DOI:
10.1109/JSAC.2017.2724442

Document Version
Peer reviewed version

On the Fundamental Limits of Random Non-orthogonal Multiple Access in Cellular Massive IoT

Mahyar Shirvanimoghaddam, *Member, IEEE,*
Massimo Condoluci, *Member, IEEE,* Mischa Dohler, *Fellow, IEEE,*
Sarah J. Johnson, *Member, IEEE*

Abstract

Machine-to-machine (M2M) constitutes the communication paradigm at the basis of Internet of Things (IoT) vision. M2M solutions allow billions of multi-role devices to communicate with each other or with the underlying data transport infrastructure without, or with minimal, human intervention. Current solutions for wireless transmissions originally designed for human-based applications thus require a substantial shift to cope with the capacity issues in managing a huge amount of M2M devices. In this paper, we consider the multiple access techniques as promising solutions to support a large number of devices in cellular systems with limited radio resources. We focus on non-orthogonal multiple access (NOMA) where, with the aim to increase the channel efficiency, the devices share the same radio resources for their data transmission. This has been shown to provide optimal throughput from an information theoretic point of view. We consider a realistic system model and characterize the system performance in terms of throughput and energy efficiency in a NOMA scenario with a random packet arrival model, where we also derive the stability condition for the system to guarantee the performance.

Index Terms

Internet of Things, Machine-to-machine, Machine-type communication, non-orthogonal multiple access, NOMA.

M. Shirvanimoghaddam is with the School of Electrical and Information Engineering, The University of Sydney, NSW, Australia (email: mahyar.shirvanimoghaddam@sydney.edu.au).

Sarah. J. Johnson is with School of Electrical Engineering and Computer Science, The University of Newcastle, NSW, Australia (e-mail: sarah.johnson@newcastle.edu.au).

M. Condoluci and M. Dohler are with the Centre for Telecommunications Research, Department of Informatics, King's College London, UK (email: {massimo.condoluci; mischa.dohler}@kcl.ac.uk).

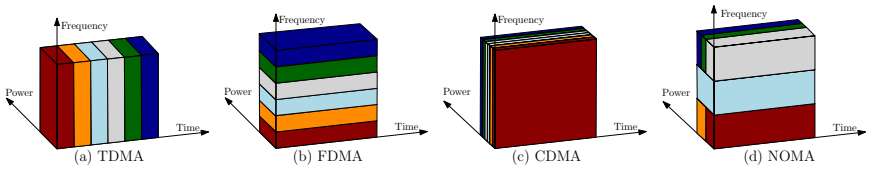

Fig. 1. Multiple access techniques.

I. Introduction

Internet of Things (IoT) represents a major technology trend, which is revolutionizing the way we interact with our surrounding physical environment as our everyday physical objects will be transformed into information sources [1, 2]. The basic enabler for IoT is the massive connectivity between devices, e.g. sensors and actuators, and with the underlying data transport infrastructure without, or with limited, human interaction. Machine-to-machine (M2M) aims at providing this communication infrastructure for the emerging IoT applications and services in the near future [3, 4]. The most promising solution proposed for M2M communications is wireless cellular, e.g., GSM, GPRS, 3G, WiMAX, as well as Long Term Evolution (LTE) and LTE-Advanced (LTE-A), due to their excellent coverage, mobility and scalability support, good security features, and the availability of the infrastructure almost everywhere [5, 6]. The focus of this study is on cellular M2M communications for massive IoT.

A. Background and Motivations

The third generation partnership project (3GPP) has already initiated several task groups to standardize several low-power solutions for emerging M2M communications, which are referred to as machine-type communications (MTC) in the 3GPP terminology. Such solutions include extended coverage GSM (EC-GSM), LTE for machine-type communication (LTE-M), and narrow band IoT (NB-IoT) [5, 7]. These standards have been proposed on top of existing cellular standards by exploiting new control and data channels to increase capacity per cell and power saving functionality to extend battery life [5].

Moving into the future, major improvements in system performance will require a more substantial shift from current protocols and designs originally proposed for human based communications. The fact that the data channels are orthogonally allocated to the devices in current cellular systems makes it a potential bottleneck for future M2M applications, where a large number of devices want to communicate with the base station (BS) and there are not enough radio resources to be orthogonally allocated to the devices [8]. We foresee that new multiple

access (MA) techniques are essential for future cellular systems to enable multiple M2M devices to effectively share radio resources. This is particularly interesting for M2M applications with small message sizes, where the same radio resource can be shared between multiple devices to deliver their messages at the base station.

Focusing in more details on multiple access techniques, they can be divided into two broad categories [9]: *(i) uncoordinated*, where the devices transmit data using slotted random access and there is no need to establish dedicated resources; *(ii) coordinated*, where devices transmit on separate resources pre-allocated by the base station. In coordinated MA, the BS knows a priori the set of devices that have data to transmit [10]. The BS can also acquire the channel state information (CSI) of these devices based on which it allocates resources to optimize system throughput.

Multiple access techniques can be also divided into orthogonal and non-orthogonal approaches (see Fig. 1). In orthogonal MA (OMA), radio resources are orthogonally divided between devices, where the signals from different devices are not overlapped with each other. Instances of OMA are time division multiple access (TDMA), frequency division multiple access (FDMA), orthogonal frequency division multiple access (OFDMA), and single carrier FDMA (SC-FDMA) [11]. OMA approaches have no ability to combat inter-cell interference; therefore careful cell planning and interference management techniques are required to solve the interference problem. *Non-orthogonal multiple access (NOMA)* allows overlapping among the signals from different devices by exploiting power domain, code domain (such as code division multiple access), and interleaver pattern. In NOMA, signals from multiple users are superimposed in the power-domain and successive interference cancellation (SIC) is used at the BS to decode the messages. NOMA has been shown to achieve the multiuser capacity region both in the uplink and downlink and provides better performance than OMA [11].

B. Contributions and Paper Organization

In uplink NOMA multiple devices simultaneously perform transmission in a shared radio resource; therefore, their transmissions are overlapping [11–13]. NOMA has been already studied for multiple access in both uplink and downlink of wireless cellular networks, where the number of devices is usually assumed to be very small, e.g., 2 or 3 users, and the channel state information is available to optimize the transmit power. However, these assumptions are not valid anymore in massive MTC. We propose a NOMA-based multiple access strategy for massive MTC with

random packet arrivals. In the proposed approach, referred to as random NOMA, each MTC device which has data to transmit randomly chooses a sub-band and encodes its message along with its terminal identity (ID) and sends the encoded packet over the selected sub-band. As multiple devices may have selected the same sub-band, their transmissions interfere with each other. Each device randomly generates a unique seed to encode its message, therefore, the base station can then decode each device and remove its interference on the remaining devices; thus performing successive interference cancellation (SIC). We find the optimal resource allocation strategy, where the optimal number of sub-bands is found for a given available bandwidth to maximize the throughput of random NOMA. We also find the optimal bandwidth allocation in order to minimize the energy consumption of the devices.

We derive the necessary condition for the stability of the system under the proposed random NOMA strategy. We find the maximum arrival rate for a system with an initial backlog such that the number of devices which are attempting to transmit to the BS in the next time slot does not increase in time. We derive a weak stability condition, where we find the maximum arrival rate as a function of the expected number of devices which were not successfully transmitted in the previous time slot and the expected number of newly generated packets. We also find the strong stability condition, where we find the maximum arrival rate such that the probability that the system remains stable (i.e., the number of devices in the next time slot does not increase) is higher than a threshold value. Moreover, we consider an M2M system with different QoS requirements and derive the stability condition. We show that random NOMA can support high packet arrival rates and can simultaneously satisfy the diverse QoS requirements of all devices.

The remainder of the paper is organized as follows. The system model and some existing results on OMA and NOMA are presented in Section II. The random NOMA strategy is proposed in Section III. The proposed NOMA strategy is analyzed in Section IV, where we derive the stability condition for the system and characterize the maximum packet arrival rate at the base station. System parameters are optimized in Section V. In Section VI, some practical considerations of massive NOMA in massive IoT are presented. Finally, Section VII concludes the paper.

II. System Model and Some Existing Results

A. System Model

We consider a single-cell wireless network consisting of one BS located at the centre and MTC devices are uniformly distributed around the BS in an angular region with inner and outer

TABLE I

NOTATION SUMMARY

Notation	Description
R_{i}	Inner cell radius
R_{o}	Outer cell radius
N_s	Number of frequency sub-bands
W	Total available bandwidth (Hz)
W_s	The bandwidth of a frequency sub-band (Hz)
α	path loss exponent
g_i	channel gain of the i^{th} device to the BS
r_i	distance between the i^{th} device and the BS
μ_{ref}	reference SNR
P_t	transmit power of MTC device
$P_{\max}$	maximum transmit power at an MTC device
μ_r	Received SNR at the BS from a device at distance r
G	antenna gain
χ	large scale shadowing gain
h	small scale fading gain
λ	New packet arrival rate at the BS
L	MTC message size (bits)
M_s	Number of available seeds
$T(n)$	duration of a time slot when n devices are transmitting
$q(c,n)$	probability that the maximum number of devices over all the sub-bands is c when the number of active devices in n.
$t(k)$	time slot duration of a sub-band containing k devices
P_c	collision probability
$(x)_{k_1}^{k_2}$	$(x_{k_1}, x_{k_1+1}, \cdots, x_{k_2}) \in \{0,1,\cdots\}^{(k_2-k_1+1)}$ for $k_1 \leq k_2$

radii R_{i} and R_{o}. For simplicity, we assume the BS and MTC devices are equipped each with a single antenna. We assume that radio resources are divided into N_{s} frequency sub-bands each with bandwidth $W_s = W/N_s$, where W is the total available bandwidth. Table I summarizes the notations commonly used in this paper.

Following [9, 10, 13], the channel between each MTC device and the BS is modeled by path loss, shadowing and small scale fading. The received power at the BS from an MTC device located at distance r with transmit power P_{t} is given by:

$$P_{\mathrm{r}} = P_{\mathrm{t}} \chi h G r^{-\alpha}, \tag{1}$$

where α is the path loss exponent, χ is the large scale shadowing gain, h is the small scale fading gain, and G is the antenna gain. Similar to [10], we introduce the term reference signal-to-noise ratio (SNR), μ_{ref}, which is defined as the average received SNR from a device transmitting at maximum power $P_{\max}$ over the whole bandwidth W located at the cell edge, i.e. at distance R_{o}.

The received SNR can then be expressed as follows [10]:

$$\mu_r = \frac{P_{\mathrm{t}}}{P_{\max}} \mu_{\mathrm{ref}} \chi h \left(\frac{r}{R_{\mathrm{o}}} \right)^{-\alpha}. \tag{2}$$

As information symbols might be transmitted over a smaller bandwidth W_{s}, the effective noise power will be reduced by a factor W/W_{s}. Therefore, the received SNR from an MTC device located at distance r from the BS and transmitting over bandwidth W_{s} can be expressed as follows:

$$\mu_r = \frac{P_{\mathrm{t}}}{P_{\max}} \frac{W}{W_{\mathrm{s}}} \mu_{\mathrm{ref}} \chi h \left(\frac{r}{R_{\mathrm{o}}} \right)^{-\alpha}. \tag{3}$$

We assume that the channel gain $\chi h (r/R_{\mathrm{o}})^{-\alpha}$ varies very slowly in time and is known at the MTC device. This is particularly advantageous for M2M communications as the device location is usually fixed and the MTC device can obtain accurate channel information in a timely manner. Moreover, the devices can perform the channel estimation by using regular pilot signals transmitted by the BS. This assumption will significantly reduce the complexity at the BS as it does not need to estimate the channel to a very large number of MTC devices usually involved in M2M communications. Unless otherwise specified in the paper, each MTC device is assumed to control its transmit power using the channel information, such that the received SNR at the BS is μ_0. Therefore, the transmit power required for an MTC device located at distance r from the BS to achieve a received SNR μ_0 over bandwidth W_{s} is given by:

$$P_{\mathrm{t}} = P_{\max} \frac{\mu_0}{\chi h \mu_{\mathrm{ref}}} \left(\frac{r}{R_{\mathrm{o}}} \right)^{\alpha}. \tag{4}$$

For simplicity, we ignore small-scale fading and shadowing, thus the channel gain and the transmit power is mainly characterized by the distance of the MTC device to the BS.

Finally, we assume that the packet arrival rate at the BS follows a Poisson distribution with mean λ packets per second. More specifically, the number of packet transmission requests in a time interval of duration t is given by $\mathrm{Poiss}(\lambda t)$. Each MTC device is assumed to have a message of length L bits, including the device unique ID. Moreover, we consider slotted transmission and each device requests for a transmission only at the beginning of a time slot.

B. Non-orthogonal Multiple Access

In this section, we briefly overview the general uplink NOMA strategy and report some upper bounds on the maximum throughput of a cellular M2M system in coordinated and uncoordinated scenarios. For further details please refer to [10, 13].

1) The Basic Concept of Uplink NOMA: In NOMA, the devices transmit their messages over the same frequency band in the same time slot; therefore, their transmissions interfere with each other. To better understand the basic concept of NOMA, we consider a system consisting of only two devices which are communicating with the same BS. Let x_1 and x_2 denote the unity power signals transmitted from device 1 and device 2, respectively. The received signal at the base station is represented by:

$$y = \sqrt{P_{r_1}} x_1 + \sqrt{P_{r_2}} x_2 + z, \tag{5}$$

where z denotes additive white Gaussian noise (AWGN) with variance σ_z^2. At the BS, successive interference cancellation (SIC) is implemented, where first x_1 is decoded by treating x_2 as interference. Once the receiver correctly decodes x_1, it subtracts x_1 from the received signal y and then decodes x_2. The receiver decides the order of decoding according to the effective signal to interference plus noise ratio (SINR) of the devices. For device 1, the SINR at the BS can be calculated as follows:

$$\mathrm{SINR}_1 = \frac{P_{r_1}}{P_{r_2} + \sigma_z^2} = \frac{\frac{P_{t_1}}{P_{\max}} \left(\frac{r_1}{R_o}\right)^{-\alpha} \mu_{\mathrm{ref}}}{\frac{P_{t_2}}{P_{\max}} \left(\frac{r_2}{R_o}\right)^{-\alpha} \mu_{\mathrm{ref}} + 1}, \tag{6}$$

where we used (2) and ignored shadowing and small-scale fading for simplicity. After the BS decodes device 1, it removes its signal from y, and tries to decode x_2. The SINR for device 2 at the BS is then given by:

$$\mathrm{SINR}_2 = \frac{P_{r_2}}{\sigma_z^2} = \frac{P_{t_2}}{P_{\max}} \left(\frac{r_2}{R_o}\right)^{-\alpha} \mu_{\mathrm{ref}}. \tag{7}$$

The rate achieved at the BS for device 1 and 2 are respectively calculated as follows:

$$R_1 = \log_2\left(1 + \mathrm{SINR}_1\right), \quad R_2 = \log_2\left(1 + \mathrm{SINR}_2\right). \tag{8}$$

In NOMA, the performance gain compared to orthogonal multiple access schemes increases when the difference in channel gains or path loss between the users and the base station is large [12]. For example, when $r = 200$m and $r_2 = 800$m, $\mu_{\mathrm{ref}} = -3$dB, and $\alpha = 3$, using NOMA device 1 and device 2 can achieve rates $R_1 = 5.0294$ bits/s/Hz and $R_2 = 0.9847$ bits/s/Hz, respectively, when they are transmitting with full power $P_{\max} = 1$W. However, when they use FDMA and each user transmits over half of the bandwidth with full power, rates $R_1 = 3.4903$ bits/s/Hz and $R_2 = 0.7823$ bits/s/Hz can be achieved for device 1 and 2, respectively, which are clearly much less than those achieved using NOMA. The same rates as FDMA can be achieved

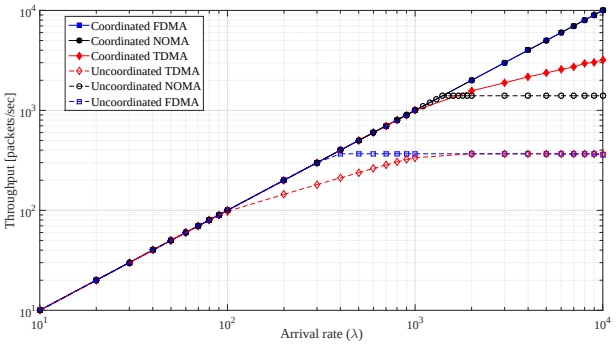

Fig. 2. Average throughput versus the arrival rate for different uncoordinated MA techniques. Total available bandwidth is $W = 1$ MHz, time slot duration is $\tau_s = 1$ sec, $P_{\max} = 1$ W, $\mu_0 = 0$ dB, and the packet length is $L = 1000$ bits. The minimum time slot duration for uncoordinated TDMA is considered to be 1 ms and the minimum subchannel bandwidth in uncoordinated FDMA is considered to be 1 kHz.

for both devices using NOMA, when device 1 and device 2 use only 28% and 74% of their maximum power, respectively. This shows one the most important advantages of NOMA over orthogonal multiple access techniques, such as FDMA, where better energy efficiencies can be achieved using NOMA which is of utmost importance for massive IoT applications.

Fig. 2 shows the maximum throughput versus arrival rate for different multiple access techniques for both coordinated and uncoordinated scenarios. See [13] for the derivation of the throughput as a function of arrival rate for FDMA, TDMA and general NOMA. In the coordinated scenario, we assume that the BS has already identified the devices and knows their channels. We also consider a single radio resource of length τ_s seconds and bandwidth W Hz. We assume that the BS performs SIC, where it starts the decoding with the device with the largest channel gain and treats the signals from other devices as additive noise [10]. In uncoordinated scenario, we consider that each device performs power control such that the received SNR at the BS for each device is μ_0. A device will only transmit if and only if the transmit power required to achieve the SNR μ_0 at the BS is less than $P_{\max}$. For TDMA and FDMA we have assumed that the BS calculates the optimal frequency and time allocation as discussed in [13]. As can be seen in Fig. 2, NOMA significantly outperforms TDMA and FDMA strategies in the uncoordinated scenario. However, FDMA can achieve the same throughput as NOMA in coordinated scenarios as the devices have been identified at the BS which can determine the optimal bandwidth allocation to accommodate more devices; thus achieving the same throughput performance as NOMA.

III. The Proposed Random NOMA Strategy for M2M Communications

In the proposed random NOMA strategy, the devices use the same radio resources for their transmissions. That is a device randomly chooses a sub-band for its data transmission and sends its data through the selected sub-band. The details of the proposed random NOMA strategy are given below:

1. When all the sub-bands allocated for M2M communications are available (i.e., they have all been released from the previous transmission), the BS broadcasts a pilot signal over each sub-band.
2. All the MTC devices which have data to transmit will listen to these pilot signals. Each MTC device will then choose a sub-band with the highest channel gain (Fig. 3 and Fig. 4) and will randomly select a seed for its random number generator from a set of M_s available seeds.
3. Each active device attaches its ID to its message and encodes it using the selected seed and transmits over the selected sub-band.
4. The BS performs load estimation (Fig. 5-a) and successive interference cancellation (SIC) to recover the message of each active device (Fig. 5-b).

As the devices perform power control such that their received power at the BS is the same, the BS can effectively estimate the number of devices over each sub-band by calculating the received power as it would be proportional to the number of devices. For further details on load estimation algorithms, please refer to [14]. We assume that the BS broadcasts pilot signals over all the sub-bands. Using these pilot signals, each device will estimate its channel to the BS over the different sub-bands and choose the sub-band which has the highest channel gain for its data transmission. This is particularly interesting for M2M applications where energy efficiency is very important as most of the devices are battery operated. For the simplicity of analysis in this paper, we assume that all the sub-bands have the same channel gain to each device; thus each device will randomly choose a sub-band for its data transmission. This assumption is valid due to the fact that the channel gains over different sub-bands are independent and devices are uniformly distributed in the cell.

In existing random access strategies, the device is identified in the random access phase through exchanging several messages between the device and the BS. This is however inefficient in M2M communications due to the large number of MTC devices and frequent preamble collisions due

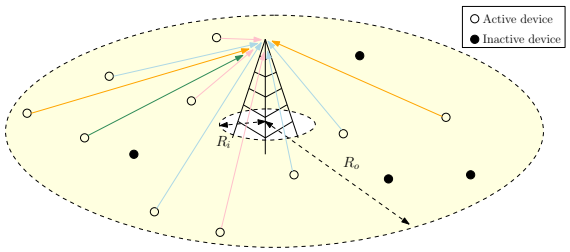

Fig. 3. Cellular M2M system.

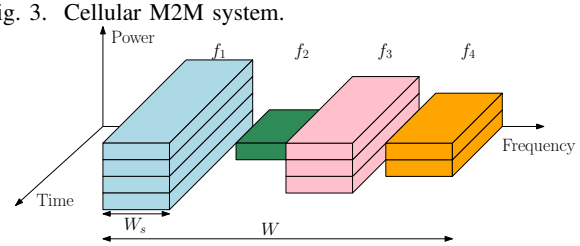

Fig. 4. Each MTC device randomly chooses a sub-band for its transmission, so the number of devices in each sub-band in random.

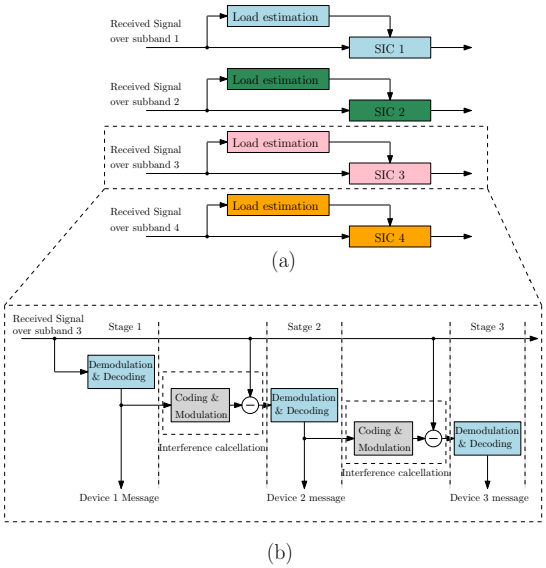

Fig. 5. (a) Load estimation and (b) successive interference cancellation at the BS for the proposed random NOMA strategy.

to multiple retransmissions of the access requests by the devices. In our proposed scheme, the BS cannot identify all the devices before the data transmission phase. This makes the decoding at the BS more challenging as the BS does not know which channel code each device is using and what are the other parameters. To solve this problem, we assume that all the devices use the same channel code for the data transmission. Due to the random number of active devices in each sub-band, the channel code rate cannot be fixed in order to adapt to random activities of devices. For this aim, Raptor codes [15], can be used which are rateless and can generate as many coded symbols as required by the BS. In particular, very low rate Raptor codes have been shown to perform well in an M2M scenario employing successive interference cancellation [14]. The code structure is random and can be represented by a bipartite graph; then the BS can reproduce the same bipartite graph using a pseudo random generator with the same seed. When more than one device selects the same seed and transmits over the same sub-band, they will be transmitting using exactly the same code structure; thus the BS cannot differentiate between them as there is no structural difference between the received codewords. We call this event a *collision*.

As the number of devices over each sub-band is random, the maximum achievable rate in each sub-band is also random. This means that the number of coded symbols that need to be transmitted over each sub-band is random. Fig. 6-b shows the length of each sub-band in two consecutive time slots. It is important to note that the duration of each time slot will be mainly

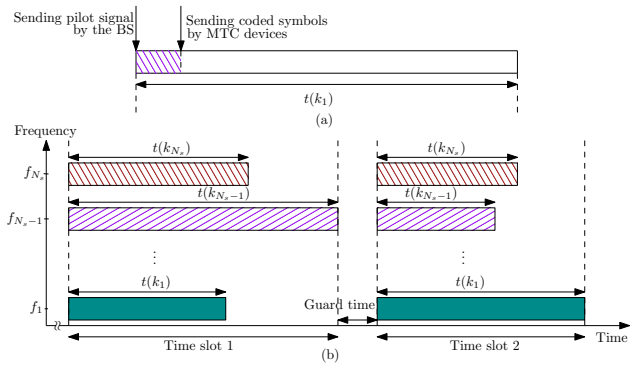

Fig. 6. Time slot duration in the proposed random NOMA strategy (a) over a single sub-band (b) in two consecutive time slots.

determined by the sub-band with the highest number of active devices as its maximum achievable rate would be lower than the rest.

The BS performs SIC to decode the message of each MTC device (Fig. 5). As the devices perform power control such that their received SNR at the BS is μ_0, the BS can almost accurately estimate the number of devices from the received signal. It can also determine the maximum achievable rate in each sub-band. This will be discussed in more details in the next section. The BS initiates the decoding with the first seed. If it cannot decode any message, it changes the seed and reattempts the decoding. In this way, those devices which have selected non-collided seeds will be decoded and later will be acknowledged by the BS. This means that the BS may need to reattempt the decoding M_s times per sub-band.

IV. Analysis of the Proposed Scheme

In this section, we analyze the proposed random NOMA strategy in terms of throughput and derive the stability condition, where we find the maximum arrival rate such that the system with an initial backlog remains constant. For this aim, we first need to characterize the time slot duration, as it is random due to the random number of devices over the sub-bands, and then find the collision probability in order to find the number of devices which reattempt their transmissions in the next time slot due to collision. It is important to note that for the purposes of our analysis in this paper, we ignore the pilot signal transmission time and guard time.

A. Time slot duration

Let k_i denote the number of devices which have selected the i^{th} sub-band, where we have dropped the time index for the ease of notation. Let $q(c,n)$ denote the probability that the

maximum number of devices over all the sub-bands is c when the number of active devices is n. It is then easy to shows that $q(c,n)$ is given by:

$$q(c,n)=\frac{\left|\left\{(k)_1^{N_s}\left|\sum_{j=1}^{N_s}k_i=n\ ,\max_j k_j=c\right.\right\}\right|}{N_s^n}. \tag{9}$$

For sufficiently large n and N_s, we can approximate the number of devices in each sub-band by a binomial distribution. This is due to the fact that each device randomly and independently selects among N_s available sub-bands with equal probability. More specifically, the probability mass function (p.m.f.) of k_i for $i=1,\cdots,N_s$ is given by:

$$\mathbb{P}[k_i|n]=\binom{n}{k_i}\left(\frac{1}{N_s}\right)^{k_i}\left(1-\frac{1}{N_s}\right)^{n-k_i}, \tag{10}$$

which can be further approximated by the normal distribution as follows [16]:

$$\mathbb{P}[k_i|n]\approx\frac{1}{\sigma_s}\varphi\left(\frac{k_i-\frac{n}{N_s}}{\sigma_s}\right), \tag{11}$$

where $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ and $\sigma_s=\sqrt{\frac{n}{N_s}(1-\frac{1}{N_s})}$. We aim at finding the p.m.f. of the maximum number of devices over all the available sub-bands. We first derive its cumulative mass function (c.m.f.) as follows:

$$\begin{aligned}\mathbb{P}\left[\max_i\{k_i\}\le\ell\middle|\,n\right]&=\mathbb{P}\left[k_1\le\ell,k_2\le\ell,\cdots,k_{N_s}\le\ell|\,n\right]\\&=\prod_{i=1}^{N_s}\mathbb{P}\left[k_i\le\ell|n\right]=\mathbb{P}[k_1\le\ell|n]^{N_s}\approx\left[1-\Phi\left(\frac{\ell-\frac{n}{N_s}}{\sigma_s}\right)\right]^{N_s},\end{aligned}$$

where $\Phi(x)=\int_x^\infty\varphi(t)dt$ is the cumulative distribution function (c.d.f.) of the standard normal distribution. The p.m.f. of the maximum number of devices over all the sub-bands can then be derived as follows:

$$q(c,n)\approx\frac{N_s\varphi\left(\frac{c-\frac{n}{N_s}}{\sigma_s}\right)\left[1-\Phi\left(\frac{c-\frac{n}{N_s}}{\sigma_s}\right)\right]^{N_s-1}}{\sigma_s} \tag{12}$$

Fig. 7 shows the p.m.f. of the maximum number of devices over all the sub-bands for different number of active devices and sub-bands. As can be seen in this figure the approximation of $q(c,n)$ in (12) is in a close agreement with the simulation results.

The duration of a time slot is determined by the sub-band with the highest number of active devices transmitting in it. The time slot duration can be calculated from Shannon's capacity formula as follows:

$$t(c)=\frac{L}{W_s\log_2\left(1+\frac{\mu}{1+(c-1)\mu}\right)},\quad c=\max_i k_i, \tag{13}$$

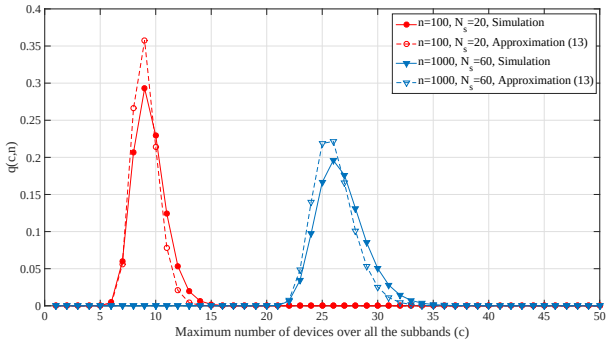

Fig. 7. The probability mass function of the maximum number of devices over all the available sub-bands.

where $\mu = \frac{W}{W_s}\mu_0 = N_s\mu_0$, as each device transmits over a sub-band with bandwidth W_s rather than W. The average time slot duration can then be calculated as follows:

$$\overline{T}(n) = \sum_{c=\lceil n/N_s \rceil}^{n} q(c,n) \frac{L}{W_s \log_2\left(\frac{1+c\mu}{1+(c-1)\mu}\right)}, \tag{14}$$

where $\lceil . \rceil$ is the ceil operator.

B. Collision probability

It is clear that the BS cannot detect the devices which have selected the same seed and are transmitting at the same sub-band. This is because there is no structural difference between the transmitted codewords from two devices which have selected the same seed and are transmitting over the same sub-band. As the number of sub-bands is N_s, each device can randomly select a sub-band and a seed among $M_s \times N_s$ different options. The collision probability, defined as the probability that a given device selects a preamble and seed which have already been selected by one or more other devices, can be approximated as follows when the number of active devices is n:

$$\mathrm{P_c}(n) \approx 1 - \left(1 - \frac{1}{M_s N_s}\right)^{n-1}, \tag{15}$$

which follows from the fact that a given device selects a specific preamble and seed with probability $1/(M_sN_s)$, and it is not in collision if other devices select different preambles and seeds, which happens with probability $(1-1/(M_sN_s))^{n-1}$. As the given device can select any of the M_sN_s configurations of preambles and seeds, the probability that a given device is not in collision is simply $M_sN_s(1/(M_sN_s))(1-1/(M_sN_s))^{n-1}$. The probability of collision is then easily derived as (15).

We determine the minimum number of seeds required for the system to have a collision probability at most $\mathrm{P_c}(n)$ as follows:

$$M_s \geq \frac{1}{N_s\left(1-(1-\mathrm{P_c}(n))^{\frac{1}{n-1}}\right)}. \tag{16}$$

This can be further simplified for $\mathrm{P_c}(n) \to 0$ as follows [10]:

$$M_s \geq \frac{n-1}{N_s \mathrm{P_c}(n)}, \tag{17}$$

where the number of required seeds to have a collision probability less than a given value linearly increases with the number of active devices.

C. The Number of Active Devices in each Time Slot

Let n_i denote the number of active devices in the i^{th} time slot, where $i \geq 1$. The probability that n_{i+1} devices are attempting to deliver their messages to the BS in the $(i+1)^{th}$ time slot is given by:

$$\mathbb{P}[n_{i+1}|n_i] = \sum_{j=0}^{\min\{n_{i+1},n_i\}} e^{-\lambda\overline{T}(n_i)} \frac{(\lambda\overline{T}(n_i))^{n_{i+1}-j}\binom{n_i}{j}\mathrm{P_c}(n_i)^j}{(n_{i+1}-j)!(1-\mathrm{P_c}(n_i))^{j-n_i}}. \tag{18}$$

In fact, j out of n_{i+1} devices might be those packets which have collided in the i^{th} time slot, while the remaining $(n_{i+1}-j)$ packets are newly generated packets. The number of collided packets is a random variable which follows a binomial distribution with success probability $\mathrm{P_c}(n_i)$, as each device is independently in collision with probability $\mathrm{P_c}(n_i)$, which is true when the number of devices is sufficiently large. It can be further approximated by a normal distribution (see (4-35) in [16]) with mean $n_i\mathrm{P_c}(n_i)$ and variance $n_i(1-\mathrm{P_c}(n_i))$ [16, equation 4-95]. The number of newly generated packets is also a random variable which is assumed to follow a Poisson distribution with mean λ packets/sec, which can be also approximated by a normal distribution with mean and variance $\lambda\overline{T}(n_i)$, when $\lambda\overline{T}(n_i)$ is sufficiently large [16, equation 4-107]. These random variables are mutually independent, therefore, the probability that n_{i+1} devices are transmitting in the $(i+1)^{th}$ time slot can be calculated by multiplying the probability of j collided devices and $n_{i+1}-j$ newly generated packets and taking the summation over j.

Using normal approximations for the number of collided devices and newly generated packets, (18) can be simplified as follows:

$$\mathbb{P}[n_{i+1}|n_i] \approx \frac{\exp\left(-\frac{(n_{i+1}-\mu_i)^2}{\sigma_i^2}\right)}{\sqrt{2\pi\sigma_i^2}}, \tag{19}$$

where $\mu_i = \lambda\overline{T}(n_i) + n_i \mathrm{P_c}(n_i)$ and $\sigma_i^2 = \lambda\overline{T}(n_i) + n\mathrm{P_c}(n_i)(1-\mathrm{P_c}(n_i))$. The average number of devices in the $(i+1)^{th}$ time slot is given by:

$$\mathbb{E}[n_{i+1}|n_i] = \lambda\overline{T}(n_i) + n\mathrm{P_c}(n_i). \tag{20}$$

D. Stability Condition without Delay Constraint

We define the stability condition such that in the steady state, the number of active devices in the next time slot, including the collided devices in the previous time slot and the newly generated packets, is not larger than the number of devices in the previous time slot. This way we make sure that the system can support all the active devices in each time slot and the number of active devices does not increase in time; otherwise the system will be quickly saturated. As the number of devices in each time slot is a random variable, we can define two stability conditions as follows.

1) Weak Stability Condition: The first stability condition, which we refer to as the *weak stability condition*, is defined based on the steady state average number of devices that can be supported by a system with initial backlog n as follows:

$$\lambda\overline{T}(n) + n\mathrm{P_c}(n) \leq n, \tag{21}$$

and the maximum arrival rate can be easily characterized by:

$$\lambda_{\max}^{(\mathrm{weak})}(n) = \frac{n(1-\mathrm{P_c}(n))}{\overline{T}(n)}. \tag{22}$$

The maximum arrival rate that can be supported by the BS is then found by maximizing $\lambda_{\max}^{(\mathrm{weak})}(n)$ over n, i.e.,

$$\lambda_{\max}^{(\mathrm{weak})} = \max_n \{\lambda_{\max}^{(\mathrm{weak})}(n)\}. \tag{23}$$

2) Strong Stability Condition: The second stability condition, referred to as the *strong stability condition*, takes into account the random behavior of the system and is given by:

$$\mathbb{P}\left[\mathrm{Poiss}(\lambda\overline{T}(n)) + \mathrm{Binom}(n, \mathrm{P_c}) > n\right] \leq \epsilon, \tag{24}$$

where $\epsilon > 0$ is the target probability of the system stability, which is a system design parameter. By using (18), the strong stability condition can be derived as follows:

$$1 - \sum_{n'=0}^{n} \sum_{i=0}^{n'} e^{-\lambda\overline{T}(n)} \frac{(\lambda\overline{T}(n))^{n'-i} \binom{n}{i} {\mathrm{P_c}}^i}{(n'-i)!(1-\mathrm{P_c})^{i-n}} \leq \epsilon, \tag{25}$$

which can be also written as follows by using (19):

$$\Phi(\frac{n-\mu}{\sigma}) \leq \epsilon, \tag{26}$$

where $\mu = \lambda\overline{T}(n) + n\mathrm{P_c}(n)$ and $\sigma^2 = \lambda\overline{T}(n) + n\mathrm{P_c}(n)(1-\mathrm{P_c}(n))$. One could easily find the maximum arrival rate using (26) as follows:

$$\Phi^{-1}(\epsilon) \leq \frac{n - \lambda\overline{T}(n) - n\mathrm{P_c}(n)}{\sqrt{\lambda\overline{T}(n) + n\mathrm{P_c}(n)(1-\mathrm{P_c}(n))}}, \tag{27}$$

and by solving this inequality with respect to λ we have:

$$\lambda_{\max}^{(\mathrm{strong})}(n) = \frac{1 + 2n\ell_\epsilon(1-\mathrm{P_c}(n)) - \sqrt{1 + 4n\ell_\epsilon(1-\mathrm{P_c}(n))^2}}{2\ell_\epsilon\overline{T}(n)}, \tag{28}$$

where

$$\ell_\epsilon := \frac{1}{[\Phi^{-1}(\epsilon)]^2}. \tag{29}$$

The maximum arrival rate is then given by

$$\lambda_{\max}^{(\mathrm{strong})} = \max_n \{\lambda_{\max}^{(\mathrm{strong})}(n)\}. \tag{30}$$

Fig. 8 shows the maximum arrival rate versus the initial backlog n, for different numbers of seeds, M_s, when the total available bandwidth is $W = 1$ MHz, the number of sub-bands is $N_s = 20$, and each device attempts to deliver a message of length $L = 1000$ bits to the BS. As can be seen, the system with higher M_s can support more devices as the collision probability decrease with M_s; so fewer devices will reattempt their transmissions in the following time slot.

Fig. 9 shows the stability regions for different number of sub-bands when the number of seeds is $M_s = 200$. With increasing number of sub-bands, the collision probability decreases but on the other hand the bandwidth of each sub-band will also decrease. This results in longer time slots as the transmission of the devices takes longer due to smaller bandwidth and lower achievable rate over each sub-band. As can be seen in Fig. 9, with increasing N_s, the maximum arrival rate decreases but the system can support a larger initial backlog.

E. Stability Condition with QoS Guarantee

We consider the delay requirement as the QoS metric for the system. More specifically, we assume that each packet has a delay constraint of d_p, which means that the packet must be delivered at the BS no later than d_p seconds after it has been generated at an MTC device.

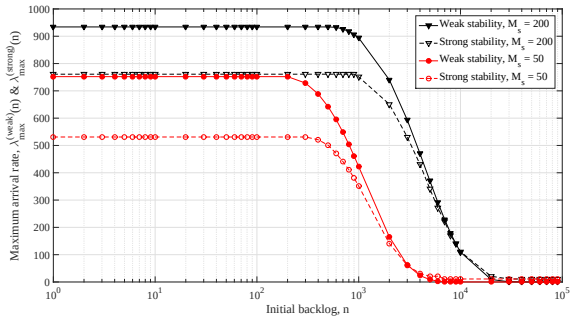

Fig. 8. The maximum arrival rate versus the initial backlog obtained from the weak and strong stability conditions for different M_s, when $N_s = 20$, $W = 1$ MHz, $L = 1000$, and the threshold probability for the strong stability condition is $\epsilon = 0.01$.

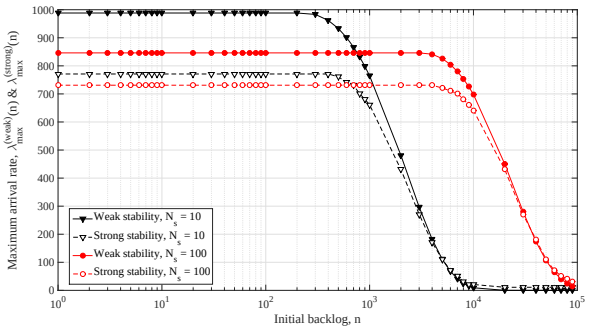

Fig. 9. The maximum arrival rate versus the initial backlog obtained from the weak and strong stability condition for different N_s, when $M_s = 200$, $W = 1$ MHz, $L = 1000$, and the threshold probability for the strong stability condition is $\epsilon = 0.01$.

We derive the maximum arrival rate as a function of the delay constraint such that the system remains stable.

Let n_1 denote the number of active devices at the first time slot. An active device can then deliver its message with delay $T(n_1)$ with probability $1 - \mathrm{P_c}(n_1)$; otherwise it will reattempt the transmission in the next time slot. More specifically, the device's message can be delivered at the BS at the j^{th} time slot with the probability given below:

$$\mathbb{P}[I = j|n_1] = \sum_{(n)_2^j} (1 - \mathrm{P_c}(n_j)) \prod_{i=1}^{j-1} \mathrm{P_c}(n_i) \mathbb{P}[n_{i+1}|n_i]. \tag{31}$$

Let us assume that the BS can change the number of seeds such that the collision probability is always less than p_c regardless of the number of active devices. Then (31) is reduced to:

$$\mathbb{P}[I = j] \approx (1 - p_c) p_c^{j-1}, \tag{32}$$

which is a decreasing function of j. The delay can then be characterized as follows:

$$\mathbb{P}[d|\lambda, n_1] = \sum_j \mathbb{P}[d|I = j]\mathbb{P}[I = j] = \sum_j \mathbb{P}[I = j] \sum_{(n)_2^j} \mathbb{P}[d|n1, \cdots, n_j] \prod_{i=1}^{j-1} \mathbb{P}[n_{i+1}|n_i]$$

$$= \sum_j \mathbb{P}[I = j] \sum_{(n)_2^j} \bigotimes_{i=1}^{j} \mathbb{P}[d_i|n_i] \prod_{i=1}^{j-1} \mathbb{P}[n_{i+1}|n_i], \tag{33}$$

where $\bigotimes$ is the convolution operator,

$$\mathbb{P}[d_i|n_i] = \begin{cases} q(c_i, n_i), & d_i = \frac{L}{W_s \log_2\left(1 + \frac{\mu}{1+(c_i-1)\mu}\right)}, \\ 0, & \text{otherwise}, \end{cases}$$

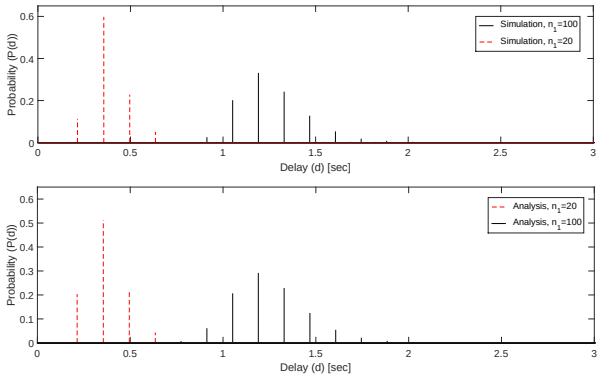

Fig. 10. The histogram of the delay at the BS, when $W = 100$ kHz, $N_s = 20$, $M_s = 10$, $L = 1000$, and the initial backlog $n_1 = 100$. The packet arrival rate is $\lambda = 100$.

and $\mathbb{P}[n_i|n_{i-1}]$ is given in (18). By using (32) and considering only the first few terms of (33), i.e., $j = 1, 2$, (33) can be simplified as follows:

$$\mathbb{P}[d|n_1] \approx (1-p_c)\mathbb{P}[d|n_1] + p_c(1-p_c)\sum_{n_2} \frac{\mathbb{P}[d|n_1] \otimes \mathbb{P}[d|n_2]}{\sqrt{2\pi\sigma_2^2}}\ e^{-\frac{(n_2-\mu_2)^2}{\sigma_2^2}}, \tag{34}$$

where $\mu_2 = \lambda\overline{T}(n_1) + np_c$ and $\sigma_2^2 = \lambda\overline{T}(n_1) + n_1p_c(1-p_c)$. Fig. 10 shows the histogram of the delay when the total available bandwidth is $W = 100$ kHz, the number of sub-bands is $N_s = 20$, the collision probability is $p_c = 0.01$, and the arrival rate is $\lambda = 100$. As can be seen in this figure, the results obtained from (34) are in a close agreement with the actual histogram of the delay.

The weak stability condition is defined as follows:

$$\mathbb{E}[d|\lambda, n_1] \leq d_p. \tag{35}$$

As shown in (22), the maximum arrival rate under the weak stability condition is given by $n(1-p_c)/\overline{T}(n)$. Therefore, the maximum initial backlog under the weak stability condition to satisfy the delay requirement d_p is given by:

$$n_{\max}^{(\text{weak,delay})}(d_p) = \max_n \left\{ n \left| \mathbb{E}\left[d \left| n, \frac{n(1-p_c)}{\overline{T}(n)} \right.\right] \leq d_p \right.\right\} \tag{36}$$

and by using (22), the maximum packet arrival rate under the weak stability condition is given by:

$$\lambda_{\max}^{(\text{weak,delay})}(d_p) = \frac{n_{\max}^{(\text{weak})}(d_p)(1-p_c)}{\overline{T}(n_{\max}^{(\text{weak,delay})}(d_p))}. \tag{37}$$

Similarly, the strong stability condition can also be found as follows:

$$1 - \int_0^{d_p} \mathbb{P}[d = \tau|n, \lambda]d\tau < \epsilon, \tag{38}$$

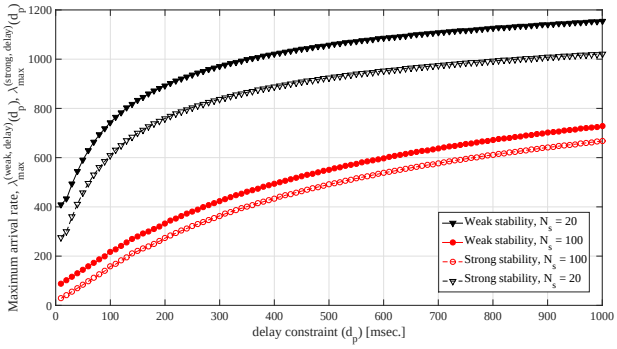

Fig. 11. The maximum packet arrival rate under weak and strong stability conditions, when $W = 100$ kHz, $N_s = 20$, $M_s = 10$, $L = 1000$, and collision probability is set to $p_c = 0.01$. The packet arrival rate is $\lambda = 100$ and the threshold probability for the strong stability condition is $\epsilon = 0.01$.

the maximum initial backlog under the strong stability condition is given by

$$n_{\max}^{(\text{strong,delay})}(d_p) = \max_n \left\{ n \left| 1 - \int_0^{d_p} \mathbb{P}\left[d = \tau \left| n, \lambda_{\max}^{(strong)}(n) \right. \right] d\tau < \epsilon \right. \right\}, \tag{39}$$

where $\lambda_{\max}^{(strong)}(n)$ is given in (28). The maximum packet arrival rate under the strong stability condition is then given by:

$$\lambda_{\max}^{(strong,delay)}(d_p) = \lambda_{\max}^{(strong)}(n_{\max}^{(\text{strong,delay})}(d_p)). \tag{40}$$

Fig. 11 shows the maximum arrival rate versus the delay constraint under weak and strong stability conditions. As can be seen, in delay sensitive condition, i.e., short delay, the number of devices which can be supported by the proposed NOMA strategy is small, and by increasing the tolerable delay, the supported arrival rate increases. Also, with increasing the number of sub-bands, the maximum arrival rate decreases which is due to the fact that by increasing the number of sub-bands, the available bandwidth per sub-band decreases, which results in a lower achievable rate over each sub-band.

V. System Optimization

A. Number of sub-bands

The number of sub-bands and seeds available will determine the overall system performance as the collision probability and the maximum achievable rate for the proposed random NOMA strategy will be mainly determined by these two parameters. The base station then needs to find the optimal values for these parameters to maximize the system throughput or satisfy the QoS requirements of the devices. It is clear that the collision probability decreases as M_s, the number of seeds, increases. One could adaptively change the number of seeds according to the incoming

traffic at the BS to fix the collision probability. However, it is also clear that increasing the number of seeds adds extra complexity at the BS as the BS should consider a larger number of seeds while performing SIC.

We first consider the optimization of the supported arrival rate when there is no delay constraint. The maximum supported arrival rate according to the strong stability condition without delay constraint, i.e., (28), for a given ϵ, M_s, W and L, is given by:

$$\max_{\{N_s,n\}} \frac{1+2n\ell_\epsilon(1-\mathrm{P_c}(n))-\sqrt{1+4n\ell_\epsilon(1-\mathrm{P_c}(n))^2}}{2\ell_\epsilon\overline{T}(n)}, \tag{41}$$

where $\overline{T}(n)$, $\mathrm{P_c}$, and ℓ_ϵ are respectively given by (14), (15), and (29).

Fig. 12 shows the maximum packet arrival rate versus the number of sub-bands without a delay constraint. As can be seen in this figure, the maximum packet arrival rate is achieved when the number of sub-bands is either 3 or 4 for different numbers of seeds. This means that to support a large number of devices using the proposed random NOMA strategy, the available bandwidth does not need to be divided into too many sub-bands, only a few sub-bands is sufficient. This is because when the number of sub-bands increases, the available bandwidth for each sub-band decreases, which also decreases the maximum achievable rate over each sub-band; therefore, fewer packets will be delivered over each sub-band.

A similar optimization problem can be defined to find the optimal number of sub-bands to maximize the supported arrival rate for a given delay constraint. Fig. 13 shows the maximum packet arrival rate versus the number of sub-bands for different delay constraints d_p. As can be seen, for a given d_p, the maximum packet arrival rate can be supported when the whole bandwidth is used as only one sub-band. In other words, dividing the bandwidth into several sub-bands degrades the performance of the proposed random NOMA in terms of the packet arrival rate which can be supported at the BS within a given delay requirement. Therefore, to satisfy the QoS requirements of a large number of devices using the proposed random NOMA strategy, the devices should use the whole bandwidth and the BS should control the collision probability by choosing a larger seed pool.

B. MTC Device Fairness

In the proposed random NOMA strategy, we have assumed that the signals received from all the devices have the same power at the BS. This way, the devices which are far from the BS should transmit with higher power to maintain the same received power at the BS. In other

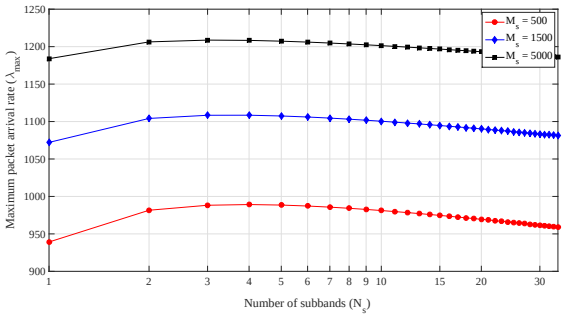

Fig. 12. The maximum packet arrival rate under the strong stability condition without delay constraint versus the number of sub-bands, N_s, when $W = 1$ MHz, $L = 1000$, and the threshold probability for the strong stability condition is $\epsilon = 0.01$.

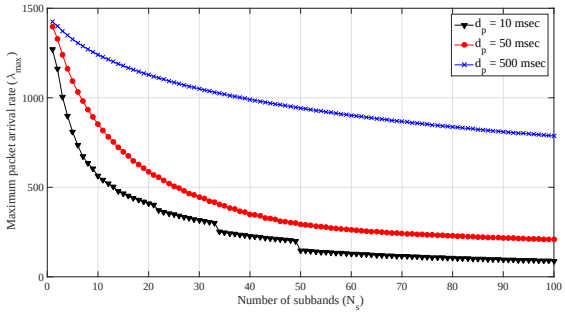

Fig. 13. The maximum packet arrival rate under the strong stability condition with delay constraint versus the number of sub-bands, N_s, when $W = 1$ MHz, $L = 1000$, and the threshold probability for the strong stability condition is $\epsilon = 0.01$.

words, the devices which are far from the BS should spend more energy to achieve the same throughput performance as the devices close to the BS. One approach to solve this problem is to allocate bandwidth to the devices according to their distances to the BS, so they can achieve the same throughput performance with the same energy consumption. For this aim, we divide the cell into N_s partitions, such that the area covered in each partition is the same. This way the average number of devices in each partition is the same due to the fact that the devices are randomly distributed in the cell. Let r_i denote the radius of the outer edge of the i^{th} partition for $i = 1, \cdots, N_s$, where $r_{N_s} = R_o$. Then, it is easy to show that r_i is given by:

$$r_i = \sqrt{\frac{i}{N_s}} R_o. \tag{42}$$

Unlike the original random NOMA presented in Section III, where the devices randomly choose among N_s available sub-bands of the same bandwidth and their received power at the BS is the same, here we assume that the total bandwidth is divided into N_s sub-bands, where the devices in the i^{th} cell partition transmit in the i^{th} sub-band with bandwidth W_i such that their received SNR at the BS is μ_i. The non-uniform allocation of the bandwidth to the sub-bands allows for the derivation of a fair multiple access strategy in terms of the energy consumption, which is explained in the following.

To have a fair system, we need to guarantee the same average throughput and energy consumption for all the devices regardless of their distances to the BS. To achieve the same average throughput over all the sub-bands and accordingly all the cell partitions, the average duration of the sub-bands should be the same. As the devices are randomly distributed in the cell, j out of

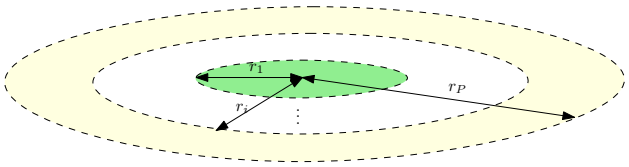

Fig. 14. Dividing a cell into N_s partitions.

n active devices belong to the i^{th} cell partition with the probability given below:

$$\mathbb{P}[n_i = j|n] = \binom{n}{j}\left(\frac{1}{P}\right)^j\left(1-\frac{1}{P}\right)^{n-j}, \tag{43}$$

and the time required for these devices to deliver their messages at the BS in the i^{th} sub-band is given by (13):

$$t_i(j) = \frac{L}{W_i \log_2\left(1+\frac{\mu_i}{1+(j-1)\mu_i}\right)}. \tag{44}$$

The average duration of the i^{th} sub-band can then be calculated as follows:

$$\overline{T}_i(n) = \sum_{j=0}^{n} \frac{L\binom{n}{j}\left(\frac{1}{N_s}\right)^j\left(1-\frac{1}{P}\right)^{n-j}}{W_i \log_2\left(1+\frac{\mu_i}{1+(j-1)\mu_i}\right)}. \tag{45}$$

This can be simplified for μ_i being sufficiently small by using the first term of the Maclaurin series $\ln(1+x) = x + \mathcal{O}(x^2)$ assuming that x is very small:

$$\begin{aligned}
\overline{T}_i(n) &= \sum_{j=0}^{n} \frac{L\ln(2)\binom{n}{j}\left(\frac{1}{N_s}\right)^j\left(1-\frac{1}{N_s}\right)^{n-j}}{W_i\frac{\mu_i}{1+(j-1)\mu_i}} \\
&= \frac{L\ln(2)}{W_i\mu_i}\sum_{j=0}^{n}\binom{n}{j}\left(\frac{1}{N_s}\right)^j\left(1-\frac{1}{N_s}\right)^{n-j}(1+\mu_i(j-1)) \\
&\overset{(a)}{=} \frac{L\ln(2)}{W_i\mu_i}\left(1+\mu_i\left(\frac{n}{N_s}-1\right)\right),
\end{aligned} \tag{46}$$

where step (a) follows from the fact that $\sum_{j=0}^{n}\binom{n}{j}(1/N_s)^j(1-1/N_s)^{n-j} = 1$ and the mean value of a Binomial distribution with success probability $1/N_s$ is $\sum_{j=0}^{n} j\binom{n}{j}(1/N_s)^j(1-1/N_s)^{n-j} = n/N_s$. In order to have the same average time duration for all the sub-bands, we need to satisfy $\overline{T}_i(n) = \overline{T}_1(n)$ for $i = 1, \cdots, N_s$, which can be rewritten as follows using (46):

$$\frac{W_i}{W_1} = \frac{\mu_i^{-1}+\left(\frac{n}{N_s}-1\right)}{\mu_1^{-1}+\left(\frac{n}{N_s}-1\right)} \approx \frac{\mu_1}{\mu_i}, \tag{47}$$

where the approximation follows from the assumption that the μ_i's are very small.

To maintain the same average energy consumption for all the devices, the average transmit power for all the devices should be the same given that we have required that the average duration of the sub-bands are the same. By using (4), the average transmit power to achieve SNR μ_i over the i^{th} sub-band is given by:

$$\overline{P}_{\mathrm{t},i} = P_{\max}\frac{\mu_i}{\mu_{\mathrm{ref}}\chi h}\int_{r_{i-1}}^{r_i}\left(\frac{r}{R_{\mathrm{o}}}\right)^{\alpha}\frac{2r}{R_o^2}dr = P_{\max}\frac{\mu_i}{\mu_{\mathrm{ref}}\chi h}\frac{r_i^{\alpha+2}-r_{i-1}^{\alpha+2}}{(2+\alpha)R_o^{\alpha+2}}. \tag{48}$$

The average energy consumption of the devices in the i^{th} cell partition, which are transmitting in the i^{th} sub-band, is given by $\overline{E}_i = \overline{T}_i(n)\overline{P}_{\mathrm{t},i}$. As we assume that the average duration of the sub-bands are the same, to have the same energy consumption for all the devices, i.e., $\overline{E}_i = \overline{E}_1$ for $i = 1, \cdots, N_s$, we need to satisfy $\overline{P}_{\mathrm{t},i} = \overline{P}_{\mathrm{t},1}$, which can be rewritten as follows:

$$\mu_i = \frac{r_1^{\alpha+2}}{r_i^{\alpha+2}-r_{i-1}^{\alpha+2}}\mu_1 = \frac{\mu_1}{i^{\frac{\alpha+2}{2}}-(i-1)^{\frac{\alpha+2}{2}}}, \tag{49}$$

where the last equality was obtained by using (42). By using (47), the bandwidth for the i^{th} sub-band can be calculated as follows:

$$W_i = \left(i^{\frac{\alpha+2}{2}}-(i-1)^{\frac{\alpha+2}{2}}\right)W_1. \tag{50}$$

As we have $\sum_{i=1}^{N_s} W_i = W$, we have

$$W = W_1\sum_{i=1}^{N_s}\left(i^{\frac{\alpha+2}{2}}-(i-1)^{\frac{\alpha+2}{2}}\right) = N_s^{\alpha+2}W_1 \tag{51}$$

and by using (50), we have:

$$W_i = \frac{i^{\frac{\alpha+2}{2}}-(i-1)^{\frac{\alpha+2}{2}}}{N_s^{\frac{\alpha+2}{2}}}W. \tag{52}$$

This shows that, to have the same energy consumption for all the devices across the cell, more bandwidth should be allocated to those devices which are far from the BS. Fig. 15 shows the bandwidth allocation versus the number of devices when the total bandwidth is 1 MHz and the number of sub-bands (or equivalently the number of cell partitions) is 3. As can be seen, with increasing the number of devices, the bandwidth will be allocated more evenly between the sub-bands, which is because n is the dominant term in (47) when n is very large. On the other hand, when n is relatively small, more bandwidth is allocated to the devices which are located far from the BS. This shows that the BS needs to have a proper load estimation strategy to allocate the bandwidth between the sub-bands so as to obtain fairness in the energy consumption and throughput.

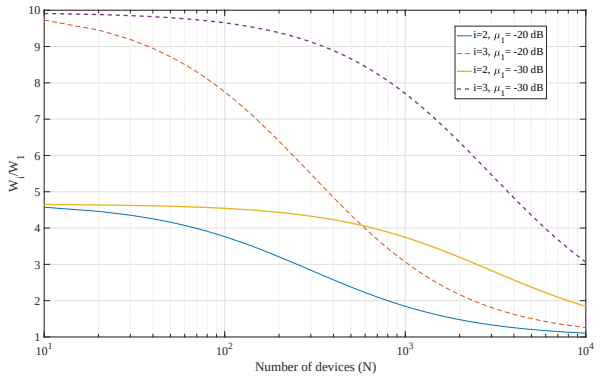

Fig. 15. Fair bandwidth allocation for energy efficient massive NOMA transmission. The total bandwidth is $W = 1$ MHz, path loss exponent is $\alpha = 3$, and $N_s = 3$. μ_i can be calculated for different i using (49).

VI. Practical Considerations of Massive NOMA for M2M Communications

NOMA is a promising solution for future wireless communications and can bring many benefits to cellular systems. This achieved through the effective use of spectrum and higher system throughput through exploiting the power domain and utilizing non-orthogonal multiplexing [12]. NOMA is compatible with OFDMA and its variants and can be applied on top of OFDMA for downlink and SC-FDMA for uplink [11]. NOMA can be combined with beamforming and multi-antenna technologies to improve the system performance [12]. NOMA can be easily combined with radio resource management and random access techniques to solve the collision and overload problem in M2M communications. Using clustering and group-based scheduling, NOMA can be used in M2M communications as the multiple access technique to deliver messages of a group of devices to the base station or the cluster head.

Although NOMA can improve spectrum efficiency and system capacity, there are many practical challenges for this technology to be potentially used in real wireless systems for M2M communications. A summary of most important challenges of NOMA has been presented in [13]. Here we emphasize two main challenges of massive NOMA and propose some solutions to effectively solve them and take an step toward developing a more practical massive IoT system using the NOMA strategy.

A. Optimal Power Allocation and Throughput

As we mentioned earlier, the duration of a sub-band is determined by the rate achieved by the device with the lowest SINR, as we assumed that devices' message are received with the same

SNR over each sub-band. The minimum rate achieved by the devices in a sub-band containing c devices is given by:

$$R_{\min} = \log_2\left(1 + \frac{\mu_0}{1 + (c-1)\mu_0}\right). \tag{53}$$

and the effective rate of the device which is decoded in the j^{th} stage of the SIC process is given by:

$$R_j = \log_2\left(1 + \frac{\mu_0}{1 + (c-j)\mu_0}\right). \tag{54}$$

However, as the devices do not know the traffic load and randomly transmit over the sub-bands in a rateless manner, the effective rate cannot be achieved by the devices as all devices must transmit at rate $R_{\min}$. This means that the optimal throughput cannot be achieved, which is mainly due to the fact that the devices are unknown to the BS, so it cannot optimally determine the power and rates.

To achieve the full potential of NOMA, the devices' messages need to be received with different powers, where the power levels are determined by the BS to optimize the throughput. However, this is only practical when the BS can identify the devices before the data transmission so it can optimally determine their received power. This was considered in [14], where the devices choose their power according to a message broadcasted by the BS. In fact, each device chooses a weighting coefficient from a pool, and multiply the transmit signal by the selected weight coefficient. This way higher throughput can be achieved, but extra control messages should be exchanged between the devices and the BS to determine which device has selected each weight coefficient. In fact, there is a tradeoff between the amount of overhead and the system throughput. That is, if the devices can exchange more information to the BS before their data transmission or ideally be identified at the BS, the BS can determine optimal transmission strategy in terms of power and rate and broadcast this information to the devices. This is, however, impractical for massive IoT applications where the message size is usually small, so the overhead must be kept as small as possible. On the other hand, it would be impractical for the BS to identify and perform channel estimation to a large number of devices in each time instant. This would incur huge delay in the system which is not acceptable for most massive IoT applications. The solution would be to minimize the control overhead by removing the device identification phase (as in the proposed scheme), and improve the system throughput by optimizing the bandwidth allocation as discussed in Section V.

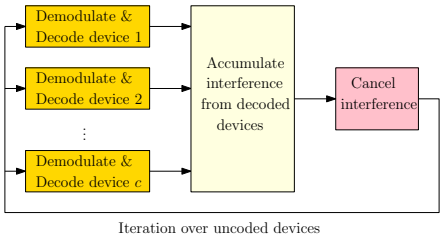

Fig. 16. Parallel interference cancellation for the proposed random NOMA strategy.

B. Delay Imposed from the SIC Process

In the proposed random NOMA strategy, we consider the successive interference cancellation at the BS, that is the BS starts the decoding of the device with the highest SINR and then removes its interference from the received signal and continues the decoding of the remaining devices. In this process, the effective SINR of the devices gradually increases assuming the BS can successfully decode the previous device. However, this imposed some delay into the system as a device should wait some time for the previous devices to be decoded by the BS before being decoded in the SIC process depending on the decoding order chosen by the BS. By considering that some enhanced IoT-oriented could be characterized by strict delay constraints (e.g., [17]), this aspect needs to be considered.

To address the delay issue in SIC, we can consider iterative parallel interference cancellation [18, 19] shown in Fig. 16. This scheme is more attractive from an implementation perspective as multiple devices are decoded and cancelled from the received signal simultaneously. If a device fails decoding in the first iteration, it will be decoded again in the subsequent iterations. The processing power is distributed among multiple parallel demodulators and decoders. Thus the tradeoff between delay and complexity is well balanced. Moreover, the performance of parallel IC can approach successive IC with a small number of iterations.

We can also assume that delay sensitive devices use specific random seeds, and accordingly group the devices based on their delay requirements; this could also be beneficial from an energy consumption point of view. Similar approaches have been considered in [20, 21] focusing on splitting random access resources to achieve energy savings. By applying this concept to our proposal, then Parallel IC is performed on high priority groups (correspond to low latency devices) to low priority groups (correspond to delay tolerant devices) successively [22]. Iterative processing can be applied to the devices which fail in the first IC iteration. Note that in any of the above schemes, even if a device fails to decode, one could attempt to cancel a portion of the users waveform by performing soft interference cancellation [23]. In this case, minimum mean square

error estimates of the data symbols can be derived from the soft output of the channel decoder to reconstruct the waveform. Moreover, most of the processing can be moved to the cloud to reduce the computing load on the BS. It is also important to note that several approaches have been proposed in the literature to reduce the complexity of the multiuser detection techniques for CDMA and WCDMA which can be used here for the proposed random NOMA strategy. The readers are referred to [18, 19, 23–25] for further details.

C. A brief comparison between Narrow-band IoT and Masive NOMA

As part of 3GPP Release 13, narrow-band IoT (NB-IoT) [26–29] has been standardized for low end massive IoT, that is the devices require relatively low data rate ($\sim$250 kbps in downlink direction, $\sim$20 kbps in uplink with the possibility to aggregate multiple tones to reach the same speed as in downlink) with relaxed delay requirements (in the order of 10 seconds). The required bandwidth for NB-IoT is 180 KHz for both uplink and downlink, which enables three different deployment options due to the small bandwidth. These are $i)$ stand alone operation, where a GSM operator can replace one GSM carrier (200 kHz) with NB-IoT, $ii)$ guard band operation, by utilizing the unused resource blocks within an LTE carrier's guard-band, and $iii)$ in-band operation, where NB-IoT can be deployed inside an LTE carrier by allocating one of the Physical Resource Blocks (PRB) of 180 kHz to NB-IoT. Although NB-IoT is an independent radio interface, it is tightly connected with LTE, which also shows up in its integration in the current LTE specifications [30].

NB-IoT promises to improve the cellular systems for massive IoT in the following aspects: 1) extended coverage, with a target Maximum Coupling Loss (MCL) of 164 dB, that is the coverage should be readily available and reliable also in challenging indoor deployments such as basements, 2) Support of massive number of low throughput devices, where up to 50000 devices can be supported per cell, for the arrival traffic of about 6 packets per second, 3) Reduced complexity in order lower the device cost, 4) Improved power efficiency in order to reach a battery life of about 10 years with 5 W/h battery, and 5) Latency, where a delay requirement of 10 seconds is appropriate for the uplink when measured from the application 'trigger event' to the packet being ready for transmission from the base station towards the core network [30].

Despite the fact that NB-IoT can support a large number of devices per cell, it is still based on a two-step procedure (i.e,. random access followed by data transmission) which is only

appropriate for low packet arrival rates as it limits the overall channel capacity. Moreover, as the delay assumption was relaxed in NB-IoT, it does not provide a solution for devices with strict delay requirements.[1]

NOMA on the other hand, provides a general framework which substantially improves the system capacity as each physical resource unit can be used by multiple devices, regardless of the bandwidth. This means that it can be deployed on top of existing standards, as it is compatible with OFDMA and SC-FDMA for downlink and uplink in LTE. With appropriate resource management and cell partitioning, delay sensitive devices can be allocated with higher bandwidth in the NOMA framework while maintaining the same overall energy per bit for all devices across the cell. Also, in NOMA devices do not need to perform random access as they can attach their ID to their messages, and later be identified by the BS. This significantly reduces the delay.

As a simple comparison with NB-IoT, we consider the total system bandwidth of $W = 180$ kHz and 12 subbands each of 15 kHz bandwidth. Our simulations show that NOMA can support arrival rates up to 100 packets per second under the strict delay requirement of 100 msec and arrival rates of up to 180 packets per second for the delay requirement of 1 sec. This is much more than what can be supported by NB-IoT which supports around 18 packets per second[2], with a worst case (i.e., for devices with very poor channel coverage) uplink latency of $\sim$2.8 s [26]. This shows that NOMA can be used for delay sensitive applications and support a larger number of devices compared to NB-IoT.

VII. Conclusions

We considered a random non-orthogonal multiple access strategy for M2M communications, where multiple devices are allowed to transmit over the same sub-band and the base station performs successive interference cancellation to decode each device's message. We derived system stability conditions, where the maximum packet arrival rate was found with and without quality of service guarantee. We then found the optimized system parameters, including the number of sub-bands under these scenarios; optimizing the throughput alone, including a delay

[1]For delay-constrained applications, 3GPP proposed a further LTE enhancement for machine-type communications, i.e., LTE-M, a.k.a. eMTC.

[2]It is worth mentioning that the target capacity for NB-IoT is about 55000 devices per cell sector, which corresponds (by considering devices transmitting one packet per hour) to a target traffic of $\sim$15 packets per second.

constraint, and ensuring user fairness in both throughput and energy consumption. We found that the optimal strategy differed for each of these conditions. More specifically, we found that without any delay constraint the whole bandwidth must be divided into only a few (i.e., 3 or 4) sub-bands to maximize the packet arrival rate which can be supported by the base station. On the other hand, when a delay constraint is imposed on the system, the whole bandwidth must be used as only one sub-band to support a large packet arrival rate and satisfy the delay requirement. This way the collision probability can be controlled by considering a larger pool of seeds, which are used to construct semi-orthogonal random codewords at different devices. Finally to ensure user fairness, the bandwidth and received SNR for each sub-band must be optimized such that more bandwidth is allocated to the devices which are far from the base station.

REFERENCES

[1] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2M: From mobile to embedded internet," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 36–43, 2011.

[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, Fourthquarter 2015.

[3] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, 2011.

[4] V. Gazis, "A Survey of Standards for Machine to Machine (M2M) and the Internet of Things (IoT)," *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, pp. 1–1, 2016.

[5] "Cellular networks for massive IoT," Ericsson, Tech. Rep. Uen 284 23-3278, January 2016. [Online]. Available: https://www.ericsson.com/res/docs/whitepapers/wp_iot.pdf

[6] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-Enabled Tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, March 2016.

[7] "LTE evolution for IoT connectivity," Nokia, Tech. Rep. Nokia White Paper, January 2016. [Online]. Available: http://resources.alcatel-lucent.com/asset/200178

[8] M. Shirvanimoghaddam, Y. Li, M. Dohler, B. Vucetic, and S. Feng, "Probabilistic rateless multiple access for machine-to-machine communication," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6815–6826, Dec. 2015.

[9] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela, "Fundamentals of throughput maximization with random arrivals for M2M communications," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4094–4109, Nov. 2014.

[10] ——, "Power-efficient system design for cellular-based machine-to-machine communications," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5740–5753, Nov. 2013.

[11] A. Li, Y. Lan, X. Chen, and H. Jiang, "Non-orthogonal multiple access (NOMA) for future downlink radio access of 5G," *China Communications*, vol. 12, no. Supplement, pp. 28–37, Dec. 2015.

[12] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. 98, no. 3, pp. 403–414, 2015.

[13] M. Shirvanimoghaddam and S. Johnson, "Multiple access technologies for cellular M2M communications: An overview," *ZTE Communications*, vol. 14, no. 4, pp. 42–49, Oct. 2016.

[14] M. Shirvanimoghaddam, M. Dohler, and S. Johnson, "Massive multiple access based on superposition raptor codes for cellular M2M communications," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1–1, 2016.

[15] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551 –2567, Jun. 2006.

[16] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*, 4th ed. Tata McGraw-Hill Education, 2002, ch. 4-5 Asymptotic Approximations for Binomial Random Variable.

[17] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IoT Machine Age With 5G: Machine-Type Multicast Services for Innovative Real-Time Applications," *IEEE Access*, vol. 4, pp. 5555–5569, 2016.

[18] P. Patel and J. Holtzman, "Performance comparison of a DS/CDMA system using a successive interference cancellation (IC) scheme and a parallel IC scheme under fading," in *IEEE Intl. Conf. Commun. (ICC)*. IEEE, 1994, pp. 510–514.

[19] D. Divsalar and M. K. Simon, "Improved CDMA performance using parallel interference cancellation," in *IEEE Military Commun. Conf. (MILCOM)*, 1994, pp. 911–917.

[20] M. Condoluci, G. Araniti, M. Dohler, A. Iera, and A. Molinaro, "Virtual code resource allocation for energy-aware MTC access over 5G systems," *Ad Hoc Networks*, vol. 43, pp. 3 – 15, 2016, smart Wireless Access Networks and Systems for Smart Cities.

[21] O. Arouk, A. Ksentini, and T. Taleb, "Group Paging-Based Energy Saving for Massive MTC Accesses in LTE and Beyond Networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1086–1102, May 2016.

[22] A. L. Johansson and L. K. Rasmussen, "Linear group-wise successive interference cancellation in CDMA," in *IEEE 5th Intl. Symp. Spread Spectrum Techniques and Applications*, vol. 1, Sep 1998, pp. 121–126 vol.1.

[23] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1046–1061, 1999.

[24] S. Verdu, *Multiuser detection*. Cambridge university press, 1998.

[25] A. L. Hui and K. B. Letaief, "Successive interference cancellation for multiuser asynchronous DS/CDMA detectors in multipath fading links," *IEEE Trans. Commun.*, vol. 46, no. 3, pp. 384–391, 1998.

[26] "Technical Specification Group GSM/EDGE Radio Access Network; Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT) (Release 13)," 3rd Generation Partnership Project, Tech. Rep., November 2015.

[27] "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, TS 36.211 (Release 13)," 3rd Generation Partnership Project, Tech. Rep., June 2016.

[28] "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding, TS 36.212 (Release 13)," 3rd Generation Partnership Project, Tech. Rep., June 2016.

[29] "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2, TS 36.300 (Release 13)," 3rd Generation Partnership Project, Tech. Rep., June 2016.

[30] Y.-P. E. Wang, X. Lin, A. Adhikary, A. Grövlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A primer on 3gpp narrowband internet of things (nb-iot)," *arXiv preprint arXiv:1606.04171*, 2016.