On the Latency-Energy Performance of NB-IoT Systems in Providing Wide-Area IoT Connectivity

Azari, Amin; Stefanovic, Cedomir; Popovski, Petar; Cavdar, Cicek

Published in:
IEEE Transactions on Green Communications and Networking

DOI (link to publication from Publisher):
10.1109/TGCN.2019.2948591

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

On the Latency-Energy Performance of NB-IoT Systems in Providing Wide-Area IoT Connectivity

Amin Azari*, Čedomir Stefanović+, Petar Popovski+, and Cicek Cavdar*

*KTH Royal Institute of Technology, +Aalborg University

Email: {aazari,cavdar}@kth.se, {cs,petarp}@es.aau.dk

Abstract

Narrowband Internet-of-Things (NB-IoT) offers a significant link budget improvement in comparison with the legacy networks by introducing different coverage classes, allowing repeated transmissions, and tuning the repetition order based on the path-loss in communications. However, those repetitions necessarily increase energy consumption and latency in the NB-IoT system. The extent to which the whole system is affected depends on the scheduling of the uplink and downlink channels. We address this question, not treated previously, by developing a tractable model of NB-IoT connectivity, comprising message exchanges in random-access, control, and data channels. The model leads to the derivation of the expected latency and battery lifetime. This is used to analyze the impact of channel scheduling and the coverage class on the performance of IoT devices. These results are subsequently employed in determining the optimized operation points: (i) scheduling of data and control channels for a given set of users and respective coverage classes, or (ii) determining the optimal set of coverage classes and served users per coverage class for a given scheduling strategy. Simulations results show the validity of the analysis and confirm that channel scheduling and coexistence of coverage classes significantly affect latency and battery lifetime performance of NB-IoT devices.

Index Terms

NB-IoT, Channel scheduling, Battery lifetime, Latency-energy tradeoff.

I. Introduction

Internet of Things (IoT) is behind two of the three major drivers of next-generation wireless networks, which are massive machine-type communications (mMTC), ultra-reliable low latency communications (URLLC), and enhanced mobile broadband (eMBB) [1]. Massive IoT connectivity, related to mMTC, has fundamentally different characteristics and requirements compared to the legacy traffic in cellular networks. This is reflected through the massive number of connected devices, short packet sizes, and long battery lifetimes. Hence, massive IoT has given rise to revolutionary connectivity solutions in the wireless industry [2, 3]. The most prominent examples are SigFox, introduced in 2009, and Long-Range wide area network (LoRaWAN), introduced in 2015, both implemented in the unlicensed 868 MHz in Europe [3, 4]. In a separate activity, the accommodation of IoT traffic over cellular networks

has been investigated by the 3rd generation partnership project (3GPP), proposing evolutionary solutions like LTE Category-1 and LTE Category-M [5, 6]. Recently, these efforts have been also complemented by the introduction of revolutionary cellular solutions like Narrowband Internet of Things (NB-IoT) [7].

NB-IoT represents a big step towards the realization of massive IoT connectivity over cellular networks. Communication in NB-IoT systems takes place in a narrow, $200\,\mathrm{KHz}$ bandwidth, resulting in more than $20\,\mathrm{dB}$ link budget improvement over the legacy LTE [8]. Furthermore, NB-IoT introduces a set of coverage classes, each associated with a number of signal repetitions, which are assigned to users based on their experiencing path loss in communications with the base station (BS). The narrow communication bandwidth and signal repetitions allow the BS to communicate reliably with smart devices deployed in remote and/or isolated areas, such as rural areas and basements. As the legacy signaling and communication protocols were designed for large bandwidths, NB-IoT introduces a solution with five new narrowband physical (NP) channels [9, 10], see Fig. 1: random access channel (NPRACH), uplink shared channel (NPUSCH), downlink shared channel (NPDSCH), downlink control channel (NPDCCH), and broadcast channel (NPBCH). NB-IoT also introduces four new physical signals: demodulation reference signal (DMRS) that is sent with user data on NPUSCH, narrowband reference signal (NRS), narrowband primary synchronization signal (NPSS), and narrowband secondary synchronization signal (NSSS).

A. Radio Resource Management for NB-IoT Systems

In this paper, we study an important and so far untreated problem: when and how many resources to allocate to NPRACH, NPUSCH, NPDCCH, and NPDSCH when the BS serves NB-IoT devices that belong to different coverage classes and feature random activations. The coexistence of multiple coverage classes makes this radio resource management problem challenging, as the resource allocation to different channels faces inherent tradeoffs. The essence of the tradeoff can be explained as follows. On the one hand, if random access opportunities (NPRACH) occur frequently, less uplink radio resources remain for uplink data channel (NPUSCH), which increases the latency in data transmissions. On the other hand, if NPRACH is scheduled infrequently, latency and energy consumption in access reservation increase due to the extended idle-listening time and increased collision probability. Further, as device scheduling for uplink/downlink channels is performed over NPDCCH, infrequent scheduling of this channel may lead to wasted uplink resources in NPUSCH and increased latency in data transmissions. Conversely, if NPDCCH occurs frequently, the latency and energy consumption of transmissions over NPUSCH will increase. While the channel scheduling itself is a complicated problem, the introduction of coexisting coverage classes, and adapting channel scheduling to their diverse quality of service requirements pose further challenges to the problem.

B. Literature Study

A set of prior works on NB-IoT investigated preamble design for access reservation of devices over NPRACH [11, 12], uplink resource allocation to the connected devices [13], coverage and capacity analysis of NB-IoT systems in rural areas [14], coverage of NB-IoT with consideration of external interference due to deployment in guard band [15], and impact of channel coherence time on coverage of NB-IoT systems in [16]. Furthermore, in [17],

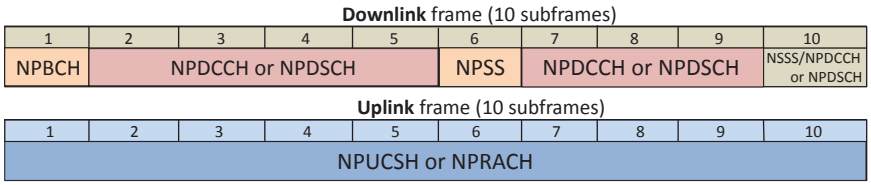


Fig. 1: NB-IoT features frequency-division duplex (FDD) for uplink and downlink [10]. Downlink/uplink NP channels and signals are time multiplexed, as depicted in the figure.

the energy consumption of IoT devices in data transmission over NB-IoT systems in normal, robust, and extreme coverage scenarios has been investigated. The results obtained in [17] illustrate that NB-IoT significantly reduces the energy consumption compared to the legacy LTE, due to the existence of the deep sleep mode for the devices that are registered to the BS.

The literature on latency and energy analysis and optimization for LTE networks is mature, see [18–20]. Specifically, in [18, 19], the latency and energy consumption of users have been modeled in downlink and uplink communications, respectively. In [20], the authors derive an analytical model of battery lifetime in IoT connectivity over LTE networks as a function of communications parameters of devices, such as transmit power, as well as radio resource scheduling of the BS, reflected in the number of radio resource blocks and the objective used for prioritization of different traffic streams. While the NB-IoT has been significantly inspired by the LTE, the revolutionary features of NB-IoT, including coexistence of coverage classes, time-multiplexing of physical channels, and non-saturated buffer of devices, mandate a detailed analysis of user's experienced latency and consumed energy. This analysis could be subsequently employed to optimize the operation of NB-IoT networks. Furthermore, one must note that there are plenty of prior works focusing on energy-efficient communications, such as [18, 19, 21]. Nevertheless, they have been mainly designed for data-hungry applications, e.g. video streaming, where there is a need for big chunks of radio resources for assuring the quality of service. There are a number of fundamental differences between machine-type communications and the legacy communication; notably, having a massive number of short-lived sessions in the former while having a small number of long-lived sessions in the latter. Due to those differences, most of the available legacy solutions could not be applied to IoT communications [22, Section 1]. In [23], inter-cell interference aware radio resource allocation to devices in uplink and downlink NB-IoT connectivity has been investigated, where the aim is to maximize the achievable data rate of each cell. A simulation-based performance analysis of NB-IoT systems, including delay and energy consumption, could be found in [24]. Energy consumption of IoT devices with power-saving mode and extended discontinuous reception (eDRX) in NB-IoT connectivity has been investigated in [25]. In [26], a predictive packet scheduler for NB-IoT has been proposed to enhance the quality of service in uplink communications. In order to decrease energy consumption in IoT communications, small data communications without connection setup in NB-IoT has been investigated in [27].

C. Paper Contributions and Structure

In this paper, we incorporate the NB-IoT channel multiplexing problem in modeling the energy consumption and experienced latency of IoT communications, while assuming coexistence of devices from a diverse set of coverage classes in the same cell. Furthermore, instead of maximizing the overall energy efficiency, we focus on minimizing the consumed energy in sending a given data packet. Specifically, the main contributions of this work are:

- Derivation of a tractable analytical model of the channel scheduling problem in NB-IoT systems that considers message exchanges on both downlink and uplink channels, from synchronization to service completion.
- Derivation of closed-form expressions for service latency and energy consumption, and derivation of the expected battery lifetime model for devices connected to the network.
- Formulating the control/data channel scheduling problem in NB-IoT systems as an optimization problem related to energy-delay minimization. Characterizing the energy-delay tradeoff in the system performance that is tuned by the channel scheduling.
- Presenting the interactions among the coverage classes offered by the system. Characterizing the performance loss for devices served in one coverage class by an increase in the number (or traffic volume) of devices from another coverage class.
- Characterizing the performance loss for devices experiencing a low-to-medium path loss when serving devices experiencing a huge path loss (so-called extreme coverage[1]).
- Elaborating a scheduling-based solution for compensating the performance loss incurred by provision of the extreme coverage in NB-IoT systems. The proposed solution adapts the scheduling of data/control channels in uplink/downlink directions for each coverage class based on its impact on other coverage classes.

The preliminary results of our research have been first presented in [32]. In this extended version, we have: (a) Added the state-of-the-art literature on IoT connectivity over cellular networks; (b) Extended the queuing model of NB-IoT connectivity, which is used in deriving analytical expressions for latency and battery lifetime; (c) Added a new section where we introduced and elaborated the set of tradeoffs related to enabling extreme coverage in NB-IoT systems; and (d) Presented the future directions of research for enhancing IoT connectivity over cellular networks, such as leveraging machine learning for activating coverage classes in NB-IoT networks that are adaptable to the actual traffic.

The remainder of the paper is structured as follows. The next section is devoted to the system model. Section III presents the modeling of key performance indicators (KPIs) of interest. The operational tradeoffs are elaborated in Section IV. The performance evaluation results are presented in Section V. Concluding remarks are given in Section VI.

[1]Extreme coverage in NB-IoT systems refers to providing connectivity for devices that experience maximum coupling loss of 164 dB [28–31].

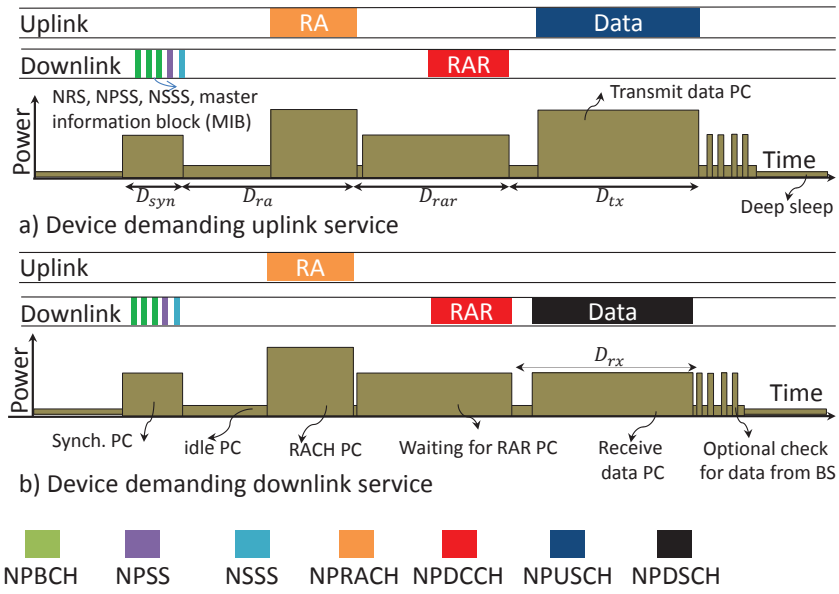


Fig. 2: Communications exchanges and power consumption in NB-IoT access networking. Note: Reference signals, including NRS, NPSS, NSSS, and master information block (MIB), are broadcasted regularly; here we show only a single realization.

II. System Model

A. NB-IoT Access Networking

Assume an NB-IoT cell with a BS located in its center, and N devices uniformly distributed in it. In general, there are C coverage classes defined in an NB-IoT cell, where the BS assigns a device to a class based on the estimated path loss between them and informs the device of its assignment. Class j, $\forall j$, is characterized by the number of replicas c_j that must be transmitted per original data/control packet. For example, based on the specifications in [10], each device belonging to group j shall repeat the preamble transmitted over NPRACH $c_j \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ times. Furthermore, let us denote the fraction of devices belonging to class j by f_j, the number of communication sessions that a typical IoT device performs *daily* by S and the probability that a device requests uplink service by p. The arrival rates of uplink/downlink service requests, per day, to the system are, respectively

$$G_u = N\,S\,p\ ,\ G_d = N\,S\,(1-p)\ . \tag{1}$$

Initially, when an NB-IoT device requires an uplink/downlink service, it first listens for the cell information, i.e., NPSS and NSSS, through which it synchronizes with the BS. Then, the device performs access reservation, by sending random access (RA) request to the BS over NPRACH. The BS answers to a successfully received RA by sending the random access response (RAR) message over NPDCCH, indicating the resources reserved for serving the device. Finally, the device sends/receives data to/from the BS over NPUSCH/NPDSCH channels, which, depending on the application, may be followed by an acknowledgment (ACK) [10]. In contrast to LTE, a device that is connected to the BS can go to the *deep sleep* state [7, Section 7.3], which does not exist in LTE and from

which the device can become reconnected just by transmitting an RA request accompanied by a random number [7, Fig. 7.3.4.5-1]. This new functionality aims to address the inefficient handling of IoT communications by LTE, as it saves a significant amount of energy because IoT devices do not need to restart all steps of the connection establishment procedure [17, 33]. Fig. 2 represents the access protocol exchanges for NB-IoT, as described in [7, Section 7.3].[2] A complete list of frequently used symbols throughout the paper and their definitions could be found in Section V.

B. Problem Formulation

Based on the model presented in Fig. 2, the expected latencies in uplink/downlink communication in class j are, respectively

$$\begin{aligned} D_{u_j} &= D_{\text{sy}_j} + D_{\text{rr}_j} + D_{\text{tx}_j}, \\ D_{d_j} &= D_{\text{sy}_j} + D_{\text{rr}_j} + D_{\text{rx}_j}, \end{aligned} \tag{2}$$

where D_{sy_j}, D_{rr_j}, D_{tx_j}, D_{rx_j} are the expected time spent in synchronization, resource reservation, data transmission in uplink service, and data reception in downlink service, respectively. Similarly, the models of expected energy consumption of an uplink/downlink communication in class j are

$$\begin{aligned} \mathcal{E}_{u_j} &= E_{\text{sy}_j} + E_{\text{rr}_j} + E_{\text{tx}_j} + E_s, \\ \mathcal{E}_{d_j} &= E_{\text{sy}_j} + E_{\text{rr}_j} + E_{\text{rx}_j} + E_s, \end{aligned} \tag{3}$$

where E_{sy_j}, E_{rr_j}, E_{tx_j}, E_{rx_j}, and E_s are, respectively, the expected device energy consumption in synchronization, resource reservation, data transmission in uplink service, data reception in downlink service, and optional communications, such as acknowledgment. Since the energy consumption of a typical reporting IoT device can be modeled as a semi-regenerative Poisson process with regeneration point at the end of each reporting period [20], one may define the expected battery lifetime as the ratio between stored energy and energy consumption per reporting period. In this case, the expected battery lifetime can be derived as

$$L_j = \frac{E_0}{Sp\mathcal{E}_{u_j} + S(1-p)\mathcal{E}_{d_j}} \text{ [day]}, \tag{4}$$

where E_0 is the energy storage at the device battery. To see how channel scheduling affects latency and battery lifetime, let us for example focus on the NPDCCH. If NPDCCH is scheduled frequently, less downlink radio resources remain for NPDSCH, and hence, D_{rx} and E_{rx} increase. On the other hand, D_{rar} and E_{rar} increase if NPDCCH is not scheduled frequently because it means devices must listen for a longer time to receive the RAR message from the BS. The increase in both E_{rx} and E_{rar}, achieved by over-scheduling and under-scheduling of NPDCCH, results in shorter battery lifetimes, and hence, one observes the crucial need for finding the optimized

[2]For the sake of completeness, we also mention another novel reconnection scheme designed for NB-IoT, in which a device can request to resume its previous connection after receiving the random access response (RAR) [9, Section III]. Towards this end, it needs to respond to the RAR message by the transmission of its previous connection ID as well as the cause for resuming the connection.

operation points. In order to derive closed-form latency and energy consumption expressions, e.g., model E_{rr_j} and D_{rr_j}, in the sequel we investigate analytically the performance impacts of channel scheduling, arrival traffic, and coexisting coverage classes on the performance indicators of interest.

III. Modeling of KPIs

As mentioned in Section II, in NB-IoT systems the control, data, random access, and broadcast channels are multiplexed on the same set of radio resources. Thus, their mutual impact in both uplink and downlink directions are significant, which is not the case in legacy LTE due to the wide set of available radio resources. In the following, we propose a queuing model of NB-IoT access networking, which captures these interactions.

A. Queuing Model of NB-IoT Access Protocol

Recall the communications exchanges presented in Fig. 2. Based on these exchanges, one observes that 5 sets of signals/physical-channels are scheduled over the uplink/downlink radio resources, including: (i) reference signals; (ii) access reservation resources; (iii) control signaling; (iv) uplink data; and (v) downlink data. Then, one can model the uplink/downlink radio resources as two servers that visit and serve their respective traffic queues, as depicted in Fig. 3. In this figure, the radio resource performs as polling server, which serves several queues. Regarding the fact that the status of queues and the radio resource management among different queues are interconnected, one cannot leverage the existing results in the literature on queuing systems with polling servers [34]. Thus, there is a need for a closer look at the dependencies of these queues, the order of user's presence in each queue, and the impact of radio resource management policy. By leveraging the definition of narrowband physical data and control channels, as defined in Section II, the abstract model in Fig. 3 can be transformed to the detailed model in Fig. 4. This figure depicts the queuing model of NB-IoT access networking, consisting of NP random access, control, and data channels. The left circle represents the uplink server serving two channel queues, NPRACH and NPUSCH, while the right circle represents the downlink channel serving three channel queues, NPDCCH, NPDSCH, as well as the reference signals, such as NPSS. Let t_j be the average time interval between two consecutive scheduling of NPRACH of class j and M_j the number of orthogonal random access preambles available in it. The duration of scheduled NPRACH of class j is $c_j\,\tau$, where τ is the unit length, equal to the NPRACH period for the coverage class with $c_j = 1$. The inter-arrival times between two NPRACH periods in NB-IoT can vary from $40\,\text{ms}$ to $2.56\,\text{s}$ [10]. Further, b denotes the fraction of time in which reference signals are scheduled in a downlink radio frame, e.g., NPBCH, NPSS, and NSSS. Five subframes in every two consecutive downlink frames are allocated to reference signals [10], implying $b = 0.2$. Finally, semi-regular scheduling of NPDCCH has been proposed by 3GPP to prevent waste of resources in the uplink channel when BS serves another device with poor coverage in the downlink [35]; we denote by d the average time interval between two consecutive NPDCCH instances. In the next section, we derive closed-form expressions for components of latency and battery lifetime models, given in (2)-(3).

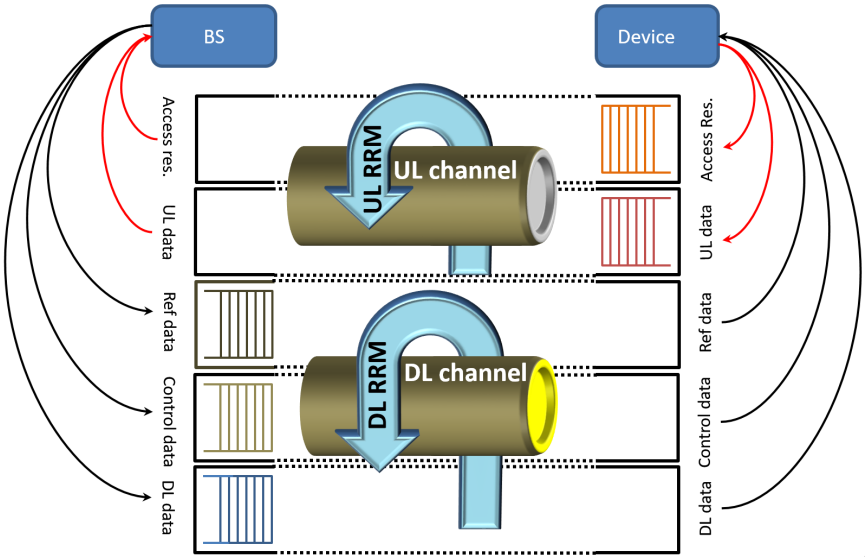


Fig. 3: The abstract queuing model of FDD NB-IoT system. Radio resource is seen as a polling server serves several interdependent queues, and radio resource management (RRM) is seen as the service policy.

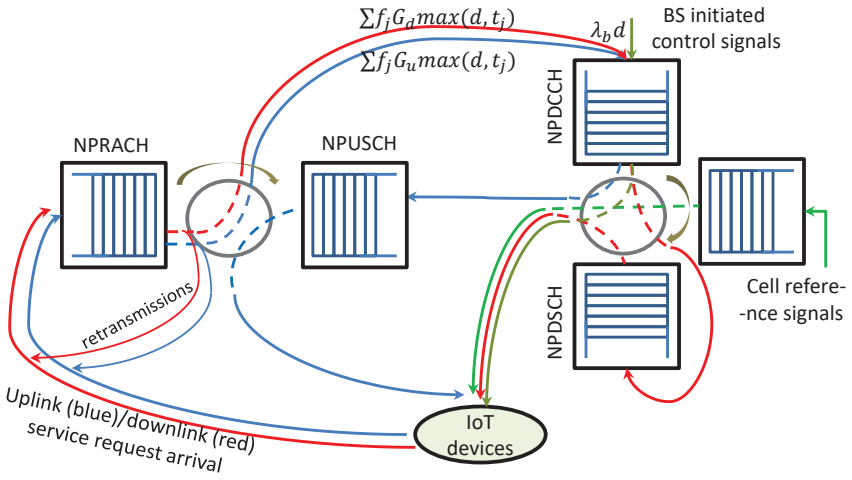


Fig. 4: Queuing model of the NB-IoT access networking. The right and left circles represent servers for downlink and uplink channels, respectively.

B. Derivations

D_{sy_j} in (2) is a function of the coverage class j. Its average value has been reported in [7, Sec. 7.3]. D_{rr_j} is given by

$$D_{\mathrm{rr}_j} = \sum_{\ell=1}^{N_{r_{\max}}} (1-\mathcal{P}_j)^{\ell-1} \mathcal{P}_j \ell (D_{\mathrm{ra}_j} + D_{\mathrm{rar}_j}), \tag{5}$$

in which $N_{r_{\max}}$ denotes the maximum allowed number of attempts, $\mathcal{P}_j$ denotes the probability of successful resource reservation in an attempt that depends on the number of devices in the class attempting the access, D_{ra_j} denotes the expected latency in sending a RA message, and D_{rar_j} denotes the expected latency in receiving the RAR message.

D_{ra_j} is a function of time interval between consecutive scheduling of NPRACHs and is equal to $0.5\,t_j + c_j\tau$, while D_{rar_j} depends on the operation of NPDCCH. NPDCCH can be seen as a queuing system in which the downlink server (see Fig. 4) visits the queue every d seconds and serves the existing requests. Thus, D_{rar_j} consists of i) waiting for NPDCCH to occur, which happens on average $d/2$ seconds, ii) time interval spent waiting to be served when NPDCCH occurs[3], denoted by D_w, and iii) transmission time, denoted by D_{t_j}.

We first characterize D_w. When the server visits the NPDCCH queue, on average there are

$$\mathcal{Q} = \sum\nolimits_{j=1}^{C} f_j(G_u + G_d)\max\{d, t_j\} + \lambda_{\text{bs}}\, d, \tag{6}$$

requests waiting to be served, where $\max\{\mathbf{x}\}$ returns the maximum value in vector $\mathbf{x}$, the first term in $\mathcal{Q}$ corresponds to NPRACH-initiated random access requests, see (1), and $\lambda_{\text{bs}}\, d$ models the arrival of BS-initiated control signals, see Fig. 4. Thus, the average waiting time before the service of a newly arrived RA message starts is

$$D_w = 0.5\,\mathcal{Q}\,\mathcal{D}_t,$$

where $\mathcal{D}_t$ is the average service time in NPDCCH[4]. Using u as the average control packet transmission time, the average transmission time for class j is $D_{t_j} = c_j\, u$. Thus:

$$\mathcal{D}_t = \sum\nolimits_{j=1}^{C} f_j \mathcal{D}_{t_j} = \sum\nolimits_{j=1}^{C} f_j c_j u, \tag{7}$$

and D_{rar_j} becomes

$$D_{\text{rar}_j} = 0.5\,d + 0.5\,\mathcal{Q}\,\mathcal{D}_t + c_j\, u. \tag{8}$$

Resource reservation of a device over NPRACH is successful if its transmitted preamble does not collide with other nodes' preambles, which happens with probability $\mathcal{P}_{j_{\text{RACH}}}$, and the RA response is received within period T_{th}, which happens with probability $\mathcal{P}_{j_{\text{RAR}}}$. Thus, the probability of successful resource reservation can be approximated as $\mathcal{P}_j = \mathcal{P}_{j_{\text{RACH}}}\,\mathcal{P}_{j_{\text{RAR}}}$. For a device belonging to class j, there are M_j orthogonal preambles available every t seconds, during which it contends on average with $\mathcal{N}_j$ devices, where

$$\mathcal{N}_j = f_j(G_u + G_d)t_j.$$

Then, $\mathcal{P}_{j_{\text{RACH}}}$ is derived as

$$\mathcal{P}_{j_{\text{RACH}}} = \sum\nolimits_{k=2}^{N} \frac{(N_j)^k e^{-\mathcal{N}_j}}{k!} \left(\frac{M_j - 1}{M_j}\right)^{k-1}. \tag{9}$$

[3]This is the queuing time during which other users are served.

[4]Note that $\mathcal{D}_t$ corresponds to the delay in serving a typical request over NPDCCH, when the BS is transmitting response to other devices, and should not be misunderstood with D_{tx}, i.e., the length of data transmission by a granted device over NPUSCH.

The cumulative distribution function of service time for a device and sum of service times for $n > 1$ devices are

$$\mathcal{F}_1(x) = \sum\nolimits_{j=1}^{C} f_j H(x - c_j u), \tag{10}$$
$$\mathcal{F}_n(x) = \sum\nolimits_{j=1}^{C} f_j \mathcal{F}_{n-1}(x - c_j u)$$

respectively, where $H(x)$ is the unit step function. Then, $\mathcal{P}_{j_{\text{RAR}}}$, which is the probability that RAR is received within T_{th}, is

$$\mathcal{P}_{j_{\text{RAR}}} = 1 - \sum_{K=2}^{\infty} \sum_{k=1}^{K-1} \frac{k}{K} \frac{\mathcal{Q}^K e^{-\mathcal{Q}}}{K!} \left(1 - \mathcal{F}_{K-k}(T_{\text{th}})\right) \mathcal{F}_{K-k-1}(T_{\text{th}}), \tag{11}$$

where in this expression, K represents the potential number of requests in the queue to be served. D_{tx_j} is a function of scheduling of NPUSCH. Operation of NPUSCH can be seen as a queuing system in which server handles requests in a fraction of each uplink frame that is allocated to NPUSCH; this fraction is

$$w = 1 - \sum\nolimits_{j=1}^{C} c_j \tau / t_j. \tag{12}$$

The arrival of service requests to the NPUSCH can be modeled as a batch Poisson process (BPP), as resource reservation happens only in NPRACH periods. The mean batch-size is

$$\mathcal{G} = \frac{1}{C} \sum\nolimits_{j=1}^{C} f_j G_u t_j,$$

and the rate of batch arrivals is $\sum_{j=1}^{C} 1/t_j$. The uplink transmission time is determined by the packet size and coverage class j. We assume that the packet length follows a general distribution with the first two moments equal to l_{a} and l_{b}. Then, the transmission (i.e., service) time for the uplink packet follows a general distribution with the first two moments

$$s_{\text{a}} = \sum\nolimits_{j=1}^{C} \frac{f_j c_j l_{\text{a}}}{\mathcal{R}_j w} \text{ and } s_{\text{b}} = \sum\nolimits_{j=1}^{C} \frac{f_j c_j^2 l_{\text{b}}}{\mathcal{R}_j^2 w^2} \tag{13}$$

where $\mathcal{R}_j$ is the average uplink transmission rate for class j. This queuing system is a BPP/G/1 system, hence, using the results from [36], one can derive the latency in data transmission for class j as

$$D_{\text{tx}_j} = \frac{\rho s_{\text{b}}}{2 s_{\text{a}} (1 - \rho)} + \frac{\mathcal{G} s_{\text{a}}}{2(1 - \rho)} + \frac{c_j l_{\text{a}}}{\mathcal{R}_j w} \tag{14}$$

where

$$\rho = \sum\nolimits_{j=1}^{C} \mathcal{G} s_{\text{a}} / t_j.$$

Similarly, performance of NPDSCH can be seen as a queuing system in which server visits the queue in a fraction of frame time and serves the requests. This fraction comprises to subframes in which NPDCCH, NPBCH, NPSS,

and NSSS are not scheduled, and can be derived similarly to (8) as

$$y = 1 - b - \frac{\mathcal{Q}}{d} \sum\nolimits_{j=1}^{C} f_j c_j u. \tag{15}$$

The arrival of downlink service requests to the NPDSCH queue can be also seen as a BPP, as they arrive only after NPRACH has occurred. The mean batch-size is

$$\mathscr{G} = \frac{1}{C} \sum\nolimits_{j=1}^{C} f_j G_d t_j,$$

and the arrival rate is $\sum_{j=1}^{C} 1/t_j$. The downlink transmission time is determined by the packet size and coverage class j. Assuming that packet length follows a general distribution with moments m_{a} and m_{b}, then first two moments of the distribution of the packet transmission time are

$$h_1 = \sum\nolimits_{j=1}^{C} \frac{f_j c_j m_{\text{a}}}{\mathscr{R}_j y} \text{ and } h_{\text{b}} = \sum\nolimits_{j=1}^{C} \frac{f_j c_j^2 m_{\text{b}}}{\mathscr{R}_j^2 y^2} \tag{16}$$

where $\mathscr{R}_j$ is the average downlink data rate for coverage class j. Defining $\nu = \sum_{j=1}^{C} \frac{\mathscr{G} h_{\text{a}}}{t_j}$, the latency in data reception D_{rx_j} becomes

$$D_{\text{rx}_j} = \frac{0.5\nu h_{\text{b}}}{h_1(1-\nu)} + \frac{\mathscr{G} h_{\text{a}}}{2(1-\nu)} + \frac{c_j m_{\text{b}}}{\mathscr{R}_j y}. \tag{17}$$

Finally, we derive the average energy consumption of an uplink/downlink service. Denote by ξ, P_I, P_c, P_l, and P_{t_j} the power amplifier efficiency, idle power consumption, circuit power consumption of transmission, listening power consumption, and transmit power consumption for class j. Then,

$$E_{\text{sy}_j} = P_l D_{\text{sy}_j}, \tag{18}$$

$$E_{\text{rar}_j} = P_l D_{\text{rar}_j}, \tag{19}$$

$$E_{\text{rr}} = \sum\nolimits_{l=1}^{N_{r\max}} (1 - \mathcal{P}_j)^{l-1} \mathcal{P}_j (E_{\text{ra}_j} + E_{\text{rar}_j}), \tag{20}$$

$$E_{\text{ra}_j} = (D_{\text{ra}} - c_j \tau) P_I + c_j \tau (P_c + \xi P_{t_j}), \tag{21}$$

$$E_{\text{tx}_j} = (D_{\text{tx}_j} - \frac{c_j l_{\text{a}}}{\mathcal{R}_j w}) P_I + (P_c + \xi P_{t_j}) \frac{c_j l_{\text{a}}}{\mathcal{R}_j w}, \tag{22}$$

$$E_{\text{rx}_j} = (D_{\text{rx}_j} - \frac{c_j m_{\text{a}}}{\mathscr{R}_j y}) P_I + P_l \frac{c_j m_{\text{a}}}{\mathscr{R}_j y}, \tag{23}$$

from which the battery lifetime model (4) is derived as

$$L_j = E_0 \Big(Sp[E_{\text{sy}_j} + E_{\text{rr}_j} + E_{\text{tx}_j} + E_s] + S(1-p)[E_{\text{sy}_j} + E_{\text{rr}_j} + E_{\text{rx}_j} + E_s] \Big)^{-1}, \tag{24}$$

where its parameters have been defined in (18)-(23). One observes that, in order to maximize the expected battery lifetime of a device, one should minimize energy consumption in both uplink and downlink communication exchanges. Scheduling of uplink and downlink resources necessarily creates coupling between them. For example, the use of uplink resources is governed by thje control signals transmitted over downlink channel. Due to this,

the optimal scheduling that aims to maximize the expected battery lifetime is not only challenging but also has side effects on the latency and other performance indicators. These performance tradeoffs are analyzed in the next section.

IV. Extreme Coverage in NB-IoT Systems: The performance tradeoffs

A. Tradeoff analysis

The analysis in the previous section could be leveraged in order to shed light on the dark side of enabling extreme coverage over NB-IoT systems and try to compensate such side effects. In order to ease following the discussion, let us exemplify the analysis and assume we have two coverage classes in the network, where the first one and second ones correspond to devices experiencing normal and extreme path loss in communications with the BS.

From the battery lifetime expression in (24), one observes that battery lifetime for class 1 of devices increases by a decrease in energy consumption in the resource reservation and data transmission/reception modes, i.e., E_{rr_1}, E_{tx_1}, and E_{rx_1} respectively. From (20)-(23), it is clear that this could be achieved by minimizing the experienced delay in receiving the RAR message and data transmission and reception, i.e., minimizing D_{rar_1}, D_{tx_1} and D_{rx_1} respectively. In the following, we highlight the interplay between these latency expressions and their corresponding expressions for the second class of devices.

- First, the expression in (8) represents that D_{rar_1} increases by a decrease in d. If one aim at decreasing d, it will result in over-scheduling of radio resources for control signaling, and hence, fewer resources will remain for data transmission in the downlink direction, which on the other hand increases D_{rx_1}. Then, the **first** tradeoff exists between the data transmission/reception latency and the latency in receiving the RAR message. As we observed, this tradeoff is tuned by the number of allocated resources to control signaling in the downlink channel. In the following section, we will conduct a wide set of analyses on this tradeoff by considering d, the inter-arrival time between two scheduling epochs of the downlink control channels, as the design parameter.
- Second, the expression in (8) and (7) represent that D_{rar_1} increases by a decrease in $\mathcal{D}_t$. On the other hand in Section III-B we observed that $\mathcal{D}_t$ increases by increasing the number of coverage classes served in the network, and the degree of increase is a function of (a) fraction of devices belong to the new coverage class, i.e., f_2; and (b) the repetition order of devices in the new class, i.e., c_2. Then, the **second** tradeoff exists between the latency in receiving the RAR message for class 1 and serving devices of coverage class 2. Then, we expect that by serving devices of the second coverage class, which require extreme coverage, the expected latency in receiving the control signals for the first class will increase, and hence, the expected battery lifetime will decrease.
- Third, the expression in (14)-(17) represent that the expected latency in data transmission and reception decrease by a an increase in ω and y and a decrease in s_{a} and l_{a}. From (12), (13) and (16) it is clear that ω and y are decreased by an increase in the number of coverage classes served in the system and the repetition order of each class; and (b) s_{a} and h_{a} increase by a decrease in ω and y respectively, and further increase by with the increase of the number of coverage classes served in the system and the repetition order of each class.

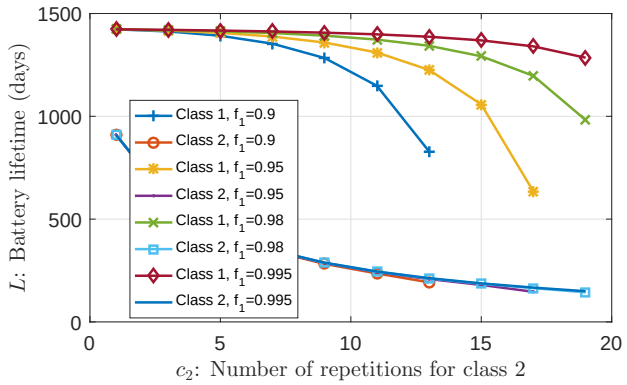


Fig. 5: Mutual impact among two coexisting classes in a cell changing with with the number of repetitions for the second class ($C = 2$, $c_1 = 1$, $f_2 = 1 - f_1$, $\tau = 2\,\text{ms}$, $d = 10\,\text{ms}$, $t = 65\,\text{ms}$). Other simulation parameters could be found in Table I.

Then, one observes the **third** strong tradeoff between the latency in data transmission/reception for class 1 and serving devices of coverage class 2.

While the overall ambition is to maximize the battery lifetime for devices of all coverage classes, the KPIs of devices belonging to different classes are interconnected, as seen above. Then, an increase in one objective results in a decrease in the other(s). Hence, finding the best set of coverage classes to be served in the network, as well as density of authenticated devices from each coverage class, has a significant impact on the overall quality of service. This problem can be addressed by leveraging the derived results in the previous section.

B. Evaluation of the tradeoffs

In order to highlight the crucial impact of the derived expression and tradeoffs within this work in the proper design of an NB-IoT system, in this subsection, we carry out analysis for an exemplary network serving devices of two coverage classes. The corresponding parameters for our analysis could be found in Table I. Fig. 5 shows the mutual impact of two coexisting coverage classes in a cell, i.e., class 1 and class 2. The y-axis represents the expected battery lifetime for both classes, while the x-axis represents the number of repetitions for class 2, i.e., c_2. Increase in c_2 increases the number of radio resources which are used for signal repetitions (i.e., coverage extension) of devices in class 2. This results in an increased latency both for class 1 and class 2 devices, and hence, increases the energy consumptions per reporting period and decreases the battery lifetime. Also, it can be seen that an increase in the fraction of nodes belonging to class 2, adversely impacts the battery lifetime performance for class 1 devices. For instance, increasing c_2 from 11 to 13 decreases the average battery lifetime of class 1 nodes for $6\,\%$ when $f_1 = 0.95$ (i.e., $f_2 = 0.05$) and for $28\,\%$ when $f_1 = 0.90$ (i.e., $f_2 = 0.1$). Nevertheless, the extended coverage enables devices in class 2 to become connected to the BS, i.e., provides a deeper coverage to indoor areas.

TABLE I: Parameters for performance analysis. Indices 1 and 2 refer to coverage class 1 and 2 respectively.

category: parameters	symbols	values
Traffic: number of devices, packet generation frequency	N, S	20000, $0.5\,\mathrm{h}^{-1}$
Traffic: probability of uplink (res. DL) service request	p (res. 1-p)	0.8 (res. 0.2)
Traffic: moments of uplink and downlink packet lengths	l_{a}, m_{a}	500, 5 Kbit
Traffic: average length of control and RA signaling	u, τ	2 ms, 10 ms
Traffic: frequency of arrival of BS-initiated control data	λ_{bs}	1/CF
Traffic: fraction of devices belongs to each coverage class	f_1, f_2	0.5, 0.5
Coverage: repetition order	c_1, c_2	1, 2
Coverage: uplink data rate	$\mathcal{R}_1, \mathcal{R}_2$	$5, 5\,\mathrm{Kbit/s}$
Coverage: downlink data rate	$\mathscr{R}_1, \mathscr{R}_2$	$15, 15\,\mathrm{Kbit/s}$
Coverage: synchronization delay	$D_{\mathrm{sy}_1}, D_{\mathrm{sy}_2}$	0.33 s, 0.66 s
RRM: length of communication frame	CF	10 ms
RRM: fraction of each frame occupied by ref. signals	b	0.2
RRM: maximum waiting for receiving RAR message	T_{th}	2 s
RRM: number of RA resources	M_1, M_2	16, 16 preambles
RRM: time interval between two scheduling of NPRACH	t	design parameter
RRM: time interval between two scheduling of NPCCCH	d	design parameter
Other: Device's battery capacity	E_0	1 KJ
Other: Device's power consumption in transmission, idle, and listening	P_t, P_I, P_l	0.2, 0.01, 0.01, 0.1 W
Other: Device's power consumption in electronic circuits	P_c	0.01 W

V. Performance Evaluation

A. Simulation setup

In this section, we validate the derived expressions, highlight performance tradeoffs in channel scheduling, find optimized system operation points, and identify the mutual impact among the coexisting coverage classes. The simulator has been developed in Matlab and is publicly available online for cross validation[5]. In the simulations, we consider a circular service area with a single BS at the center and a multitude of IoT devices deployed in the service area following a Poisson Point Process. The other system parameters are presented in Table I.

B. Validation of the analytical results

Fig. 6 compares the analytical up/downlink latency expressions derived in Section III-B (dashed curves) against the simulation results (solid curves) for class 1 devices. The abscissa represents t, the average time between two scheduling of random access resources. It is obvious that the simulations results, including service latency in uplink and downlink, match well with the respective analytical results. The comparison of analytical and simulation results for average energy consumption per day of activity could be found in Fig. 7. By comparing Fig. 6 and Fig. 7, one may observe that by a decrease in t, i.e., an increase in the amount of allocated resources to the random access channel, the amount of remaining resources for uplink data transmission will decrease. This, in turn, results in a significant increase in latency in data transmission over the uplink channel, and hence, results in an increase in

[5]https://github.com/AminAzari/NB-IoT

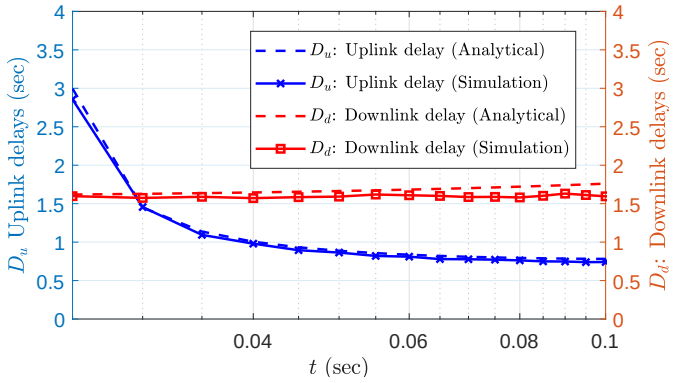


Fig. 6: Comparison of analytical and simulation results for experienced up/downlink latency versus t for class 1. $d = 10\,\text{ms}$, $S = 0.25\text{h}^{-1}$, $l_\text{a} = 200$ bits, and $m_\text{a} = 5$Kbits.

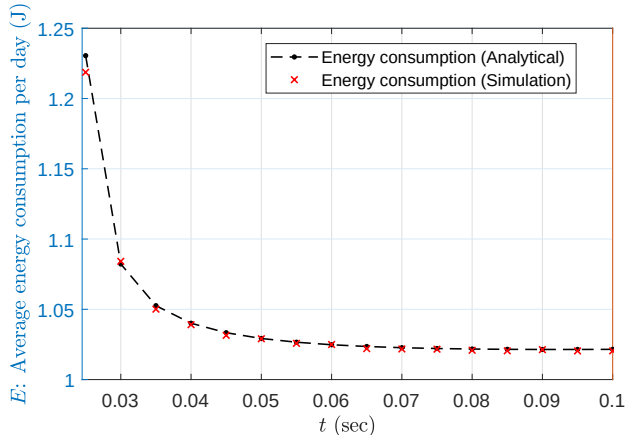


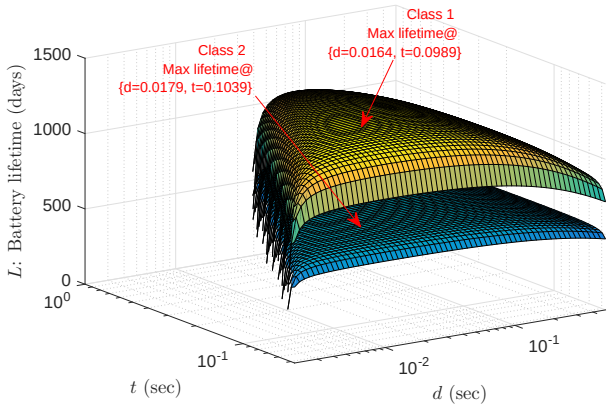
Fig. 7: Comparison of analytical and simulation energy consumption results versus t for class 1. $d = 10\,\text{ms}$, $S = 0.25\text{h}^{-1}$, $l_\text{a} = 200$ bits, and $m_\text{a} = 5$Kbits.

the energy consumption accordingly. By increasing t, i.e., decreasing the amount of allocated resources to random access, devices need to wait longer to become connected to the network and contention over the random access resources will be more intense. Hence, more collisions are expected to happen for large values of t.

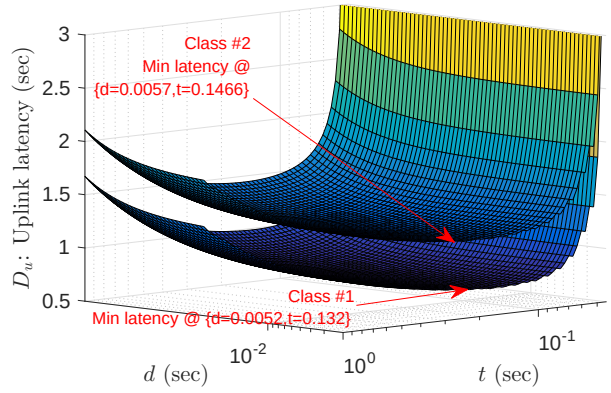
C. Optimizing physical channel scheduling

Fig. 8a shows the expected battery lifetime changing with t and d, i.e., the time intervals between two consecutive scheduling of NPRACH and NPDCCH, respectively, for the same coexistence scenario. Increasing t at first increases the lifetime of devices in both classes, as it provides more resources for NPUSCH scheduling and decreases time spent in data transmission, i.e., D_{tx}. After a certain point, increasing t reduces the lifetime due to the increase of the expected time in resource reservation. Similarly, increasing d at first increases the lifetime by providing more resources for NPDSCH, decreasing the time spent in data reception, D_{rx} (refer to Fig. 2 for more details), while after a certain point it decreases the lifetime by increasing the expected time in resource reservation.

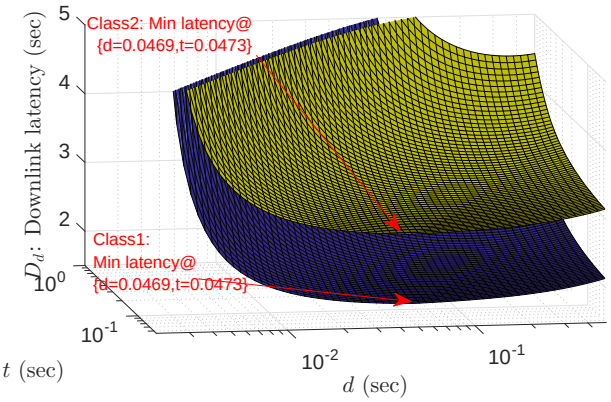
The impact of t and d on latency in uplink/downlink services is shown in Fig. 8b/Fig. 8c. If the uplink/downlink



(a) Battery lifetime L versus t and d.



(b) Uplink latency D_u versus t and d.



(c) Downlink latency D_d versus t and d.

Fig. 8: Performance as function of t and d, which are time intervals between two scheduling of NPRACH and NPDCCH, respectively.

latency, or the battery consumption represents the only optimization objective, it is straightforward to derive the optimized operation points. However, Figs. 8a-8c show that overall optimization of the objectives is coupled in conflicting ways. Furthermore, Figs. 8a-8c show that the latency- and lifetime-optimized resource allocation strategy differ on a class basis; thus, selecting the optimized values of t and d depends on the required quality of service

(lifetime and/or latency) for each class. In Fig. 8c, we observe that the optimized t and d values for minimizing downlink latency of users belonging to coverage class 1 and 2 are the same. The reason for this regarding t is in the weak dependence of the downlink latency on t. For d, one must note that while the extra communications demands of coverage class 2's users may call for extra radio resources, and hence a higher d value, increasing d increases the latency for downlink communications. Then, the optimized d value for both coverage classes are the same. Fig. 9 illustrates normalized lifetime and latency for class 1 when d is fixed. For a given d, from the analytical derivations, we expect that the downlink and uplink latency increase and decrease by an increase in t, respectively. This is due to the fact that latency in downlink data transmissions increases when the inter-arrival time between two consecutive random access opportunities increase, and hence, downlink packet needs to wait longer before channel access. Then, we have the minimum downlink latency at the left side of Fig. 9. On the other hand, while uplink latency is also coupled with the random access opportunity in the same way as the downlink communications, random access and uplink data transmission also share the same set of radio resources. Thus, an increase in t, i.e., decrease in the fraction of resources dedicated to random access, decreases the experienced uplink latency up to some extent. Finally, regarding the fact that device has both uplink and downlink communications and both affect the battery lifetime of the device, we expect that the battery lifetime maximizing t value must be between the derived values during minimization of uplink and downlink latency expressions. For instance, when $d = 4.4$ ms, the downlink and uplink latency are minimized for $t = 50\,\text{ms}$ and $t = 110\,\text{ms}$, and lifetime is maximized for $t = 70\,\text{ms}$.

Fig. 10 illustrates normalized lifetime and latency for class 1 when t is fixed. Here, we expect that uplink latency increases by an increase of d. This is due to the fact that increasing d, and hence delaying the occurrence of a downlink control channel, increases the latency for an uplink packet to access the channel because radio resources are governed based on the control signaling over the downlink control channel. On the other hand, due to the fact that control and downlink data share the same set of radio resources, an increase in d decreases the downlink latency up to some extent, beyond which, downlink latency may also increase due to waiting for the control signal. Finally, the battery lifetime maximizing d value resides in the interval between uplink and downlink latency minimizing d values, as depicted in Fig. 10. For example, we observe in this figure that when $t = 100$ ms, the downlink and uplink latency values are minimized for $d = 350\,\text{ms}$ and $d = 40\,\text{ms}$, and lifetime is maximized for $d = 80\,\text{ms}$.

VI. Conclusions and Future Work

A. Summary

In this paper, the side effects of enabling extreme coverage over NB-IoT systems have been studied and channel-scheduling based solutions have been presented aiming at compensating the side effects. First, a tractable analytical framework has been proposed to analyze the impact of the scheduling of control and data channels, as well as the coexistence of coverage classes, on the experienced latency and battery lifetime of IoT devices. Using the derived model, it has been found that the experienced latency and consumed energy over different physical channels, e.g.,

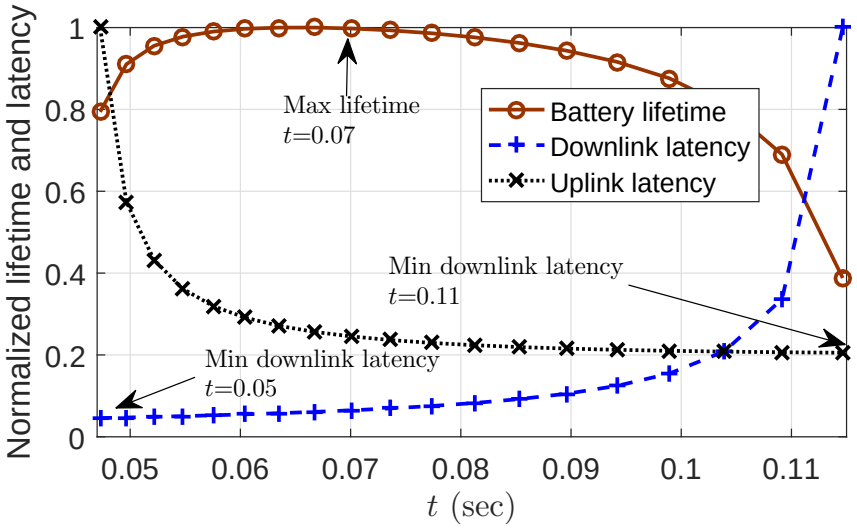


Fig. 9: Overall performance analysis for class 1 changing with t, when $d = 0.0044$. t and d are the average time intervals between two scheduling of NPRACH and NPDCCH, respectively.

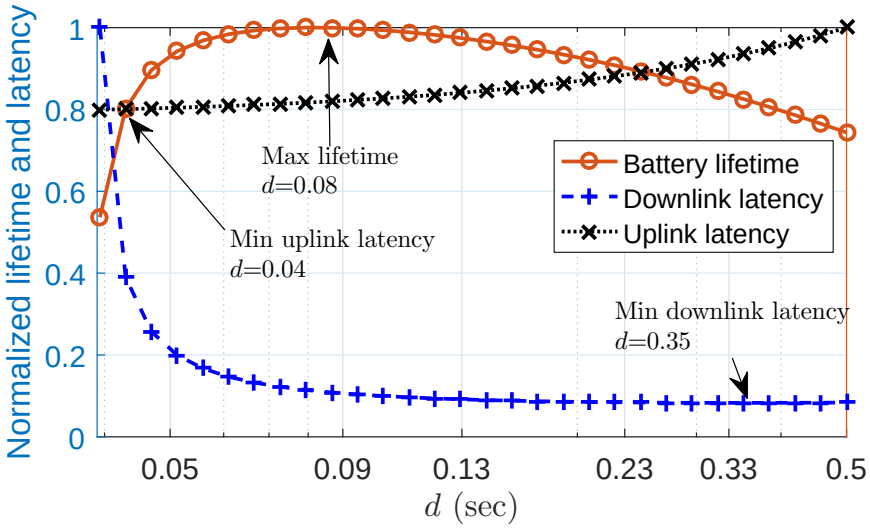


Fig. 10: Overall performance analysis for class 1 changing with d, when $t = 1$. t and d are the average time intervals between two scheduling of NPRACH and NPDCCH, respectively.

random access, data and control channels, are coupled in conflicting ways, as an improvement in one channel adversely affects the other channels. Hence, scheduling of physical channels cannot be optimized separately. The proposed analytical framework has been further employed to analyze performance trade-offs when offering different coverage classes in NB-IoT systems. The results show strong coupling between the support of extreme coverage for devices experiencing huge path loss and degradation of energy- and latency-performance for devices experiencing lower path loss. Through performance evaluation, we have presented the way channel scheduling could be used to tune the performance of devices belonging to different coverage classes. Finally, given the set of radio resources in an NB-IoT system and the traffic arrival statistics, optimized scheduling policies minimizing the experienced

latency and maximizing the expected battery lifetime have been proposed.

B. Future research directions

In the following, we present three research directions that can follow the aspects of NB-IoT systems presented in this paper.

1) Consideration of an extended set of KPIs: In this paper, we have focused on the impact of coexistence of different coverage classes on the latency and energy consumption of IoT devices and have neglected other KPIs, such as fairness, outage probability, and throughput. In this respect, a more unifying approach that considers all these KPIs may be of interest in order to provide insights on the optimized set of coverage classes that could be offered in each service area. This study may also consider the impact of resource allocation, the number of allocated carriers to NB-IoT, and the resource scheduling policy over the allocated resources for a more comprehensive analysis.

2) Learning-powered network management: In recent years there has been a profound interest in applying machine learning algorithms in communications networks for automating the management processes [37]. Application of these algorithms not only reduces the operating costs but also enables the network to react rapidly to the internal changes, e.g., operational anomalies or burst arrival. An interesting direction of future study consists in extending the channel scheduling policies developed in this work to be run in a learning and self-configured manner by leveraging machine learning algorithms, especially the reinforcement learning algorithms [37, 38]. Using such schemes, the network will be able to select the best policy for scheduling and allocating resources to different physical channels, based on the updated status of the network, e.g., set of present devices and their communications needs.

3) Novel solutions for compensating the side effects of providing extreme coverage: Until now, we have leveraged scheduling of uplink/downlink radio resources in NB-IoT systems in order to compensate the side-effects of serving devices with extreme coverage requirement in addition to IoT devices with less demanding coverage requirements. Here, we present some novel approaches that may be useful in addressing the side effects of enabling extreme coverage. The first approach is related to grant-free radio access. While the legacy cellular systems only allow data transmission after radio resource reservation, grant-free radio access is an alternative solution with a confirmed position in the new radio (NR) of 5G networks [39, 40]. The state of the art analysis shows that in the low-to-medium traffic load regimes, grant-free access can significantly decrease the access delay and increase the battery lifetime [39]. In this respect, it may be beneficial to configure the NB-IoT devices with the lowest repetition order(s) to send their short packets in a pool of radio resources dedicated to grant-free access. Along these lines, investigations of the operation regions in which grant-free access is beneficial in NB-IoT and which are the suitable modulation and coding schemes that enable decoding of signals with potential time/frequency overlaps should be explored [41]. The second approach is to keep the control signaling of devices with extreme coverage requirement in the NB-IoT bandwidth and to perform their data transmissions over a standard LTE-carrier. In this case, the device will still leverage signal repetitions in time for range extension over NB-IoT control link, but the data link budget could be less than the one achieved in NB-IoT connectivity. On the other hand, the side-effects on the other devices will be minimized with this solution. Here, investigation of radio resource management solutions for traffic steering between NB-IoT and LTE resources will be crucial.

ACKNOWLEDGMENT

The research presented in this paper was supported in part by the EIT Digital Project ACTIVE (Advanced Connectivity Platform for Vertical Segments), in part by the Celtic Plus Project SooGreen (Service Oriented Optimization of Green Mobile Networks), and in part by the European Research Council (ERC Consolidator Grant Nr. 648382 WILLOW) within the Horizon 2020 Program.

REFERENCES

[1] C. Mavromoustakis, G. Mastorakis, and J. M. Batalla, *Internet of Things (IoT) in 5G mobile technologies*. Springer, 2016, vol. 8.

[2] É. Morin, M. Maman, R. Guizzetti, and A. Duda, "Comparison of the device lifetime in wireless networks for the internet of things," *IEEE Access*, vol. 5, pp. 7097–7114, 2017.

[3] W. Yang *et al.*, "Narrowband wireless access for low-power massive internet of things: A bandwidth perspective," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 138–145, 2017.

[4] A. Azari and C. Cavdar, "Self-organized low-power IoT Networks: A distributed learning approach," in *IEEE Globecom*, Nov 2018.

[5] M. E. Soussi, P. Zand, F. Pasveer, and G. Dolmans, "Evaluating the Performance of eMTC and NB-IoT for Smart City Applications," *arXiv preprint arXiv:1711.07268*, 2017.

[6] R. Ratasuk, N. Mangalvedhe, A. Ghosh, and B. Vejlgaard, "Narrowband LTE-M System for M2M Communication," in *IEEE VTC-Fall)*, Sept 2014, pp. 1–5.

[7] 3GPP TR 45.820, "Technical Specification Group GSM/EDGE Radio Access Network; Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)," 2015.

[8] R. Ratasuk, B. Vejlgaard, N. Mangalvedhe, and A. Ghosh, "NB-IoT system for M2M communication," in *IEEE WCNC*, 2016, pp. 1–5.

[9] Y. P. E. Wang *et al.*, "A primer on 3GPP narrowband internet of things," *IEEE Communications Mag.*, vol. 55, no. 3, pp. 117–123, March 2017.

[10] J. Schlienz and D. Raddino, "Narrowband internet of things," Rohde and Schwarz, Tech. Rep., 08 2016.

[11] X. Lin, A. Adhikary, and Y. P. E. Wang, "Random access preamble design and detection for 3GPP Narrowband IoT systems," *IEEE Wireless Communications Letters*, vol. 5, no. 6, pp. 640–643, Dec 2016.

[12] T. Kim, D. M. Kim, N. Pratas, P. Popovski, and D. K. Sung, "An enhanced access reservation protocol with a partial preamble transmission mechanism in NB-IoT systems," *IEEE Communications Letters*, June 2017.

[13] C. Yu *et al.*, "Uplink scheduling and link adaptation for narrowband internet of things systems," *IEEE Access*, vol. 5, pp. 1724–1734, 2017.

[14] M. Lauridsen *et al.*, "Coverage and capacity analysis of LTE-M and NB-IoT in a rural area," in *IEEE VTC Fall*, 2016, pp. 1–5.

[15] A. Adhikary, X. Lin, and Y. P. E. Wang, "Performance Evaluation of NB-IoT Coverage," in *IEEE VTC-Fall*, Sept 2016, pp. 1–5.

[16] Y. D. Beyene, R. Jantti, K. Ruttik, and S. Iraji, "On the Performance of Narrow-Band Internet of Things (NB-IoT)," in *2017 IEEE WCNC*, March 2017.

[17] P. A. Maldonado *et al.*, "Narrowband IoT data transmission procedures for massive machine-type communications," *IEEE Network*, vol. 31, no. 6, pp. 8–15, November 2017.

[18] K. Hammad, A. Moubayed, S. L. Primak, and A. Shami, "QoS-aware energy and jitter-efficient downlink predictive scheduler for heterogeneous traffic LTE networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1411–1428, June 2018.

[19] D. Triantafyllopoulou, K. Kollias, and K. Moessner, "QoS and energy efficient resource allocation in uplink SC-FDMA systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3033–3045, June 2015.

[20] A. Azari and G. Miao, "Network lifetime maximization for cellular-based M2M networks," *IEEE Access*, vol. 5, pp. 18 927–18 940, 2017.

[21] G. Miao, N. Himayat, G. Y. Li, and S. Talwar, "Low-complexity energy-efficient scheduling for uplink OFDMA," *IEEE Transactions on Communications*, vol. 60, no. 1, pp. 112–120, 2012.

[22] A. Azari, "Serving IoT communications over cellular networks: Challenges and solutions in radio resource management for massive and critical IoT communications," Ph.D. dissertation, KTH Royal Institute of Technology, 2018.

[23] H. Malik *et al.*, "Radio resource management scheme in NB-IoT systems," *IEEE Access*, vol. 6, pp. 15 051–15 064, 2018.

[24] B. Martinez *et al.*, "Exploring the performance boundaries of NB-IoT," *arXiv preprint arXiv:1810.00847*, 2018.

[25] A. K. Sultania *et al.*, "Energy modeling and evaluation of NB-IoT with PSM and eDRX," in *2018 IEEE Globecom Workshops*, 2018, pp. 1–7.

[26] J. Lee and J. Lee, "Prediction-based energy saving mechanism in 3GPP NB-IoT networks," *Sensors*, vol. 17, no. 9, p. 2008, 2017.

[27] S. Oh and J. Shin, "An efficient small data transmission scheme in the 3GPP NB-IoT system," *IEEE Communications Letters*, vol. 21, no. 3, pp. 660–663, March 2017.

[28] G. T. R. WG1, "Nb-iot ad-hoc meeting," Tech. Rep., March 2016, sophia Antipolis, France.

[29] M. Chafii, F. Bader, and J. Palicot, "Enhancing coverage in narrow band-IoT using machine learning," in *IEEE Wireless Communications and Networking Conference*, 2018, pp. 1–6.

[30] ——, "Enhancing coverage in narrow Band-IoT using machine learning," in *IEEE Wireless Communications and Networking Conference*, 2018.

[31] S. Ullerstig, A. Zaidi, and C. Kuhlins, "Know the difference between NB-IoT vs. Cat-M1 for your massive IoT deployment," Tech. Rep., 2019, www.ericsson.com.

[32] A. Azari, G. Miao, C. Stefanovic, and P. Popovski, "Latency-energy tradeoff based on channel scheduling and repetitions in nb-iot systems," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2018.

[33] K. Wang, J. A. Zarate, and M. Dohler, "Energy-efficiency of LTE for small data machine-to-machine communications," in *2013 IEEE ICC*, June 2013, pp. 4120–4124.

[34] O. J. Boxma, O. Kella, and K. Kosiński, "Queue lengths and workloads in polling systems," *Operations Research Letters*, vol. 39, no. 6, pp. 401–405, 2011.

[35] 3GPP TSG- RAN1 AdHoc NB-IoT, "NPDSCH resource allocation," Tech. Rep., March 2016.

[36] H. Akimaru and K. Kawashima, *Teletraffic: theory and applications*. Springer Science and Business Media, 2012.

[37] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Transactions on Communications*, March 2019.

[38] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[39] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," in *2017 IEEE Globecom*, Dec 2017, pp. 1–7.

[40] R1-1808304, "Discussion on the reliability enhancement for grant-free transmission," Tech. Rep., Aug. 2018, 3GPP TSG-RAN1 Meeting 94, Gothenburg, Sweden.

[41] M. Masoudi *et al.*, "Grant-free radio access IoT networks: Scalability analysis in coexistence scenarios," in *IEEE ICC*, 2018, pp. 1–7.