Quantifying NB-IoT Performance in 5G Use-Cases With Mixture of Regular and Stochastic Traffic

Pavel Masek, *Member, IEEE*, Dmitri Moltchanov, Martin Stusek, Radek Mozny, Yevgeni Koucheryavy, *Senior Member, IEEE*, and Jiri Hosek, *Senior Member, IEEE*

***Abstract*—The increasing demand for power distribution systems in terms of control with nearly immediate response requires deploying a new type of user equipment (UE) that demands permanent connectivity. In Narrowband Internet of Things (NB-IoT) systems, the traffic generated by such UEs may constitute a large part of the overall load. In this article, we first propose a detailed 2-D Markov chain model to capture the system's behavior with the mixture of conventional stochastic and regular traffic types. To provide a computationally efficient solution, we then apply the state aggregation technique to reduce it to a 1-D model and develop approximations and associated numerical algorithms for assessing the mean delay when transmitting the considered traffic. Our results show that a single NB-IoT cell remains stable for up to 72×10^4 conventional UEs and 9×10^3 UEs demanding permanent connectivity. The presence of the latter UEs type has a linear effect on their delay, but affects conventional UEs more drastically. A delay bound of 10 s specified in ITU-R M.2410 is met for the conventional UEs, even under a high number of permanently connected UEs (10^3). However, the delay on the side of the latter UEs is violated even for 100 permanently connected UEs requiring redesigning the NB-IoT channel access mechanism or expanding resources.**

***Index Terms*—Delay performance, fifth generation (5G) mobile networks, massive machine-type communication scenario (mMTC), mixture of traffic types, Narrowband Internet of Things (NB-IoT).**

I. Introduction

THE FIFTH generation (5G) mobile wireless systems reformed the way of wireless data transmissions. It not only brings enhanced mobile broadband (eMBB) and ultrareliable low-latency communications (URLLCs) but also targets Internet of Things (IoT) scenarios as the main direction. Narrowband IoT (NB-IoT), recently recognized as a 5G solution, enables the deployment of smart devices for massive machine-type communication scenarios (mMTCs). The NB-IoT technology was standardized back in 2014, and since that moment, it clearly stands as the first widely adopted technology for IoT data transmissions [1]. Using the opportunity to integrate the NB-IoT in the guard band of already deployed long-term evolution (LTE) systems, network operators have updated the network infrastructure at a never-seen rate.

The NB-IoT network infrastructure was ready for the first trials in 2016, and current NB-IoT installations already meet 5G ITU-R M.2410 requirements of delay and density [2]. However, newly introduced use cases induce communication patterns that may significantly affect NB-IoT performance, i.e., the capacity of the NB-IoT base station (BS) (hundreds of connected devices) could be fully utilized only by one customer. Eventually, this would lead to a significant performance drop or even denial of service, resulting in delayed data transmissions and failure to fulfill the schedule of remote reading [3]. The NB-IoT technology was mainly designed for specific purely stochastic traffic patterns, as detailed in the evaluation scenarios provided in ITU-R M.2412 [4]. However, these patterns are mostly true for "legacy IoT applications," i.e., those without the need for permanent network connectivity.

The industry segment defines the recently introduced group of IoT applications, where the emerging mMTC services demand stable, secure, and permanent connection to user equipment (UE). An excellent representative of a new group of devices that brings the need for permanent connectivity is the so-called "smart metering." Such use-case currently includes already deployed electricity meters either without communication modules or equipped with legacy 2G modules, enabling data communication via the general packet radio service (GPRS). Even more demanding is the smart metering deployment strategy in the European Union [5]. Starting from 2024, every deployed electricity meter for locations (houses, factories, warehouses, etc.) with an annual consumption above 6 MWh needs to have a smart meter capability, i.e., an electricity meter with a communication module for remote metering and management.

Recent smart metering enhancements demand that UEs have permanent connectivity through sustainable wireless communication technology. The reason is the utilization of connection-oriented protocols, such as DLMS/COSEM and

Received 22 December 2023; revised 23 July 2024 and 24 October 2024; accepted 5 November 2024. Date of publication 11 November 2024; date of current version 7 March 2025. This work was supported in part by the Technology Agency of the Czech Republic under Grant TN02000067, and in part by the Program National Competence Centre. The work of Dmitri Moltchanov was supported by the Business Finland Ultra Scalable Wireless Access (USWA) Project within the CELTIC-NEXT Programme and "Machine Learning Algorithms for Energy Efficient and QoS Aware Communications in Heterogeneous 6G mmWave/sub-THz Networks" (ML6GThz) funded by the Academy of Finland. *(Corresponding author: Dmitri Moltchanov.)*

Pavel Masek, Martin Stusek, and Jiri Hosek are with the Faculty of Electrical Engineering and Communications and the Department of Telecommunications, Brno University of Technology, 61600 Brno, Czech Republic.

Dmitri Moltchanov and Yevgeni Koucheryavy are with the Unit of Electrical Engineering, Tampere University, 33720 Tampere, Finland (e-mail: moltchanov.dmitri@tuni.fi).

Radek Mozny is with the Faculty of Electrical Engineering and Communications and the Department of Telecommunications, Brno University of Technology, 61600 Brno, Czech Republic, and also with the Unit of Electrical Engineering, Tampere University, 33720 Tampere, Finland.

Digital Object Identifier 10.1109/JIOT.2024.3495698

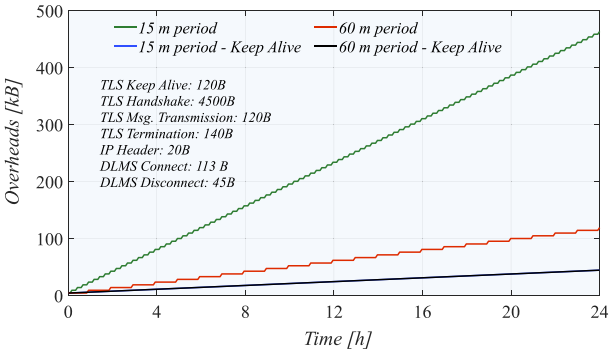


Fig. 1. TLS communication overheads for 24 h.

TCP/TLS, which are de facto standards in electricity metering. These protocols involve initialization procedure that requires significant overhead prior to data transmission [6]. For example, the reinitialization of the connection every 15 min leads to 4 times higher overheads than permanently connected UE in just 24 h, as depicted in Fig. 1, where the query periodicity is set to 15 min. Due to this, smart metering devices remain in "RRC Connected" state and queried over the regular time intervals, which results in bursty traffic load.

Modern low-power wide area network (LPWAN) technologies have been designed with asynchronous purely stochastic traffic in mind. Furthermore, in their current stage, they cannot distinguish between traffic with different arrival patterns and Quality of Service (QoS) guarantees. Thus, new bursty traffic coexists at the air interface with conventional asynchronous traffic. As discussed in Section II, the question of coexistence between bursty and stochastic traffic in NB-IoT has not received much attention thus far, and little is known about the performance of these types of traffic patterns and their impact on each other.

A. Structure of This Article

To provide insights into this growing problem, in this article, we develop a model capturing the specifics of conventional UEs' service process and requiring permanent connectivity at the NB-IoT air interface. We first propose a detailed model based on a 2-D Markov chain that tracks the number of both types of UEs in the system. To reduce the computational complexity, we then utilize state-space aggregation techniques. Further, we introduce a numerical evaluation procedure and accurate approximation for mean delays of the considered traffic types. Using the developed model, we investigate its accuracy and quantify the delay performance as a function of system parameters. The developed model can be utilized to assess the amount of resources that need to be provided to the NB-IoT system to satisfy the prescribed delay guarantees. The two main features of the proposed model are: 1) joint analysis of coexistence between two different types of traffic in NB-IoT system and 2) bursty nature of one of the traffic types. To the best of our knowledge no such models have been considered in the past.

The main contributions of our study are as follows.

1) A mathematical model allowing to capture the system behavior when the conventional and new types of UEs demanding permanent connectivity are coexisting and sharing the NB-IoT air interface.
2) Simplified approximate model based on state aggregation technique and associated numerical algorithm for assessment of the mean communication delay of considered traffic types.
3) Numerical results showing that: a) NB-IoT system is stable for 72×10^4 and 9×10^3 UEs in the systems; b) required 10 s delay performance of permanently connected UEs is violated already for 10^2 such UEs in the system; and c) delay performance of conventional UEs is met for up to 9×10^3 permanently connected UEs.

The remainder of this article is organized as follows. First, in Section II, we overview the new NB-IoT use cases requiring permanent connectivity. The system model is introduced in Section III. Then, we analyze it for performance metrics of interest in Section IV. Numerical results are provided in Section V. Finally, conclusions are drawn in the last section.

II. Background and Related Work

This section first introduces new use cases emerging recently for mMTC communication scenarios requiring synchronous transmission from continuously connected devices. Then, we outline the related work done in the mathematical modeling of random access systems under the superposition of batch and purely random arrivals.

A. New mMTC Use-Cases for 5G

Throughout the last couple of years, newly introduced industrial communication scenarios significantly differ from those the NB-IoT technology was designed for back in the previous decade. As the 3GPP standardization body mainly aimed to cover the Industrial IoT (IIoT) communication scenarios for UEs performing asynchronous data transmissions resulting in purely stochastic traffic patterns, the discussion between the telecommunication operators and the industry companies has begun [7], [8].

The leading industry sector that pushes the boundaries of the "legacy communication schemes" while utilizing the recently adopted Cellular IoT (CIoT) technologies is the one of energy distribution over smart grids (SGs). Not only the photovoltaics were installed rapidly over the last decade, but an entirely new segment has emerged. Electromobility and intelligent (smart) electricity meters both require compliance with defined requirements, where the crucial one is permanent connectivity. To this end, intelligent microenergy grids are being constructed as part of the 5G platform. We point out that not only the communications between the electric vehicles (EVs) and the local charging station (wall boxes, charging stations) are defined. On top of what we see while charging the EVs, the intelligent grid communicates with the electricity distributor to 1) verify the user; 2) approve the charging process; and 3) finalize the charging and process with payment [9]. In addition to charging stations across the highways, there are locations at home or work, e.g., the underground garages,

TABLE I
COVERAGE CLASSES FOR THE ECL ASSIGNMENT

Coverage Class	SINR and RSRP Conditions
Class 0	SINR >7 and RSRP >-110
Class 1	-3 <SINR <7 and RSRP >-133
Class 2	SINR <-3 or RSRP <-133

where the charging stations will need to have permanent connectivity to the distributors' networks [10].

Thus, deployment of the smart meters brings new communication paradigms to the play. The game-changing information is that the reading scenarios defined by the electricity distributors target communication ranging from 5 to 15 min (even more often for specific situations, e.g., on-demand management of the electricity meter in case of critical failure, such as power grid overload), contrary in less demanding cases the periodicity is up to 1 h. Notably, these nonstandard situations requiring fast response times (communication in the downlink (DL) direction) are the primary culprits for permanent (two-way) connectivity needs [11], [12], [13].

B. Smart Metering Infrastructure in the Czech Republic

Consider the whole situation with a concrete example in the Czech Republic, as the research team does have long-term cooperation with the network operators across Europe. The network operator Vodafone Czech Republic finished with the installation of the NB-IoT technology, and with more than 4000 BSs (eNodeBs), they claim the 100% outdoor coverage and the 95% indoor communication coverage. Having in mind the extreme use cases known as the "deep indoor" scenarios, the NB-IoT technology is at this moment the only representative working in the licensed frequency band [band no. 20 (800 MHz)] enabling the communications with the smart meters in locations with poor signal coverage, where the values of the reference signal receive power (RSRP) parameter go below −120 dBm. The selection of the coverage class based on both the RSRP and signal-to-interference plus noise ratio (SINR) is shown in Table I [3].

Having this information, the network operators work on the network capacity planning studies, where the communication patterns and network load are elaborated. Notably, our partner, Vodafone Czech Republic, conducted a thorough study in which they derived the expected performance and capacity of the network. All information in this study stems from the actual NB-IoT network deployment around the city of Hradec Králové in East Bohemia, Czech Republic. Due to the symmetric nature of the DLMS protocol, as it is strictly a request–response system, a "100 B / 100 B" traffic model for the rough estimation of network capacity was considered. Current findings lead to the two scenarios for two extended coverage levels (ECLs) distributions (ECL 0, ECL 1, ECL 2): 1) 60%, 30%, and 10% and 2) 80%, 15%, and 5%. Unfortunately, these findings stem from the tight cooperation of the research team with the network operator, and the results are not publicly available. However, similar studies verify our findings, such as [14], where the distribution of ECL for water meters is nearly identical to the first scenario.

TABLE II
RESULTS FOR 100 B / 100 B (UL / DL) TRAFFIC MODEL FOR TWO ECL DISTRIBUTIONS [20], [21], [22], [23]

Percentage of served users	ECL distribution (60 %, 30 %, 10 %)	ECL distribution (80 %, 15 %, 5 %)
In 5 min	68.9%	86.6%
In 10 min	89.5%	97.7%
In 15 min	96.1%	99.6%

Considering the NB-IoT deployment in the guard-band frequency spectrum (180 kHz carrier and 20 kHz guard band), the theoretical cell capacity is ∼ 10150 devices per hour in case of the 60%, 30%, and 10% distribution. As expected, in the case of the second distribution, where the number of smart meters in the ECL 0 increases, i.e., 80%, 15%, and 5%, the cell capacity is significantly higher; ∼ 19253 devices per hour can be served. For both groups of ECL distributions, the calculations of the delivery time were performed. Let us differentiate between two situations. When the smart meters send data asynchronously, the data from all smart meters will be delivered with an end-to-end delay ranging from 250 ms to 10 s, which does not exceed the 15 s threshold defined by the electricity distributors. Alternatively, when the smart meters send data synchronously, the data will be delivered with a delay in the range from 250 ms to x (in units of minutes) where x depends on the number of simultaneously connected devices in the cell [15], [16].

Continuing with the description of the new communication scenario for the smart meters, we focus on the approximate number of smart meters connected to the eNodeBs. The average number of the locations equipped with the smart meter and covered by one BS (eNodeB) is 800 in the case of the mid-size city and 350 if one targets the rural area [17], [18], referring to the conducted study. At this point, we draw the calculations of how many users (smart meters) can be served within the defined period of time; see Table II. In our calculations, the NB-IoT channel is estimated based on the traffic model and UE distribution model defined in 3GPP TR 45.850, without the use of the power saving mode (PSM) by the smart meters so they stay in the IDLE mode with active radio reception all the time. The reason for omitting the PSM is the need for remote control of the load of the smart electricity meter and individual relays of the meter [19].

Even if the communication scenarios mentioned above seem strict, the electricity distributors define even more challenging on-demand use cases or situations of communication recovery after the blackout. For example, the recovery scenario after a blackout in the case of the NB-IoT networks adds the 80 B in the case of the uplink (UL) attach procedure and 30 B for the DL direction as the synchronization procedure and status update. The results for this scenario are shown in Table IV.

To conclude this part, we can state the NB-IoT network capacity meets the 3GPP TR 45.820 NB-IoT cell capacity model requirements [24], [25]. However, in the case of the on-demand scenarios or the recovery situations, these requirements are no longer satisfied. The communication with the smart meter must be successfully managed within 15 min for

TABLE III
RESULTS FOR THE RECOVERY SCENARIO [ATTACH + 80 B (UL) + 30 B (DL)]–FOR TWO ECL DISTRIBUTIONS [15], [16]

Percentage of served users	User ECL 0:1:2 distribution	
	(60 %, 30 %, 10 %)	(80 %, 15 %, 5 %)
In 5 min	43.2%	62.9%
In 10 min	67.1%	85.2%
In 15 min	80.3%	93.9%
In 20 min	88.3%	97.2%
In 25 min	93.0%	98.7%

both the regular readings and blackout situations. To meet this threshold, the network capacity expansion based on: 1) the number of deployed smart meters; 2) radio conditions; and 3) forecasted traffic patterns can be made either by logical channel optimization or by adding another NB-IoT channel within the frequency spectrum (2nd guard band or even the stand-alone carrier). To provide a realistic assessment of the required capacity of the NB-IoT system servicing both conventional stochastic traffic and new regular traffic, detailed performance evaluation models are needed [13].

C. Related Work

Systems with random access have been known since the 70s of the last century. The first wave of studies considered systems with many UEs, each having low-message intensity. By tending the number of UEs to infinity while the message intensity per UE to zero such that the overall message intensity is constant, the message arrival process in such a system is approximately Poisson in nature. Such a system with random access, an infinite number of UEs, and a Poisson arrival process has been first studied in [26] and [27]. In [28], two methods suitable for such systems have been proposed: 1) Markov chain approach and 2) diffusion approximation.

The first wave of mMTC communications technologies, including NB-IoT, has been developed by utilizing the assumption of purely random asynchronous message arrivals from UEs as specified in the reference test model in ITU-R M.2410. Only recently, the research community's attention has been attracted to random access systems under batch traffic arrival patterns. These studies have been inspired by pure research interest in accessing the response of random access mechanisms to synchronous access attempts. In contrast, the second wave of interest resulted from emerging first-generation mMTC technologies, including EC-GSM, SigFox, LoRaWAN, IEEE 802.11ah, and NB-IoT. Notably, Van Houdt and Blondia [29], [30] considered different random access mechanisms under batch arrivals and discovered that the message transmission delay increases exponentially with the size of the batch for the same arrival traffic intensity.

Stefanovic et al. [31] considered the contention-resolution random access algorithms under the batch arrivals of messages. The developed model is based on direct application of the Markov chain modeling approach and shows a good match with computer simulations. However, such random access algorithms are not utilized in mMTC communications. The authors have studied a similar random access mechanism in [32], where guidelines for designing access probabilities and analytical calculation of the error probability have been proposed. A logical continuation in this work is provided in [33], where the authors consider random access in LTE advance (LTE-A) technology with nonorthogonal multiple access (NOMA) enhancements. They compare the reception of pilot signals for successive joint decoding (SJD) and interference cancelation (SIC) at the random access phase.

An accurate treatment of the batch arrivals appears in [34], where the authors explicitly assumed batch arrivals from several independent sources. The effect of batch arrivals in IEEE 802.11ah systems has been discussed in [35] and [36], while a particular case of batch arrivals in satellite communications is considered in [37]. In the context of access barring, LTE-M and NB-IoT technologies under batch arrivals have been analyzed in [7]. Notably, most of these studies assume the presence of only one type of traffic at the air interface, and often, the number of UEs is considered to be finite. This assumption allows us to formulate the performance evaluation models in terms of Markov chains with finite transition probability matrices and then resort to the standard analysis methods.

Recently, with the emergence of new applications, the interest in mMTC access schemes has revived. Specifically, Chowdhury and De [38] proposed a delay-aware priority access classification (DPAC)-bases access class barring (ACB) scheme introduced for LTE advanced (LTA-A) for LTE random access preamble congestion control aiming at delay-constrained mMTC devices. The proposed method with further RL enhancements showed an increase in preamble success probability compared to the ACB approach. However, in terms of constrained scenarios with limited preambles and a large number of preamble arrivals, the tradeoff came with higher delays and increased device drops compared to ACB and other traditional static approaches. In terms of power consumption optimization, Rostami et al. [39] investigated wake-up-based DL access for delay-constrained devices in 5G. The proposed wake-up scheme (WuS) can coexist with DRX schemes and achieve desired power savings in DRX periods. This approach seems more suitable for mMTC nonperiodic traffic nature by being more robust and agnostic by the traffic type. On the other hand, Sun et al. [40] introduced power and resource-saving by offloading methods based on mobile edge computing (MEC), implementing a digital twin (DT) edge network aided by deep reinforced learning (DLR). This approach is promising in terms of resource offloading optimization and delay optimization during device mobility.

Metzger et al. [41] demonstrated that even conventional LPWAN traffic may indeed exhibit batch behavior owing to aggregation. This is related to the presence of bursts/clustering even in conventional Poisson processes [42] as well as traffic aggregates with not fully asynchronous sources. While the model considered in their study is rather simple and does not account for specifics of the random access procedure (RAP), their study actually extends the applicability of the model proposed in this article. Differently from [41], we consider burstiness happening as a result of special operation

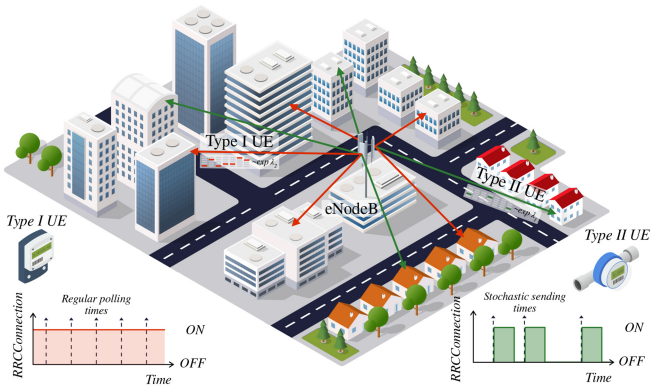


Fig. 2. Two types of UE devices in NB-IoT network: *Type I* UEs are characterized by regular polling times and *Type II* UEs are conventional UEs with stochastic arrival pattern.

of application layer protocols performing regular polling of UEs. Also, we account for random access phase and actually demonstrate that it serves as the bottleneck when mixture of conventional Poisson traffic and bursy traffic is served at the air interface.

The abovementioned review illustrates that the systems under mixed stochastic and regular arrivals have not been considered in detail yet. Furthermore, there are no studies addressing the performance of the NB-IoT system, recently accepted by ITU-R as an enabling technology for 5G mMTC under such load conditions.

III. System Model

In this section, we formulate our system model by specifying its components, including deployment, traffic, and access procedures. Then, we introduce our assumptions and define metrics of interest.

A. Deployment Model

We consider a single NB-IoT cell with one resource block (RB) allocated for operation. The NB-IoT operation is assumed to be in guard-band or stand-alone mode. The network deployment is considered to be well provisioned by the network operator, i.e., there are no UEs in the cell with borderline channel conditions. Such kind of well-provisioned deployment is typical for cities. For this reason, in what follows, we assume that the loss of messages due to incorrect reception is negligible. In such conditions, the access phase becomes the dominant factor for performance degradation.

B. Traffic Types

We consider two types of UEs utilizing the resources of the NB-IoT system; see Fig. 2. The first, Type I UEs, are those explicitly controlled from the remote application server and, thus, they are always awake and keep their radio interface active. For this reason, they are in a "RRC connected" state during the whole operation lifetime, as discussed later. These types of devices represent an evolution of current smart metering devices predominantly utilizing the communication channel asymmetrically. These new devices allow

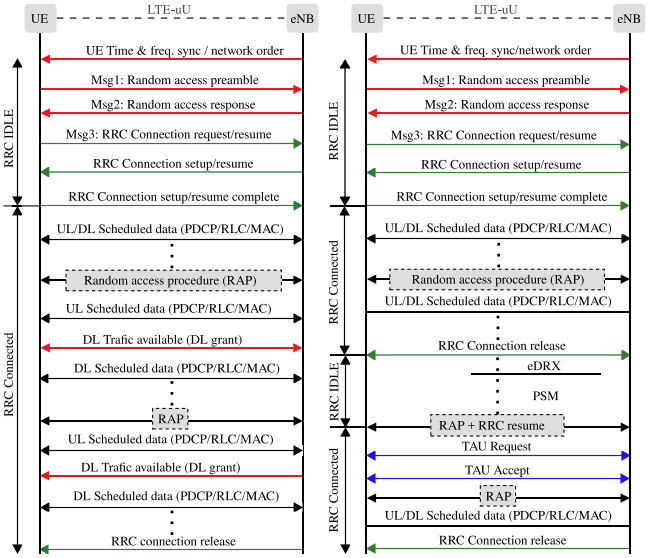


Fig. 3. NB-IoT access: Left–*Type I* UEs and right–*Type II* UEs.

for proper two-way communication with minimal response time [12], [43], [44].

The second type, Type II UEs, is a conventional UEs going through the "sleep-awake-transmit" cycle. These UEs may represent sensors performing periodic measurements, and their operational cycles are assumed to be asynchronous. When the number of UEs tends to infinity and the probability of message generation at each UE tends to zero, as is the case for conventional asynchronous sensors in CIoT systems, the aggregate Type II UEs traffic can be approximated by the homogeneous Poisson process with intensity λ_2 messages/s. [26], [27], [45]. In this article, λ_2, represents the aggregated load from Type II UEs that, depending on the interpretation, may reflect different densities of UEs in the coverage area of BS generating messages according to ITU-R M.2410 recommendation or may represent different message generation intensities at a single UE for a given density of UEs.

C. Random Access Phase in NB-IoT

By following NB-IoT specification [46], see Fig. 3, a UE is assumed to determine NB-IoT carrier by measuring the power of the received synchronization signals in the DL direction and performing time and frequency synchronization together with Cell ID decoding. The time interval between synchronization information repetition may vary between 24 and 2604 ms [47]. Next, the narrowband physical broadcast channel (NPBCH) that carries the master information block (MIB) for 640-ms transmission time interval (TTI) is decoded. Also, additional information about the cell characteristics is transmitted to the SIB1-NB for 2560 ms and other SIB2-NB information from the BS. More details can be found in [48].

Once synchronized, UE can configure the narrowband physical random access channel (NPRACH) and perform UL transmission of preambles according to the network settings so that both the number of repetitions and the transmit power are sufficient. The number of preamble repetitions can vary between 1 and 128. One preamble with the deterministic

hopping pattern within a repetition unit consists of four groups of characters. Each group consists of five characters and a cyclic prefix (66.67 or 266.7 μs for 10 or 40 km cell radius, respectively). For this reason, the random access attempt duration ranges between 5.6–819.2 ms [49]. Upon reception at the BS side, it can correct the frequency and time offset and estimate timing advance (TA) for upcoming transmissions. NB-IoT specifies the minimum number of orthogonal preambles (subcarriers) to be $l = 12$ out of 48 available. The data transfer phase is initiated in the narrowband physical DL control channel (NPDCCH), which is utilized to transmit the DL control information (DCI). Repetitions of this signal can range from 1 to 2048 times [50].

Following the above-mentioned access phase, the initial transition from RRC IDLE to RRC connected state followed by message transmission is identical for both Type I and Type II devices. Notably, in this case, both device types undergo a RAP in the initial connection setup as well as in subsequent message transmission. However, the different nature of the device types arises after this point. In contrast with the Type I UE, the Type II device releases an RRC connection after each message transmission. Thus, if Type II UE wants to transmit the data, it has to complete the transition from RRC IDLE to RRC connected state by performing an RRC resume procedure, which includes another RAP. Then, the Type II UE can request UL/DL grants by completing RAP followed by message transmission. On the other hand, Type I devices maintaining a permanent RRC connection undergo only a single RAP to acquire UL/DL grants with actual message transmission [50].

D. Modeling Assumptions

We consider a slotted system with time divided into slots, as illustrated in Fig. 4. The slot duration is $\delta = 10$ ms, which equals the NB-IoT frame and specifies the minimum random-access time without repetitions. The remaining parameters are measured using δ as the basic unit. As described previously, the message transmission process in modern CIoT systems, including the NB-IoT, consists of synchronization, random access, and data transmission phases. The former phase has a fixed duration t_S. Modern CIoT systems that typically transmit small UL data, as specified in ITU-R M.2410, are designed such that there is no bottleneck in the data transmission phase. Taking this fact into account and considering that we concentrate on a mixture of regular and stochastic traffic, we assume that the random-access phase is the main bottleneck. Thus, we assume that the data transmission time, t_{data}, is fixed, whereas the random access duration t_R is random. To this end, Fig. 4 only shows the parameters related to the random-access phase. Note that we also account for the message processing time t_H. The concrete values of these parameters are aligned with the NB-IoT standard and our measurements [51] and are provided later in Section V.

There is a fixed number of Type I UEs under cell coverage. These UEs are queried by the applications at the beginning of the regular time interval (period) of duration Δ, measured in the number of NB-IoT frames. This time interval depends on the application type, and can be on a timescale of minutes to hours. Upon request, all Type I UEs attempt message transmission by performing the NB-IoT RAP. Type II messages arrive according to a Poisson process with an intensity λ_2.

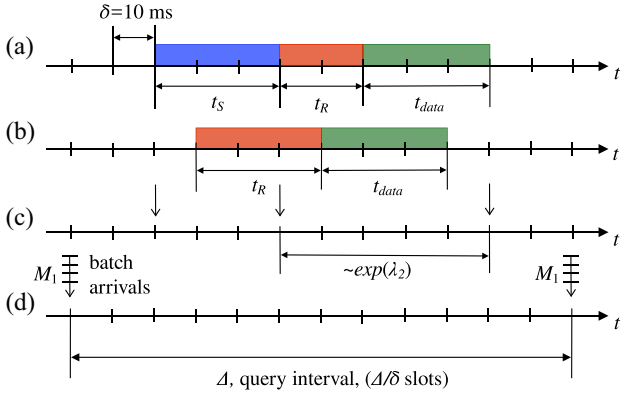


Fig. 4. Timing diagram of system operation: (a) actions performed by Type II UEs, (b) actions performed by Type I UEs, (c) arrival pattern of Type II UEs, and (d) arrival pattern of Type I UEs.

Upon generating a message, both types of UEs initiate the random-access procedure in the next time slot. We assume that the probability of starting the random-access procedure is $p = \min(n/k, 1)$, where n is the number of preambles and k is the number of UEs that compete in a slot. As shown in [52] and [53] this approach minimizes the channel access delay of UEs. As shown in Fig. 4 Type II messages do not need to perform the synchronization procedure, as the connection is always up. Despite arriving in batches at the beginning of period Δ, the actual time when they manage to send their messages, depends on the collision resolution.

E. Metrics of Interest

We are interested in assessing the mean delay of Type I and Type II traffic, $E[\tau_1]$ and $E[\tau_2]$, respectively. We characterize the full delay, including the time required for synchronization, random access, and data transmission.

IV. Performance Analysis

In this section, we first introduce the 2-D Markov chain framework for modeling the coexistence of Type I and Type II UEs in the NB-IoT cell. Then, to reduce the dimension of the model and, thus, the computational complexity, we utilize the state aggregation technique. Further, we present the solution algorithm and finally provide an assessment of the metrics of interest.

A. Exact Model

The model introduced in Section III can be described by 2-D Markov chain $\{N_t^1, N_t^2, t = 1, 2, \ldots, \delta\}$, where N_t^1 and N_t^2 are the number of Type I UEs and Type II UEs in the system at time t, respectively. The index set of the process is discrete with slot duration Δ/δ, where Δ is the number of

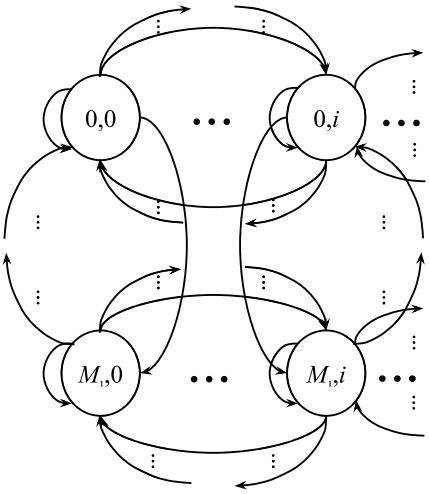


Fig. 5. State transition diagram of 2-D model.

slots between message arrivals from Type I UEs, δ is the time slot duration.[1] The process is defined on the state-space

$$N_t^1 \times N_t^2 = \{0, 1, \ldots, M_1\} \times \{0, 1, \ldots\} \tag{1}$$

where M_1 is the number of Type I UEs in the system.

The structure of the state transition diagram of the 2-D model is shown in Fig. 5. Note that the transitions are limited by l states corresponding to the maximum number of available preambles. The reason is that no more than l messages can be processed simultaneously due to restrictions on the number of orthogonal channels in NB-IoT. We can define the following relation for Type I UEs:

$$N_{t+1}^1 = N_t^1 - T^1\left(N_t^1, N_t^2\right) \tag{2}$$

where $T^1(N_t^1, N_t^2)$ is the number of transmitted Type I UE messages and N_t^2 is the number of Type II UEs that remain in the system at the end of the previous period Δ.

Similarly, for Type II UEs we have

$$N_{t+1}^2 = N_t^2 - T^2\left(N_t^1, N_t^2\right) + V_t^2 \tag{3}$$

where N_t^2 is the number of Type II UEs that remain in the system at the end of the previous period Δ, V_t^2 is the number of Type II UEs that become active in the slot, and $T^2(N_t^1, N_t^2)$ is the number of successfully transmitted Type II UE messages.

Note that (2) establishes the relationship between the number of Type I UEs and the previous and next Markov points required for the parameterization of the Markov chain according to the standard process. One further needs to estimate probabilities $\Pr\{N_t^1 = n | N_{t-1}^1 = m\}$ using the relation (2). The same applies to (3). When finding these probabilities determining the values of $T^1(N_t^1, N_t^2)$ and $T^2(N_t^1, N_t^2)$ specifying the number of Type I and Type II messages served between two Markov time points is the most complex thing as both of them depends on the number of Type I and Type II UEs at the previous Markov point.

[1]Note that in NB-IoT and LTE-M systems δ is the frame duration.

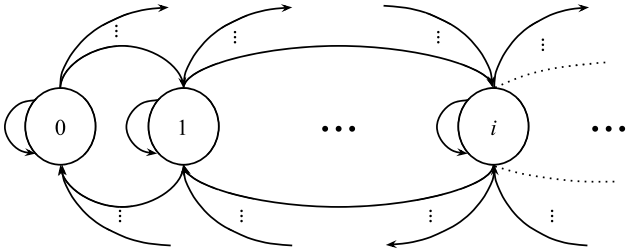


Fig. 6. State transition diagram of 1-D model.

To solve the introduced model, one needs to construct a transition probability matrix, use it to solve a system of linear equations, and finally find stationary probabilities of states of a 2-D Markov chain. Stationary probabilities will further allow us to estimate system parameters, such as the mean number of active UEs of both types in the system and the associated mean delays. However, the state space of such a process is very large, while transition probabilities between states of the process involve combinatorial calculations that are impossible to estimate for practical numbers of Type I and Type II UEs in the system. Thus, while it is possible to describe this model mathematically, obtaining the final characteristics with a reasonable number of Type I UEs in the system is difficult. Thus, we will apply state aggregation technique [54], [55] to reduce the dimension of the process and provide a simple numerical algorithm that practitioners can utilize in their work.

Note that the state aggregation technique might be applied in the following two ways: 1) by analyzing the structure of the state transition diagram and identifying sets of states with similar properties and then replacing them with a macro state or 2) by utilizing different granularity of the system, i.e., by omitting some of the details of the original system. In our study, we utilize the second approach by considering the aggregated amount of Type I and Type II UEs in the system in Section IV-B as compared to tracking individual types of UEs in this section. Thus, the constructed Markov process defined in what follows is the aggregated version of the exact process in this section. We then utilize it to develop approximations for mean delays Type I and Type II UEs.

B. Approximation via State Aggregation

Consider the 1-D Markov chain model $\{N_t, t = 1, 2, \ldots, \delta\}$, where N_t is the overall number of UEs of both types in the system at time t. Thus, the state-space of the process is simply $N_t = \{0, 1, \ldots\}$. Recalling our assumptions, it is easy to see that the choice of the next state depends only on the current one, implying that the process $\{N_t, t = 1, 2, \ldots, \delta\}$ is Markov in nature.

The structure of the state transition diagram of the approximate model is provided in Fig. 6. The main difference compared to the 2-D model in Fig. 5 is that the derivation of transition probabilities is now simpler as one should not differentiate between two types of UEs. Such behavior, however, complicates the derivation of the metrics of interest. Notably, the latter may only be obtained in terms of bounds and approximations, as discussed below.

We now proceed with parameterizing the proposed model. One may observe that the number of active Type I and Type II UEs at the time slot $t+1$ is related to the number of active UEs in the time slot t as follows:

$$N_{t+1} = N_t - T(N_t) + V_t \tag{4}$$

where V_t is the number of Type II UEs that become active in a single slot and $T(N_t)$ is the number of successful message transmissions when there are N_t messages ready for transmission.

Using (4), one may calculate the transition probabilities of $\{N_t, t = 1, 2, \ldots, \delta\}$ by accounting for all the possible transitions from state i to state j. By accounting for the fact that the system enters state i if the number of successfully served UEs in the slot is exactly $i-j$, but no more than the number of preambles l, we have the transition probabilities in the following form:

$$p_{ij} = p\{j \mid i\} = \sum_{m=\max(0,i-j)}^{l} \frac{\lambda_2^{m-(i-j)}}{(m-(i-j))!} \times e^{-\lambda_2} \Pr\{T(i) = m\} \tag{5}$$

where m is the number of successfully transmitted messages, the maximum value of m is $\max(0, i-j)$; in this case, $m-(i-j)$ new active UEs enter the system. The upper sum limit is specified by l, as no more messages than the number of orthogonal channels l can be successfully transmitted in a slot. We also note that the expression under the sum in (5) consists of two terms. The first term, unconditional probability, determines the probability of a certain number of new UEs entering the system in a slot. The second term, conditional probability, determines the probability of the number of successfully transmitted messages.

Note that in (5) there is the term $\Pr\{T(i) = m\}$ specifying the probability of successfully served m UEs in the slot given that there are i active UEs with a message ready for transmission in the system. This quantity takes on 0 when $N_t = 0$. Otherwise, we have

$$\Pr\{T(i) = m\} = \sum_{k=0}^{i} C_i^k p^k (1-p)^{i-k} P(l, k, m) \tag{6}$$

where C_i^k is the shortcut for binomial coefficient, $C_i^k = \binom{k}{i}$.

In (6), p is the probability that the active UE transmits in the current slot that is obtained as follows. Observe that each UE can choose a specific preamble with the probability $1/l$. Hence, if there are more preambles than active UEs in a slot, each UE transmits with a probability 1. Contrarily, if there are more active UEs than preambles, then the likelihood of message transmission from UE will be less than 1. By following [53], we assume that UE follows an optimal scheme with the following transmission probability:

$$p = \begin{cases} 0, & \text{if } i = 0 \\ \min(l/i, 1), & \text{otherwise.} \end{cases} \tag{7}$$

The final term in (6), $P(l, k, m)$, is the probability of the distribution of k messages over l channels, such that m channels are selected by exactly one UE. We derive this unknown in the next section as a part of the solution algorithm.

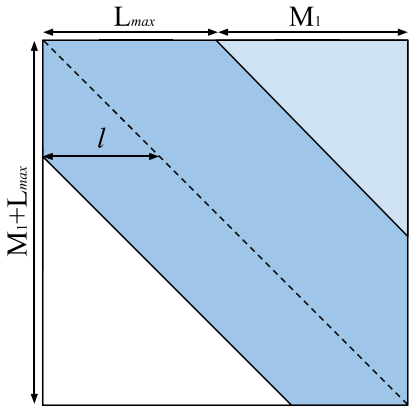


Fig. 7. Structure of the transition probability matrix.

C. Solution Algorithm

Even relying upon the approximate 1-D model, the direct calculation of transition probabilities in (5) is still complicated. The reason is the presence of complex binomial coefficients of high order in $P(l, k, m)$. In what follows, we will develop a simple computational algorithm for these probabilities.

One may observe that the probability $\Pr\{T(i) = m\}$ can be decomposed into the following components. First, for the case of $i = 0$, we have $P(l, k, m) = 0$, since k is the total number of messages posted on l channels for any number of m channels that the UEs select. In the case of $i = 1$, we have only one component $P(l, k, m)$ different from zero. Other components $C_i^k p^k (1-p)^{i-k}$ define a binomial probability that k out of i messages are successfully transmitted. In order to go through all the i states, we take the sum over all values from 0 to i. Furthermore, when $i \to \infty$ the following approximation holds:

$$\Pr\{T(i) = m\} = C_l^m e^{-m} \left(1 - e^{-m}\right)^{l-m} \tag{8}$$

where m is the number of successful transmissions. The rationale for (8) is that when the number of UEs tends to infinity, the probability that a particular preamble will be chosen by only one UE is e^{-1}. Accordingly, the probability that m preambles from l are selected by only one UE follows the binomial law with the success probability e^{-1}.

The only missing parameter is now $P(l, k, m)$, i.e., the probability of the distribution of k messages over l channels, such that m channels are selected by exactly one UE. By following [56], we develop the solution for estimating exact values of stationary probabilities $P(l, k, m)$ in Appendixes A-C. This solution is readily provided by

$$P(l, k, m) = \frac{(-1)^m l! k!}{l^k m!} \times \sum_{f=m}^{\min(l,k)} \frac{(-1)^f (l-f)^{k-f}}{(f-m)!(l-f)!(k-f)!}, \quad m \le k \tag{9}$$

and $P(l, k, m) = 0$, $m > k$.

The solution of the model involves finding stationary probabilities of the 2-D Markov chain. It is convenient to represent the probabilities obtained by (5) in the form of a square matrix of transition probabilities over the states of the Markov chain as shown in Fig. 7. For practical calculations, the number of

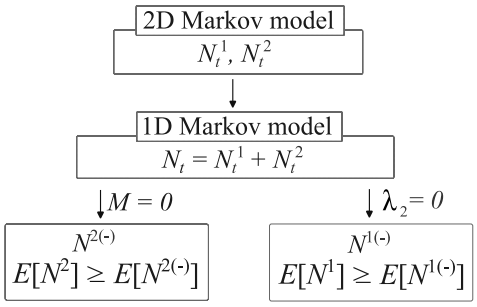


Fig. 8. Overall analysis procedure.

rows and columns must be truncated to, i.e., $M_1 + L_{\max}$. Here, M_1 represents the number of Type I UEs in the system, $L_{\max}$ is the values chosen such that the sum of elements in the first row is approximately 1 with a given precision, e.g., 10^{-6}.

D. Metrics of Interest

Recall that we started with a 2-D Markov model defined over two variables – the number of Type I and Type II UEs, N_t^1 and N_t^2, respectively. Then, as this model is associated with high-computational complexity, we utilized the state aggregation and defined a simplified 1-D Markov chain model that keeps track of the evolution of the joint variable, $N_t = N_t^1 + N_t^2$. This model, however, allows us to explicitly determine the mean aggregated number of UEs in the system $E[N]$ and the associated delay $E[D]$. Thus, from now on, we develop bounds on the number of Type I and Type II UEs in the system. The approach is to first consider the system without Type II UEs, calculate the mean number of Type I UEs, and then analyze the system without Type I UEs to determine the mean number of Type II UEs in the system. Both systems form a separate 1-D Markov chain. The overall approach is illustrated in Fig. 8.

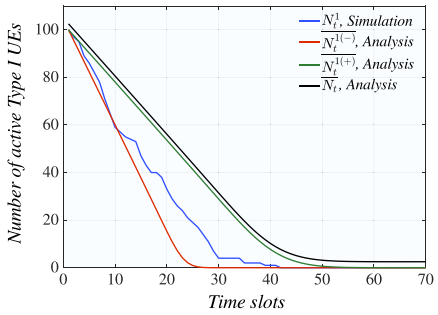
Consider the transition probability matrix of the 2-D Markov chain, which is entirely determined by the parameters λ_2 and M_1. Assume that the distribution of the number of Type I and Type II UEs (N_1^1, N_1^2) is known in the initial time slot when all Type I UEs become active. Throughout this section, we utilize Fig. 9 to show the behavior of the investigated intermediate metrics. Specifically, it compares the intermediate metrics of interest for the mean number of Type I and Type II UEs in the system as a function of time to a single run of computer simulations for $\lambda_2 = 2.5$, $M_1 = 100$.

By utilizing the transition probability matrix, one may obtain the distribution of pair (N_t^1, N_t^2) in each subsequent time slot, $t = 2, 3, \ldots, \delta$ and, thus, determine their mean values, $E[N_t^1]$ and $E[N_t^2]$, respectively. Thus, the number of active UEs over the time interval Δ can be provided as

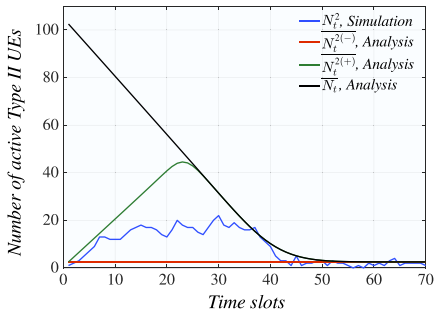
$$E\left[N^1\right] = \sum_{t=1}^{\delta} \frac{E\left[N_t^1\right]}{\delta}, \; E\left[N^2\right] = \sum_{t=1}^{\delta} \frac{E\left[N_t^2\right]}{\delta}. \tag{10}$$

Now, by utilizing Little's result, we have the following for Type I and Type II UEs delays $E[D_1]$ and $E[D_2]$:

$$E[D_1] = \frac{E\left[N^1\right]\delta}{M_1}, \; E[D_2] = \frac{E\left[N^2\right]}{\lambda_2}. \tag{11}$$



(a)



(b)

Fig. 9. Number of active UEs in the system during Δ. (a) Mean number of active UEs Type I in the system. (b) Mean number of active UEs Type II in the system.

Taking into account the specifics of the 2-D Markov model described in Section III, the metrics of interest, $E[\tau_1]$ and $E[\tau_2]$, measured in seconds, are calculated as

$$\begin{aligned} E[\tau_1] &= (E[D_1]t_R + t_{\text{data}})10^{-2} \\ E[\tau_2] &= (t_S + E[D_2]t_R + t_{\text{data}})10^{-2} \end{aligned} \tag{12}$$

where the factor 10^{-3} accounts for the slot duration (NB-IoT frame of duration $\delta = 10$ ms), t_S is the synchronization duration, t_R is the random access duration, and t_{data} is the data transmission time. Note that the principal difference between Type I and Type II traffic in (12) is that the former do not require synchronization phase as the connection is always up.

Note that the method based on a 2-D Markov chain is associated with rather high-computational complexity. To this aim, we propose upper bounds on the mean delay performance of Type I and Type II UEs. Similar to the 2-D case, consider the state of the transition probability matrix in the first slot of the interval. This state depends on the intensity of messages from Type II UEs, λ_2, and the number of Type I UEs, M_1. Also, similarly to the 2-D Markov chain case, we can determine the mean number of active UEs of both types in all the subsequent slots, $E[N_t]$. The trajectory of the process $\{N_t, t = 1, 2, \ldots, \delta\}$ is shown in Fig. 9 by black lines. Note that at the beginning of the interval, N_t combines both types of UEs. However, the number of Type I UEs decreases as time progresses, and only Type II UEs eventually start contributing to N_t. Thus, starting from a specific time slot, the distribution of Type II UEs would coincide with the distribution of both types of UE in the remainder of the interval Δ. Therefore, the

mean value $E[N]$ is obtained as

$$E[N] = \sum_{t=1}^{\delta} \frac{E[N_t]}{\delta}. \quad (13)$$

Note that since the 1-D Markov model includes a 2-D one, in the sense of state aggregation, the following holds [57]:

$$E[N] = E\left[N^1\right] + E\left[N^2\right]. \quad (14)$$

Consider now a system where there are no Type II UEs, i.e., $\lambda_2 = 0$. The mean number of active UEs in this system is bounded from below by $\overline{N_t^1}$. The behavior of this metric, denoted by $E[N_t^{1(-)}]$, is illustrated in Fig. 9 by red lines. The value of $E[N^{1(-)}]$ averaged over the period Δ provides the lower bound for $E[N^1]$.[2] This metric can be obtained by substituting $E[N^{1(-)}]$ into (14) to get

$$E[N] > E\left[N^{1(-)}\right] + E\left[N^2\right] \quad (15)$$

and implying that $E[N^2] < E[N] - E[N^{1(-)}] = E[N^{2(+)}]$, that is also illustrated in Fig. 9. Rearranging the terms in (15) and applying the Little's law [58] we arrive at

$$E[D_2] < \frac{E[N] - E\left[N^{1(-)}\right]}{\lambda_2}. \quad (16)$$

Contrarily, consider a system without Type I UEs, i.e., $M_1 = 0$. The mean number of UEs in this system is a lower bound for $E[N_t^2]$, which is also shown in Fig. 9. Similar to the previous case, the corresponding mean values over Δ, $E[N^{2(-)}]$, is a lower bound for $E[N^2]$. As Type I UEs are absent, the distribution of N_t over the whole interval Δ is provided by the defined Markov model and can be found for any given λ_2. By substituting the lower bound for $E[N^2]$ into (14) we arrive at

$$E[N] > E\left[N^1\right] + E\left[N^{2(-)}\right] \quad (17)$$

leading to $E[N^1] < E[N] - E[N^{2(-)}] = E[N^{1(+)}]$ which is also illustrated in Fig. 9. Thus, we finally obtain the mean delay approximation for Type I UEs in the following form:

$$E[D_1] < \frac{\delta\left(E[N] - E\left[N^{2(-)}\right]\right)}{M_1}. \quad (18)$$

Substituting the right-hand sides of (18) and (18) into (12), instead of $E[D_2]$ and $E[D_1]$, provides the upper bounds for the metric of interests, i.e.,

$$E[\tau_1] < \left(\frac{\delta\left(E[N] - E\left[N^{2(-)}\right]\right)}{M_1} t_R + t_{\text{data}}\right) 10^{-3}$$
$$E[\tau_2] < \left(t_S + \frac{E[N] - E\left[N^{1(-)}\right]}{\lambda_2} t_R + t_{\text{data}}\right) 10^{-3}. \quad (19)$$

The computational algorithm is presented in Algorithm 1.

[2]Due to the utilization of a single simulation run in Fig. 9, the simulated behavior of the metrics can cross an analytical lower bound.

Algorithm 1: Calculation of E[N], $E[N^{1(-)}]$ and $E[N^{2(-)}]$

Input: $n \times n$ transition probability matrices $P(\lambda_2)$ and $P(0)$ with given parameter λ_2 and $\lambda_2 = 0$, where $n = M_1 + L_{\max}$.
Output: $E[N]$, $E[N^{1(-)}]$ and $E[N^{2(-)}]$
First stage: obtaining stationary distribution and $E[N^{2(-)}]$.
Step 1. Based on $P(\lambda_2)$ obtain the stationary distribution $P(\lambda_2) \rightarrow \overline{\pi} = \left(\pi^0, \pi^1, \ldots \pi^{n-1}\right)$
Step 2. Based on stationary distribution, calculate $E[N^{2(-)}] = \sum_{i=0}^{n-1} i\pi^i$.
Second stage: calculation of $E[N]$.
Step 1. Form the $(n + M_1) \times 1$ initial state vector $\overline{p_0} = \left(\underset{0}{0}, \underset{1}{0}, \ldots \underset{M_1}{\pi^0}, \pi^1, \ldots \pi^{n-1}, 1 - \sum_{i=0}^{n-1} \pi^i\right)$.
Step 2. **for** $t = 1, 2, \ldots \delta$ **do**
  Calculate the distribution
  $\overline{p_t} = \overline{p_{t-1}} P(\lambda_2)$
  Obtain the mean number of active UEs
  $E[N_t^{2(-)}] = \sum_{i=0}^{n-1} i p_t^{(i)}$
end
Step 3. Calculate $E[N^2] = \sum_{t=1}^{\delta} \frac{E[N_t^{2(-)}]}{\delta}$
Third stage: calculation of $E[N^{1(-)}]$.
Step 1. Form the $(n + M_1) \times 1$ initial state vector $\overline{p_1} = \left(\underset{0}{0}, \underset{1}{0}, \ldots \underset{M_1}{\pi^0}, 0, \ldots 0\right)$.
Step 2. **for** $t = 2 \ldots \delta$ **do**
  Calculate the distribution
  $\overline{p_t} = \overline{p_{t-1}} P(0)$
  Calculate the mean number of active UEs
  $E[N_t^{1(-)}] = \sum_{i=0}^{n-1} i p_t^{(i)}$
end
Step 3. Calculate $E[N^1] = \sum_{t=1}^{\delta} \frac{E[N_t^{1(-)}]}{\delta}$

TABLE IV
DEFAULT SYSTEM PARAMETERS

Parameter name	Symbol	Value
Time slot duration	δ	10 ms
Number of preambles	l	12
The time to synchronization	t_S	30 slots
Temporary collision-free RAP	t_R	4 slots
Message handling time	t_H	1 slot
Data transfer time	t_{data}	35 slots
Interval between Type I UE arrivals	Δ	15 minutes
Number of Type I UEs	M_1	100-10000
Arrival rate of messages from Type II UEs	λ_2	1-100 per sec

V. NUMERICAL RESULTS

In this section, we elaborate on the presented model by assessing the impact of the coexistence of two traffic types in the same NB-IoT cell. We start this section by evaluating the accuracy of the proposed model. Then, we continue with demonstrating the stability region of the considered system. Further, we assess the accuracy of the proposed approximation for the mean delay and, finally, proceed with evaluating and discussing the response of the performance metrics to input system parameters. The parameters utilized in our model are

provided in Table IV, and are derived from typical settings of the operator-grade mobile network [51] and the environment of the NB-IoT simulator [59].

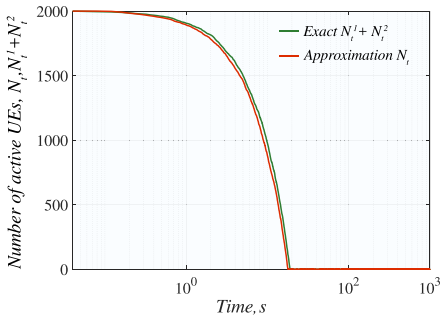
A. Accuracy Assessment

We start by assessing the accuracy of the proposed model. Recall, that there might be two sources of errors due to our assumptions required for analytical tractability and simplicity of numerical evaluation. The first one is related to replacing the detailed 2-D Markov model with a 1-D approximation. The utilization of delay bounds instead of exact delays causes the second one. Here, we address the first one, while in what follows, we assess the accuracy of approximations for the mean delay of messages.

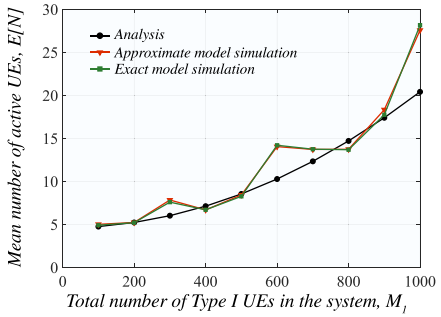
To assess the accuracy of approximating a 2-D model by a 1-D one, it is sufficient to consider the overall number of active UEs in the system. Recall, that we track Type I and Type II UEs separately in the former case, while in the latter one, the overall number of UEs of both types is captured. We utilize the following two-step procedure for benchmarking. First, to ensure that we do not make the mistake of applying the state-aggregation technique, we compare the evolution of the overall number of active UEs during a cycle in "exact" 2-D Markov model ($N_t^1(t) + N_t^2(t)$) and the aggregated number of active UEs in a 1-D model ($N_t(t)$). Then, in the second stage, we already work with time-averaged metrics that are directly related to the final metric–delay. Here, we compare the mean number of UEs in the system $E[N]$ obtained using the simplified 1-D model to those obtained in simulating "exact" 2-D and approximate 1-D model.

Our simulation tool is written in MATLAB. It implements the system model described in Section III, and follows the timings illustrated in Fig. 4. The minimum time granularity is δ–NB-IoT frame duration. Type II messages arrive at the system according to a Poisson process with an intensity λ_2. As the system time is slotted, the arrival time is rounded off to the nearest slot boundary. The beginning of the simulations was also aligned with the first query interval of the duration Δ slots. Type I messages that arrive are batched in the first slot. For each message, the RAP is explicitly simulated, including the random choice of preambles, whereas synchronization and data transmission times are implicitly accounted for by skipping the appropriate number of slots. At each slot, a separate variable is utilized to keep track of the numbers of Type I and Type II messages in the system. Statistical data were collected only in the steady state. To detect the beginning of the steady-state, the exponentially weighted moving average (EWMA) statistic was utilized [60]. Finally, to gather statistical data, we utilized the batch-means method [60].

Fig. 10(a) presents the overall number of active UEs during the single interval Δ obtained by simulating the considered model and averaging over 100 of realizations for 500 Type I UEs and intensity of $\lambda_2 = 0.1$ messages/s from Type II UEs. As one may observe, the approximate 1-D model precisely captures the number of UEs. This finding confirms the hypothesis that the distributional characteristics of the number of UEs in the system for two considered models coincide. Note that we consider the mean delay of Type I and Type II messages as



(a)



(b)

Fig. 10. Numerical assessment of model accuracy. (a) Active UEs during a single cycle. (b) Mean number of active UEs in the system.

the main metric of interest. However, they are unambiguously related to the mean number of active messages or, alternatively, UEs having messages ready for transmission. Thus, the delay will have a qualitatively similar shape.

To assess whether the above-mentioned hypothesis holds for other system parameters, in Fig. 10(b), we show the analysis performed for $E[N]$ using the proposed 1-D model as a function of different numbers of Type I UEs and compare it to the computer simulations of the exact and 1-D model. Note that although we averaged the results of the computer simulations over a number of runs [similar runs have been utilized, and thus exact and simplified models follow the same pattern in Fig. 10(b)], stochastic factors are still involved. As a result, there are crossings between the analysis curve and simulated ones. Importantly, they follow the same trend. When the number of Type I UEs becomes large (e.g., ≥ 1000), the variance of the simulations increases, and deviations become visually larger in Fig. 10(b). Nevertheless, the presented simulation data coincide well, implying that no error is produced by replacing the complex 2-D model with the simpler one.

Note that for large values of the time between Type I UEs messages arrivals to the system, there are long periods of time when the system is almost empty. In fact, there are only Type II UE messages there time after time. However, the overall QoS in terms of mean delay (averaged by both time and UEs having messages ready for transmission) for both types of traffic is severely affected by what happens at the beginning of these intervals. Thus, the metric we assess is the so-called "ergodic" one, averaging the mean number of active UEs (and mean delay) over time as well. At the beginning of the intervals, the instantaneous delay for Type II UEs is drastically higher than at the end of them. However, on average, the delay experienced

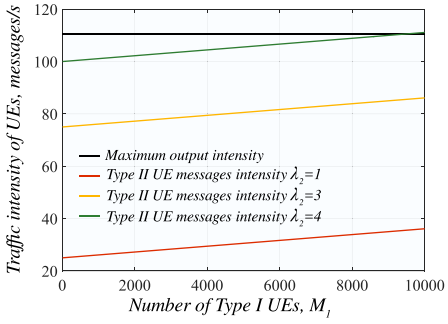


Fig. 11. Stability region of the considered system.

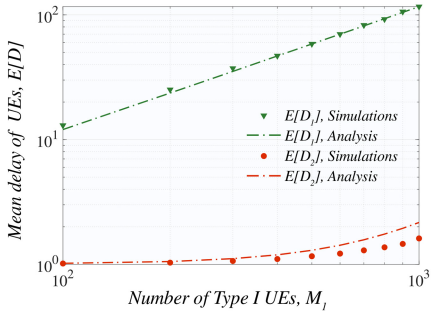


Fig. 12. Delay approximations for considered UE types.

by Type II UEs will be as presented in the steady state. For Type I UEs, the logic is similar, but the delay is higher than that for Type II UEs as they leave the system much earlier than the end of the time period between Type I UEs' arrivals.

B. System's Stability Region

Note that for practical deployments, one has to understand the performance region of the system when the message transmission delay does not tend to infinity. We note that in our system, all the Type I UE messages that are not transmitted by the end of the interarrival period of Type I UE messages are removed and replaced by new ones. However, Type II UE messages are different, and they can be accumulated infinitely. Still, Type I messages add to the overall intensity of Type II messages when they coexist in the system. For this reason, we now continue with the assessment of the system stability region. Recall that the considered system is stable if the maximum output traffic intensity that can be supported by the system is higher than the input traffic intensity. To this aim, Fig. 11 shows output traffic intensity in messages per second versus the number of Type I UEs in the system for multiple Type II UEs message intensities (also in messages per second) and interval between Type I UE arrivals set to 15 m. It also shows the maximum output traffic intensity that can be supported by the system (black line). We see that for approximately 9000 Type I UEs and $\lambda_2 = 0.4$ messages/s, we cross the maximum output traffic intensity, implying that the system starting from this point becomes unstable. For lower values of λ_2 there are also crossing points, but they correspond to much higher numbers of Type I UEs.

Recall that for the transmission time of 40 ms, the Type II message intensity of $\lambda_2 = 4$ messages/s with standard message interarrival time of 2 h specified in ITU-R M.2410 [2] leads to 72×10^4 Type II UEs in the coverage area of a single NB-IoT cell. Thus, the proposed model can be utilized for performance assessment of 5G mMTC services based on NB-IoT having the target of 10^6 UEs/km^2 [61] with up to 9000 Type I UEs demanding permanent connectivity. The stability region for other traffic parameters can be estimated using the proposed model. When planning the network, these constraints need to be accounted for by the network operator. However, we emphasize that system stability does not imply that other performance indicators stay within the specified bounds. We now proceed with delay analysis.

C. Delay Approximations

Having revealed an accurate approximation of the overall number of active UEs in the system by 1-D model and identified the system stability region, we now assess the accuracy of Type I UE message delay approximation based on the proposed model. To this aim, Fig. 12 illustrates the developed approximation for the mean delay of both types of UEs as a function of the number of Type I for $\lambda_2 = 1$ messages/s and $\Delta = 15$ min. By analyzing the presented results, one may observe that the developed approximation accurately matches the actual mean delay experienced by Type I UEs across the whole considered range of the number of UEs. For Type II UEs, the match is good until relatively high values of the number of Type I UEs in the system, e.g., 10^3. For higher values of the number of Type I UEs, the model starts to overestimate Type II UEs delay. We also note that as the number of Type I UEs becomes bigger, the accuracy of delay approximation for Type I UEs becomes better while the accuracy for Type II UEs worsens. The rationale is that under this condition, Type I UEs start dominating traffic aggregation. In fact, a similar property holds for increasing intensity of Type II UEs arrival intensity, see, e.g., Fig. 13(a). However, the impact is not that noticeable as the delay experienced by both types of UEs is mainly affected by the number of Type I UEs.

D. Performance Analysis

Having identified the suitable delay approximation and the system stability region, we proceed with a performance assessment of the delay performance of traffic types. To this aim, Fig. 13 shows the mean delay of both traffic types as a function of the number of Type I UEs and intensity of Type II messages in the system for $\Delta = 15$ m. By analyzing the presented data, one may observe that the intensity of Type II UEs does not produce any noticeable impact on the performance of both traffic types. The rationale is that the type of load produced by these UEs is well distributed in time. Contrarily, by increasing the amount of Type I UEs, the mean delay of both traffic types is negatively influenced. Expectedly, the self-impact is linear, while the mean delay of Type II UEs increases exponentially with the number of Type I UEs. Noticeably, this increase starts to be evident from approximately 500 of UEs; however, this particular value depends on the number of Type I UEs and the cycle time values Δ.

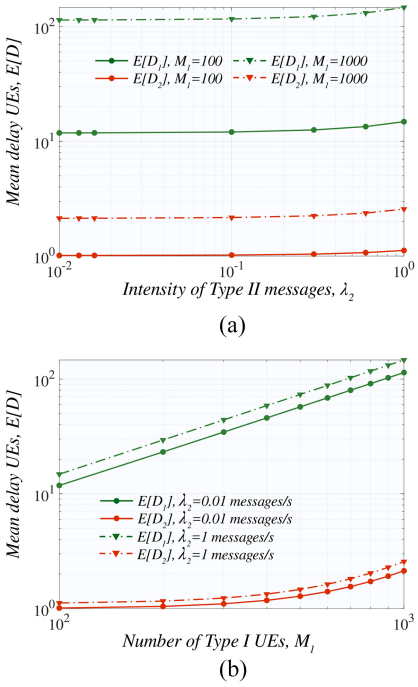


Fig. 13. Delay performance as a function of traffic load. (a) As a function of Type II message intensity. (b) As a function of the number of Type I UEs.

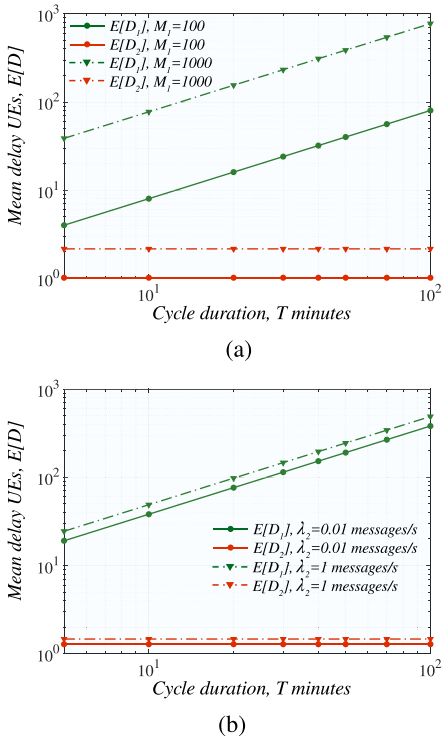


Fig. 14. Delay performance as a function of cycle time. (a) Multiple number of Type I UEs, $\lambda_2 = 0.1$ messages/s. (b) Multiple Type II message arrival intensities, $M_1 = 500$.

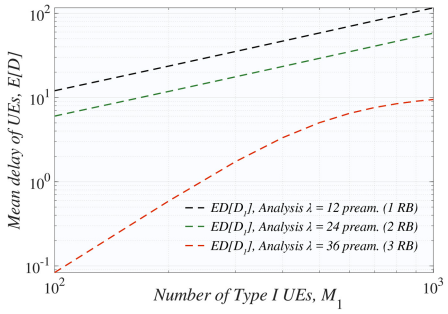


Fig. 15. Mean Type I UE delay for different number of RBs.

Another critical observation is that the presence of Type I UEs affects the compliance of NB-IoT technology with 5G mMTC requirements formalized in ITU-R M.2410 [61]. Recall that according to this recommendation, the delay bound is set to 10 s for 10^6 UEs, assuming that each UE generates at most 1 message in two hours. Furthermore, recall that we have demonstrated that the system is stable for approximately up to 10^4 and 72×10^4 Type I and Type II UEs, respectively. However, by observing the data presented in Fig. 13(b), we see that the mean delay of Type I UEs is already higher than 10 s, even for 100 Type I UEs in the system. At the same time, it stays well below conventional UEs even for 10^3 Type I UEs in the system. Thus, we may conclude that the presence of UEs demanding permanent connectivity questions the applicability of NB-IoT technology in its current form for environments with a mixture of considered traffic types.

Another parameter that might impact the delay performance of considered traffic types is the cycle time Δ. To this aim, Fig. 14 shows mean delays of considered traffic types as a function of the cycle duration, Δ [$\lambda_2 = 1$ message/s in Fig. 14(a) and $M_1 = 500$ in Fig. 14(b)]. By analyzing the presented data, one may observe that the cycle interval duration does not affect the delay performance of considered traffic types. The rationale is that the arrival of Type I UEs for service is well localized in time for considered system parameters. In other words, the instantaneous delay upon arrival of Type I UEs explodes drastically. However, due to the system being in stable conditions and significantly underloaded for considered input parameters, the period when Type I UEs exist in the system is much smaller than the cycle duration Δ. Smaller, unrealistic values of Δ lead to delay the explosion, but the system is generally unresponsive to any sensible duration of Δ.

Finally, we consider the way to satisfy guarantees for Type I UEs. To this aim, Fig. 15 shows the mean delay of Type I UEs as a function of their number in the system for multiple RBs allocated for NB-IoT network, $\Delta = 15$ m and Type II UE arrival intensity of $\lambda_2 = 1$ messages/s. Here, we assume that both types of UEs are equally divided between NB-IoT RBs. As we may observe, when one, two, and then three RBs are utilized, the delay for Type I UEs decreases. Furthermore, the considered trend is, in fact, nonlinear, and the mean delay drops much higher when more RBs are added to the system. By utilizing 3 RBs, the mean Type I UE delay satisfies M.2410 of 10 s for all the considered number of Type I UEs.

VI. Conclusion

Motivated by the emergence of novel mMTC applications requiring permanently connected UEs, in this article, we considered the performance of a mixture of conventional stochastic and regular traffic types at a single NB-IoT cell. To

this aim, we have developed exact and approximate models, efficient numerical algorithms, and delay approximations. The developed model can be utilized to assess the delay performance of UEs for different mixtures of considered traffic types.

Our results demonstrate that NB-IoT remains stable for up to 72×10^4 conventional UEs per NB-IoT cell (sectoral antenna) and 9×10^3 UEs demanding permanent connectivity. The presence of the latter UEs has a linear effect on their delay performance but affects conventional UEs more drastically. The delay performance of 10 s specified in ITU-R M.2410 is met for conventional UEs even under a relatively high number of permanently connected UEs reaching 10^3. However, the delay of the latter UEs is violated even for 10^2 permanently connected UEs, requiring redesigning the NB-IoT channel access mechanism or increasing the amount of available resources.

Conventional traffic does not significantly impact the delay of UEs requiring permanent connectivity. Contrarily, the mean delay behavior of both traffic types is almost solely affected by the number of Type I UEs and is independent of the cycle time Δ. Overall, the emergence of UEs requiring permanent connectivity questions the appropriateness of NB-IoT technology for 5G mMTC in its current form. Finally, we note that in the system considered in this article, one might be interested in evaluating the probability that the delay is higher than a certain threshold for both traffic types. These metrics can be estimated numerically by utilizing the exact model introduced in Section IV-A.

Appendix A

In this Appendix, we derive $P(l,k,m)$ utilized in (6) to parameterize transition probabilities of the approximate Markov model. Consider k indistinguishable messages from active UEs to be transmitted over l orthogonal channels. Denote by $\chi(k)$ the number of channels that contain exactly one message. In general, this is a random variable that can take values from 0 to $\min(l,k)$, and we are interested in its probability mass function (pmf) $P(l,k,m)$. The analysis presented below heavily relies upon the approach proposed in [56] and utilizes the apparatus of the probability generating functions (PGFs). To this aim, we will sketch it briefly in a step-by-step manner.

Step 1: Let $G(z)$ be the PGF of a random variable $\chi(k)$ having discrete distribution, i.e.,

$$G(z) \triangleq \sum_{m=0}^{l} \Pr\{\chi(k)=m\} z^m. \tag{20}$$

Step 2: Expanding the introduced PGF in a Taylor series around the point $z=1$ we obtain

$$G(z) = \sum_{m=0}^{l} \frac{1}{m!} \frac{d^m G(z)}{dz^m} \Big|_{z=1} (z-1)^m. \tag{21}$$

Using the results of Appendix B, we can show that

$$\frac{d^m G(z)}{dz^m} \Big|_{z=1} = E[\chi(k)(\chi(k)-1)\dots(\chi(k)-m+1)]. \tag{22}$$

By introducing the following notation:

$$\phi^{[m]}(\cdot) \triangleq \begin{cases} \phi(\cdot)[\phi(\cdot)-1]\dots[\phi(\cdot)-m+1] & \phi(\cdot) \geq m \\ 0 & \phi(\cdot) < m \end{cases} \tag{23}$$

we can write (21) in the following form:

$$G(z) = \sum_{m=0}^{l} \frac{1}{m!} E\left[\chi^{[m]}(k)\right](z-1)^m. \tag{24}$$

Step 3: Now, we are to calculate the raw moments of the considered random variable, $E[\chi^{[m]}(k)]$ as discussed in Appendix C leading to

$$E\left[\chi^{[m]}(k)\right] = l^{[m]} \frac{k^{[m]}}{l^m} \left(1 - \frac{m}{l}\right)^{n-m}. \tag{25}$$

Step 4: Substituting (25) into (24), we obtain an expression for the generating function

$$G(z) = \sum_{m=0}^{l} \frac{l^{[m]}}{l^m} k^{[m]} \left(1 - \frac{m}{l}\right)^{k-m} \frac{(z-1)^m}{m!}. \tag{26}$$

Step 5: By following [56], we now invert the obtained PGF using a partial fraction expansion approach to obtain the pmf $P(l,k,m)$ of the random variable $\chi(k)$ as:

$$\begin{cases} \frac{(-1)^m l! k!}{l^k m!} \sum_{f=m}^{\min(l,k)} \frac{(-1)^f (l-f)^{k-f}}{(f-m)!(l-f)!(k-f)!} & m \leq k \\ 0 & m > k. \end{cases} \tag{27}$$

Appendix B

Proposition 1: For the PGF of the discrete random variable $\chi(k)$, the following expansion holds true:

$$\frac{d^m G(z)}{dz^m} \Big|_{z=1} = E[\chi(k)(\chi(k)-1)\dots(\chi(k)-m+1)]. \tag{28}$$

Proof: We prove the result in (28) by induction. For $m=1$, we can write

$$\begin{aligned} \frac{d^1 G(z)}{dz^1} \Big|_{z=1} &= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} n (1)^{n-1} \\ &= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} n = E[\chi(k)]. \end{aligned} \tag{29}$$

Similarly, for $m=2$ we can write

$$\begin{aligned} \frac{d^2 G(z)}{dz^2} \Big|_{z=1} &= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} n(n-1)(1)^{n-2} \\ &= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} n(n-1) \\ &= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} n^2 - \sum_{n=0}^{l} \Pr\{\chi(k)=n\} n \\ &= E\left[\chi^2(k)\right] - E[\chi(k)] = E\left[\chi^2(k) - \chi(k)\right] \\ &= E[\chi(k)(\chi(k)-1)]. \end{aligned} \tag{30}$$

Now, we state the hypothesis that the following holds true for some $m > 2$:

$$\frac{d^m G(z)}{dz^m} = \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \frac{d^m (z^n)}{dz^m}$$

$$= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \prod_{j=0}^{m} (n-j) z^{n-j} \,|_{z=1}$$
$$= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \prod_{j=0}^{m} (n-j)$$
$$= E\left[\prod_{j=0}^{m} (\chi(k)-j)\right]. \tag{31}$$

Now, for some $m+1$ we see

$$\frac{d^{m+1}G(z)}{dz^{m+1}} = \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \frac{d^{n+1}(z^m)}{dz^{n+1}}$$
$$= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \prod_{j=0}^{m+1} (n-j) z^{n-j} \,|_{z=1}$$
$$= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \prod_{j=0}^{m+1} (n-j)$$
$$= \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \prod_{j=0}^{m} (n-j) n$$
$$- \sum_{n=0}^{l} \Pr\{\chi(k)=n\} \prod_{j=0}^{m} (n-j)(m+1)$$
$$= E\left[\prod_{j=0}^{m} (\chi(k)-j)\chi(k)\right]$$
$$- E\left[\prod_{j=0}^{m} (\chi(k)-j)(m+1)\right]$$
$$= E\left[\prod_{j=0}^{m+1} (\chi(k)-j)\right] \tag{32}$$

proving (28). ■

APPENDIX C

Define an indicator random variable as

$$\Theta_i = \begin{cases} 1, \text{ exactly one message occupies a preamble} \\ 0, \text{ otherwise.} \end{cases} \tag{33}$$

By using (33) $\chi(k)$ can be written as

$$\chi(k) = \sum_{i=0}^{l} \Theta_i. \tag{34}$$

We are now to state the following proposition.

Proposition 2: Powers of random variable $\chi^{[m]}(k)$ can be written using the following expansion:

$$\chi^{[m]}(k) = \sum_{A_m} \Theta_{i,1}\Theta_{i,2}, \ldots, \Theta_{i,m} \tag{35}$$

where the sum is calculated over all possible sets of m elements from the set $\{1, 2, \ldots, l\}$, denoted by A_m.

Proof: The proof is provided in [56]. ■

By using Proposition 1, we have the following proposition.

Proposition 3: The mean values of the sequence of random variables $\chi^{[m]}(k)$ are provided by

$$E\left[\chi^{[m]}(k)\right] = l^{[m]} \frac{k^{[m]}}{l^m} \left(1 - \frac{m}{l}\right)^{k-m}. \tag{36}$$

Proof: The mean value of $\chi^{[m]}(k)$ can be written as

$$E\left[\chi^{[m]}(k)\right] = E\left[\sum_{i,1\neq i,2\ldots i,m} \Theta_{i,1}\Theta_{i,2}\ldots\Theta_{i,m}\right]$$
$$= \sum_{A_m} \Pr\{\Theta_{i,1}=1, \Theta_{i,2}=1\ldots\Theta_{i,m}=1\}$$
$$= l^{[m]} \Pr\{\Theta_1=1, \ldots, \Theta_m=1\}. \tag{37}$$

By utilizing the assumption of independent choice of preambles, (37) can be rewritten as

$$\Pr\{\Theta_1=1, \ldots \Theta_m=1\} = \frac{k^{[m]}}{l^m} \left(1 - \frac{m}{l}\right)^{k-m}. \tag{38}$$

Finally, substituting (38) into (37) we obtain

$$E\left[\chi^{[m]}(k)\right] = l^{[m]} \frac{k^{[m]}}{l^m} \left(1 - \frac{m}{l}\right)^{k-m} \tag{39}$$

that finalizes the proof of the proposition. ■

REFERENCES

[1] M. Stusek et al., "LPWAN coverage assessment planning without explicit knowledge of base station locations," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4031–4050, Mar. 2022.

[2] *Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface*, Int. Telecommun. Union, Geneva, Switzerland, Rec. M.2410-0, Jul. 2017.

[3] P. Masek et al., "Tailoring NB-IoT for mass market applications: A mobile operator's perspective," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–7.

[4] *Guidelines for Evaluation of Radio Interface Technologies for IMT-2020*, Int. Telecommun. Union, Geneva, Switzerland, Rec. M.2412-0, Jul. 2017.

[5] C. Alaton and F. Tounquet, *Benchmarking Smart Metering Deployment in the EU-28*, Tractebel Antwerpen, Antwerpen, Belgium, 2020.

[6] N. Myoung, Y. Kwon, M. Park, and C. Eun, "Data interworking model and analysis for harmonization of smart metering protocols in IoT-based AMI system," *Sensors*, vol. 23, no. 6, p. 2903, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/6/2903

[7] O. Vikhrova, S. Pizzi, A. Molinaro, and G. Araniti, "Paging group size distribution for multicast services in 5G networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2020, pp. 484–489.

[8] "Minimum requirements related to technical performance for IMT-2020 radio interface (s)," Int. Telecommun. Union, Geneva, Switzerland, Rep. 2410-0, 2017.

[9] *Advanced Technologies for Industry: Product Watch IoT Components in Connected and Autonomous Vehicles*, Eur. Comm., Brussels, Belgium, 2020.

[10] R. Germanà, E. De Santis, F. Liberati, and A. Di Giorgio, "On the participation of charging point operators to the frequency regulation service using plug-in electric vehicles and 5G communications," in *Proc. IEEE Int. Conf. Environ. Elect. Eng. Proc. IEEE Ind. Commer. Power Syst. Europe (EEEIC/I CPS Europe)*, 2021, pp. 1–6.

[11] M. Kemal, R. Sanchez, R. Olsen, F. Iov, and H.-P. Schwefel, "On the tradeoff between timeliness and accuracy for low voltage distribution system grid monitoring utilizing smart Meter data," *Int. J. Elect. Power Energy Syst.*, vol. 121, Oct. 2020, Art. no. 106090. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061519316801

[12] G. Giaconi, D. Gunduz, and H. V. Poor, "Privacy-aware smart metering: Progress and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 59–78, Nov. 2018.

[13] Y. Perez and W. Arowolo, "Integration of electromobility with the electric power systems: The key challenges," *Annales des Mines-Enjeux Numériques*, vol. 15, pp. 60–66, Sep. 2021.

[14] X. Chang, J. Zhan, G. Xing, J. Huang, B. Chen, and L. Zhou, "Measurement-based optimization of cell selection in NB-IoT networks," *ACM Trans. Sens. Netw.*, vol. 18, no. 4, Nov. 2022, Art. no. 3544017. [Online]. Available: https://doi.org/10.1145/3544017

[15] "Energy performance of buildings directive," European Commission. Accessed: Aug. 16, 2022. [Online]. Available: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en

[16] "Benchmarking smart metering deployment in the EU-28," European Commission. Accessed: Aug. 16, 2022. [Online]. Available: https://energy.ec.europa.eu/benchmarking-smart-metering-deployment-eu-28_en

[17] Y.-M. Kim, D. Jung, Y. Chang, and D.-H. Choi, "Intelligent micro energy grid in 5G era: Platforms, business cases, Testbeds, and next generation applications," *Electronics*, vol. 8, no. 4, p. 468, 2019.

[18] Y. Shen, W. Fang, F. Ye, and M. Kadoch, "EV charging Behavior analysis using hybrid intelligence for 5G smart grid," *Electronics*, vol. 9, no. 1, p. 80, 2020.

[19] O. Bularca, M. Florea, and A.-M. Dumitrescu, "Smart metering deployment status across EU-28," in *Proc. Int. Symp. Fund. Elect. Eng. (ISFEE)*, 2018, pp. 1–6.

[20] "Narrowband-IoT for massive IoT: Capacity study for selected areas," Vodafone Czech Republic. Accessed: Nov. 15, 2023. [Online]. Available: https://www.vut.cz/www_base/vutdisk.php?i=324298a408

[21] "Vodafone is going to use NB-IoT to extend implementation of the Internet of Things," Vodafone Czech Republic. Accessed: Nov. 15, 2023. [Online]. Available: https://www.vodafone.cz/en/about-vodafone/press-releases/message-detail/vodafone-si-pro-rozsireni-sluzby-Internet-veci-vyb/

[22] "Vodafone to light up a nationwide NB-IoT network," Vodafone Czech Republic. Accessed: Nov. 15, 2023. [Online]. Available: https://www.vodafone.cz/en/about-vodafone/press-releases/message-detail/vodafone-letos-rozsviti-celonarodni-sit-pro-nb-iot/

[23] "Vodafone's IoT network coverage of the Czech Republic completed," Vodafone Czech Republic. Accessed: Nov. 15, 2023. [Online]. Available: https://www.vodafone.cz/en/about-vodafone/press-releases/message-detail/vodafone-dokoncil-pokryti-ceske-republiky-siti-pro/

[24] A. Abou El Hassan, A. El Mehdi, and M. Saber, "NarrowBand-IoT and eMTC toward massive MTC: Performance evaluation and comparison for 5G mMTC," in *Networking, Intelligent Systems and Security*. Singapore: Springer, 2022, pp. 177–195.

[25] A. Chehri, H. Chaibi, R. Saadane, E. M. Ouafiq, and A. Slalmi, "On the performance of 5G narrow-band Internet of Things for industrial applications," in *Networking, Intelligent Systems and Security*. Singapore: Springer, 2022, pp. 275–286.

[26] B. S. Tsybakov and V. A. Mikhailov, "Free synchronous packet access in a broadcast channel with feedback," *Problemy Peredachi Informatsii*, vol. 14, no. 4, pp. 32–59, 1978.

[27] J. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 505–515, Sep. 1979.

[28] L. Kleinrock and S. Lam, "Packet switching in a multiaccess broadcast channel: Performance evaluation," *IEEE Trans. Commun.*, vol. 23, no. 4, pp. 410–423, Apr. 1975.

[29] B. Van Houdt and C. Blondia, "Throughput of Q-ary splitting algorithms for contention resolution in communication networks," *Commun. Inf. Syst.*, vol. 4, no. 2, pp. 135–164, 2005.

[30] B. Van Houdt and C. Blondia, "Stability and performance of stack algorithms for random access communication modeled as a tree structured QBD Markov chain," *Commun. Stat. Stochastic Models*, vol. 17, no. 3, pp. 247–270, 2001.

[31] C. Stefanovic, P. Popovski, and D. Vukobratovic, "Frameless ALOHA protocol for wireless networks," *IEEE Commun. Lett.*, vol. 16, no. 12, pp. 2087–2090, Dec. 2012.

[32] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Random access for M2M communications with QoS guarantees," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2889–2903, Jul. 2017.

[33] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Mar. 2019.

[34] G. C. Madueno, N. K. Pratas, Č. Stefanović, and P. Popovski, "Massive M2M access with reliability guarantees in LTE systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 2997–3002.

[35] M. S. Haghighi, "Critical study of Markovian approaches for batch arrival modeling in IEEE 802.15.4-based networks," 2020, *arXiv:2003.00829*.

[36] B. Yu, X. Chi, and H. Sun, "Delay analysis for aggregate traffic based on martingales theory," *IET Commun.*, vol. 14, no. 5, pp. 760–767, 2020.

[37] D. T. C. Wong, Q. Chen, X. Peng, and F. Chin, "Multichannel pure collective aloha MAC protocol with decollision algorithm for satellite uplink," in *Proc. IEEE 4th World Forum Internet Things (WF-IoT)*, 2018, pp. 251–256.

[38] M. R. Chowdhury and S. De, "Delay-aware priority access classification for massive machine-type communication," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13238–13254, Dec. 2021.

[39] S. Rostami, S. Lagen, M. Costa, M. Valkama, and P. Dini, "Wake-up radio-based access in 5G under delay constraints: Modeling and optimization," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1044–1057, Feb. 2020.

[40] W. Sun, H. Zhang, R. Wang, and Y. Zhang, "Reducing offloading latency for digital twin edge networks in 6G," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12240–12251, Oct. 2020.

[41] F. Metzger, T. Hoßfeld, A. Bauer, S. Kounev, and P. E. Heegaard, "Modeling of aggregated IoT traffic and its application to an IoT cloud," *Proc. IEEE*, vol. 107, no. 4, pp. 679–694, Apr. 2019.

[42] C. Correia, A. C. M. Freitas, and J. M. Freitas, "Cluster distributions for dynamically defined point processes," *Physica D Nonlinear Phenomena*, vol. 457, Jan. 2024, Art. no. 133968.

[43] S. Lee et al., "Anomaly detection of smart metering system for power management with battery storage system/electric vehicle," *ETRI J.*, vol. 45, no. 4, pp. 650–665, 2023.

[44] Office of Electricity Delivery and Energy Reliability, "Advanced metering infrastructure and customer systems," 2016. [Online]. Available: https://www.energy.gov/sites/prod/files/2016/12/f34/AMI

[45] N. Stepanov, A. Turlikov, and V. Begishev, "Balancing the data transmission and random access phases in 6G mMTC radio technologies," *IEEE Commun. Lett.*, vol. 27, no. 12, pp. 3419–3423, Dec. 2023.

[46] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Physical Layer; V14.1.0*, ETSI Standard TS 36.201, Apr. 2017.

[47] A. Adhikary, X. Lin, and Y.-P. E. Wang, "Performance evaluation of NB-IoT coverage," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, 2016, pp. 1–5.

[48] O. Liberg, M. Sundberg, E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of Things: Technologies, Standards, and Performance*. London, U.K.: Academic, 2017.

[49] L. Feltrin et al., "Narrowband IoT: A survey on downlink and uplink perspectives," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 78–86, Feb. 2019.

[50] M. Kanj, V. Savaux, and M. Le Guen, "A tutorial on NB-IoT physical layer design," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2408–2446, 4th Quart., 2020.

[51] R. Mozny et al., "Characterizing optimal LPWAN access delay in massive multi-RAT smart grid deployments," *Internet Things*, vol. 25, Apr. 2024, Art. no. 101001.

[52] M. Koseoglu, "Lower bounds on the LTE-a average random access delay under massive M2M arrivals," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2104–2115, May 2016.

[53] O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, "Stabilizing multichannel slotted aloha for machine-type communications," in *Proc. IEEE Int. Symp. Inf. Theory*, 2013, pp. 2119–2123.

[54] B. N. Feinberg and S. S. Chiu, "A method to calculate steady-state distributions of large Markov chains by aggregating states," *Oper. Res.*, vol. 35, no. 2, pp. 282–290, 1987.

[55] G. Franceschinis and R. R. Muntz, "Bounds for quasi-lumpable Markov chains," *Perform. Eval.*, vol. 20, nos. 1–3, pp. 223–243, 1994.

[56] W. Szpankowski, "Statistic analysis of multiaccess systems with random access and feedback," Ph.D. dissertation, Dept. Electr. Eng., Univ. Gdańsk, Gdańsk, Poland, 1980.

[57] J. G. Kemeny and J. L. Snell, *Finite Markov Chains: With a New Appendix 'Generalization of a Fundamental Matrix'*. New York, NY, USA: Springer, 1983.

[58] U. N. Bhat, *An Introduction to Queueing Theory: Modeling and Analysis in Applications*, vol. 36. Boston, MA, USA: Birkhäuser, 2008.

[59] P. Jörke, T. Gebauer, and C. Wietfeld, "From LENA to LENA-NB: Implementation and performance evaluation of NB-IoT and early data transmission in NS-3," in *Proc. Workshop Ns-3*, 2022, pp. 73–80. [Online]. Available: https://doi.org/10.1145/3532577.3532600

[60] H. G. Perros, *Computer Simulation Techniques: The Definitive Introduction*! Raleigh, NC, USA: Harry Perros, 2009.

[61] S. A. Gbadamosi, G. P. Hancke, and A. M. Abu-Mahfouz, "Building upon NB-IoT networks: A roadmap toward 5G new radio networks," *IEEE Access*, vol. 8, pp. 188641–188672, 2020.