In this overloaded case, P-CTMBM outperforms CTMBM. We see that with priority scheduling, we enhance the connection throughput of high-

TABLE I
KEY SYSTEM SIMULATION PARAMETERS

Parameter	Value
UE Deadline (t_d)	Random Uniform $U(2, 11)$ time frames
Carrier Frequency	900 MHz
RB Bandwidth	180 kHz
Sub-Carrier Bandwidth (B)	3.75 kHz
Path-Loss (PL)	$120.9 + 37.6 \log \frac{D}{1000} - G + L_P$ dB
UE Antenna Gain (G)	4 dB
Indoor Penetration Loss (L_P)	20 dB
UE Max Transmit Power (P)	23 dBm
AWGN Spectral Density (N_0)	-174 dBm/Hz
Noise Figure (F)	5 dB
Total Resource Grants (RG)	480
System SIC Limit (M)	$M \in \{1, 2, 3\}$
Percentage of Indoor UEs	80%

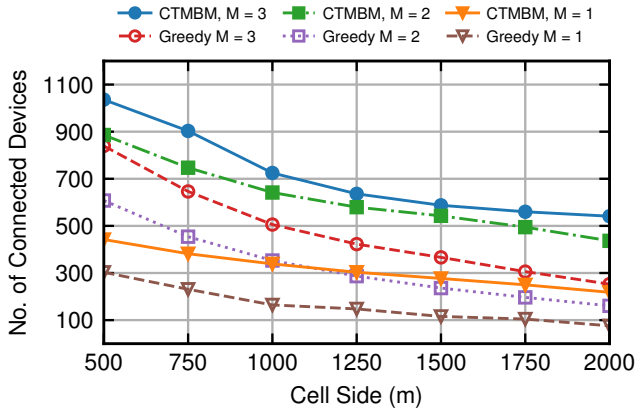

Fig. 4. Performance with different cell size

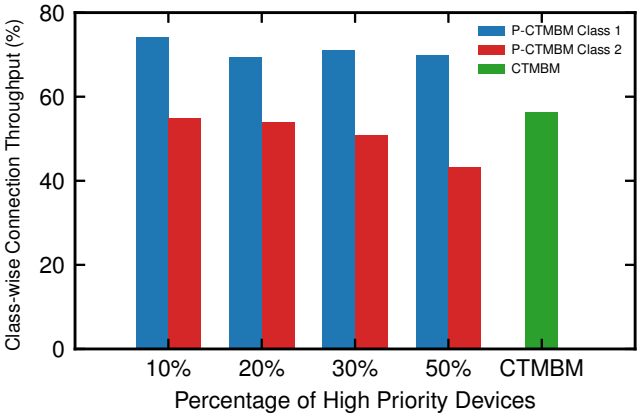

Fig. 5. Performance of CTMBM and P-CTMBM with two service classes

priority devices while maintaining the same overall connection throughput for the system. Compared to equal-opportunity scheduling in CTMBM, P-CTMBM provides 15% higher connection throughput for the priority devices of class 1, averaged over different device-class distributions. This gain is higher when the number of contending priority devices is small and it stagnates as the number of class 1 devices increases as almost all eligible devices get connected. P-CTMBM maintains the better connection throughput of class 1 devices by suppressing the connection throughput of class 2 devices.

VI. Conclusion

We address the problem of connection throughput maximization in the uplink of mMTC networks supported through NB-IoT using NOMA. We formulate the resource allocation optimization problem using a bipartite graph matching approach and present the CTMBM algorithm for solving this problem efficiently. Through computer simulations, we study the performance of the proposed scheme with varying target service rates and cell sizes. Additionally, we demonstrate the versatility of our proposed framework in accommodating different device classes with varying degrees of scheduling priority. We show that in all the evaluated scenarios, CTMBM can steadily outperform OMA, connecting up to 95% and 149% more devices than OMA with $M = 2$ and $M = 3$, respectively. Furthermore, we show that CTMBM outperforms the connection throughput of the greedy allocation strategy-based single-tone scheme in [5], with a significantly lower computational complexity.

References

[1] Q. Bi, "Ten trends in the cellular industry and an outlook on 6G," *IEEE Communications Magazine*, vol. 57, no. 12, pp. 31–36, 2019.

[2] N. H. Mahmood, G. Berardinelli, E. J. Khatib, R. Hashemi, C. de Lima, and M. Latva-aho, "A functional architecture for 6G special purpose industrial iot networks," *IEEE Trans. on Ind. Informat.*, pp. 1–11, 2022.

[3] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys & Tutorials*, vol. 22, no. 3, pp. 1805–1838, 2020.

[4] R. B. Di Renna and R. C. de Lamare, "Joint channel estimation, activity detection and data decoding based on dynamic message-scheduling strategies for mMTC," *IEEE Trans. on Commun.*, pp. 2464–2479, 2022.

[5] A. E. Mostafa, Y. Zhou, and V. W. S. Wong, "Connection density maximization of narrowband IoT systems with NOMA," *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4708–4722, 2019.

[6] 3GPP, "Evolved Universal Terrestrial Radio Access; Medium Access Control Protocol Specification," Tech. Rep. 36.321, Jan. 2018, version 15.0.0. [Online]. Available: http://www.3gpp.org/DynaReport/36321.htm

[7] M. E. Tanab and W. Hamouda, "Efficient resource allocation in fast-uplink grant for machine-type communications with NOMA," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 18 113–18 129, 2022.

[8] H. Jiang, Q. Cui, Y. Gu, X. Qin, X. Zhang, and X. Tao, "Distributed layered grant-free non-orthogonal multiple access for massive MTC," in *IEEE PIMRC*, 2018, pp. 1–7.

[9] S. Mishra, L. Salaün, J.-M. Gorce, and C. S. Chen, "Maximizing downlink user connection density in NOMA-aided NB-IoT networks through a graph matching approach," in *IEEE VTC-Fall*, 2022, pp. 1–7.

[10] 3GPP, "Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)," Tech. Rep. 45.820, Dec. 2015. [Online]. Available: http://www.3gpp.org/DynaReport/45820.htm

[11] D. Malak, H. Huang, and J. G. Andrews, "Throughput maximization for delay-sensitive random access communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 709–723, 2019.

[12] 3GPP, "Evolved Universal Terrestrial Radio Access; Radio Resource Control; Protocol specification," Tech. Rep. 36.311, Mar. 2018, version 15.3.0. [Online]. Available: http://www.3gpp.org/DynaReport/36311.htm

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

[14] Z. Mlika and S. Cherkaoui, "Competitive algorithms and reinforcement learning for NOMA in IoT networks," in *IEE ICC*, 2021, pp. 1–6.

[15] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2736–2743, 2017.

[16] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, pp. 325–340, 2005.

[17] B. Anass, Z. Zhang, Y. Li, and Y. Chi, "3GPP defined 5G requirements and evaluation conditions," *NTT DOCOMO Technical Journal*, vol. 19, no. 3, pp. 13–23, 2018. [Online]. Available: https://www.docomo.ne.jp/english/binary/pdf/corporate/technology/rd/technical_journal/bn/vol19_3/

[18] *NB-IoT & LTE Cat-M1 Field Measurements and SLA Verification*, 2018. [Online]. Available: https://www.keysight.com/fr/en/assets/7018-06022/application-notes/5992-2747.pdf

[19] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[20] C. Bockelmann *et al.*, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28 969–28 992, 2018.