Impact of MTC on Energy and Delay Performance of Random-Access Channel in LTE-Advanced

Mikhail Gerasimenko, Vitaly Petrov, Olga Galinina, Sergey Andreev, and Yevgeni Koucheryavy

February 28, 2014

Abstract

Machine-Type Communications (MTC) are a rapidly growing technology, which is expected to generate significant revenues to mobile network operators. In particular, smart grid is predicted to become one of the key MTC use cases that involves unattended meters autonomously reporting information to a grid infrastructure. With this research, we consider a typical smart metering MTC application scenario in the context of 3GPP LTE-Advanced wireless cellular system featuring a large number of devices connecting to the network near-simultaneously. The resulting overload of the random access channel (RACH) requires a novel evaluation methodology based on comprehensive analysis and simulations. In this paper, we target to complement a validated evaluation framework fully compatible with the 3GPP test cases with a thorough analysis of RACH performance in overloaded MTC scenarios. We also look at the regular MTC operation, when the devices are sending their data after initial network entry has been performed. By including energy consumption into our methodology together with the conventional performance metrics, we aim at providing a complete and unified insight into MTC device operation, including its energy efficiency.

Introduction

Motivation and scope

Machine-Type Communications (MTC) also known as machine-to-machine have recently developed into a critical technology that is expected to generate significant revenues. Industry reports indicate the considerable potential of the MTC market, with millions of devices connected within the following years resulting in predicted revenues of up to $300 billion [1]. According to [2], the concept of MTC broadly enables a device (smart meter, actuator, or sensor) to capture a specific event and relay it through the underlying network to the associated application, which in turn translates it into meaningful data for the service consumer.

As traditional voice service revenues continue to shrink, mobile network operators are increasingly interested in MTC-based applications to bridge in the growing revenue gap [3]. Consequently, ETSI has started new activities with the goal of defining an end-to-end MTC architecture [4], whereas emerging IEEE 802.16p proposals address enhancements for IEEE 802.16m technology to support MTC applications [5]. Our recent analysis in [6] indicates that smart grid may become one of the key MTC use cases that involves meters autonomously reporting usage and alarm information to grid infrastructure to help reduce operational cost, as well as to regulate a customer's utility use based on load-dependent pricing signals received from the grid [7].

We expect that cellular technologies, such as 3GPP LTE and IEEE 802.16, will play a pivotal role in enabling smart metering applications. 3GPP LTE has recently defined several work items on MTC communications, primarily with respect to RAN overload control [8], [9]. The 3GPP Services group is also interested in MTC-related improvements for LTE Release 12 within the context of mobile data applications [10].

Summarizing the latest developments in 3GPP, ubiquitous smart grid deployments were shown to be hindered by many technical challenges. For instance, the use of random access mechanisms adds extra complexity to the evaluation of the target system [11]. With this research, we consider a typical smart metering MTC application scenario in 3GPP LTE-Advanced wireless cellular system featuring a large number of devices connecting to the network near-simultaneously and then sending their data through the network. As a starting point, we target comprehensive analysis of the random access channel (RACH) within the LTE-Advanced technology with respect to the congested MTC scenario and discuss some related research in what follows.

Research background

Thorough evaluation of RACH capacity, both with simulations and analytically, has been a popular research direction around 10 years ago for the legacy 3G cellular networks based on CDMA technology [12]. Originally, RACH served as an uplink contention-based channel to carry control information from client devices to the base station [13].

More specifically, a transmission of a random access request from a network client has been decomposed into two stages. At the preamble transmission stage, the power ramping technique was used to adjust the transmit power to particular channel conditions (see the related analysis with respect to the blocking, throughput, and delay in [14]). The basic principle of the power ramping procedure is that

a user starts sending its preambles with lower power and then gradually increases its transmit power in case a transmission failed at the previous attempt. As a result, less interference is caused to other network nodes and actual data transmission begins with already adjusted power. Further, a meaningful message was transmitted to the base station for the purposes of initial network access or bandwidth requesting.

The improved version of RACH within the 4G LTE-Advanced system has also attracted significant research attention. Considering a superior OFDMA technology, the successful transmission probability and throughput of RACH were studied in [15]. An alternative approach to the throughput and access delay evaluation of RACH has been pursued in [16] also providing several options for enhanced RACH resource utilization.

Importantly, all the aforementioned research efforts have only considered the lighter loads from human-oriented traffic and thus the related results are not directly applicable to MTC scenarios where a large population of devices attempts to access the network within a very short period of time.

Accounting for a surge in initial network entry, many recent works focus on overloaded RACH performance. Reflecting some initial discussions in 3GPP identifying the key impact of RAN overload, the work in [17] reviewed potential solutions and technology options to enhance the capability of LTE-Advanced to handle numerous requests from MTC devices. Alternatively, [18] compared the two most probable (as per ongoing 3GPP discussions) candidate solutions for random access preamble allocation and management.

However, previous work on RAN overload has rather been a set of candidate proposals while 3GPP was evaluating those identifying the minimal required changes to LTE specification. Most recently, the research in [19] concludes on some of these efforts by detailing the officially approved 3GPP evaluation methodology produced within the work item on RAN overload control [20].

Summarizing, the existing frameworks for RACH evaluation are mostly simulation-based. Furthermore, the obtained simulation results are often disjoint and even contradictory due to the lack of a unified methodology. As long as the recent calibration data approved by 3GPP has not been accounted for, many older findings may not be trustworthy for the community anymore.

In this paper, we develop a novel RACH evaluation methodology building upon the calibrated baseline and conduct thorough analysis and simulations of the RACH performance under MTC overload. We also give our prediction for the regular MTC operation, when the network is not experiencing a congestion.

Due to the fact that the MTC devices are typically small-scale and battery-powered, accounting for their energy consumption is of paramount importance [7]. In what follows, we seek to extend a validated evaluation methodology fully compatible with the 3GPP test cases with an in-depth analysis of RACH performance in overloaded MTC scenarios. By including energy consumption into our framework together with the traditional performance metrics (such as access delay and success probability), we aim at providing a complete and harmonized insight into MTC device operation.

*

Technology Background

*

Review of RACH signaling

Random access (RA) procedure of 3GPP LTE-Advanced is briefly summarized in Figure 1. Firstly, a User Equipment (UE) sends a random access preamble (Msg 1) to the base station via the Physical Random Access Channel (PRACH) by choosing it randomly out of the maximum of 64 preamble sequences [21]. Note that fewer preambles may actually be available, depending on the network configuration. A collision can occur at the base station when two or more UEs choose identical preamble sequences and send them at the same time. Preamble transmission may also fail due to insufficient transmission power.

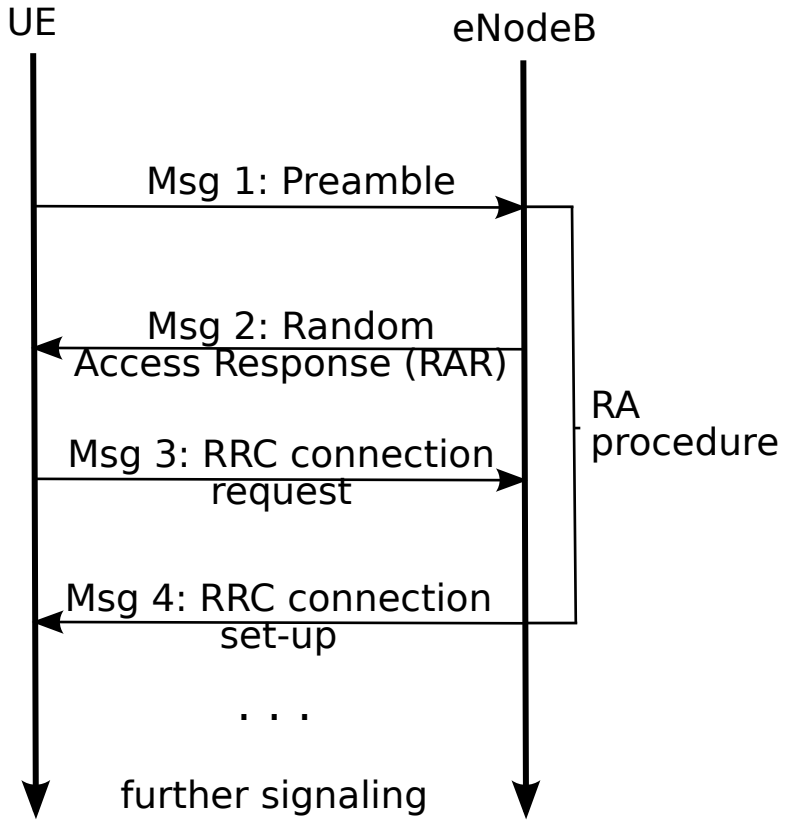

Figure 1: RA procedure signaling.

If a preamble has been received correctly, the base station (eNodeB) acknowledges it by a random access response (RAR or Msg 2) within the response window. An indicator of the resource in Physical Downlink Shared Channel (PDSCH) where RAR can be received is sent over Physical Downlink Control Channel (PDCCH) [22].

As eNodeB needs to establish which UE sent which preamble, collision resolution process is required. After some RAR processing time, UE transmits RRC connection request message (Msg 3) via the Physical Uplink Shared Channel (PUSCH) using the resources granted by Msg 2. RA procedure ends with a successful reception of RRC connection set-up message (Msg 4) from eNodeB.

When more than one UEs send a similar Msg 3 (due to a preceding preamble collision), eNodeB will at best respond only to one of these requests. Otherwise, if any of signaling messages has not been

received by UE, it restarts the RA procedure after some backoff time chosen randomly within a window given by the backoff indicator.

The bottleneck of the reviewed signaling procedure, especially when there are many requesting UEs, may be the growing collision probability (see Figure 2(a)). However, RAR delivery within the response window may also fail because of limited PDCCH resources (not considered in this paper). Furthermore, Msg 3 and Msg 4 may also have some probability of unsuccessful reception.

As a summary, several negative events may be the cause of a failed RA procedure and hence higher network access delays. Naturally, collision probability increases with the number of requesting UEs (or, in our case, MTC devices) and also depends on their traffic patterns.

For overloaded RACH scenarios, the number of contending devices per cell may reach the astonishing number of 30 000 (30K), as originally estimated by Vodafone in [23], borrowed by [24], and reused by 3GPP in [20] to conclude on the expected device densities. Such high numbers of competitors may lead to prohibitive collision probabilities and quickly deteriorate system resources. Therefore, 3GPP has recently been very active on evaluating the causes and consequences of such overloads.

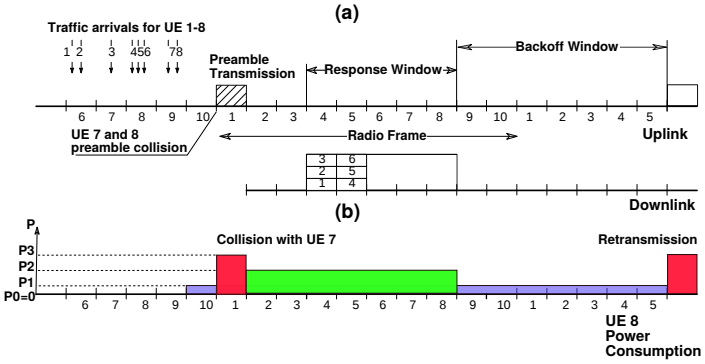

Figure 2: Example RA procedure: time evolution (a) and power consumption of e.g. UE 8 (b).

Recent standardization efforts

As mentioned in the previous section, a comprehensive evaluation methodology for LTE RA procedure has recently been sketched in [20]. The motivation behind this document was to identify the parameters of a verification scenario, as well as to present calibration data providing a trustworthy baseline for various 3GPP member companies. As such, it is very important that existing RACH-related simulation frameworks are harmonized with respect to the results therein.

Table 1 reviews parameters from several simulation methodology documents [20], [25]. In particular, cell bandwidth does not directly impact the RACH parameters and performance, while the value of 5 MHz only serves here for maintaining the consistency with 3GPP. In fact, the major parameter is the configuration of the Msg 1, which is mostly based on the PRACH configuration index. The latter defines subframe numbers, where a UE can attempt preamble transmissions, as well as the preamble length. The mac-ContentionResolutionTimer is the maximum number of subframes the UE waits after Msg 3 transmission and before considering the RA procedure as failed. Other settings have mostly been explained in the previous subsection, while some additional parameters (out of the methodology scope) will be detailed further on.

Table 1: Core simulation parameters

	Parameter	**Value**
-	Cell bandwidth	5 MHz
-	PRACH Configuration Index	6
s	Total number of preambles	54
L_1	Max. number of preamble transmissions	10
-	Number of UL grants per RAR	3
-	Number of CCEs allocated for PDCCH	16
-	Number of CCEs per PDCCH	4
-	Ra-ResponseWindowSize	5 ms
-	mac-ContentionResolutionTimer	48 ms
W	Backoff Indicator	20 ms
π_3/π_4	Probability of successful delivery for Msg 3/Msg 4	0.9/0.9
L_3	Max. number of HARQ Tx for Msg 3 and Msg 4 (non-adaptive HARQ)	5
M	Number of MTC devices	5K, 10K, 30K
N	Number of available subframes for device activation	10K, 60K
b	Periodicity of PRACH opportunities	5 ms
K	RAR response window	5 ms
K_1	Preamble transmission time	1 ms
K_0	Preamble processing time at eNodeB	2 ms
t_{pr}	Processing time before Msg 3 transmission	5 ms
t_{tx}	Time of transmission of Msg 3, waiting, and reception of Msg 4	6 ms
P_0	Power consumption in inactive state	0.0 mW
P_1	Power consumption in idle state	0.025 mW [26]
P_2	Power consumption of processing and Rx	50 mW [26]
P_3	Power consumption during Tx	50 mW [26]

System model and assumptions

We continue with more detailed system assumptions. One cell of 3GPP LTE-Advanced is considered featuring M identical MTC devices. A device randomly chooses a subframe for its uplink activation following the uniform distribution (traffic type 1) or beta distribution (traffic type 2) over $[1, N]$. A preamble, which takes 1 subframe to be sent, may be attempted for transmission at each b-th subframe, i.e. at slots 1, $b + 1$, ..., $b \cdot i + 1$, $i \in Z^+$.

Whenever activated, the MTC device is said to be backlogged until the completion of its RA procedure. Otherwise, the device is inactive. At subframes of service (when there is a PRACH opportunity), every backlogged MTC device uniformly chooses one of s preambles and sends it. According to [20], we assume a collision when two or more MTC devices select the same preamble, and all the collided preambles are considered failed (ignoring the power capture effect) after some service time. Otherwise,

the preamble is successful with the probability $1 - e^{-i}$ due to the power ramping, where i is the number of the transmission attempt [20]. The maximum allowed number of preamble transmission attempts is L_1.

If a transmission fails due to the collision or insufficient power, the MTC device uniformly selects a backoff counter within W. After K_0 subframes of pausing, the response window of size K starts (see Figure 2(a)). Within the response window, eNodeB sends RAR messages in the subframe uniformly distributed over $[1, K]$. If the MTC device does not receive RAR, the preamble transmission attempt is considered failed and the device backoffs.

When the MTC device receives RAR successfully, it starts processing Msg 3 for transmission during t_{pr}. Further, it sends Msg 3 and waits for $t_{tx} - 1$ to receive Msg 4 (see Figure 1). Msg 3 and Msg 4 are delivered successfully with the probabilities π_3 and π_4 respectively. The maximum allowed number of Msg 3 transmission attempts is L_3.

Simulation Methodology

Limitations of 3GPP methodology

Some initial parameters and assumptions related to simulations of LTE RACH for MTC scenarios have been proposed in [20]. The MAC layer parameters are borrowed from [21] and for the most part detail the RA procedure which was considered above. According to the proposed methodology, most PHY layer features are abstracted away to simplify performance evaluation. It is assumed that out of those, the power ramping procedure has the most impact on the metrics of interest. The ramping procedure is meant for power control and has been detailed in [27]. In [20], this procedure has been reduced to a simple function e^{-i} that defines the probability of failure.

Another important aspect of the methodology is the considered traffic patterns. The document [20] is focused on the overloaded scenarios, which could theoretically cause abnormal system loads, high collision probabilities, and prohibitive RA procedure delays. In fact, as our subsequent analysis shows, only traffic type 2 (beta distribution) is causing actual overloads. This overloaded scenario, however, is difficult to evaluate analytically and we analyze it mostly based on the simulation results. Traffic type 1 scenario is used primarily for calibration purposes and, in contrast to the other one, could be approximated and verified with our analytical approach.

Complementing the 3GPP methodology, which is already considering delay, collision probability, and the average number of preamble transmissions, we propose an extended analysis of energy-related metrics. In this paper, we also consider some overload control mechanisms and regular system operation conditions, which, in combination with a detailed analytical model, is intended to complete evaluation of LTE RACH in MTC scenarios.

Simulator description

In what follows, we detail our advanced protocol-level simulator of RACH operation and the related improvements. For the purpose of conducting extensive evaluations, existing network simulation tools were considered to be either inadequately slow or lacking the necessary signaling support. As such, a novel simulator has been developed taking advantage of extensible modular structure for improved

scalability. The benefit of our simulator is its flexibility in the choice of the parameters of interest, including number of devices, signaling timings, processing mechanisms, and system settings such as number of preambles, backoff window size, etc.

More importantly, our simulation tool allows for simple integration of the extended components, such as overload control mechanisms and power consumption measurements. Finally, the software is supplied with flexible statistics collecting and processing functions while is able to evaluate various parameters of interest ranging from access latency/probability to fine-grained energy-related metrics. All the messages transmitted over the same channel are multiplexed and processed jointly with explicit modeling of collision behavior. The operation of RACH accounts for all the necessary features discussed previously.

In Figure 3, the simplified structure of the simulator is captured. There are three core classes implemented in C++: traffic generator, UE, and eNodeB. Traffic generator has support for three basic patterns: Uniform, Poisson, and Beta, which are configured for all the UEs at the beginning of a simulation run. Full buffer (saturated) model is also available as a separate option. Each device has a dedicated traffic generator implementing the chosen traffic pattern.

The UE class is supporting operation related to Msg 1 and Msg 3 transmission, as well as Msg 2 and Msg 4 reception. Several supplementary functions, such as power ramping, are also maintained at the UE side. The eNodeB class is responsible for detection of Msg 1 failures due to a collision or insufficient transmission power. After the detection procedure, a decision on whether to send Msg 2 is made.

In our event-driven simulator, each event is processed by the event handler and could trigger another event of the same or different type. For example, a traffic arrival event triggers the Msg 1 transmission mechanism at the appropriate UE, which in turn schedules Msg 2 transmission at the eNodeB if Msg 1 has been successful. At the same time, a traffic arrival event causes the formation of another traffic arrival event at the same device based on the traffic arrival patterns discussed above. After Msg 2 reception, Msg 3 transmission is scheduled. This process is repeated until the successful reception of Msg 4, which is enabling the statistics collector. Finally, sorted results are saved into a file that is delivered to a Matlab parser for the purposes of visualization.

Simulator validation

In order to validate our simulation tool against the trustworthy and reliable 3GPP test cases, we have conducted in-depth calibration. In particular, we used the recent reference data approved by 3GPP in the technical report TR 37.868 [20] described above. In Table **??**, a comparison between the results from [20] (see e.g. Table 6.2.2.1.1) and our simulation/analysis results are shown for traffic type 1 (uniform activation pattern). The details of our analytical approach will be given in the following section and we include them here only for consistency.

The CDFs of initial network entry delays for 30K MTC devices (both uniform and beta traffic patterns) are presented in Figure 4. Importantly, 90% and 10% quantiles agree with the reference values from [20] with less than 15% of difference.

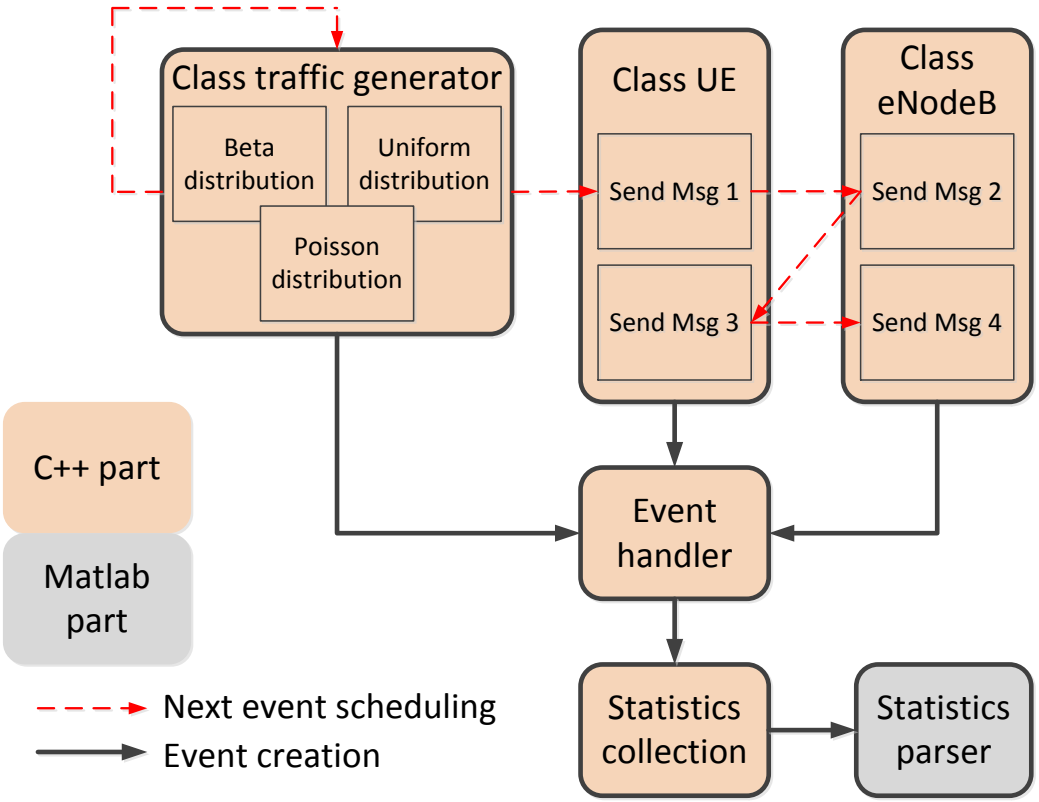

Figure 3: Simplified simulator structure.

Analytical approach

Delay analysis

In this section, we concentrate on the overloaded RACH scenario with the traffic type 1 (uniform activation pattern) according to the 3GPP methodology [20], and detail our approach to the analytical evaluation of the RACH performance in terms of, primarily, average network access delay. We split the overall delay into two components, corresponding to the Msg 1-2 and Msg 3-4 processing:

$$E[\tau] = E[\tau^{(1)}] + E[\tau^{(2)}], \tag{1}$$

where $E[\tau^{(1)}]$ is the time interval between the device activation and the RAR response reception and $E[\tau^{(2)}]$ is the time interval between the end of the subframe when RAR was received and the end of the Msg 4 processing.

The calculation of the distribution and the mean value of the random variable $\tau^{(2)}$ is nearly trivial and the final expression is given as follows:

$$E[\tau^{(2)}] = t_{pr} + t_{tx} \cdot \bar{n}_3, \tag{2}$$

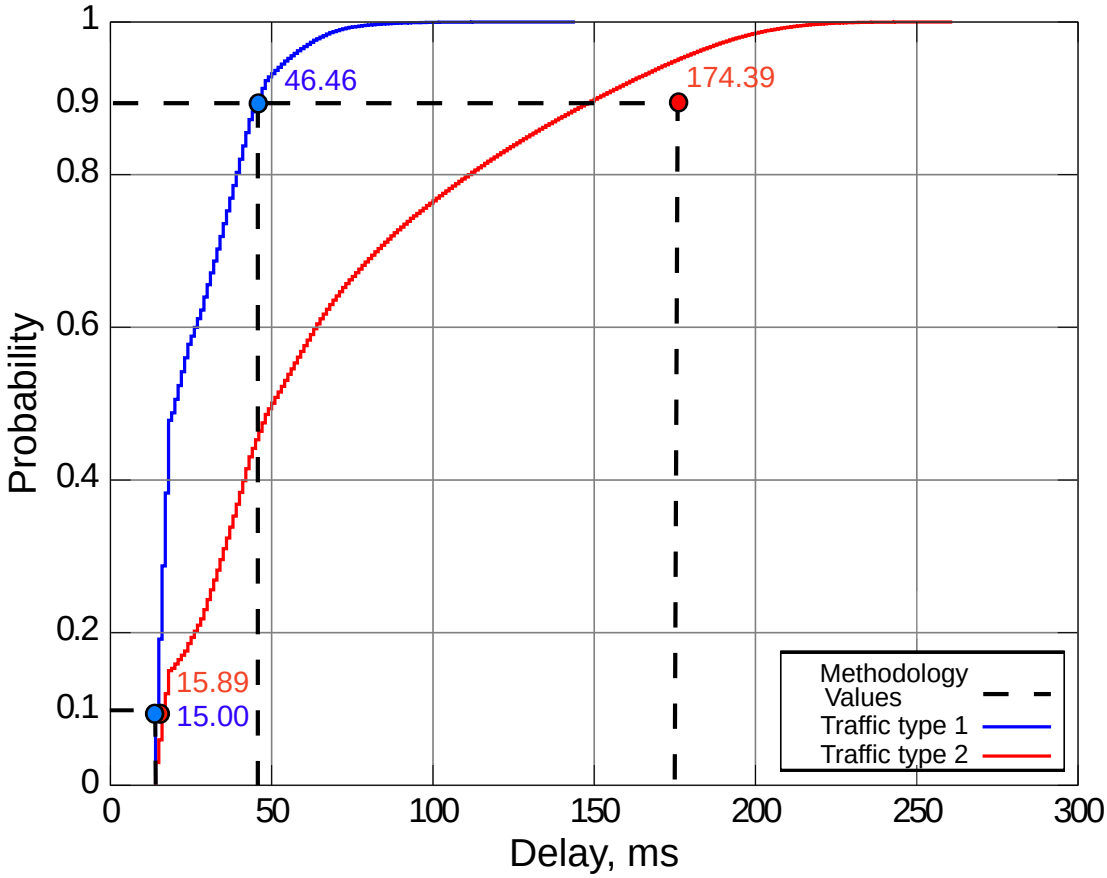

Figure 4: Access delay CDF for 30K MTC devices (simulation results).

where $\bar{n}_3$ is the average number of Msg 3 and Msg 4 transmissions addressed below, while parameters t_{pr} and t_{tx} are the processing and Tx timings respectively.

The distribution of the number of Msg 3 and Msg 4 transmissions is given as follows:

$$Pr\{n_3 = 1\} = \pi_{tx},$$

$$Pr\{n_3 = 2\} = (1 - \pi_{tx})\pi_{tx}, ...$$

$$Pr\{n_3 = L_3\} = (1 - \pi_{tx})^{L_3 - 1}\pi_{tx},$$

where $\pi_{tx} = \pi_3\pi_4$ is the probability that both Msg 3 and Msg 4 are transmitted successfully (complementary, $(1 - \pi_3\pi_4)$ is the probability that either Msg 3 or Msg 4 is lost), and L_3 is the maximum number of allowed Msg 3 and Msg 4 transmission attempts. Here, we take into account only successful transmissions. Due to the fact that the loss probability is negligibly small, we disregard lost preambles and assume that the expectation of the number of transmissions over all preambles approximately equals the conditional expectation over successfully transmitted preambles.

Therefore, the average number of Msg 3 and Msg 4 transmissions can be established as follows:

$$\bar{n}_3 = \pi_{tx} \sum_{n=1}^{L_3} n(1-\pi_{tx})^{n-1} =$$

$$= \frac{1}{\pi_{tx}} \left[1 - (1-\pi_{tx})^{L_3}(1+L_3\pi_{tx})\right].$$

System without collisions

To analyze $\tau^{(1)}$, we first consider the original random-access system without any collisions. Hence, retransmissions occur solely due to the power ramping. In case of successful preamble transmission at the first attempt, the service time consists of preamble transmission time, preamble processing, and RAR response time. Also we take into account the averaged time between the device activation and the first preamble transmission attempt $b/2$, i.e.

$$E[\tau^{(1)}|\text{success at the 1st attempt}] =$$

$$= b/2 + K_1 + K_0 + (K+1)/2, \tag{3}$$

where K_1 is the preamble transmission time, K_0 is the pausing time, and K is the RAR response window size (in ms). Here, $(K+1)/2$ stands for the average RAR response time since we assume that the processing starts immediately after receiving the RAR response; it is obtained as the expectation of discrete uniform distribution over $[1, K]$.

As mentioned above, the probability of a successful preamble transmission at the attempt i is $(1 - e^{-i})$ and the complementary probability of a failed transmission is e^{-i}, correspondingly. Further, we average the sum of the backoff time and additional waiting time until the next b-th slot denoting the aggregate value as $\bar{w}$. The distribution of the service time for Msg 1-2 can be given as:

$$Pr\left\{E[\tau^{(1)}] = \frac{b}{2} + K_1 + K_0 + \frac{K+1}{2}\right\} = (1 - \frac{1}{e^1}),$$

$$Pr\left\{E[\tau^{(1)}] = \frac{b}{2} + (K_1 + K_0 + K + \bar{w}) + K_1 + K_0 + \frac{K+1}{2}\right\} =$$

$$= \frac{1}{e^1}(1 - \frac{1}{e^2}), ...$$

$$Pr\left\{E[\tau^{(1)}] = \frac{b}{2} + (n-1)(K_1 + K_0 + K + \bar{w}) + K_1 + K_0 + \frac{K+1}{2}\right\} =$$

$$= \left(1 - \frac{1}{e^n}\right) \prod_{i=1}^{n-1} \frac{1}{e^i}, ...,$$

where $b/2$ stands for the time between the arrival and the beginning of the first preamble transmission attempt, $(K_1 + K_0 + K + \bar{w})$ is the component, which is added every time when transmission fails. Then we average the service time and, as such, the mean service time before the beginning of Msg 3 Tx

is obtained as:

$$E[\tau^{(1)}] = (K_1+K_0+K+\bar{w}) \sum_{n=1}^{L_1} n \left(1 - \frac{1}{e^n}\right) \prod_{i=1}^{n-1} \frac{1}{e^i} +$$

$$+\frac{b-K+1}{2} - \bar{w} = c_1(K_1+K_0+K+\bar{w}) + \frac{b-K+1}{2} - \bar{w}, \quad (4)$$

where $\bar{w} = c_2(c_2+1)+(c_2+b+bc_3)(W-bc_3-c_2)+bc_3c_2$, $c_1 \cong 1.42$, $c_2 = b\lceil K/b \rceil - K$, and $c_3 = \lfloor (W-c_2)/b \rfloor$. This expression presents a lower bound of $E[\tau^{(1)}]$ for the studied system.

System with collisions

Analysis of the system with collisions constitutes a more challenging task, and an accurate solution is difficult to obtain due to the property of memory as long as the system features random backoff time, constant timings and, especially, large number of preambles. For example, in the classical multi-user system with one preamble, the approximate delay values can easily be obtained as has been done for ALOHA in [28]. For our system, however, the use of that popular technique does not give a good approximation and we thus extend the approach from [29]. In order to abstract away the memory property and establish an estimate for $E[\tau^{(1)}]$ for the system with collisions, we adopt the following equivalent model.

(i) We assume Bernoulli activation flow with the rate of π, when a device generates a new connection request per subframe with the equivalent probability $\pi = 1/N$, where N is the number of subframes in the original system.

(ii) We omit explicit consideration of the waiting interval and the backoff window replacing them by an assumption that at every subframe a backlogged device activates with a certain probability $\pi_0 = 1/(K_0 + K_1 + K + \bar{w})$. Basically, this means that the device activates once over the period $(K_1 + K_0 + K + \bar{w})$ if the first transmission fails due to a collision or insufficient power.

(iii) The probability of successful departure is μ, i.e. the request is served with a certain probability μ in the current subframe, otherwise the device attempts to access the channel in the next available subframe.

(iv) Finally, we abstract away the maximum number of preamble transmission attempts.

Within the simplified equivalent system model, an approximation of the mean network entry delay may be obtained as follows. For the system without collisions, the probability of being served ($\tilde{\mu}$) can be calculated from the equation $E[\tilde{\tau}^{(1)}] = E[\tau^{(1)}]$ as:

$$\tilde{\mu} = \frac{1}{E[\tau^{(1)}]} = \frac{1}{c_1(K_1+K_0+K+\bar{w}) + \frac{b-K+1}{2} - \bar{w}}, \quad (5)$$

where $E[\tau^{(1)}]$ is the mean time interval between the device activation and the RAR reception, whereas $E[\tilde{\tau}^{(1)}]$ is the respective interval in the equivalent model. We will refer to the expression (5) in what follows, when calculating the system load for the devices that avoid collisions.

We continue by actually accounting for collisions. Let us consider one subframe and assume that a particular device i has generated a request and also selected a preamble. Let the system be in the state

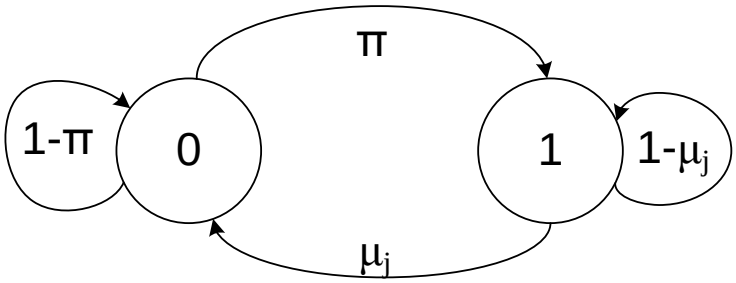

Figure 5: Two-state Markov chain describing the number of requests at a device.

j, where j is the number of backlogged devices including the device i. In the state j, the behavior of the device i can be described by a simple two-state Markov chain, where a state represents the number of pending requests Q_i at the device, which can be equal to either 0 or 1 (see Figure 5). The transition matrix for the considered chain is given as:

$$\Pi = \begin{pmatrix} 1-\pi & \pi \\ \mu_j & 1-\mu_j \end{pmatrix}. \tag{6}$$

As such, the steady-state distribution $\omega = \{\omega_0, \omega_1\}$ can be obtained from the matrix equation $\Pi^T \omega = \omega$, when $\omega_0 + \omega_1 = 1$. Hence, the average number of requests Q_i is expressed as:

$$E[Q_i] = 1 \cdot \omega_1 = \frac{\pi}{\pi + \mu_j}, \tag{7}$$

where Q_i is the number of requests at the considered device i and μ_j is the probability of successful preamble transmission.

By the Little's law, we obtain the average time spent by the system in the state j as:

$$E[\tau_j^{(1)}] = \frac{E[Q_i]}{\pi} = \frac{1}{\pi + \mu_j}. \tag{8}$$

In the state j, for $j-1$ backlogged devices, the probability of accessing the channel and selecting the same preamble as the device i had is $\pi_0 \cdot 1/s$ (the probability to activate times the probability to select the same preamble). For the inactive $M-j$ devices, the corresponding probability is $\pi \cdot 1/s$ (the probability of arrival in a subframe times the probability to select the same preamble).

Thus, the probability π_j^* to avoid collision for the device i in the state j can be calculated as follows:

$$\pi_j^* = \left(1 - \pi_0 s^{-1}\right)^{j-1} \left(1 - \pi s^{-1}\right)^{M-j}. \tag{9}$$

Further, we account for the power ramping effect. The probability to avoid collision at the attempt n is given as follows:

$$Pr\{\text{1st successful}\} = \left(1 - \frac{1}{e}\right)\pi_j^*,$$

$$Pr\{\text{2nd successful}\} = \left(1 - \left(1 - \frac{1}{e}\right)\pi_j^*\right)\left(1 - \frac{1}{e^2}\right)\pi_j^*,$$

...

$$Pr\{n\text{th successful}\} = \left(1-\frac{1}{e^n}\right)\pi_j^* \prod_{i=1}^{n-1}\left(1-\pi_j^*\left(1-\frac{1}{e^i}\right)\right),$$

....

Here, we also neglect all the lost preambles as we did before, averaging by successful transmissions and replacing the sought expectation with the conditional one. The average number of attempts can be obtained as:

$$\bar{n}_j = \pi_j^* \sum_{n=1}^{L_1} n\left(1-\frac{1}{e^n}\right)\prod_{i=1}^{n-1}\left(1-\pi_j^*\left(1-\frac{1}{e^i}\right)\right). \tag{10}$$

Taking into account the effect of power ramping, we establish the probability μ_j of successful request i transmission:

$$\mu_j = \left(\bar{n}_j \cdot (K_1+K_0+K+\bar{w}) + \frac{b-K+1}{2} - \bar{w}\right)^{-1}, \tag{11}$$

where $\bar{n}_j$ is given by (10).

The average service time can then be calculated as:

$$E[\tau^{(1)}] = \sum_{j=1}^{M}\theta_j E[\tau_j^{(1)}] = \sum_{j=1}^{M}\theta_j \frac{1}{\pi+\mu_j}, \tag{12}$$

where $\{\theta_j\}_{j=1}^{M}$ is the steady-state distribution, θ_j is the steady-state probability of being in the state j.

In order to obtain the stationary distribution defined above, we need to consider all the state transitions and solve the corresponding matrix equation of dimension M. To reduce the complexity of such calculations, we omit more complicated transitions between the states and average θ_j, using binomial distribution, by:

$$\theta_j = \binom{M-1}{j-1}\rho^{j-1}(1-\rho)^{M-j}, \tag{13}$$

where ρ is the device load, and $\binom{M-1}{j-1} = \frac{(M-1)!}{(j-1)!(M-j)!}$.

Here, we disregard all the collisions between other devices by assuming that only the considered device i falls into a collision. Thus, we can calculate the system load $\rho = \pi/\tilde{\mu}$ using the expression (5) for the probability of being served μ, derived for the system without collisions.

The resulting expression for the approximate mean service time is:

$$E[\tau^{(1)}] = \sum_{j=1}^{M} \frac{\binom{M-1}{j-1}\rho^{j-1}(1-\rho)^{M-j}}{\frac{1}{N} + \left(a_j(K_1+K_0+K+\bar{w}) + \frac{b-K+1}{2} - \bar{w}\right)^{-1}}, \tag{14}$$

where π_j^* and a_j are given above.

Applicability discussion

In this subsection, we emphasize that the proposed analytical approach is applicable only for the practical systems, which can be reduced to a stationary system. This, obviously, can be done when considering the uniform distribution of the device activation time over a fixed time interval. Otherwise, for instance, in case of beta distribution (traffic type 2), one should take into account dynamic changes of

the parameters $\pi(t)$, $\pi_0(t)$, $\mu(t)$, and $\tilde{\mu}(t)$, which is rather tedious and is thus left out of scope of this paper.

However, our approach allows for a broad range of important practical extensions. In particular, we can easily analyze a regular MTC operation scenario described further on in Section 0.0.13, where inter-arrival time follows exponential distribution with a certain parameter $1/\lambda$. Due to the stationarity of this process, we exploit the same approach and, literally, the same formulas, while only replacing the probability π with the probability that at least one packet arrives in a particular subframe:

$$\pi = 1 - \Pr\{X(t, t + t_0) = 0\}, \tag{15}$$

where t_0 is the size of the subframe, $t_0 = 1$ ms.

In more detail, the data arrival flow constitutes a stationary, ordinary, and memoryless process $\{X(0,t), t \geq 0\} = \{X(t), t \geq 0\}$ representing the number of data arrivals occurred until the moment t:

$$\Pr\{X(t) = k\} = \frac{\lambda^k t^k}{k!} e^{-\lambda t}, k = 0, 1, 2, ..., \tag{16}$$

where λ is the arrival flow rate.

Hence, due to the property of stationarity, the probability p_0 that the number of arrivals within a slot of length t_0 equals 0, is given by:

$$\pi = 1 - \Pr\{X(t, t + t_0) = 0\} = 1 - e^{-\lambda t_0}, \tag{17}$$

where $X(t, t + t_0)$ is the number of arrivals over the time interval $[t, t + t_0)$, t is an arbitrary time moment, and t_0 is the subframe length.

We finally note that the proposed analytical framework can also incorporate some overload control mechanisms, such as e.g. initial backoff (see Section 0.0.12 for details). It will produce changes to the equation (9) and derivations above concerning the probability to avoid collisions, i.e. the probability to collide should be set to $\pi \cdot \pi_0 \cdot 1/s$ for all $(M - j)$ inactive devices due to the device activation before its first transmission attempt.

Energy consumption analysis

As mentioned previously, our methodology is powerful enough to be extended for the energy-related analysis of the MTC device behavior. Therefore, we introduce new important parameters (out of scope of [20]), which represent power consumption levels of a typical MTC device. In particular, we consider the maximum of four different device power states (see Figure 2(b)):

(i) P_0 – Inactive State. In this state, the device consumes minimum power. The buffer is empty, no data to transmit.

(ii) P_1 – Idle State. The device is activated, but it does not transmit in the current subframe.

(iii) P_2 – Rx State. The device is expecting Msg 2/Msg 4 or is processing the related responses.

(iv) P_3 – Tx State. The device is transmitting Msg 1/Msg 3. The maximum power is consumed.

We estimate the total energy consumption of a device per subframe as a sum of fractions of time spent in every power state multiplied by the power consumption in the corresponding state. Therefore, we analytically establish the time spent by the device in every of four possible power states by calculating the corresponding time proportions as follows.

Tx state time is given by:

$$q_3 = K_1\bar{n} + \frac{1}{\pi_{tx}}\left[1-(1-\pi_{tx})^{L_3}(1+L_3\pi_{tx})\right], \tag{18}$$

where $\bar{n}$ is the estimation for the mean number of preamble transmission attempts and $K_1\bar{n}$ corresponds to the time of preamble transmission, while the second part accounts for the average number of Msg 3 transmissions.

Rx state time is given as:

$$q_2 = K(\bar{n}-1) + \frac{K+1}{2} + t_{pr} + \\ + \frac{t_{tx}-1}{\pi_{tx}}\left[1-(1-\pi_{tx})^{L_3}(1+L_3\pi_{tx})\right], \tag{19}$$

where $K(\bar{n}-1)$ is the time spent expecting the RAR response, $(K+1)/2$ is the mean index of the response from eNodeB at the successful attempt, and the remainder corresponds to the processing and receiving of Msg 3 and Msg 4.

Idle state time can be calculated as:

$$q_1 = \frac{b}{2} + K_0\bar{n} + (\bar{n}-1)\bar{w}, \tag{20}$$

where $K_0\bar{n}$ is the time for the eNodeB to process the preamble after its reception and $\frac{b}{2}$ is the idle time between the activation and the beginning of the preamble transmission. The approximate average number of preamble transmission attempts is given by the formula:

$$\bar{n} = \sum_{j=1}^{M}\binom{M-1}{j-1}\rho^{j-1}(1-\rho)^{M-j}\bar{n}_j. \tag{21}$$

Finally, the estimated total energy expenditure of an MTC device can be calculated as:

$$\epsilon = P_0(1-q_3-q_2-q_1)+P_1q_1+P_2q_2+P_3q_3. \tag{22}$$

The analytical and simulation results for the MTC device power consumption (traffic type 1, uniform) are summarized in Figure 6. We notice that the provided analytical approximation is extremely accurate even when the population of MTC devices is high.

Numerical results

General remarks

In this section, we detail important numerical results obtained with our methodology. It further comprises two main subsections, which touch upon overload control mechanisms and energy efficiency of

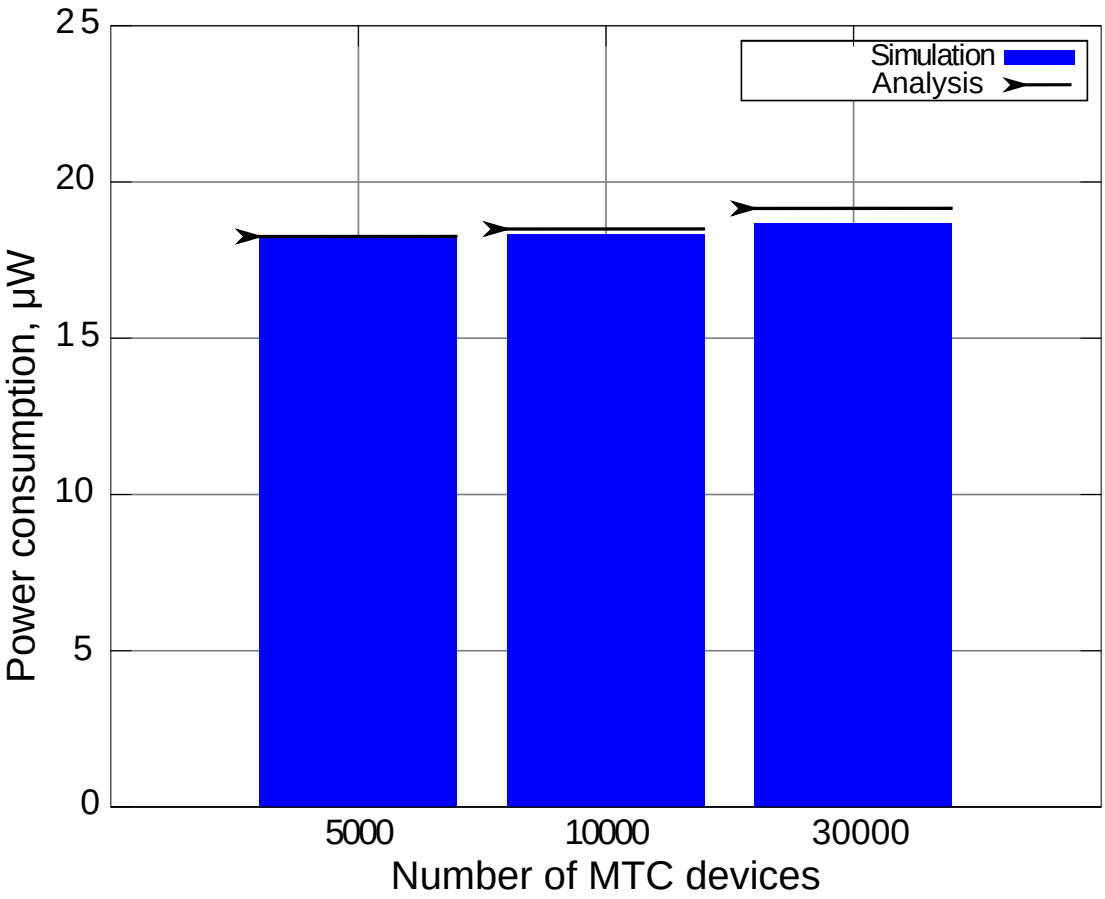

Figure 6: RA power consumption.

regular MTC operation, respectively.

The former subsection considers the overloaded conditions when the UEs are activated according to beta distribution (traffic type 2). This practically means that the arrival process is non-stationary and, due to the reasons mentioned in subsection 0.0.9, our analytical approach cannot be used. Therefore, we concentrate on simulation results therein.

The latter subsection features both analytical and simulation data due to the stationarity of the considered traffic arrival process.

Overload control mechanisms

Our evaluation framework detailed above may be used to conclude on the feasibility of the candidate RAN overload control solutions (see e.g., [19]). In particular, we consider a combination of initial backoff (pre-backoff) proposed by [30] and MTC-specific backoff described in [31]. The main idea is that the backoff time is invoked not only after any unsuccessful preamble transmission attempt, but also at the very beginning of every RA procedure to de-correlate the surge in channel access attempts from many MTC devices. As a result, with a large enough backoff indicator (BI) value chosen, the network entry peaks can be smoothened and the collision probability may be decreased. Given that many MTC applications are delay-tolerant, some increase in the mean access delay is often acceptable.

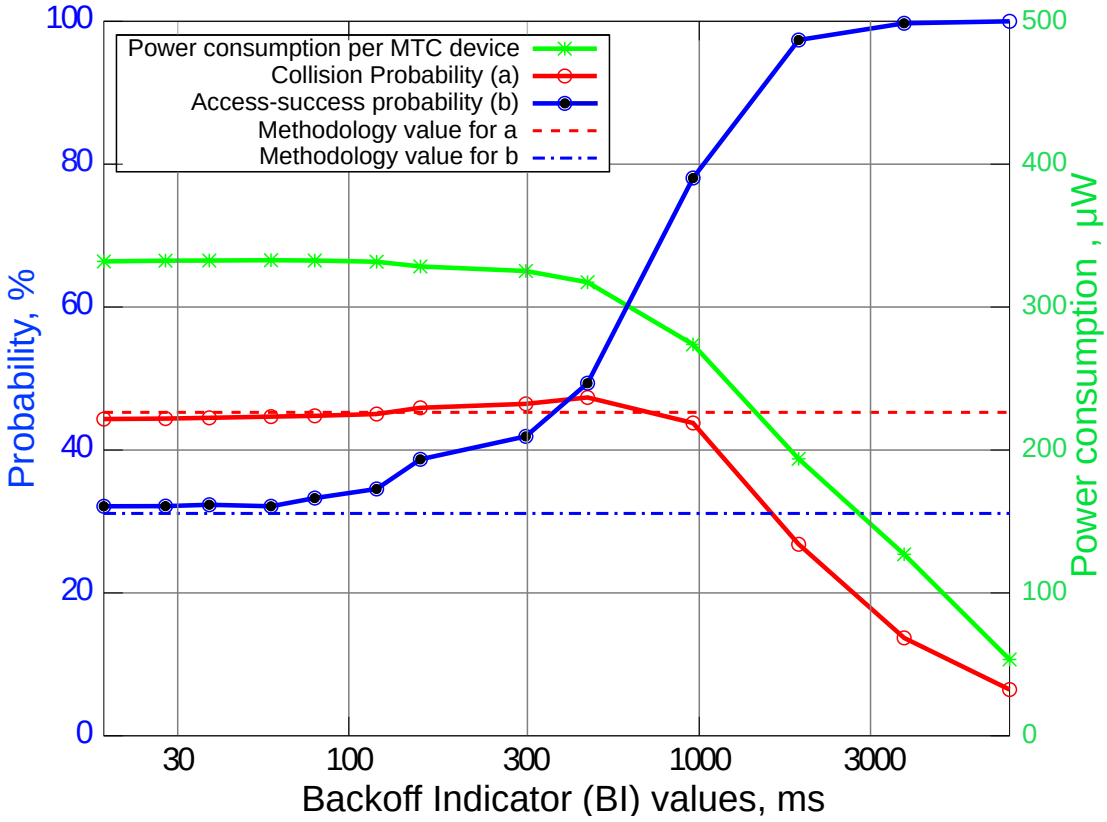

Figure 7: Overload control performance (simulation results).

To conclude on this research vector, we have analyzed the behavior of the MTC system under both traffic patterns proposed in [20] for the RAN overload scenario. The heavier traffic type 2 (beta distribution of device activation times) yields more correlated network entry attempts (see Section 0.0.4). As mentioned above, analytical tractability of this traffic pattern is very limited. Therefore, we focus on simulation to obtain important numerical results.

In particular, Figure 7 details MTC device power consumption, collision probability, and access success probability for different BI values (which may also be larger than those currently defined by the LTE specification). In this figure, the BI starts from 20 ms and is increased up to its maximum standardized value of 960 ms [21]. As can be seen from the plot, access success probability after all the available retries is about 80%, which may not be acceptable for many MTC-aware scenarios. Therefore, we consider the use of three reserved options for the BI in [21] with larger values: 1920, 3840, and 7680 ms. As a result, additional delay is compensated by a considerably higher (up to 100%) reliability level of the network access.

Summarizing, in contrast to original 3GPP expectations that both traffic patterns (type 1 and type 2) will overload the MTC system, it appears that traffic type 1 does not cause any significant network congestion (see Table **??**). Moreover, the considered simplistic overload control mechanisms, such as pre-backoff and MTC-specific backoff, can alone alleviate congestion for traffic type 2. However, when designing overload control mechanism to handle the correlated network entry attempts, we should not negatively impact the regular MTC operation. Therefore, below we pay attention to the regular MTC

traffic and the values of key system parameters when serving it.

Energy efficiency of regular MTC operation

By contrast to the previous (sub)sections focusing on the case of MTC overload, this subsection concentrates on regular MTC operation when all the devices have already performed their initial network entry. The LTE specification allows the use of PRACH for scheduling request transmission, whenever the device does not have the resources allocated over the default control channel [21]. However, the methodology in [20] defines only overloaded network entry patterns leaving open the actual device traffic model. Therefore, below we consider the reference MTC uplink traffic model in accordance with the recent 3GPP technical report [32].

In order to predict the MTC network load in the regular case, and with respect to [33], the expected density of MTC devices was estimated to be around $5K/km^2$. As such, according to [23], for the general cell radius of $0.5km$, the expected number of transmitting MTC devices should not exceed $7K$. As such, we observe the network behavior starting from a considerably low number of connected MTC devices and up to the predicted maximum for the non-overloaded network.

Further, the document [32] suggests that the packet inter-arrival time distribution is exponential with the constant mean value of 30 seconds and alternative packet sizes of 256 and 1024 bits. Given these assumptions, the data packet delay CDFs for different numbers of MTC devices are demonstrated in Figure 8. Here, the packet delays were accumulated starting from the moment of traffic arrival, and not from the beginning of the RA procedure (as was suggested by the methodology [20]).

More importantly, since in the considered scenario the MTC devices send actual data, we may explicitly account for their energy efficiency. Energy efficiency may be calculated as the number of data packets that were transmitted successfully by a device, weighted with the packet size, and related to the total energy (in Joules) spent by this device. Consequently, the dimension of this important metric is bits per Joule (bpJ).

Both simulation results and our analysis (see subsection 0.0.9) are shown in Figure 9. Noteworthy, the device energy efficiency is changing insignificantly with the overall population of the MTC devices. This is due to the fact that the actual MTC device energy efficiency is quite low when compared to e.g., a typical mobile device [34], and has significant potential for further improvement. Finally, we conclude that the analytical results are very close to the simulated values which confirms the practical usefulness of the proposed methodology.

Conclusions and future work

In this paper, we emphasized the lack of comprehensive evaluation frameworks for the performance assessment of the RACH mechanism within the 3GPP LTE-Advanced technology that would rely on both simulation and analysis components. Moreover, previous evaluation results are often disjoint and contradictory due to the fact that the unified 3GPP calibration methodology has only been finalized very recently. As such, we have accounted for the latest reference data approved by 3GPP while validating our own advanced protocol-level RACH simulator.

Further, we conducted an in-depth analysis of the case when the RAN is facing a surge in near-simultaneous network entry attempts from an excessive number of MTC devices. To add even more

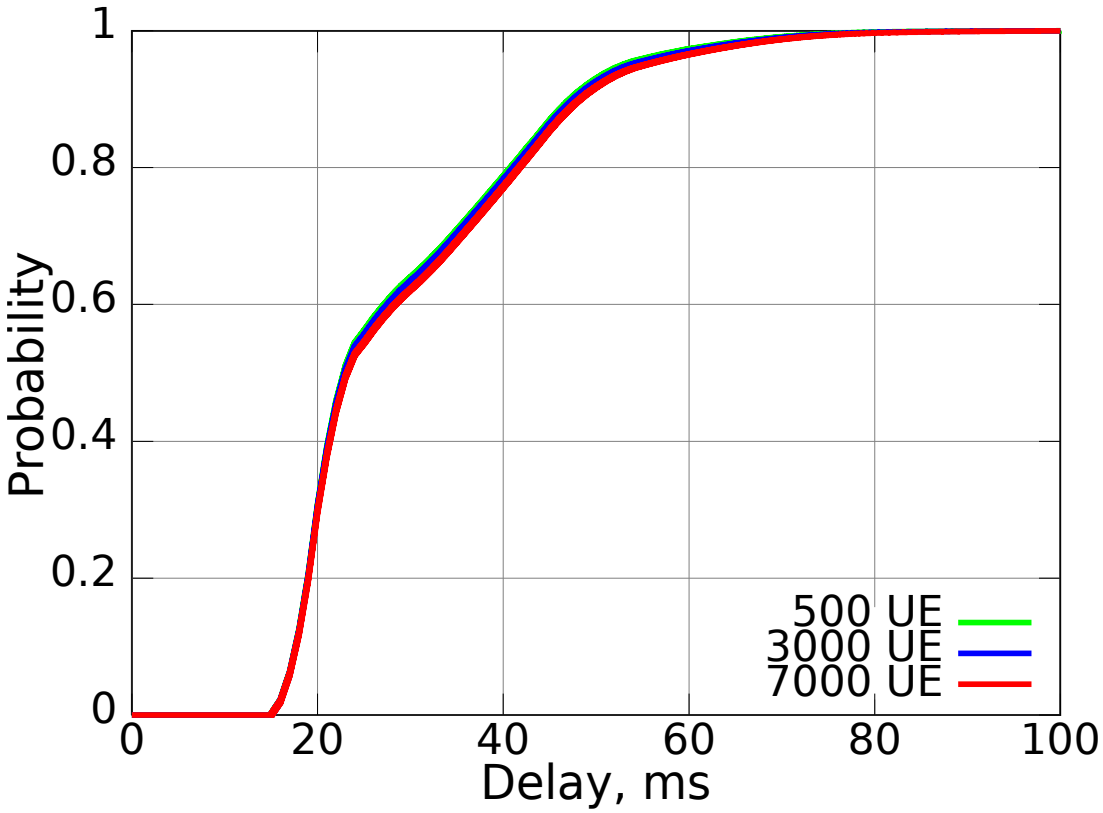

Figure 8: Data delay CDF for regular MTC traffic (simulation results).

insight to the MTC device behavior, we also considered the regular device operation, when initial network entry has already been performed and the device is sending its actual uplink data.

Our approach allows to investigate the performance of MTC devices, the impact of RACH settings, and the overload control mechanisms in terms of conventional metrics, such as access success probability and medium access delay. In particular, the limitation of existing access protocol in case of correlated network entry attempts has been indicated and the benefits of several potential enhancements were highlighted. These modifications do not require major protocol change and feature the pre-backoff technique complemented by the usage of larger MTC-specific backoff values. Moreover, the analytical technique presented in this paper is a powerful tool that can be used to extend the 3GPP RACH calibration methodology [20]. One such improvement accounts for the power-related metrics of an MTC device to conclude on all the aspects of the random access procedure, including its energy efficiency.

Our estimation of delay and power consumption has been found to be very accurate and we plan to work on it further considering additional realistic features of RACH performance. Our analytical approach is also applicable for studying other MTC-related enhancements within LTE-Advanced, such as sending scheduling requests via PUCCH, Extended Access Barring (EAB, [35]) scheme, and Extended Wait Timer (eWaitTimer, [36]) mechanism. Another challenging research direction is the consideration of scheduling-based approaches [37] for MTC, which can also be incorporated into our framework.

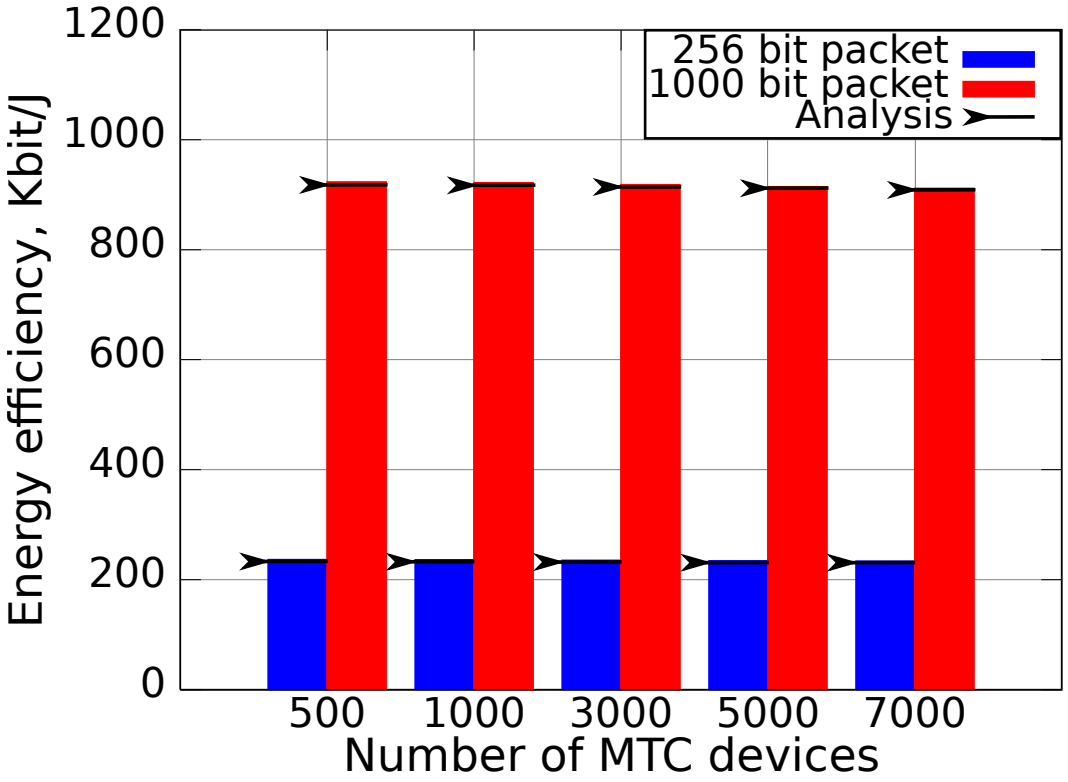

Figure 9: Regular MTC operation performance.

*

Acknowledgment

This research was conducted within the Internet of Things program of Tivit, funded by Tekes. The authors are also grateful to Dr. Nageen Himayat (Wireless Communications Laboratory, Intel Corporation, USA) for initializing their interest in this topic, as well as to Prof. Andrey Turlikov (St. Petersburg State University of Aerospace Instrumentation, Russia) and Dr. Zsolt Saffer (Budapest University of Technology and Economics, Hungary) for many helpful discussions and insightful comments in the course of this work.

Bibliography

[1] Machine-To-Machine (M2M) & Smart Systems Forecast 2010-2014. *Harbor Research Report*, 2009.

[2] Emmerson B. M2M: the Internet of 50 billion devices. *M2M Magazine*, 2010.

[3] Wu G., Talwar S., Johnsson K., Himayat N., Johnson K. M2M: From Mobile to Embedded Internet. *IEEE Communications Magazine*, 2011.

[4] ETSI. Machine-to-Machine communications (M2M); M2M service requirements. *TS 102 689*, 2010.

[5] Cho H., Puthenkulam J. Machine to Machine (M2M) Communication Study Report. *IEEE 802.16p-10/0005*, 2010.

[6] Himayat N., Talwar S., Johnsson K., Mohanty S., Wang X., Wei G., Schooler E., Goodman G., Andreev S., Galinina O., Turlikov A. Informative text on Smart Grid applications for inclusion in IEEE 802.16p Systems Requirements Document (SRD). *IEEE C802.16p-10/0007r1*, 2010.

[7] Andreev S., Galinina O., Koucheryavy Y., Energy-Efficient Client Relay Scheme for Machine-to-Machine Communication, *Global Telecommunications Conference (GLOBECOM)*, 2011.

[8] 3GPP. System Improvements for Machine-Type Communications. *TR 23.888*, 2011.

[9] 3GPP. Study on enhancements for MTC. *TR 22.888*, 2012.

[10] 3GPP. Machine-Type and other Mobile Data Applications Communications Enhancements. *TR 23.887*, 2012.

[11] Sanguinetti L., Morelli M., Marchetti L. A random access algorithm for LTE systems. *European Transactions on Telecommunications*, 2012.

[12] Vukovic I., Throughput Comparison of Random Access Schemes in 3GPP, *Vehicular Technology Conference (VTC)*, 2003.

[13] Yang Y., Yum T. Analysis of random access channel in UTRA-TDD on AWGN channel. *International Journal of Communication Systems*, 2004.

[14] Yang Y., Yum T. Analysis of Power Ramping Schemes for UTRA-FDD Random Access Channel. *IEEE Transactions on Wireless Communications*, 2005.

[15] Koo I., Shin S., Kim K., Performance Analysis of Random Access Channel in OFDMA Systems, *Systems Communications*, 2005.

[16] Zhou P., Hu H., Wang H., Chen H. An Efficient Random Access Scheme for OFDMA Systems with Implicit Message Transmission. *IEEE Transactions on Wireless Communications*, 2008.

[17] Lien S., Chen K. Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications. *IEEE Communications Magazine*, 2011.

[18] Lee K., Kim S., Yi B., Throughput Comparison of Random Access Methods for M2M Service over LTE Networks, *International Workshop on Machine-to-Machine Communications*, 2011.

[19] Cheng M., Lin G., Wei H. Overload Control for Machine-Type-Communications in LTE-Advanced System. *IEEE Communications Magazine*, 2012.

[20] 3GPP. Study on RAN Improvements for Machine-type Communications. *TR 37.868*, 2011.

[21] 3GPP. Evolved Universal Terrestrial Radio Access (EUTRA); Medium Access Control (MAC) protocol specification. *TS 36.321*, 2007.

[22] Johnson C. *Long Term Evolution In Bullets*. 2010.

[23] 3GPP. TSG RAN WG2. RACH intensity of Time Controlled Devices. *R2-102296*, 2010.

[24] Maeder A., Staehle D., Rost P. The Challenge of M2M Communications for the Cellular Radio Access Network, *11th Wurzburg Workshop on IP: Joint ITG and Euro-NF Workshop*, 2011.

[25] 3GPP. Feasibility study for Further Advancements for E-UTRA (LTE-Advanced). *TR 36.912*, 2011.

[26] Dohler M., Alonso-Zrate J., Watteyne T. Machine-to-Machine: An Emerging Communication Paradigm, *Wireless World Research Forum*, 2010.

[27] 3GPP. Evolved Universal Terrestrial Radio Access (EUTRA); Physical layer procedures. *TS 36.213*, 2011.

[28] Kleinrock L., Lam S. Packet-Switching in a Multi-Access Broadcast Channel: Performance Evaluation. *IEEE Transactions on Communications*, 1975.

[29] Sidi M., Segall A. Two Interfering Queues in Packet-Radio Networks. *IEEE Transactions on Communications*, 1983.

[30] 3GPP. Access barring for delay tolerant access in LTE. *TSG RAN WG2 Meeting 74. R2-113013*, 2011.

[31] 3GPP. Backoff enhancements for RAN overload control. *TSG RAN WG2 Meeting 73bis. R2-112863*, 2011.

[32] 3GPP. Study on provision of low-cost MTC UEs based on LTE. *TR 36.888*, 2012.

[33] Office for National Statistics, 2011 Census: Population and household estimates for England and Wales, published in 2012.

[34] Andreev S., Gonchukov P., Himayat N., Koucheryavy Y., Turlikov A. Energy efficient communications for future broadband cellular networks. *Computer Communications Journal (COMCOM)*, 2012.

[35] Cheng J., Lee C., Lin T. Prioritized Random Access with dynamic access barring for RAN overload in 3GPP LTE-A networks, *GLOBECOM Workshops*, 2011.

[36] 3GPP. Discussion on the UE behaviour when receiving the eWaitTime in LTE. *TSG RAN WG2 Meeting 73bis. R2-112202*, 2011.

[37] Elias Y., Zaher D. Uplink scheduling in LTE systems using distributed base stations, *European Transactions on Telecommunications*, 2010.