An Analytical Model for Evaluating the Interplay Between Capacity and Energy Efficiency in NB-IoT

Romain Barbau[‡§], Vincent Deslandes[‡], Gentian Jakllari[§] and André-Luc Beylot[§]
[‡]Airbus Defence and Space - France
surname.name@airbus.com
[§]IRIT-INPT/ENSEEIHT - France
surname.name@toulouse-inp.fr

***Abstract*—The Narrowband Internet Of Things, better known by the acronym NB-IoT, is spearheading 3GPP's effort to address the massive Machine Type Communication (mMTC) segment of 5G. With 103 deployments in more than 40 countries, NB-IoT's performance in terms of cell capacity and coverage is well established. However, the performance it can offer in terms of battery life, a foundational requirement of IoT, has received little attention.**

In this paper, we introduce the first theoretical model based on the M/D/H/K queues capable of evaluating the energy performance of NB-IoT in advanced cell scenarios defined by a coverage distribution, a payload uplink and downlink size, and an incoming communication rate. After using the values computed by 3GPP for single-terminal cells to demonstrate its accuracy, we use our model to evaluate the NB-IoT's energy performance. Our analysis reveals that a) Early Data Transmission, originally designed for improving the latency and connection density, can also significantly improve the energy performance, and b) When the load in multi-terminal cell settings increases, the terminal battery lifetime collapses. Leveraging our model, we introduce a solution requiring no changes to the standard and show that it significantly improves the terminal battery life.

I. Introduction

The recent surge in the number of Internet of Things (IoT) terminals [1] is underpinned by networks that provide better performance compared to traditional cellular technologies in terms of: a) cost, b) lifetime, c) coverage and d) capacity. The majority of these low-power wide-area networks (LPWANs) use the unlicensed ISM frequency bands [2], reducing operating cost but at the expense of Quality of Service (QoS) [3]. Without operator coordination, interference is inevitable and real questions arise as to the viability of these technologies as the number of terminals continues to grow [4]. Recognizing the value of developing a technology for IoT applications that require low latency and reliability, 3GPP standardized Narrowband IoT (NB-IoT) in Release 13. NB-IoT is well positioned to meet the IoT demands because it features a) simple modcods that reduce the cost of module manufacturing, b) power saving methods (eDRX and PSM), c) narrowband transmissions that improve coverage, and d) reduced network overhead [5], [6].

3GPP introduced in Release 15 a new optimization of the NB-IoT communication process: Early Data Transmission (EDT) [7], aimed at meeting the objectives related to the capacity of a cell [8]. EDT is a procedure that allows a terminal to send its uplink data directly after Random Access (RA), in the message 3 of the traditional procedure (RRC Connection Setup/Resume). The end of the communication can, therefore, be triggered as soon as message 4 [9].

Unfortunately, the objectives regarding the lifetime of the terminal (10 years with a capacity of 5Wh) are not met in certain scenarios, in particular in extreme coverage areas (164 dB coupling loss) and frequent transmissions [10], [11]. Part of the reason for this disappointing result is the common assumption that traditional networking performance metrics and energy efficiency can and should be evaluated separately. Nevertheless, it is straightforward to realize that there is a correlation between coverage conditions and energy consumption [12]. And on closer inspection, it becomes clear that cell load could also impact energy consumption. This is especially true for NB-IoT since, unlike other LPWAN protocols, it implements reliability. Studies have shown that when the cell load increases, it leads to higher collisions on the random access channel [13] and longer waiting times for resource allocation. The question, however, of how exactly these neworking mechansims and behaviour impact energy consumption remains open.

In this paper, we introduce the first analytical model based on the M/D/H/K queues capable of shedding light on the interplay between performance and energy consumption, especially under high-load scenarios. Our model is capable of taking as input an elaborate NB-IoT cell setting – defined by a coverage distribution, a payload uplink and downlink size, the ratio between random, control and data channels, and an incoming communication rate – and evaluating the lifetime of an NB-IoT terminal.

Our main contributions may be summarized as follows:

- We develop a model that computes the energy consumption of an NB-IoT communication and can account for three NB-IoT communication procedures: legacy, Resume and EDT.
- We introduce a new analytical approach for modeling the resource allocation on the different channels using deterministic service queues, bulk arrivals/departures and finite capacities (M/D/H/K). Our model is able to take into account the effect of the cell load through the calculation of waiting times during allocation, collision rates, congestion rates and the number of attempts of a communication.
- We leverage our analytical model to study the battery life as a function of the load on a cell using 3GPP defined coverage distributions. Our study paints a mixed picture of the interplay between performance and energy efficiency in

NB-IoT. On one hand EDT, while designed for improving performance, it also improve energy efficiency. On the other hand, the NB-IoT terminal energy efficiency collapses under high loads, in part due to mechanisms designed for improving performance.

- We introduce a simple solution that requires no changes to the standard and use our model to show its effectiveness at improving the energy efficiency.

The rest of the paper is organized in three parts. The first part, Section II, gives a global view on how the NB-IoT works, its energy consumption and the EDT procedure. In the second part, Section III, we introduce our analytical model. Finally, in Section IV, we analyze the effect of the EDT and the load on the energy consumption of NB-IoT.

II. Primer on NB-IoT

NB-IoT is organized in carriers of 180 kHz bandwidth, called Physical Resource Blocks (PRB), which can be stacked for scaling purposes. The interest of operating on such a small bandwidth is on one hand to allow its deployment as a replacement for GSM channels (200 KHz) or LTE PRBs (also 180 kHz) and on the other hand to reduce energy consumption.

In this section, we describe in detail the operation of a NB-IoT cell and the parameters we will use to model it later, and we explain the Early Data Transmission (EDT) procedure.

A. NB-IoT Architecture

1) **Uplink Channel:** The uplink is divided into two channels. The Narrowband Physical Random Access CHannel (**NPRACH**) offering periodic connection setup opportunities called RAOs for Random Access Opportunities. These RAOs are specific to a coverage area (Normal, Hard and Extreme) referred to as CE in the literature. Each RAO within a CE has an associated number of repetitions, a period and a number of opportunities (subcarriers), see fig. 1. When a user wishes to transmit, they expect a RAO corresponding to their CE. They then randomly choose a preamble index from the RAO, and collisions may occur.

The rest of the uplink resources is dedicated to the NPUSCH, which is a contention-free channel dedicated to sending data. Users only send messages there if the base station has specifically allocated resources to them. Resources are organized in Resource Units (RU) which can have several time/frequency forms illustrated in fig. 1. The more a terminal is in poor coverage, the more RUs with small bandwidths but long durations are allocated to it by the base station. When making an allocation, the base station specifies a modcod corresponding to a number of useful bits for each RU, a number of RUs (up to 10) and a number of repetitions to increase the coverage. This allocation represents a Transport Block (TB) that varies according to the terminal's coverage and payload. Depending on the shape of the TBs, a different number of terminals can transmit simultaneously without collision. For modeling purposes (§ III), we denote with H the number of terminals that can be served simultaneously during a block period d. Thus for the NPUSCH, H varies between 1 and 48 terminals and d between 1 and 40,960 ms.

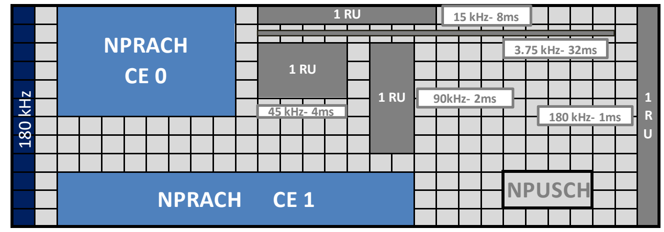

Fig. 1: Uplink Frame- Overview

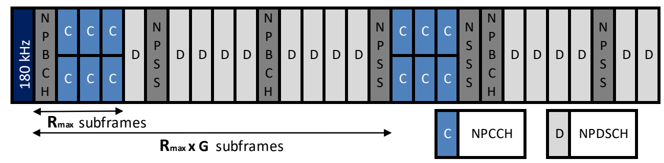

Fig. 2: Downlink Frame- Overview

2) **Downlink Channel:** The downlink is divided into subframes that last one millisecond across the entire bandwidth fig. 2. There are synchronization signals (NPSS and NSSS) that allow the terminals to demodulate the channels correctly. The NPSS occures one subframe out of ten while the NSSS occludes one out of twenty. The NPBCH occupies one subframe in ten and carries general information about the NB-IoT cell. The remaining 75% is shared between the NPDCCH control channel and the NPDSCH data channel.

This sharing is based on two parameters specific to each CE: R_{max} the number of NPDCCH subframes in a row and G the sharing ratio between NPDSCH and NPDCCH (see illustration in fig. 2). We have $G \in \{1.5, 2, 4, 8 \ldots, 64\}$ and $R_{max} \in \{1, 2, 4, 8 \ldots, 2048\}$. For the downlink, the duration of a downlink block is therefore $d = R_{max} \cdot G$. And H depends on the number of repetitions per user for the NPDCCH and the size of the TBs for the NPDSCH.

Finally, the size of the downlink block allows to know what is the maximum delay before allocating a block. The standard allows a delay of 10 downlink periods, and since the maximum delay that can be allocated in the Downlink Control Information (DCI) is 64 ms for uplink TBs and 128 ms for downlink TBs, we have the following timers:

$$\begin{cases} T_c = 10pp \\ T_u = 10pp + 64ms \\ T_s = 10pp + 128ms \end{cases} \quad (1)$$

These values are used to determine the size of the buffers in our model, see Equation (5).

B. Early Data Transmission Procedures

The Rel.13 of the 3GPP was accompanied by an optimization called "User Plane" (UP) allowing to shorten a connection and the optimization called "Control Plane" (CP) allowing to send data through the control plane without instantiating radio bearer user plane. These two optimizations have enabled LTE to achieve performances in terms of latency, capacity and lifetime that LTE had not been able to achieve. However, ITU expectations for the 5G IoT segment are higher than those of LTE IMT-2020. This is why 3GPP has developed an enhancement called "Early Data Transmission" (EDT) that allows to send useful data directly in messages 3 and 4 for uplink and downlink, respectively. EDT can be applied in UP optimization as CP. In the rest of this section we summarize the operation of the EDT as explained in the report [9].

1) **Message 1 et 2:** The first two messages of the communication procedure remain unchanged. The first message is a preamble sent by the terminal that notifies the base station that it wishes to communicate. This preamble must be sent on a dedicated RAO (NPBCH information). The second is the "Random Access Response" (RAR) which is a message from the base station that allocates a TB to the terminal so that it can transmit its message 3.

2) **Message 3:** For the UP-EDT, the terminal sends an RRC Connection Resume message allowing the base station to retrieve the security and radio bearer information previously established. It also attaches to this message 3 its uplink data.

For CP-EDT, the terminal sends the RRC Connection Request for carrying UL NAS-PDU. The terminal then uses the control plan to send data as in CP optimization. A way to retrieve the data from the MME is then necessary on the core network side.

3) **Message 4:** In both cases, the terminal can indicate in message 3 that it has no more data to send and wishes to return to idle mode after message 4. In the case of UP-EDT the base station will send an RRC Connection Release as message 4. For CP-EDT the base station will send an end of connection message. This message 4 is an opportunity for the network to send an application message, for example an acknowledgement of message 3.

In case the terminal wants to transmit more data or stay connected, the rest of the communication procedure is standard: RRC Connection Setup Complete for message 5, UL and DL data and then RRC Connection Release. We can deduce that if the payload of a terminal can be contained in a TB (<2536 bits in Rel.15) then only two round trips are necessary thanks to the EDT.

C. Questions regarding NB-IoT

Two important questions arise regarding NB-IoT and its latest amendments:

1) How does EDT, designed for improving performance, impact energy efficiency?
2) What is the interaction between performance, key focus of the NB-IoT design, and energy efficiency, key requirement of many IoT application in an NB-IoT cell?

III. Analytical Models

The objective of this paper is to allow the evaluation of the battery life of an NB-IoT terminal as a function of the cell parameters, the communication procedure, the distribution of the coverage conditions and the load. To this end we have modeled the steps that the terminal goes through from an energy point of view in the fig. 3. Basically, we distinguish three types of energy usage:

Obligatory : Constant and unavoidable.
Coverage : Depending on coverage conditions.

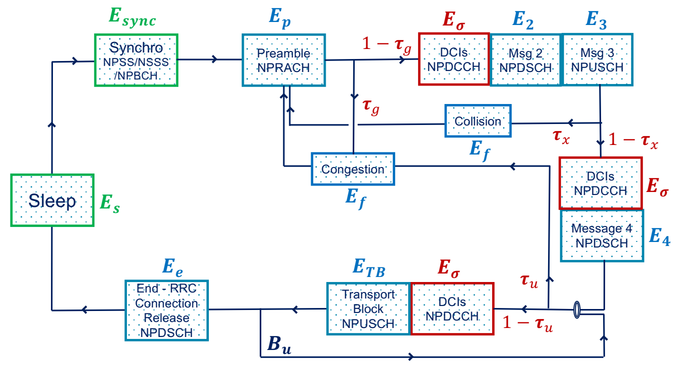

Fig. 3: Model for energy consumption evaluation - Overview.

Load : Depending on the cell load.

In the first part of this section we will briefly explain how we take into account in our model the first two types of energy consumption that have been widely studied in the literature [12], [14]. However, energy consumption as a function of cell load has not yet been studied to the best of our knowledge. The third type of energy consumption will be detailed in a second part. Finally, in a last part, we will determine the collision rates τ_x and the congestion rates τ_g necessary to calculate the energy consumption when considering the load in the cell.

At the end we will be able to calculate the battery life eq. (21) through the determination of each energy component that takes part in a communication attempt eq. (2) derived from the model of Figure 3:

$$
\begin{aligned}
E_a &= E_{sync} + E_p + \tau_g \cdot E_f + \\
&\quad + (1-\tau_g)(E_\sigma + E_2 + E_3 + \tau_x \cdot E_f + (1-\tau_x)E_\Delta) \\
E_\Delta &= E_\sigma + E_4 + \sum_{j=0}^{B_u-1} \left[(1-\tau_u)^j (\tau_u) \cdot (j(E_\sigma + E_{TB}) + E_f) \right] \\
&\quad + B_u (1-\tau_u)^{B_u} (E_\sigma + E_{TB}) + E_e
\end{aligned} \tag{2}
$$

A. Mandatory and coverage-related energy consumption

First there are the mandatory components that are outside the communication procedure. Between two communications the terminal is asleep and expends an energy E_s equal to the multiplication of the time interval between the communications and the power used in this state of the order of μW (deep sleep of table III). Next, before sending the preamble, the terminals must re-synchronize with the cell in time and frequency, and retrieve the information contained in the broadcast channel (MIB) [15]. The energy consumed during synchronization E_{sync} is the multiplication of the time needed for synchronization table I and the power needed for reception P_{Rx} table III.

Finally, there is the energy consumption on the different channels related to the terminal coverage, i.e. the transmission of preambles on the NPRACH, the transmission of TBs on the NPUSCH, the reception of subframes on the NPDSCH and the demodulation of the NPDCCH in case of failure. Not all terminals consume the same amount of energy to perform these same tasks. Depending on the coverage conditions: the number of repetitions, the quantity of RUs/SFs (respectively for the NPUSCH and the NPDSCH) and the modcods are adjusted. E_p, $E_{2/3/4}, E_{TB}$ and E_f are computed by multiplying the duration of the received or sent message and the power used in reception or transmission table III. When there is a collision or congestion the terminal demodulates the NPDCCH for the duration of the timer (§ II-A). It only demodulates the downlink control frames, so it is only active a fraction of the time equal to $\frac{1}{G}$ and is in light sleep mode the rest of the time:

$$E_f = T_f \left(P_{Rx} \cdot \frac{1}{G} + P_{ls} \cdot \frac{G-1}{G} \right) \tag{3}$$

Finally, the nominal transmission power of an NB-IoT terminal is 23dBm, but it is drastically lowered for terminals with good coverage [12], [16]. By noting P the power with which the terminal must transmit to reach a block error rate of 10^{-1}, the total power P_{Tx} spent by the terminal during transmission will be calculated in the model as follows:

$$P_{Tx} = \frac{P}{\eta} + P_c \tag{4}$$

with η the power amplifier efficiency and P_c the power needed for support circuity.

B. Load-related energy consumption

Adding more load to the cell will result in a) longer allocation wait times and b) having to make multiple communication attempts.

1) Computation of allocation times: A weakness of the connected mode is that one has to wait for a resource allocation to know when to transmit. Since the terminal does not know when the resource allocation will arrive, it must demodulate all NPDCCH opportunities. Since in the NB-IoT the downlink band to be demodulated is generally wider than the one used in uplink and since the 3GPP waveform is complex, the energy spent in reception is not negligible [15], [17]. It is expected that this problem of active waiting will be exacerbated as the load in the cell increases.

Principle of resource allocation: On a channel the station serves a number of blocks, H, periodically, see § II. If the station has to allocate more than H blocks, it keeps in memory the surplus of blocks in order to allocate them over the next period, thus they are delayed. When the load increases, the station's response time also increases. Beyond a certain timer, T, the UE gives up to try again in the random access after a back-off (new attempt), leading to congestion. Since service times are deterministic, one can know when a request for a block arrives if it will be served on time or not. This is like limiting the buffer to a size K, see Figure 4.

This description of how channels work corresponds to M/D/H/K queues with bulk starts and arrivals.

Principle of the MDHK queues: We considered the embedded Markov Chain at departure epochs. At each allocation period of duration d, H users at most are served by the queue (H servers). When new users arrive in the queue, they are added to those already present if the size of the buffer allows it. To model the abandonment, we simply take a buffer size, K, such that:

$$K = H \cdot \frac{T}{d} \tag{5}$$

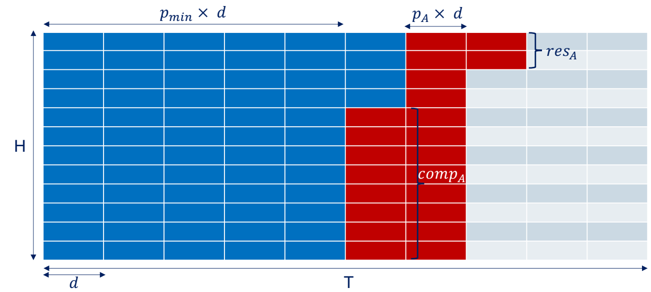

Fig. 4: Visualisation of allocation process - MDHK queue

with T the duration of the abandonment timer eq. (1).

Transition matrix and stationary state of the MDHK queue:

$$M = \begin{bmatrix} a_0 & a_1 & \dots & a_H & a_{H+1} & \dots & a_{K-1} & b_K \\ a_0 & a_1 & \dots & a_H & a_{H+1} & \dots & a_{K-1} & b_K \\ \vdots & & & & & & & \vdots \\ a_0 & a_1 & \dots & a_H & a_{H+1} & \dots & a_{K-1} & b_K \\ 0 & a_0 & \dots & a_{H-1} & a_H & \ddots & a_{K-2} & b_{K-1} \\ \vdots & & \ddots & & \ddots & & \ddots & \\ 0 & 0 & \dots & 0 & 0 & \dots & a_{H-1} & b_H \end{bmatrix}$$

With a_n the probability that n users arrive in the queue during a period d and $b_K = \sum_{k=K}^{+\infty} a_k$.

We can then numerically compute the stationary distribution Π of the included Markov chain, the vector noted $\Pi = (\pi_0, \pi_1, \dots, \pi_K)$. With π_j the probability that there are j users in the queue.

Determination of waiting times: We will reason in terms of the periods of the MDHK queue, depicted in the Figure 4. For the users that will be served, the waiting time T_σ depends on the number A of new users arriving during the last period and the number S of users in the queue before the new users arrive. New users will first have to wait for a minimum number of period p_{min} equals to the number of full services that can be performed with existing users. Some new arrivals c_A will then be served in the same period as the last of the users already present. Finally the base station will fill a number p_A of complete periods of newcomers up to a last period partly allocated to the r_A last arriving users. For $A = i$ and $S = j$ we obtain the following results:

$$\begin{aligned} p_{min} &= \left\lfloor \frac{j}{H} \right\rfloor & c_A &= \min\left[H - (j - p_{min}H),\ i\right] \\ p_A &= \left\lfloor \frac{i - c_A}{H} \right\rfloor & r_A &= i - p_A \cdot H - c_A \end{aligned} \tag{6}$$

We can now compute the number of users waiting k periods:

$$N_{k_{i,j}} = \begin{cases} 0 & \text{for } k < p_{min} \\ c_A & \text{for } k = p_{min} \\ H & \text{for } p_{min} < k \leq p_{min} + p_A \\ r_A & \text{for } k = p_{min} + p_A + 1 \\ 0 & \text{for } k > p_{min} + p_A + 1 \end{cases} \quad (7)$$

The total number of periods waited by the users is therefore:

$$W_{i,j} = \sum_{k=0}^{p_{max}} k \cdot N_{k_{i,j}}$$

With the maximum number of waiting period $p_{max} = \frac{T}{d}$. Let i_{max} denote the maximum number of new clients that can be served (for simplicity we omit the dependence on j in the notations):

$$i_{max} = K - (j - H) \quad (8)$$

Finally we can compute the number of periods that a served user waits on average, $\mathbb{E}(W)$:

$$\mathbb{E}(W) = \frac{1}{\mathbb{E}(A)} \sum_{j=0}^{K} \left[\sum_{i=1}^{i_{max}-1} W_{i,j} \cdot a_i + W_{i_{max},j} \cdot b_{i_{max}} \right] \cdot \pi_j$$

$$\mathbb{E}(A) = \sum_{j=0}^{K} \left[\sum_{i=1}^{i_{max}-1} i \cdot a_i + i_{max} \cdot b_{i_{max}} \right] \cdot \pi_j \quad (9)$$

, with $\mathbb{E}(A)$ the average number of users admitted per period. Thus we have the average time a user spends waiting for the resource allocation when served:

$$T_\sigma = \mathbb{E}(W) \cdot d \quad (10)$$

The energy consumption due to waiting for the allocation is:

$$E_\sigma = T_\sigma \left(P_{Rx} \cdot \frac{1}{G} + P_{ls} \cdot \frac{G-1}{G} \right) \quad (11)$$

2) Collision rates, congestion rates, and number of attempts: An attempt can fail for two reasons: either because of a collision in the first phase (random access) or because of congestion in the second phase (connected). When an attempt fails, the terminal will restart the procedure after a randomly selected period of time (backoff time). After a number of N failed attempts, the communication is considered to have failed. This behavior is modeled in fig. 5. In this part we compute τ_x, τ_g and N_a the probabilities of collision, congestion and the number of attempts of the flows in a cell in order to solve the energy computations introduced above.

Collision probability: In the model, we consider a communication flow arriving in the cell consisting of a multitude of Poisson flows, each characterized by a unique tuple (coverage area, blocks) denoted (k, b). The coverage area represents the CE in which the terminal is located based on its coupling loss (Table I). The blocks to be sent by the terminal are composed of three numbers corresponding to the contention-free channels NPUSCH, NPDSCH and NPDCCH: we have $b = (B_u, B_d, B_c)$. These numbers depend on the communication procedure, the size of the message to send on the channels and the coverage.

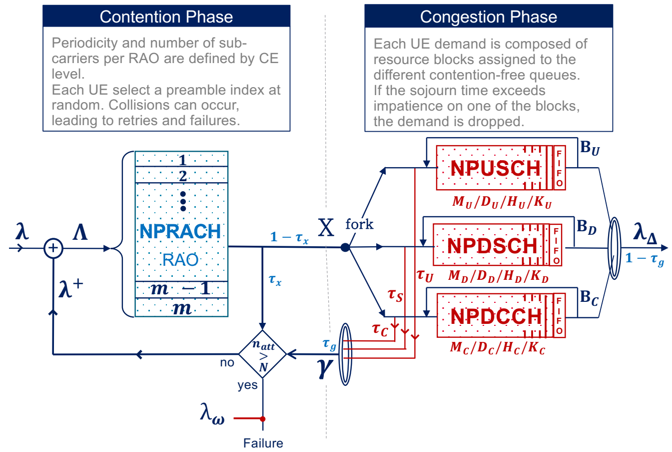

Fig. 5: Model for the determination of load variables - Overview.

Collisions between users occur in NPRACH when sending preambles that initialize the connection setup. The total flow, Λ, arriving on the NPRACH is an aggregated flow between new arrivals in the system λ and those whose previous communication attempt failed λ^+. Thus, the flow arriving on the RAO of the CE n°k is:

$$\Lambda_k = \sum_b \left(\lambda_{k,b} + \lambda^+_{k,b} \right)$$

In our model the arrivals on the NPRACH are assumed to be Poisson, an approximation introduced by Abramson [18]. The probability of collision involving the CE n°k, knowing that the period of the RAOs is δ_k and the number of subcarriers is m_k, is given by:

$$\tau_{x_k} = 1 - e^{-\frac{\Lambda_k \cdot \delta_k}{m_k}} \quad (12)$$

Congestion rate: Every non-colliding user access to the contention-free channels, the arrival flow in these is therefore $X_{k,b} = (1 - \tau_{x_k})\Lambda_{k,b}$. Congestion must be computed for the three channels. A communication will need B_q successful successive allocation on each channel, $B_q \in (B_u, B_c, B_d)$ respectively see fig. 5. Three MDHK queues whose parameters depend on the channel are thus considered. For each queue q, we apply the following analysis.

Recall that the maximum number of users that can be served among new arrivals is i_{max} (Equation (8)). The congestion rate for a block, τ_q, is given by Equation (13) where ρ_q is the queue load.

$$\tau_q = 1 - \frac{\sum_{j=0}^{K_q} \left(\sum_{i=1}^{n-1} i \cdot a_i + i_{max} \cdot b_{i_{max}} \right) \cdot \pi_j}{\rho_q \cdot H_q} \quad (13)$$

Should a congestion occur on one of the blocks of a communication, the communication is abandoned. Thus, the load actually processed in the queues is computed as a fixed point of:

$$\rho_q = \frac{d_q}{H_q} \cdot \sum_{k,b} \left(X_{k,b} \cdot \frac{1 - (1 - \tau_q)^{B_q}}{\tau_q} \right) \quad (14)$$

Π, τ_q and ρ_q are calculated using an iterative method. We can then compute the congestion rate of a communication on one of the channels - the probability that the B_q blocks have been successfully served. Note that, congestion in one of the queues is enough to cause a communication failure. Thus, we have $\tau_{g_{k,b}}$, the congestion rate of the flow (k, b), equals to:

$$\tau_{g_{k,b}} = \prod_q \left(1 - (1 - \tau_q)^{B_q}\right) \tag{15}$$

Number of attempts and Failure probability: In this section, we compute the average number of attempts per flow, which is necessary to compute the total energy consumed. Recall that, if a communication demand fails, either due to collision or congestion, the UE will retry after a randomly selected interval. The UE abandons if it exceeds the maximum number of retries, N. The failure probability for a single attempt, $\mathcal{P}_{f_{k,b}}$, concerning the flow (b,k) is the complement of the joint probability of the communication succeeding through the contention phase and the congestion phase:

$$\mathcal{P}_{f_{k,b}} = 1 - (1 - \tau_{x_k}) \cdot \left(1 - \tau_{g_{k,b}}\right) \tag{16}$$

The failure probability concerning communications, $\mathcal{P}_{f_{k,b}}$, is the probability that a communication fails N attempts in a row and determine the successful throughput $\lambda_{\Delta_{k,b}}$:

$$\begin{aligned} \mathcal{P}_{\omega_{k,b}} &= (\mathcal{P}_{f_{k,b}})^N \\ \lambda_{\Delta_{k,b}} &= \lambda_{k,b}(1 - \mathcal{P}_{\omega_{k,b}}) \end{aligned} \tag{17}$$

A key aspect of the model is that, due to the retries, the load on the random access, Λ, is higher than the input charge, λ. To calculate Λ, it is necessary to determine the average number of attempts per communication computed as follows:

$$\begin{aligned} N_{a_{k,b}} &= \left(1 - \mathcal{P}_{f_{k,b}}\right) \sum_{j=0}^{N} (j+1)\, \mathcal{P}^j_{f_{k,b}} + N(\mathcal{P}_{f_{k,b}})^N \\ &= \frac{1 - \mathcal{P}^N_{f_{k,b}}}{1 - \mathcal{P}_{f_{k,b}}} \end{aligned} \tag{18}$$

Finally, we can now compute the communication flow arriving into NPRACH by solving the following recursive equation:

$$\Lambda = \sum_{k,b} \lambda_{k,b} \cdot N_{a_{k,b}} \tag{19}$$

Every time Λ is updated, collision and congestion rates can be computed again to output new values for P_ω and N_a. When the difference between two successive values of Λ is small enough, we consider that the model has converged to a steady state producing the necessary parameters to evaluate the energy consumption of an attempt in a cell (Equation (2)). The energy consumed by communication is:

$$E_c = N_a \cdot E_a \tag{20}$$

Finally, we consider the natural battery discharge during the constant power time P_{ds} (see Table III). Thus, for a given battery capacity, C, in Wh, the lifetime in years of an NB-IoT terminal is given by:

$$L = \frac{C}{P_{ds} + \frac{E_c}{t}} \cdot \frac{1}{24 \times 365.25} \tag{21}$$

with t the time interval (in hours) between two communications.

Cell Parameters used in the Models				
Cell Bandwidth		180 kHz for both UL and DL		
Deployment Mode		Standalone		
CE Levels		CE 0	CE 1	CE 2
Coupling Loss		144 dB	154 dB	164 dB
Repetitions	NPRACH	2	8	32
	NPDCCH	1	8	128
	NPUSCH	1	2	16
	NPDSCH	1	16	128
UL Tx Bandwidth		45 kHz	15 kHz	15 kHz
$\mathbf{I_{MCS}}$	NPUSCH	7	4	2
	NPDSCH	4	6	4
G		4		
$\mathbf{R_{max}}$		8	8	128
Synchro 90^{th} percentile	NPSS/NSSS	154 ms	164 ms	354 ms
	MIB-NB	10 ms	20 ms	240 ms

TABLE I: Cell parameters according to [11], [19].

IV. PERFORMANCE EVALUATION AND ANALYSIS

The performance evaluation is organized in three parts. The first part is aimed at validating our model while the second part is aimed at addressing the questions raised in § II-C regarding NB-IoT. Finally, we analyze a solution for improving the energy performance of NB-IoT under high loads.

A. Methodology, Settings and Validation

1) **Methodology:** We detail below how to obtain the lifetime distribution in an NB-IoT cell using our model.

In a cell, we consider a distribution of terminals with different coverage levels and payload sizes. For a given coverage and payload size, we deduce flows and their characteristics (k,b) as well as those of the channels (H, d and T) (§ II). Using this information, we use our analytical model to calculate for any incoming aggregated flow λ (expressed in communications per second) the collision rates τ_x, congestion rates τ_g and communication failure probabilities P_ω. In addition, we use Equation (2) to compute the energy consumption of each flow in an NB-IoT cell based on its characteristics, τ_x and τ_g.

Finally, based on the time interval between the communications and the battery capacity, the battery life is computed for each terminal of the cell Equation (21).

2) **Settings:** Table I summarizes the general parameters of an NB-IoT cell. For the configuration of NPRACH and NPDCCH we use the information in [11], for R_{max} and G the information in [19] and finally, for the synchronization times on the values cited in [9]. The modcods and the number of repetitions used for the NPUSCH and NPDSCH are redefined for each cell stream according to the coverage conditions of the distribution provided by 3GPP for mMTC segment of 5G [8] - we have selected configuration B of the Macro-Urban scenario. The values indicated in the table I are for specific coverage cases (144 dB, 154 dB and 164 dB coupling loss) which correspond to those used for the model validation.

Note that the energy spent in a state by an NB-IoT module depends on the manufacturers and operators [16], [12]. Therefore, in an effort to provide results applicable to the widest range of settings, we use the values provided by the 3GPP in Release 13 [9] and summarized in Table III.

3) **Validation:** Table II compares results from a 3GPP study [7], [9] and our own results regarding battery life, for two traffic

		3GPP results [7] (years)		Our results (years)	
Trafic Models [7]	**CL**	**Resume**	**UP-EDT**	**Resume**	**UP-EDT**
Scenario 1	144 dB	36.8	36.4	34.7	36.0
	164 dB	9.2	12.0	7.8	12.6
Scenario 2	144 dB	29.7	31.4	26.9	30.9
	164 dB	2.6	3.8	2.0	3.7

TABLE II: Battery life comparison - single-terminal case.

Device Power Consumption		
$\mathbf{P_{Tx}}$ (P = 23 dBm) with η = 45%	Preamble	500 mW (incl. 60mW support circuity)
	Data	
$\mathbf{P_{Rx}}$	Control	80 mW
	Data	
	Synchro	
	Broadcast	
Sleep	Light - $\mathbf{P_{ls}}$	3 mW
	Deep - $\mathbf{P_{ds}}$	0.015 mW

TABLE III: Power Consumption according to Release 13 [9].

models. It considers a single terminal per cell, so we call it the single-terminal case.

The data shows that the results of our model are broadly in line with those obtained by 3GPP. In particular, in the case of EDT, for which we have exact overhead values [9], we observe only minor differences. Our evaluation of the Resume procedure is more pessimistic, which can be explained by a difference in the choice of the size of the network overhead.

B. NB-IoT Energy Questions

In this section, we analyse some of the fundamental questions regarding energy consumption in NB-IoT.

1) **Question 1:** *What is the impact of EDT on battery life?*

We consider terminals in each of the three coverage areas defined by the standard: Good (144 CL), Hard (154 CL) and Extreme (164 CL). The traffics considered are those defined by the 3GPP regarding the mMTC segment of the 5G : 50 bytes or 200 bytes to be transmitted for uplink data every 2 hours or 24h with an acknowledgement of 29 bytes in the DL (inc. IP header compression). The objective of 3GPP is to obtain a lifetime of more than 10 years on all 3 scenarios.

Findings: Table IV summarizes the results obtained using our model. The first finding is that EDT solutions use less energy than the Resume solution and that the difference

Battery Life (Years)					
Reporting Interval (Hours)		2		24	
DL packet size (bytes)		29			
UL packet size (bytes)		50	200	50	200
144 dB CL	**Resume**	27.0	19.1	36.8	35.1
	UP-EDT	30.0	25.4	37.2	36.5
	CP-EDT	30.0	25.4	37.2	36.5
154 dB CL	**Resume**	13.7	8.7	33.1	29.7
	UP-EDT	18.4	8.5	34.9	29.5
	CP-EDT	18.6	8.5	35	29.5
164 dB CL	**Resume**	2.1	1.0	15.8	9.1
	UP-EDT	4.0	1.4	22.1	11.8
	CP-EDT	3.9	1.4	22.1	11.8

TABLE IV: Battery life depending on coverage conditions and communication procedure.

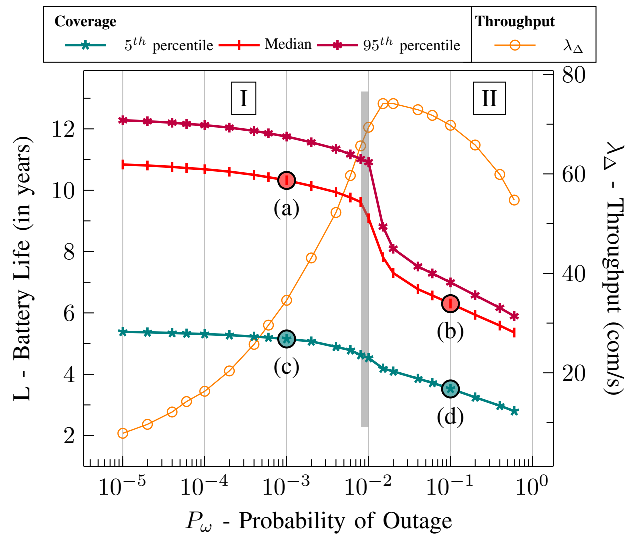

Fig. 6: Battery life and throughput as a function of cell load. Traffic scenario: 200-byte payload, uplink, 15-minute interval, and using CP-EDT procedure.

between the two EDT solutions is minimal. While EDT has been designed primarily to achieve low latency and high connection density, we show that it is also energy efficient.

The second finding is that the NB-IoT energy performance under extreme coverage conditions is mixed. It meets the 10-year lifetime objective for applications with 24-hour reporting interval, with EDT being the enabler, but it fails to do so for applications with a two-hour reporting interval (200-byte case). As a result, additional work and mechanisms are necessary.

2) **Question 2:** *How does traffic load impact energy performance?*

Inspired from the 3GPP reports, so far we have considered single-terminal cell cases. However, as detailed § III, when the cell load increases, the collision and congestion rates will increase, which may have an impact on the terminals' energy consumption. Therefore, in this section, we use our model to shed light on NB-IoT's energy performance in multi-terminal and, by consequence, high-load scenarios.

Findings: Figure 6 shows the battery life and capacity according to the cell load (considered cell described in § IV-A).

First of all we notice that the battery life is reduced when the load increases and therefore the values calculated when the cell is empty in table IV can be questioned in a real scenario. It would be welcome to specify the load in the NB-IoT battery life assessments.

Figure 6 points to two operation regimes with distinct performance:

(I) The new attempts are essentially due to collisions. When we increase P_ω both the cell capacity and the collisions increase, decreasing the battery life.

(II) As the load increases the data channels become congested, leading to a drastic increase in power consumption. To identify the underlying reasons for this behaviour, in Figure 7 we show a fine-grained energy analysis in NB-IoT. Unexpectedly, we discover that the drastic increase of power consumption under high loads is due to the demodulation

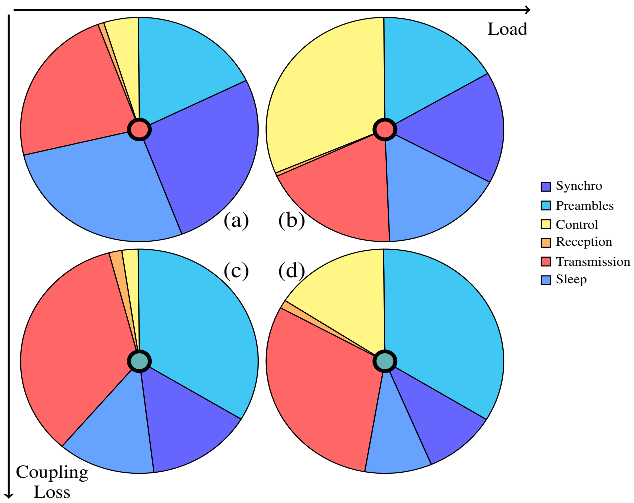

Fig. 7: Distribution of energy consumption for different scenarios. The different components and their calculations are described in the model $\mathcal{M}_2$ § III.

Median CL	G	2	4	8
$\mathbf{P}_\omega = \mathbf{10^{-3}}$	λ_Δ	34.37	34.37	34.37
	L	28.30	28.52	28.63
$\mathbf{P}_\omega = \mathbf{10^{-2}}$	λ_Δ	48.75	71.72	44.50
	L	25.25	27.23	27.48
$\mathbf{P}_\omega = \mathbf{10^{-1}}$	λ_Δ	50.05	77.70	45.01
	L	23.65	23.77	24.42

TABLE V: Throughput - λ_Δ - in com/s and Battery Life - L - in years as function of the cell load and the configuration of the control channel, G.

of the control channel. A careful analysis of NB-IoT reveals that, as the load increases and its buffers fill up, the base station takes increasingly longer to allocate resources. In the meantime, the terminal still has to demodulate all NPDCCH frames, unnecessarily increasing its energy consumption.

Figure 7, (a) vs (c), shows that this phenomenon is present regardless of the terminal's coverage, even if it has a relatively lesser impact on terminals with poor coverage since every transmission is expensive in terms of energy.

In the next two sections, we use our model to explore approaches, which do not require modification to the standard, for reducing energy consumption in high-load scenarios.

3) **Question 3:** *Control channel configuration and energy consumption under high loads.*

As shown by Equation (11) (§ III-B), the G parameter dictating the split between control and data channel in the downlink plays a role in the energy consumption. Therefore, we explore the possibility of reducing energy consumption using this parameter.

Findings: Table V shows the performance in terms of lifetime and capacity for different configurations. When we increase G we space the subframes of the control channel. Thus, for a given waiting time, the terminal demodulates fewer control subframes before allocating the resource, which, intuitively, should reduce energy consumption. The data, however, shows a marginal gain in battery life. This is due to the fact that when the buffers are full, the waiting time is close to the maximum (10 NPDCCH periods). Regardless of the spacing between NPDCCH opportunities, the terminals will demodulate

$\mathbf{P}_\omega$	$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$	$\mathbf{10^{-1}}$
Relative energy saved per communication with M=4	2.7 %	6.3 %	18.4 %
Battery Life with M=4 in years	10.53	9.54	7.45
Battery Life classic in years	10.32	9.08	6.30
Additionnal Latency - l	48 ms		

TABLE VI: Results obtained implementing our proposition which reduce the energy consumption when load increases. Results showed for the median coupling loss.

approximately 10 subframes.

As a result, the control channel configuration is not a feasible solution to reduce energy consumption in high-load scenarios.

C. Reducing energy consumption under heavy loads

Our analysis so far shows that for NB-IoT to meet 3GPP's lifetime target of 10 years it needs to address the challenge of preventing the terminals from consuming energy in the control channel when the cell load increases.

To address this challenge, we introduce a scheme that does not require changes to the standard. According to our scheme, a terminal does not demodulate all the control frames – the default behaviour leading to the battery lifetime collapse – but only one out of M, with M an input parameter. While the idea is intuitively simple, the challenge is implementing it beginning with message 2. The difficulty stems from the fact that the base station needs to know which NPDCCH opportunities the terminal will demodulate despite having no information about it. We solve this problem by leveraging k, the subcarrier index of the preamble – one of the only pieces of information the base station and the terminal have in their possession at that time. The base station and the terminal can agree that the opportunities to be demodulated will be $k + nM$ with $n \in \mathbb{N}$. The M parameter can be broadcast in the NPBCH, so the terminal will know it before initiating the communication. As a result, the equations for computing the power consumption of the terminal during demodulation of the NPBCH (Equation (11)) become:

$$E_f = T_f \cdot P_{Rx} \left(P_{Rx} \cdot \frac{1}{M \cdot G} + P_{ls} \cdot \frac{M \cdot G - 1}{M \cdot G} \right)$$
$$E_\sigma = T_\sigma \left(P_{Rx} \cdot \frac{1}{M \cdot G} + P_{ls} \cdot \frac{M \cdot G - 1}{M \cdot G} \right) \quad (22)$$

1) **Analysis of our scheme:** Table VI shows the performance of our scheme in the scenario of Figures 6 and 7 for $M = 4$. The data shows that as the load increases so does the effectiveness of our solution at reducing energy consumption. The advantage of our solution compared to the one studied previously (§ IV-B3) is that it does not modify the capacity of the cell since the quantity of NPDCCH resources remains unchanged. On the other hand, our solution will impact the communication delay since each allocation will be delayed. We have computed the average delay induced by our solution in the case of median coverage (CE 0) in the table VI.

Our solution addresses a power consumption problem identified in this work when the cell is under load. This proposal is all the more interesting since the cell parameters already present have proved to be inefficient § IV-B3.

V. Related Work

The model presented by Andres-Maldonado et al. [20] allows to evaluate the energy consumption of a NB-IoT terminal. The strength of this model is to be extremely precise in the definition of the cell parameters: scheduling gaps, short/initial synchronizations, and especially those of the eDRX. We have shown in this paper that the demodulation of control frames is a power-hungry mechanism, the problem of considering eDRX in the context of sporadic uplink data transmission is that this mechanism may drive the power consumption. Therefore we decided to consider a terminal that goes into sleep mode after the end of the communication as the standard allows (see § II). The authors of [21] have developed an energy model that works for both NB-IoT and LTE-M. The evaluation metrics compared to lifetime are interesting (latency, spectral efficiency, data rates), unfortunately the generality of the model limits its ability to be reused to evaluate new communication procedures (such as EDT) or other optimizations of the standard. In the two models mentioned, neither takes into account the load of the cell when estimating the energy consumption.

Energy experiments on real equipment have been conducted by Yang et al. [16] which have the particularity of showing that in practice energy consumption varies from one terminal to another and from one operator to another. In particular, they show that the Msg 3 repetition parameters are particularly overestimated to allow terminals with very poor coverage to be received correctly. They propose to reduce this number of repetitions in order to reduce on average the energy consumed in CE2. The drawback of this technique is that it reduces the BLER of the Msg3, which has the effect of triggering a new attempt and thus increasing the load on the NPRACH. Since the battery itself is negatively impacted by this load (§ IV), the proposed method may be counter-productive. In the same spirit of field experience, the authors of [12] investigate the energy consumption of COTS NB-IoT equipment. They find that the transmission power varies from one CE to another and from one operator to another depending on whether they favour the capacity or the lifetime of the terminal. In our model we have varied the transmission power according to the margin with respect to the BLER targeted eq. (4).

Finally, 3GPP's experts detail in [9] the EDT procedure for the control and user plans. They explain how energy consumption can be calculated without taking into account the cell load. In the article [7] the results are presented for three specific coupling loss values (144dB, 154dB and 164dB). Our model allows to take a coupling loss distribution which is necessary in the case of a study with load. Their work is a reference base for our model and allowed us to validate our results when the cell is empty.

VI. Conclusion

The NB-IoT is a complex protocol that introduces numerous mechanisms to accommodate the vast majority of 5G scenarios for the mMTC segment. Our work responds to the necessity of having a modeling tool reflecting this complexity and these new mechanisms so as to evaluate the performance of NB-IoT. First we show that the EDT mechanism, which has been designed to improve the capacity and latency of a cell, also has significant energetic virtues. However, we have shown that the cell load has a significant adverse effect on the battery life. Using our model, we first identified the source of the increase in energy consumption when the load increases: the demodulation of the control channel. Then we proposed a solution to increase the lifetime of the terminal and evaluated it with the help of our model. Our solution does not require any modification of the communication modules and can be quickly adopted by operators.

Our results show the capability of our model to evaluate complex scenarios and its versatility with regards to the implementation of new mechanisms.

References

[1] ITU, "The general development trends of iot," 2017.

[2] K. Mekki, E. Bajic, F. Chaxel, and F. Meyer, "A comparative study of lpwan technologies for large-scale iot deployment," *ICT Express*, vol. 5, no. 1, pp. 1–7, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405959517302953

[3] A. Dvornikov, P. Abramov, S. Efremov, and L. Voskov, "Qos metrics measurement in long range iot networks," in *2017 IEEE 19th Conference on Business Informatics (CBI)*, vol. 02, 2017, pp. 15–20.

[4] K. Mikhaylov, J. Petäjäjärvi, and J. Janhunen, "On lorawan scalability: Empirical evaluation of susceptibility to inter-network interference," in *2017 European Conference on Networks and Communications (EuCNC)*, 2017, pp. 1–6.

[5] R. Ratasuk, N. Mangalvedhe, Y. Zhang, M. Robert, and J. Koskinen, "Overview of narrowband iot in lte rel-13," in *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2016, pp. 1–7.

[6] A. Hoglund, X. Lin, O. Liberg, A. Behravan, E. A. Yavuz, M. Van Der Zee, Y. Sui, T. Tirronen, A. Ratilainen, and D. Eriksson, "Overview of 3gpp release 14 enhanced nb-iot," *IEEE Network*, vol. 31, no. 6, pp. 16–22, 2017.

[7] A. Hoglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3gpp release 15 early data transmission," *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 90–96, 2018.

[8] 3GPP. Connectiondensity - urbanmacro-mmtc. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/37_series/37.910/

[9] Ericsson, "Evaluation for early data transmissions - r2-1713058," 2017.

[10] 3GPP, "Cellular system support for ultra-low complexity and low throughput internet of things (ciot)," vol. Technical Report 45.820, no. v13.1.0, November 2015.

[11] O. Liberg, *Cellular Internet of Things*. Elsevier Aca- demic Press, 2017.

[12] F. Michelinakis, A. S. Al-Selwi, M. Capuzzo, A. Zanella, K. Mahmood, and A. Elmokashfi, "Dissecting energy consumption of nb-iot devices empirically," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 1224–1242, 2021.

[13] B. G. Martiradonna S, Piro G, "On the evaluation of the nb-iot random access procedure in monitoring infrastructures," 2019.

[14] M. Lauridsen, R. Krigslund, M. Rohr, and G. Madueno, "An empirical nb-iot power consumption model for battery lifetime estimation," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–5.

[15] M. Lukic, S. Sobot, I. Mezei, D. Vukobratovic, and D. Danilovic, "In-depth real-world evaluation of nb-iot module energy consumption," in *2020 IEEE International Conference on Smart Internet of Things (SmartIoT)*, 2020, pp. 261–265.

[16] D. Yang, X. Zhang, X. Huang, L. Shen, J. Huang, X. Chang, and G. Xing, "Understanding power consumption of nb-iot in the wild: Tool and large-scale measurement," 2020.

[17] Actility, "Market segmentation for lpwan technologies," 2019.

[18] N. Abramson, "The aloha system—another alternative for computer communications," 1970.

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

[20] P. Andres-Maldonado, M. Lauridsen, P. Ameigeiras, and J. M. Lopez-Soler, "Analytical modeling and experimental validation of nb-iot device energy consumption," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5691–5701, 2019.

[21] P. Jörke, R. Falkenberg, and C. Wietfeld, "Power consumption analysis of nb-iot and emtc in challenging smart city environments," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.