Toward Optimal Connection Management for Massive Machine-Type Communications in 5G System

Wen Zhan, *Member, IEEE*, Chen Xu, Xinghua Sun, *Member, IEEE*, and Jun Zou, *Member, IEEE*

***Abstract*—The massive machine-type communications (mMTC) is one of the three generic services for 5G. With the connection-based random access (CBRA) scheme, each machine-type device (MTD) establishes a connection with the base station (BS) prior to its data transmission. Due to the explosive growth of the number of MTDs, many MTDs would establish connections with the BS, which necessitates the study on how to efficiently manage the massive connections with MTDs. To address this issue, in this article, we propose a unified utility-based analytical framework for the optimal connection management of mMTC in 5G networks, where the signaling overhead for connection establishment, access delay, and the connection resource utilization ratio is included. Specifically, we first derive key performance metrics, i.e., the mean time length of each connection and resource utilization ratio, as functions of traffic input rate and inactivity timer. By further considering the signaling overheads and access delay of each MTD, the network utility is formulated and maximized by optimally choosing the inactivity timer. We then present a detailed discussion on the effect of system parameters on the optimal inactivity timer and the corresponding maximum network utility. Finally, we extend the analytical framework to the scenario in which the CBRA scheme coexists with the packet-based random access scheme, i.e., transmitting packets in the random access channel without connection establishment. The critical threshold in terms of the traffic input rate is characterized, which sheds important light on the access scheme selection issue.**

***Index Terms*—Connection, machine-type communications, radio resource control (RRC), random access.**

Manuscript received October 25, 2020; revised February 3, 2021; accepted February 28, 2021. Date of publication March 11, 2021; date of current version August 24, 2021. The work of Wen Zhan was supported in part by the National Natural Science Foundation of China under Grant 62001524, and in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant JCYJ20200828214622001. The work of Xinghua Sun was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101120003, in part by the National Natural Science Foundation of China under Grant 61801244, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515011906. The work of Jun Zou was supported by the National Natural Science Foundation of China under Grant 61701234. *(Corresponding authors: Jun Zou; Xinghua Sun.)*

Wen Zhan and Xinghua Sun are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China (e-mail: zhanw6@mail.sysu.edu.cn; sunxinghua@mail.sysu.edu.cn).

Chen Xu and Jun Zou are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: chen_xu@njust.edu.cn; jun_zou@njust.edu.cn).

Digital Object Identifier 10.1109/JIOT.2021.3065506

I. Introduction

MASSIVE machine-type communications (mMTC) is an emerging communication paradigm in which a large number of low-complexity/low-power machine-type devices (MTDs) are attached to the Internet and communicate with each other without human intervention. It has been identified as one of the three generic service types to be supported by 5G system, and serves as the indispensable foundation for a variety of Internet-of-Things (IoT) applications, e.g., home automation, intelligent transportation systems, and smart metering [1]. Fueled by the booming IoT market, the number of MTDs is now explosively growing and billions of MTDs, such as sensors, actuators, and meters, are predicted to come into existence [2]. Most of them require ubiquitous wireless connectivity service from the 5G system, which poses unprecedented challenges on the radio access network of the 5G system.

A. Connection-Based Random Access

Currently, the 4G or 5G system adopts the connection-based random access (CBRA) scheme. That is, each MTD has to establish a connection with the base station (BS) via the random access procedure, before BS allocates time–frequency resources for its data transmission. The CBRA scheme fits well for traditional human-type communications (HTC), such as video streaming and mobile games, where the number of devices is small but each device transmits a significant amount of data [3]. In mMTC, however, devices usually transmit small packets while the number of devices could be very large. As a result, when mMTC is served via CBRA scheme, two key issues arise.

1) *Congestion Issue:* A large number of MTDs would generate massive access requests, which congest the random access channel in cellular systems, leading to an intolerably low chance of successful access. Therefore, how to optimize the access efficiency for mMTC is of critical importance [4].
2) *Connection Management Issue:* Lots of MTDs may establish connections with the BS, indicating that the BS needs to manage massive concurrent connections (i.e., sessions). The signaling overhead and resource consumption due to connection management will be nonnegligible, which necessitates the study on how to efficiently control the massive connections with MTDs.

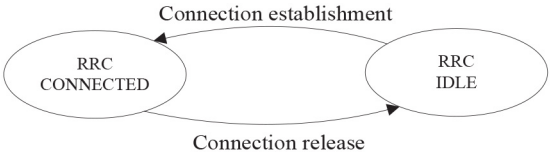


Fig. 1. State transition process of each MTD.

For the congestion issue, there have been a plethora of related works. For instance, various analytical models have been formulated in analyzing the access performance of mMTC in LTE networks [5]–[8]. It was found that to optimize the access performance, key system parameters, including the backoff parameters and the number of preambles, should be carefully adjusted. Accordingly, by periodically estimating the traffic load in the random access channel, lots of algorithms were proposed for adaptively tuning of those parameters [9]–[12]. In recent, a new analytical framework was established for characterizing the behavior of each MTD in the CBRA scheme, based on which explicit expressions of access parameters for optimizing the access throughput and access delay performance are derived [13], [14]. While the aforementioned developments have been substantial, none of above works consider the behavior of MTDs after the successful completion of the random access procedure, and therefore, the connection management issue is largely ignored.

The connection management issue, which is the focus of this article, receives much less attention in the existing literature. Before the review of related works, let us first gain a more clear understanding of this issue from the system point of view. Specifically, in radio resource control (RRC) layer, two basic radio states[1]: 1) RRC IDLE state and 2) RRC CONNECTED state, guide the RRC operation in cellular systems, as shown in Fig. 1 [15], [16]. When an MTD in RRC IDLE state successfully completes the random access procedure, then a connection is established and it shifts into the RRC CONNECTED state. Only in RRC CONNECTED state, the transmissions of data to/from MTD can take place. For each connection, the BS: 1) allocates exclusive time–frequency resources for control signaling exchange, e.g., scheduling request (SR) and 2) configures an inactivity timer. When the BS detects MTD, which does not have any traffic to send and receive during the inactivity timer, then the BS releases the connection and places MTD in the RRC IDLE state [17], [18].

In the context of mMTC, the connection management issue becomes challenging. Intuitively, if the BS releases connections often by setting a small inactivity timer, then MTDs may frequently initiate the random access procedure, which induces a significant amount of signalings and congests the random access channel. Conversely, if the BS maintains the connection with each MTD long, then a large body of the system resources would be wasted in connection management since the number of concurrent connections could be large. Therefore, how to efficiently manage the massive connections with MTDs is a challenging issue, and the key to address this issue lies in the proper setting of system parameters, such as the inactivity timer.

Related works usually take signaling overhead and energy efficiency as key performance measures. For instance, in [19] and [20], simulation results revealed that the inactivity timer crucially determines the tradeoff between the signaling overhead and energy efficiency. A smaller inactivity timer can save energy as it reduces the time length that devices stay in RRC CONNECTED state. Yet, it also boosts the signaling overhead in the random access channel. In [21], simulation results further suggested that to improve the network performance, mobile operators should tune the inactivity timer according to the traffic input rate. A large inactivity timer is preferred if the traffic becomes heavy [22]. Apart from the simulation results, analytical models were established in [23]–[25]. In [23], signaling overhead, energy consumption, and processing delay in the discontinuous reception procedure were considered, and an algorithm was proposed to jointly tune system parameters including the inactivity timer for balancing system performance. In [24], a new RRC state transition scheme was proposed for MTC traffic, where the connection is released once no traffic is detected. In [25], besides the inactivity timer, a threshold was introduced, where the connection is released once the number of transmitted packets reaches this threshold.

Most of above analytical models, however, only focus on the RRC state transition process while ignoring the signaling overhead and the access delay in the random access process. In fact, due to the congestion issue, the corresponding signaling overhead and the access delay would be nonnegligible. With the tight coupling between the RRC state transition process and the random access process, those factors should, therefore, be carefully taken into consideration in designing the connection management scheme for mMTC. Moreover, with mMTC, the number of connections would be large, indicating that lots of uplink resources will be used in connection management. Yet, in the existing works, little light was shed on how the system parameters affect the corresponding uplink resource utilization. Therefore, how to jointly consider the signaling overhead, access delay, and resource utilization in designing the connection management scheme still remains an open issue.

B. Packet-Based Random Access

Note that it has been widely argued that the CBRA scheme might be unsuitable for mMTC [26]. The fundamental reason lies in establishing a connection prior to data transmission is inefficient for small packets [27]. Therefore, in the 5G cellular system, the packet-based random access (PBRA) scheme is introduced, where MTDs can transmit one small data packet in the random access procedure [28]. Accordingly, for the PBRA scheme, the connection management issue does not exist since the data transmission is completed in the random access procedure. When the CBRA scheme and PBRA scheme coexist in the 5G cellular system, a key question naturally arises: which random access scheme should be chosen?

Intuitively, the CBRA scheme should be chosen only if establishing a connection is beneficial. For instance, although

[1]A new RRC state, i.e., RRC INACTIVE, is introduced in 5G RRC state machine. The discussion will be extended to further incorporate this new state in Section VI.

the MTD transmits small packets, it transmits packets frequently, in which case establishing a connection is beneficial. If the PBRA scheme is used, then MTD has to perform random access often as each time only one packet is transmitted. To determine which random access scheme should be chosen, it is clear that a unified analytical framework that can characterize both schemes is the key, based on which the network performance with different schemes can be properly compared. However, most of the existing theoretical models are customized only for characterizing the contention in the random access channel [5]–[8] or the RRC state transition [23]–[25]. None of existing models can be directly used for mMTC in 5G system where both CBRA scheme and PBRA scheme coexist. As a result, given the traffic characteristic of mMTC, which random access scheme should be chosen is still an open issue.

C. Our Contributions

In this article, we will address the connection management issue for mMTC in 5G networks by properly tuning the inactivity timer and developing the criteria in choosing the random access scheme. Specifically, we consider a single-cell system with n MTDs and each has a packet arrival rate of λ. By characterizing the behavior of each MTD in RRC CONNECTED state, we derive key performance measures, such as the mean time length of each connection and the utilization ratio of the uplink resource for signaling exchange, i.e., SR, as functions of λ and inactivity timer T_{in}. The analysis shows that the mean time length of each connection increases with the traffic input rate λ and inactivity timer T_{in}. Yet, the resource utilization ratio decreases if T_{in} grows.

By further incorporating the signaling overheads and access delay of each MTD in the random access process into consideration, a utility-based analytical framework is developed for the CBRA scheme. Assume the system receives incomes for packet transmissions while incurs penalty for resource underutilization, signaling overheads, and delay. Toward the maximization of the network utility, an algorithm is proposed for computing the optimal inactivity timer T_{in}^* and the maximum network utility. The analysis is validated via simulations. It is revealed that with high signaling overheads or reward for each packet transmission, the optimal inactivity timer T_{in}^* increases so that devices perform random access procedure less frequently and stay long in RRC CONNECTED to deliver more packets. On the other hand, to improve resource utilization, T_{in}^* should be reduced because the traffic input rate of MTD λ is usually small.

The analysis in this article sheds important light on the issue of random access scheme selection. The analytical framework can be extended to the PBRA scheme for deriving the corresponding network utility. Based on the comparison of utilities between PBRA and CBRA, we characterize the critical threshold in terms of the traffic input rate λ_{LB}, below which the PBRA scheme is preferred.

The remainder of this article is organized as follows. Section II presents the system model. Key performance measures are characterized in Section III. The utility-based analytical framework is established in Section IV and the analysis is validated by the simulation results in Section V. Section VI conducts a comparative study between the CBRA scheme and the PBRA scheme. Finally, concluding remarks are summarized in Section VII.

II. System Model

Consider a single-cell cellular network serving a large number of MTDs. In this article, we focus on the state transition process of each MTD in the RRC layer, where each MTD is either in RRC IDLE state or in RRC CONNECTED state, as shown in Fig. 1. According to the standard [15], an MTD is in RRC CONNECTED state when an RRC connection has been established between itself and BS. If this is not the case, i.e., no RRC connection is established, then the MTD is in RRC IDLE state.

A. MTD in RRC IDLE State

For each MTD in the RRC IDLE state, as it has no RRC connection with the BS, it cannot request time–frequency resources for uplink data transmission from BS. Therefore, once the MTD has data packets pending for transmission, it has to establish the RRC connection via the random access procedure [29]. The random access procedure consists of a four-way-handshake: 1) a preamble transmission initiated by the MTD; 2) a random access response replied by the BS; 3) an RRC connection request transmitted by the MTD; and 4) an RRC connection setup message replied by the BS. If the MTD completes the random access procedure, then the RRC connection is established and the MTD shifts into the RRC CONNECTED state, as shown in Fig. 1.

B. MTD in RRC CONNECTED State

For each MTD in RRC CONNECTED state, the transmissions of unicast data to/from MTD and the transmissions of broadcast/multicast data to MTD can take place. Unlike traditional mobile traffic, which is dominant in downlink rather than in uplink, MTC traffic is uplink dominant [4]. Therefore, in this article, we only focus on the uplink data transmissions of unicast data from MTD to BS. To facilitate the uplink data transmission resource allocation, for each MTD in the RRC CONNECTED state, the BS would configure uplink resources in physical uplink control channel (PUCCH) for its exclusive use, more specifically, for its SR transmission [17]. That is, if an MTD in the RRC CONNECTED state needs to transmit data packets but has no uplink data transmission resources, then it sends an SR to the BS via the dedicated uplink resource. Upon the reception of SR, the BS assigns uplink data transmission resources to the MTD. As the BS does not know in advance when the MTD would request uplink data transmission resources, the dedicated resources for SR transmission to each MTD are assigned periodically. The period of dedicated resources for SR transmission ranges from 1 to 80 ms, depending on the network configuration [17].

Define a time slot as the interval between two consecutive resource units for SR transmission, as shown in Fig. 2. Assume that for each MTD in each time slot, the arrivals of data packets follow a Bernoulli process with parameter $\lambda \in (0, 1)$ and

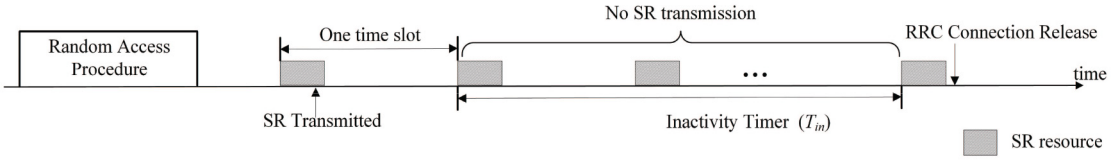


Fig. 2. Periodical SR resources in cellular system.

each MTD has an infinite buffer size. Typically, for M2M communications, the packet size is small (e.g., sensor data and location update with a size of a few bytes [30]). Therefore, we assume that the SR transmission, uplink resource allocation, and the data packet transmission can be completed in one time slot (i.e., the data buffer is cleared in one time slot). Moreover, for simplicity, we refer to the slot that contains an SR transmission as a busy slot and the slot without SR transmission as an empty slot.

Due to the randomness of the packet arrival, MTD in the RRC CONNECTED state may have no newly arrival packets, and thus do not send SRs during a period of time. In practice, the BS can consider this MTD as an inactive user equipment and release the connection for saving the time–frequency resources [18]. It is worth mentioning that the BS may release an MTD's connection for many reasons, such as network overload and handover. In this article, we only consider one case, that is, the BS detects an MTD, which does not have any traffic to send for T_{in} consecutive slots, and then initiates the RRC connection release procedure by sending the RRC connection release message to MTD, as shown in Fig. 2. Upon the reception of this message, the MTD shifts back to the RRC IDLE state. T_{in} is referred to as the inactivity timer.

C. Inactivity Timer

It is clear that the network performance is crucially determined by the inactivity timer T_{in}. Specifically, from the viewpoint of MTDs, the four-way-handshake random access procedure consumes nonnegligible signaling overheads and power, as MTDs are mostly low-power/low complexity devices. With an excessively small inactivity timer T_{in}, the connection is likely to be released in a short period of time, which may lead to frequent initiations of the random access procedure and thus does not serve MTDs' interests. From the viewpoint of BS, an excessively large inactivity timer T_{in} may reduce the utilization ratio of SR resources as inactive MTDs with light traffic do not use the assigned resources efficiently. Therefore, it is of great practical importance to study how to properly set the inactivity timer T_{in}. In the following section, we will first derive key performance measures and evaluate the effect of T_{in} on network performance.

III. Performance Evaluation

In this section, we will first evaluate the network performance by considering two key performance measures: 1) the mean time length of each connection, denoted by $E[T_C]$, which is defined as the mean time length upon the successful establishment of the connection to the connection release and 2) the utilization ratio of each connection, denoted by R, which is defined as the percentage of busy slots in each connection. Based on these two key performance metrics, we will then take a closer look at how the inactivity timer T_{in} determines the network performance.

A. Mean Time Length of Each Connection

Note that the mean time length of each connection can be written as

$$E[T_C] = \sum_{i=1}^{i=+\infty} i \cdot \Pr\{T_C = i\} \tag{1}$$

where $\Pr\{T_C = i\}$ denotes the probability that the time length of the connection $T_C = i$ for $i = 1, 2, \ldots$ To obtain $E[T_C]$ according to (1), we, therefore, focus on deriving the expression of $\Pr\{T_C = i\}$ for $i = 1, 2, \ldots$ in the following.

Recall that for each MTD's connection, it will be released if it does not request SR transmission for T_{in} slots, and in each time slot, the MTD will have an SR request with probability $\lambda \in (0, 1)$. Depending on possibilities of the length of each connection T_C, we consider following cases.

1) $T_C \in \{1, 2, \ldots, T_{in}\}$: Upon the establishment of the connection, i.e., in the first slot, the MTD transmits an SR. Accordingly, the connection will not be released during the period from first SR slot to T_{in}th slot, that is

$$\Pr\{T_C = i\} = 0, \text{ for } i \in \{1, 2, \ldots, T_{\text{in}}\}. \tag{2}$$

2) $T_C = T_{in} + 1$: There should be no SR arrival during second slot to $T_{\text{in}} + 1$th slot. Accordingly, we have

$$\Pr\{T_C = T_{\text{in}} + 1\} = (1 - \lambda)^{T_{\text{in}}}. \tag{3}$$

3) $T_C \in \{T_{in} + 2, \ldots, 2T_{in} + 1\}$: There should be one SR arrival at $T_C - T_{\text{in}}$th slot while no arrival after that. Accordingly, we have

$$\Pr\{T_C = i\} = \lambda \cdot (1 - \lambda)^{T_{\text{in}}} \tag{4}$$

for $i \in \{T_{\text{in}} + 2, \ldots, 2T_{\text{in}} + 1\}$.

4) $T_C \in \{2T_{in} + 2, \ldots\}$: There should be: a) no SR arrival during the period from $T_C - T_{\text{in}} + 1$th slot to T_Cth slot with probability $(1-\lambda)^{T_{\text{in}}}$; b) one SR arrival in the $T_C - T_{\text{in}}$th slot with probability λ; and c) from 1st slot to $T_C - T_{\text{in}} - 1$th slot, there is no consecutive T_{in} empty slots with probability p_x. To derive p_x, let us define the probability transition matrix $\mathbf{P}$ with dimension $(T_{\text{in}} + 1) \times (T_{\text{in}} + 1)$ as

$$\mathbf{P} = \begin{bmatrix} \lambda & 1-\lambda & 0 & \cdots & 0 \\ \lambda & 0 & 1-\lambda & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \\ \lambda & 0 & 0 & \cdots & 1-\lambda \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{5}$$

Denote the element at ith row and jth column after $\mathbf{P}$ multiplies by itself for x times as $\mathbf{P}^x[i,j]$. The probability of getting at least T_{in} consecutive empty slots in $T_C - T_{\text{in}} - 1$ slots can then be written as $\mathbf{P}^{T_C-T_{\text{in}}-1}[1, T_{\text{in}}+1]$, with which we have $p_x = 1 - \mathbf{P}^{T_C-T_{\text{in}}-1}[1, T_{\text{in}}+1]$. Therefore, we can conclude that

$$\Pr\{T_C = i\} = \lambda\Big(1 - \mathbf{P}^{i-T_{\text{in}}-1}[1, T_{\text{in}}+1]\Big)(1-\lambda)^{T_{\text{in}}} \quad (6)$$

for $i > 2T_{\text{in}} + 1$.

Finally, by combining (2)–(6), we can have the expression of $\Pr\{T_C = i\}$ for $i = 1, 2, \ldots$ in (7), shown on the bottom of the page, with which the mean time length of each connection $E[T_C]$ can be calculated by combining (1) and (7) given packet arrival rate λ and the inactivity time T_C.

B. Utilization Ratio of SR Slots

The utilization ratio of each connection, denoted by R, is defined as the percentage of busy slots in each connection. It can be written as

$$R = \sum_{i=T_{\text{in}}+1}^{i=+\infty} \Pr\{T_C = i\} \cdot R_i \quad (8)$$

where R_i denotes the utilization ratio of SR slots given that the time length of the connection is i slots. In the following, we focus on deriving R_i for $i \geq T_{\text{in}} + 1$ by considering the following cases.

1) For $i = T_{\text{in}} + 1$, the connection is released once the inactivity timer is reached. In this case, only the first slot is used, and therefore, the utilization ratio of SR slots is given by

$$R_i = \tfrac{1}{T_{\text{in}}+1}. \quad (9)$$

2) For $i \in \{T_{\text{in}} + 2, \ldots, 2T_{\text{in}} + 1\}$, there should be one arrival packet at slot $i - T_{\text{in}}$, while for each slot $j \in \{2, \ldots, i - T_{\text{in}} - 1\}$, there may have packet arrival with probability λ. Therefore, the utilization ratio of SR slots is given by

$$R_i = \tfrac{2+\lambda(i-T_{\text{in}}-2)}{i}, \text{ for } i \in \{T_{\text{in}} + 2, \ldots, 2T_{\text{in}} + 1\}. \quad (10)$$

3) For $i > 2T_{\text{in}} + 1$, there should be one arrival at slot 1 and $i - T_{\text{in}}$, respectively. At the meantime, from slot 2 to slot $i - T_{\text{in}} - 1$, there should be no streak of T_{in} empty slots.[2] Let $\gamma_{i,T_{\text{in}}}$ denote the number of busy slots in the period from slot 2 to slot $i - T_{\text{in}} - 1$ given no streak of T_{in} empty slots, where $\gamma_{i,T_{\text{in}}} \in \{\lfloor (i - T_{\text{in}} - 2)/(T_{\text{in}}) \rfloor, \lfloor (i - T_{\text{in}} - 2)/(T_{\text{in}}) \rfloor + 1, \ldots, i - T_{\text{in}} - 2\}$. Accordingly, the utilization ratio of SR slots for $i > 2T_{\text{in}} + 1$ can then be written as

$$R_i = \tfrac{1}{i} \cdot \left(2 + \sum_{k=\left\lfloor \frac{i-T_{\text{in}}-2}{T_{\text{in}}} \right\rfloor}^{i-T_{\text{in}}-2} k \cdot \Pr\{\gamma_{i,T_{\text{in}}} = k\} \right) \quad (11)$$

in which $\Pr\{\gamma_{i,T_{\text{in}}} = k\}$ denotes the probability mass function (PMF) of $\gamma_{i,T_{\text{in}}}$.

The Appendix reveals that

$$\Pr\{\gamma_{i,T_{\text{in}}} = k\} = \frac{\lambda^k(1-\lambda)^{i-T_{\text{in}}-2-k} \cdot M^k_{i-T_{\text{in}}-2}}{\sum_{k=\lfloor (i-T_{\text{in}}-2)/T_{\text{in}} \rfloor}^{i-T_{\text{in}}-2} \lambda^k(1-\lambda)^{i-T_{\text{in}}-2-k} \cdot M^k_{i-T_{\text{in}}-2}} \quad (12)$$

for $k \in \{\lfloor (i - T_{\text{in}} - 2)/(T_{\text{in}}) \rfloor, \lfloor (i - T_{\text{in}} - 2)/(T_{\text{in}}) \rfloor + 1, \ldots, i - T_{\text{in}} - 2\}$, where $M^k_{i-T_{\text{in}}-2}$ denotes the total number of possibilities for k nonemtpy slots within $i - T_{\text{in}} - 2$ slots given no streak of T_{in} empty slots, and can be recursively calculated by

$$M^k_{i-T_{\text{in}}-2} = M^{k-1}_{i-T_{\text{in}}-3} + M^k_{i-T_{\text{in}}-3} - M^{k-1}_{i-2T_{\text{in}}-3} \quad (13)$$

with

$$M^1_l = \begin{cases} l & l \leq T_{\text{in}} \\ 2T_{\text{in}} - l & T_{\text{in}} < l \leq 2T_{\text{in}} - 1 \\ 0 & l > 2T_{\text{in}} - 1 \end{cases} \quad (14)$$

and

$$M^k_l = \binom{l}{k}, l - k < T_{\text{in}}. \quad (15)$$

Finally, by combining (9)–(11), R_i is given in (16), shown on the bottom of the page. By substituting (7) and (12)–(16) into (8), the utilization ratio of SR slots R can be obtained.

C. Simulation Results

In this section, simulation results are presented to verify the preceding analysis. The simulation setting is the same as the system model. Specifically, the simulation of each connection

[2]If this is not the case, then the connection has been released before slot i.

$$\Pr\{T_C = i\} = \begin{cases} 0, & \text{for } i \in \{1, 2, \ldots, T_{\text{in}}\} \\ (1-\lambda)^{T_{\text{in}}}, & \text{for } i = T_{\text{in}} + 1 \\ \lambda \cdot (1-\lambda)^{T_{\text{in}}}, & \text{for } i \in \{T_{in} + 2, \ldots, 2T_{in} + 1\} \\ \lambda\big(1 - \mathbf{P}^{i-T_{\text{in}}-1}[1, T_{\text{in}}+1]\big)(1-\lambda)^{T_{\text{in}}}, & \text{for } i > 2T_{\text{in}} + 1 \end{cases} \quad (7)$$

$$R_i = \begin{cases} \frac{1}{T_{\text{in}}+1}, & \text{for } i = T_{\text{in}} + 1 \\ \frac{1}{i} \cdot (2 + \lambda(i - T_{\text{in}} - 2)), & \text{for } i \in \{T_{\text{in}} + 2, \ldots, 2T_{\text{in}} + 1\} \\ \frac{1}{i} \cdot \left(2 + \sum_{k=\left\lfloor \frac{i-T_{\text{in}}-2}{T_{\text{in}}} \right\rfloor}^{i-T_{\text{in}}-2} k \cdot \Pr\{\gamma_{i,T_{\text{in}}} = k\} \right), & \text{for } i > 2T_{\text{in}} + 1 \end{cases} \quad (16)$$

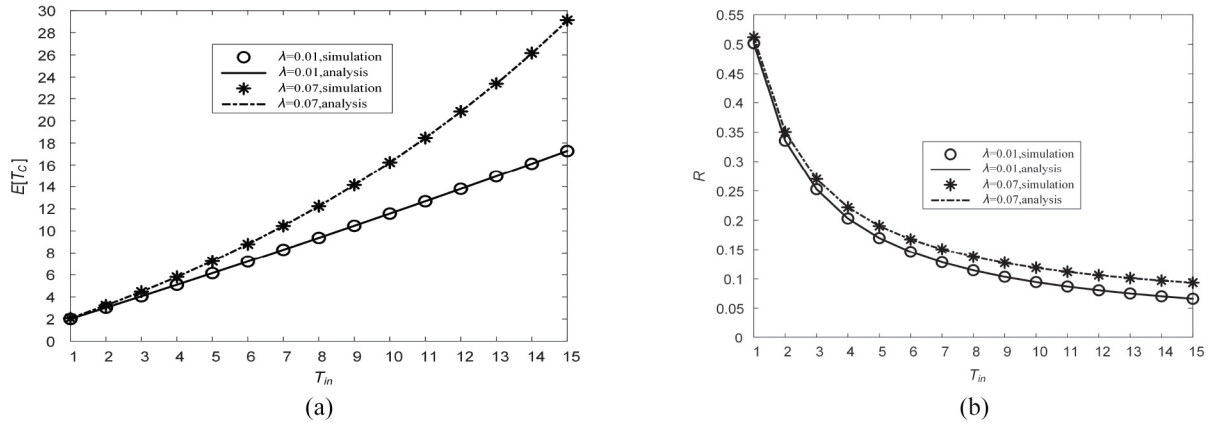


Fig. 3. (a) Mean time length of each connection $E[T_C]$ versus the inactivity timer T_{in} (in unit of slots). (b) Utilization ratio R versus the inactivity timer T_{in} (in unit of slots). $\lambda = 0.01$ or 0.07.

is on a time-slot basis, where different events (e.g., packet generation, SR transmission, uplink resource allocation, and the data packet transmission) happen and are observed at the end of a slot. Each connection starts with a busy slot and the device generates a packet in every slot based on a Bernoulli trial with probability λ. The connection ends once the inactivity timer is timeout, i.e., there are T_{in} consecutive empty slots. Each simulation is carried for 10^7 connections. We count the time length of each connection and the number of busy slots in each connection. Accordingly, the mean time length of each connection is obtained by calculating the ratio of the sum of the time length of all connections to 10^7. The utilization ratio of SR slots is obtained by calculating the ratio of the total number of busy slots in all connections to the sum of the time length of all connections.

Fig. 3 demonstrates how the mean time length of each connection $E[T_C]$ and the utilization ratio R vary with the inactivity timer T_{in} with the packet arrival rate $\lambda = 0.01$ or 0.07. A perfect match between the analysis and simulation results can be clearly observed, which validates the derivations presented in above sections.

We can see from Fig. 3(a) that the mean time length of each connection $E[T_C]$ increases with the inactivity timer T_{in} and the arrival rate λ. Recall that each connection will be released if it does not request packet transmission for T_{in} slots. It is, therefore, intuitive that the connection could be maintained longer if the device frequently sends packets or the inactivity timer T_{in} is larger. On the other hand, the utilization ratio R should be low if the device with light traffic occupies the connection long. Thus, we see from Fig. 3(b) that R declines as λ decreases or T_{in} increases.

IV. Utility Formulation and Optimization

So far, we have demonstrated that the network performance is crucially determined by the inactivity timer T_{in}, which indicates that to optimize the network performance, T_{in} should be carefully selected. In this section, we will formulate the network utility to evaluate the network performance by considering the signaling overheads of devices, the resource utilization ratio, delay, and further study how to properly tune the inactivity timer T_{in} to maximize the network utility. Without loss of generality, the network utility can be written as

$$U = U_{\text{MTD}} + U_{\text{BS}} \tag{17}$$

where U_{MTD} and U_{BS} denote the utility of MTDs and that of the BS, respectively. Table I summarizes notations and corresponding definitions used in this section.

TABLE I
Notations and Definitions

Notations	Definitions
U	Network Utility
U_{MTD}	Utility of MTDs
U_{BS}	Utility of the BS
S_{MTD}	Average total amount of signaling overheads during random access
I_{MTD}	Average total reward of SR transmissions for each connection
c	Signaling overheads of each access attempt
r	Reward for each SR transmission
q	Access request transmission probability
T_{cycle}	Mean length of each connection cycle
β	Penalty factor

A. Utility of MTDs

With the CBRA scheme, each MTD has to establish a connection with the BS via the random access procedure, before it sends SR to the BS for requesting uplink data transmission resources. Therefore, the utility of MTDs should consist of two parts: 1) signaling overheads in the random access process and 2) the reward of SR transmissions in the RRC CONNECTED state, which is given by

$$U_{\text{MTD}} = \frac{I_{\text{MTD}} - S_{\text{MTD}}}{T_{\text{cycle}}} \cdot n \tag{18}$$

where I_{MTD} denotes the average total reward of SR transmissions for each connection and S_{MTD} denotes the average total amount of signaling overheads during random access. The

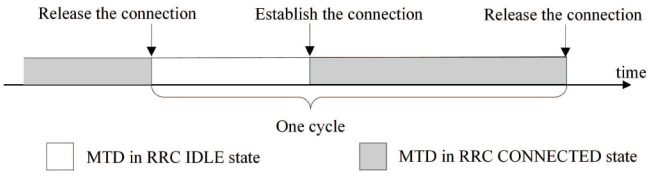


Fig. 4. Graphic illustration of one cycle.

physical meaning of I_{MTD} and S_{MTD} in the practical scenario will be presented in Section V.

Note that since the length of each connection is a random variable that may affect the total utility, in this article, we consider the average utility in unit of time slots. Specifically, define each connection cycle as the time interval between two consecutive instants that the MTD leaves the RRC CONNECTED state. In another word, it is the time interval from the instant the connection is released to the instant that next connection is released for each MTD, as shown in Fig. 4. Let T_{cycle} denote the mean time length of each connection cycle. Thus, in (18), by dividing the mean time length of each connection cycle, the average utility of MTDs qualifies as a fair criterion for comparing the utilities with different lengths of connections.

1) Derivation of S_{MTD}: As the congestion issue mentioned in Section I, when a large number of MTDs send access requests to the BS, severe congestion occurs in the random access channel, in which case many trials have to be made by each MTD until it successfully completes the random access procedure [4].

To efficiently handle the massive number of access requests, BS can tune the access request transmission probability of each MTD,[3] denoted by $q \in (0, 1]$, to reduce the chance of concurrent transmissions, and improve the probability of successful access, denoted by p. The optimal setting of the transmission probability of each MTD for optimizing the access delay[4] performance has been given in [14] as

$$q = \begin{cases} \dfrac{\hat{\lambda}}{n(\hat{\lambda}-e^{-1})}, & \text{if } \hat{\lambda} > \hat{\lambda}_0 \\ \dfrac{4\mathbb{W}_{-1}^2\left(-\frac{\sqrt{\hat{\lambda}}}{2}\right)}{n\left(-2\mathbb{W}_{-1}\left(-\frac{\sqrt{\hat{\lambda}}}{2}\right)-1\right)}, & \text{otherwise} \end{cases} \tag{19}$$

where $\hat{\lambda} = n\lambda$, $\hat{\lambda}_0 \approx 0.48$, and $\mathbb{W}_{-1}(\cdot)$ is one real-valued branch of the Lambert W function. The probability of successful access p is the single nonzero root of the following fixed point [14]:

$$p = \exp\left(-\frac{\hat{\lambda}}{\frac{\hat{\lambda}}{nq}+p}\right) \tag{20}$$

which can be numerically obtained by further combining (19).

[3]The transmission probability of each MTD is referred to as the access classing barring (ACB) factor in the 3GPP standards [31].

[4]The time spent from the generation of an access request until its successful transmission.

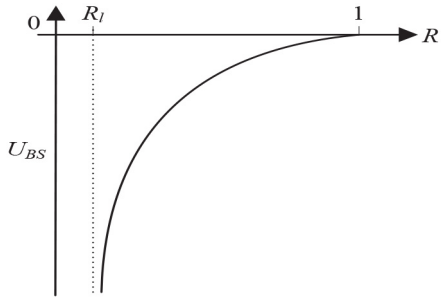


Fig. 5. Graphic illustration of U_{BS}.

Accordingly, the average number of trials that each MTD has to make until it successfully completes the four-way handshake random access procedure can be written as $\sum_{i=1}^{+\infty} ip(1-p)^{i-1} = 1/p$, with which the average total amount of signaling overheads is given by

$$S_{\text{MTD}} = \frac{c}{p} \tag{21}$$

where c denotes the signaling overhead of each access attempt.

2) Derivation of I_{MTD}: After the successful completion of the random access procedure, MTD shifts into the RRC CONNECTED state, in which it has a connection with BS and send SRs. Given the input rate λ, the average number of SRs that each MTD sends during the period of the connection can be written as $\sum_{i=T_{\text{in}}+1}^{i=+\infty} \Pr\{T_C = i\} \cdot R_i \cdot i$ according to (8), where $\Pr\{T_C = i\}$ and R_i are given in (7) and (16), respectively. Let r represent the reward for each SR transmission. Accordingly, the average total reward of SR transmissions for each connection can then be written as

$$I_{\text{MTD}} = r \cdot \sum_{i=T_{\text{in}}+1}^{i=+\infty} \Pr\{T_C = i\} \cdot R_i \cdot i. \tag{22}$$

3) Derivation of T_{cycle}: To derive the mean time length of each cycle T_{cycle}, we can decompose each cycle into three parts.

1) The time interval from the instant the connection is released to the instance that MTD has new packets in its buffer. As for each MTD in each time slot, arrivals of new packets follow a Bernoulli process with parameter λ. Then, the mean time length of this interval can be written as $1/\lambda$.
2) Access delay, i.e., the time spent from the generation of an access request until its successful transmission. It has been shown in [14] that the mean access delay of each MTD $E[D_T] = (1/qp)$, where p is the single nonzero root of fixed point (20). With the transmission probability of each MTD tuning according to (19), the mean access delay of each MTD can be explicitly written as

$$E[D_T] = \begin{cases} ne - \frac{n}{\hat{\lambda}}, & \text{if } \lambda > \frac{\hat{\lambda}_0}{n} \\ \dfrac{n\left(-2\mathbb{W}_{-1}\left(-\frac{\sqrt{\hat{\lambda}}}{2}\right)-1\right)}{4M\mathbb{W}_{-1}^2\left(-\frac{\sqrt{\hat{\lambda}}}{2}\right)p}, & \text{otherwise.} \end{cases} \tag{23}$$

3) The time interval from the instant that the connection is established to the instant that the connection is released.

The mean time length of the connection $E[T_C]$ has been given in (1).

By combining the mean time length of each of those three parts together, we can conclude that the mean time length of each cycle

$$T_{\text{cycle}} = \frac{1}{\lambda} + E[D_T] + E[T_C]. \tag{24}$$

Finally, by substituting (19)–(24) into (18), we can have the utility of MTDs as

$$U_{\text{MTD}} = \frac{r\sum_{i=T_{\text{in}}+1}^{i=+\infty} \Pr\{T_C=i\} R_i i - \frac{c}{p}}{\frac{1}{\lambda}+E[D_T]+E[T_C]} \cdot n. \tag{25}$$

B. Utility of BS

In this article, we assume that the BS puts the major concern on the utilization ratio of the SR resources R and therefore, the utility of BS U_{BS} is determined by R only. Without a loss in generality, we assume that the BS would impose a lower bound on R, denoted by $R_l \in [0, 1]$. It is intuitively clear that a high resource utilization ratio R is always preferred and R should not below the lower bound R_l. Accordingly, the following logarithmic penalty function is adopted for evaluating the utility of BS:

$$U_{\text{BS}} = \begin{cases} -\beta\left(\frac{\ln(2-R_l)}{\ln(R-R_l+1)} - 1\right), & \text{if } R > R_l \\ -\infty, & \text{otherwise} \end{cases} \tag{26}$$

where $\beta > 0$ represents the penalty factor. A graphic illustration of U_{BS} is presented in Fig. 5, from which we can see when the network achieves full utilization of the resource, i.e., $R = 1$, the penalty $U_{\text{BS}}|_{R=1} = 0$. On the other hand, when $R \le R_l$, we have $U_{\text{BS}}|_{R\le R_l} = -\infty$, which indicates that the BS would refuse the connection request.

Finally, by combining (19)–(26), the network utility in (17) can be calculated.

C. Optimal Inactivity Timer

Typically, the input rate λ, the number of MTDs n, the signaling overhead c, the reward for each SR transmission r, and the penalty factor β are system input parameters. Therefore, in this article, we aim to maximize the network utility U by optimally choosing the inactivity timer T_{in}, i.e.,

$$\begin{aligned} U_{\max} = \max_{T_{\text{in}}\ge 1} \quad & U \\ \text{s.t.} \quad & U_{\text{MTD}} > 0. \end{aligned} \tag{27}$$

Let T_{in}^* denote the optimal inactivity timer to maximize the network utility. Yet, the implicit nature of (17) makes it hard, if not impossible, to derive the explicit expression of T_{in}^*. Therefore, to obtain T_{in}^*, we propose an exhaustive search algorithm in Algorithm 1. The basic idea of Algorithm 1 is to obtain the network utility U for each value of T_{in} for $T_{\text{in}} \in \{0, \ldots, T_{W,\max}\}$, where $T_{W,\max}$ is sufficiently large, and finally, choose the value of T_{in} that maximizes U while satisfies the constraints $U_{\text{MTD}} > 0$. Note that although explicit expressions of the maximum network utility $U_{\max}$ and the corresponding T_{in}^* cannot be obtained, the numerical results presented below also shed important light on the practical system design.

Algorithm 1 Calculation of T_{in}^* and $U_{\max}$

1: Input n, λ, c, r, β and initialize $T_{\text{in}} = 1$, $U_{\max} = 0$.
2: **repeat**
3: Obtain $E[T_C]$ and R according to (1), (7), (8), (12)–(16).
4: Calculate U_{MTD}, U_{BS} and U according to (17), (25) and (26).
5: **If** $U > U_{\max}, U_{\text{MTD}} > 0$ **then**
6: $U_{\max} = U$ and $T_{\text{in}}^* = T_{\text{in}}$.
7: **endif**
8: $T_{\text{in}} = T_{\text{in}} + 1$.
9: **until** $T_{\text{in}} > T_{in,\max}$
10: Output T_{in}^* and $U_{\max}$.

V. Simulation Results and Discussion

In this section, simulation results are presented to validate the above analysis. Specifically, the simulation is divided into two parts. The first part is the simulation of the random access of MTDs and the second part is the simulation of the connection. For the first part, the simulation of random access is on a time-slot basis. For each MTD, the arrivals of data packets follow a Bernoulli process. The MTD with nonempty queue would transmit the access request with probability q. If more than one MTDs transmit the access requests over the random access channel at the same time slot, then a collision occurs and all of them fail. The access request is successful if and only if there is one single MTD transmitting at each time slot.[5] Upon successful request, a connection is established and then the second part of the simulation will be performed based on the details that have been presented in the first paragraph of Section III-C.

Note that each simulation in the random access part lasts for 10^7 time slots. To obtain the utility of MTDs and BS in simulations, we count: 1) the total number of access request transmissions of all MTDs; 2) access delay of each access request; 3) the time length of all connections; and 4) total number of busy slots in all connections. Parameter settings of simulations will be given at the caption of each figure.

In the following, we will first focus on how the inactivity timer T_{in} affects the utility, and then evaluate the effect of system parameters, including the access request transmission probability q, signaling overhead of each access attempt c, reward for each SR transmission r, and the penalty factor β on the optimal network performance.

1) Utility: Fig. 6 demonstrates how the utility of BS, utility of MTDs, and utility of the network vary with the inactivity timer T_{in}. Recall that it has been shown in Fig. 3(b) that the utilization ratio of SR resource R declines as the inactivity timer T_{in} increases. Accordingly, we can see from Fig. 6(a) that the utility of BS drops, in other words, the penalty increases, and the increasing rate sharply grows with T_{in} as the utilization ratio of SR resource R is close to its lower bound. Note that

[5]In reality, the access request may not be correctly decoded even without concurrent transmissions if the channel condition is too poor. How to take the channel condition into consideration is an important future direction of this work.

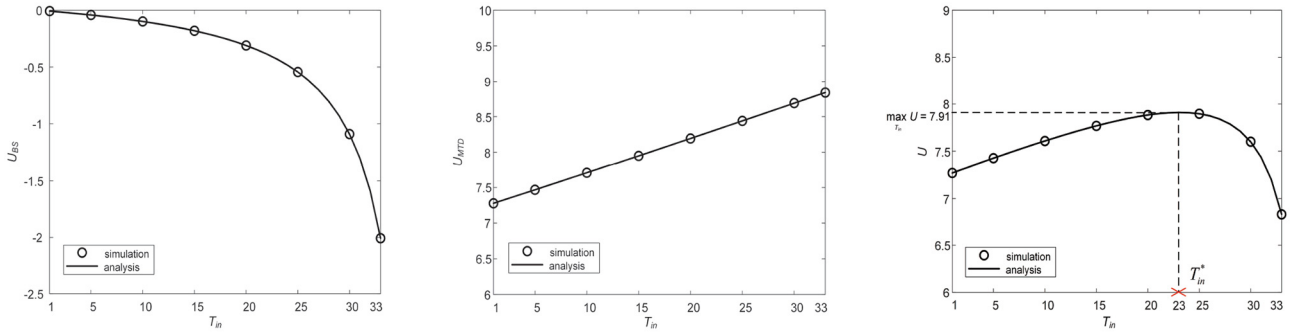


Fig. 6. Utility of BS U_{BS}, utility of MTDs U_{MTD}, and utility of the network U versus the inactivity timer T_{in}. $n = 100$, $\lambda = 0.01$, $c = 0.1$, $R_l = 0.03$, $r = 20$, and $\beta = 0.01$. (a) U_{BS} versus T_{in}. (b) U_{MTD} versus T_{in}. (c) U versus T_{in}.

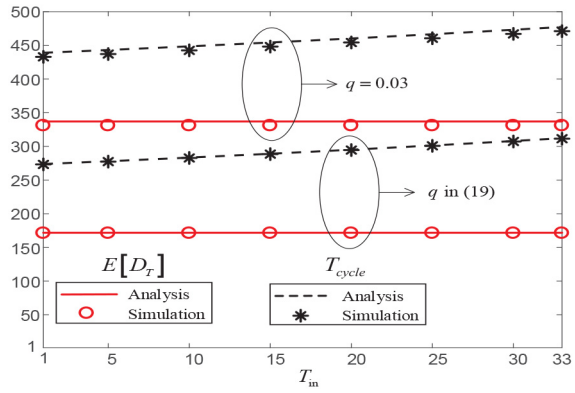


Fig. 7. Mean access delay of each MTD $E[D_T]$ (in unit of slots) and the mean time length of each cycle T_{cycle} (in unit of slots) versus the inactivity timer T_{in}. $n = 100$. $\lambda = 0.01$. $q = 0.03$ or (19).

according to Fig. 3(a), a larger T_{in} indicates a larger mean time length of each connection. In this case, MTD can transmit more packets per connection and receives more reward in average. Therefore, Fig. 6(b) shows that the utility of MTDs grows with the inactivity timer T_{in}.

A closer look at Fig. 6(a) and (b) reveals that although a small inactivity timer T_{in} can reduce the penalty, it reduces the utility of MTDs as well. Increasing T_{in} boosts the revenue that the MTD can obtain, yet leads to prohibitively high penalty. Therefore, as shown in Fig. 6(c), the network utility is crucially determined by the inactivity timer T_{in}. By optimally choosing $T_{\mathrm{in}} = T_{\mathrm{in}}^* = 23$, we can see that the network utility U is maximized at $U_{\max} = 7.91$, where T_{in}^* and $U_{\max}$ are obtained according to Algorithm 1. The perfect match between simulation results and the analysis in Fig. 6 verifies the derivations we present above.

2) Effect of Access Request Transmission Probability: Note that the analytical framework in this article not only considers the time length of each connection when the device is in the RRC CONNECTED state but also incorporates the access delay of each device when it is in the RRC IDLE state. The access delay performance is critically determined by the access request transmission probability q [10]–[13]. To minimize the mean access delay, we assume q is always optimally tuned according to (19), which depends on the number of MTDs n and the aggregate input rate $\hat{\lambda}$.

In practice, the access request transmission probability q is preselected from a certain range and might be unable to adaptively tuned according to n and λ [13]. To see the performance loss, Fig. 7 shows how the mean access delay of each MTD $E[D_T]$ and the mean time length of each cycle T_{cycle} vary with the inactivity timer T_{in} with $q = 0.03$ or q in (19). According to (24) and Fig. 7, we can clearly observe that $E[D_T]$ is a significant component of T_{cycle}. If the access request transmission probability q is improperly configured, i.e., $q = 0.03$, then the mean access delay $E[D_T]$ would be much larger than that with q in (19). As a result, T_{cycle} is boosted, which degrades the utility according to (18). It can, therefore, be concluded that adaptive tuning of access request transmission probability is indispensable for optimizing the connection management for mMTC.

3) Effect of Transmission Reward and Signaling Overhead: In this article, we introduce r to denote the reward for each SR transmission, and c to denote the signaling overhead for each access request transmission for each MTD. In the practical scenario, each SR transmission indicates at least one uplink data transmission that may contain crucial information for ensuing the safe operation of the system, particularly in mission-critical M2M applications. The coefficient r is therefore introduced for quantifying the contribution of each SR/data transmission. Fig. 8 shows how the optimal inactivity timer T_{in}^* and the corresponding maximum network utility $U_{\max}$ vary with the reward for each SR transmission r with the signaling overhead $c = 0.1$ or 1 and the input rate $\lambda = 0.01$ or 0.02. When r or λ becomes large, the system should extend the mean length of each connection, such that the MTD can transmit more data packets and receives more income. Accordingly, we can see from Fig. 8 that as r or λ grows, both the optimal inactivity timer T_{in}^* and the corresponding maximum network utility $U_{\max}$ increase.

On the other hand, the signaling overhead c can refer to, for instance, one unit of energy consumption for access request transmission. As MTDs are usually with low complexity and limited battery life, if c is large, then the cost for establishing a connection might be high. In this case, the system should extend the mean length of each connection, such that the MTD performs the access procedure less frequently for reducing the energy consumption in time average. The above intuition is

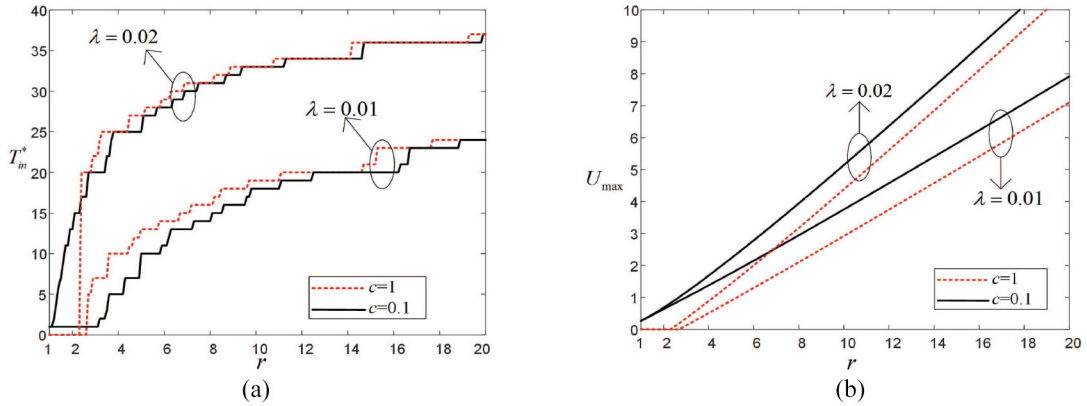


Fig. 8. Optimal inactivity timer T_{in}^* and maximum network utility $U_{\max}$ versus the reward for each SR transmission r. $n = 100$, $\lambda = 0.01$ or 0.02, $c = 0.1$ or 1, $R_l = 0.03$, and $\beta = 0.01$. (a) T_{in}^* versus r. (b) $U_{\max}$ versus r.

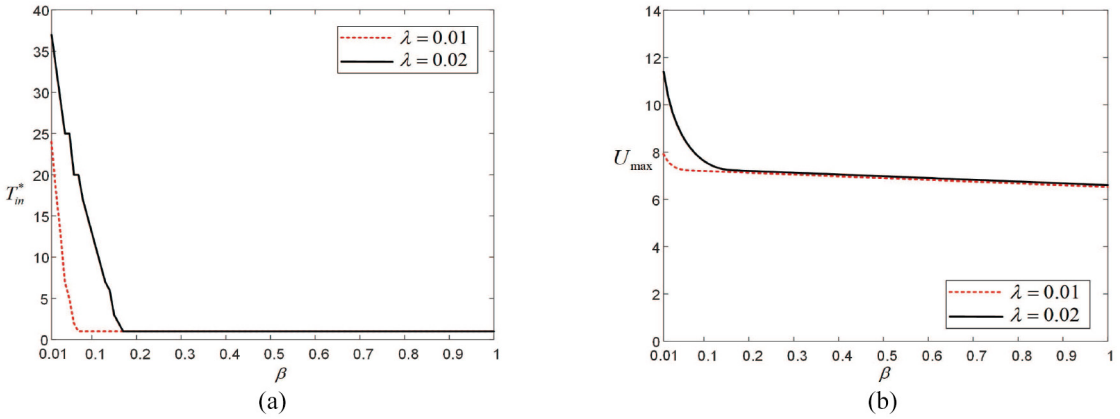


Fig. 9. Optimal inactivity timer T_{in}^* and the corresponding maximum network utility $U_{\max}$ versus the penalty factor β. $n = 100$, $\lambda = 0.01$ or 0.02, $c = 0.1$, $r = 20$, $R_l = 0.03$. (a) T_{in}^* versus β. (b) $U_{\max}$ versus β.

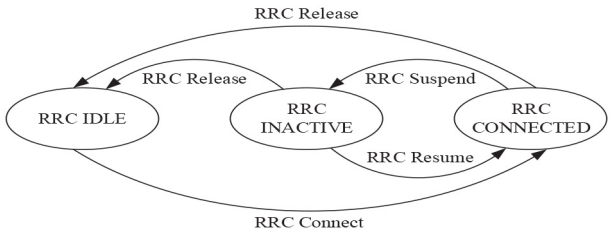


Fig. 10. RRC state machine in 5G.

confirmed by the observation in Fig. 8, where we can see that with a larger signaling overhead c, the optimal inactivity timer T_{in}^* increases as well.

4) Effect of Penalty Factor: Fig. 9 demonstrates how the optimal inactivity timer T_{in}^* and the corresponding maximum network utility $U_{\max}$ vary with the penalty factor β. Note that β indicates the level of sensitivity that the system has toward the SR resource utilization ratio. According to (26), with a large penalty factor β, the utility of BS would drop quickly as the resource utilization ratio R declines. Since R is a decreasing function of T_{in}, we can, therefore, observe from Fig. 9 that when the penalty factor β increases, both the optimal inactivity timer T_{in}^* and the corresponding maximum network utility $U_{\max}$ decrease.

When the penalty factor β is large, the resource utilization ratio R would be the dominant role in determining the network utility, such that improving R is of critical important. Note that after the successful random access procedure, each MTD certainly sends an SR at the first SR slot. If the input rate of each MTD λ is small, then the MTD would be unlikely to transmit another SR in the next few SR slots, leading to the underutilization of SR resources. Therefore, the optimal strategy for maximizing R could be to set $T_{\text{in}} = 1$, in which case the BS releases the connection after the MTD sends its first SR. Accordingly, we can see from Fig. 9(a) that when the penalty factor β is large, T_{in}^* would finally converge to 1.

VI. Insights on mMTC in 5G: Connection-Based Random Access Versus Packet-Based Random Access

In the above sections, for the CBRA scheme, we have derived the expressions of key performance metrics, based on which the network utility framework is established and optimized by properly tuning the inactivity timer. The effect of key system parameters has been investigated. The analysis above sheds important light on the access scheme selection for mMTC in 5G networks and beyond.

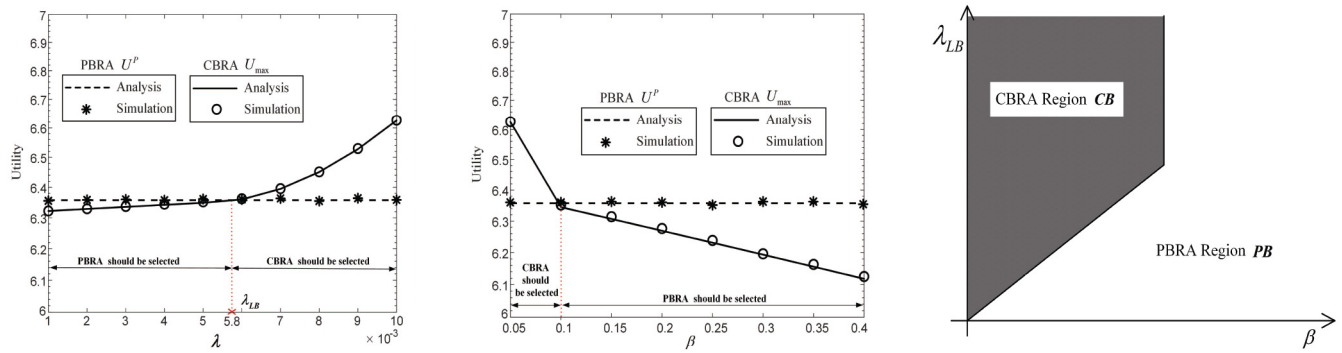


Fig. 11. (a) Network utility with CBRA scheme $U_{\max}$ and the network utility with PBRA scheme U^P versus the traffic input rate λ. $\beta = 0.05$. (b) $U_{\max}$ and U^P versus the penalty factor β. $\lambda = 0.01$. $n = 1000$. $r = 20$. $c = 1$. q in (19). (c) Graphical illustration of **CB** and **PB**.

Specifically, as we have pointed out in Section I-B, the PBRA scheme and the CBRA scheme coexist in the 5G system. From the RRC layer point of view, a new RRC state, i.e., RRC INACTIVE, is introduced in the 5G RRC state machine [32], as shown in Fig. 10. With this new RRC state, the BS may configure MTDs that intermittently transmit small packets via PBRA. That is, the MTD in RRC INACTIVE state can transmit one small packet in the third step of the random access procedure, without reestablishing a connection with BS and moving into the RRC CONNECTED state, which saves signalling and resource [33]. The PBRA scheme is suitable for sporadic small packet transmission. The CBRA scheme, on the other hand, is preferred in the scenario where MTDs frequently deliver packets. As a result, to provide efficient services for mMTC applications with different traffic characteristics, the random access scheme should be properly chosen.

To address this issue, the utility-based analytical framework proposed above will be extended to the PBRA scheme. Intuitively, the CBRA scheme should be chosen only if establishing a connection is beneficial. From the perspective of utility, establishing a connection is beneficial, which indicates that the network utility with the CBRA scheme is larger than that with the PBRA scheme.

Let us now derive the network utility of the PBRA scheme, denoted by U^P. For the PBRA scheme, the random access and data transmission process is completed in the random access procedure and does not consume the SR resource [28]. Therefore, the penalty due to SR resource underutilization $U^P_{\mathrm{BS}} = 0$. As each MTD transmits only one packet in random access procedure, we have the utility of MTDs as $U^P_{\mathrm{MTD}} = [n(r - [c/p])]/[(1/\lambda) + E[D_T]]$. Finally, the network utility of the PBRA scheme can then be written as

$$U^P = U^P_{\mathrm{BS}} + U^P_{\mathrm{MTD}} = \frac{n\left(r - \frac{c}{p}\right)}{\frac{1}{\lambda} + E[D_T]}. \tag{28}$$

We can see from (28) that U^P is determined by the number of MTDs n, the signaling overhead c, the probability of successful transmissions p, the traffic input rate λ, and the mean access delay $E[D_T]$.

In this article, we are interested in characterizing the critical threshold in terms of λ, denoted by λ_{LB}, above which the CBRA scheme is more beneficial compared to the PBRA scheme. λ_{LB} can be characterized as

$$\lambda_{\mathrm{LB}} = \min\left\{\lambda | U_{\max} > U^P\right\} \tag{29}$$

where $U_{\max}$ is the maximum network utility with the CBRA scheme, which has been given in (27).

Fig. 11(a) and (b) demonstrates how $U_{\max}$ and U^P vary with the traffic input rate λ and the penalty factor β in the massive access scenario where the number of MTDs $n = 1000$. We can clearly see from Fig. 11(a) that the network utility of the PBRA scheme U^P is insensitive to the variation of λ. Meanwhile, the maximum utility of CBRA scheme $U_{\max}$ grows with λ, because more packets are transmitted in each connection, which improves not only the utilization ratio of SR resources R but also the average total reward of SR transmissions. Accordingly, it can be seen that when $\lambda < \lambda_{\mathrm{LB}} \approx 0.0058$, we have $U_{\max} < U^P$, implying that in this case, the PBRA scheme should be chosen; when $\lambda \geq \lambda_{\mathrm{LB}}$, the CBRA scheme should be chosen. On the other hand, in Fig. 11(b), we can observe that the maximum utility of CBRA scheme $U_{\max}$ decreases with the penalty factor β. The network utility of the PBRA scheme U^P does not change because in the PBRA scheme, packets are transmitted in the random access procedure and no SR resource is needed. Therefore, the CBRA scheme should be chosen only if β is small.

To gain a deeper understanding on the the effect of β, let us define the CBRA region **CB** $= \{\lambda | \lambda \geq \lambda_{\mathrm{LB}}\}$, in which the system should choose the CBRA scheme, and the PBRA region **PB** $= \{\lambda | \lambda_{\mathrm{LB}} > \lambda > 0\}$, in which the PBRA scheme should be chosen. A graphical illustration of **CB** and **PB** in terms of β is given in Fig. 11(c). Since $U_{\max}$ declines as the penalty factor β increases. Accordingly, with a larger β, it is more likely that $U_{\max} < U^P$, which enlarges the threshold λ_{LB}, as demonstrated in Fig. 11(c). If the BS is too sensitive to SR resource utilization or the SR resource is too limited, i.e., a large β, then the penalty in the case of CBRA would be excessive such that $U_{\max} < U^P$ always hold and the system always in the PBRA region **PB**, indicating that the PBRA scheme is preferred.

VII. Conclusion

In this article, we develop a utility-based analytical framework for the optimal connection management of mMTC in 5G networks. We first derive key performance measures, including the mean time length of each connection and the resource utilization ratio, based on which a unified analytical framework is established and is capable of characterizing the network utility of the CBRA scheme and that of the PBRA scheme. For maximizing the network utility of the CBRA scheme, an algorithm is proposed to obtain the optimal inactivity timer and the corresponding maximum network utility. By comparing the utility of CBRA scheme and that of PBRA scheme, we characterize the CBRA region and PBRA region in terms of the traffic input rate for addressing the access scheme selection issue.

The analysis in this article shows that the optimal inactivity timer increases with the signaling overheads, the reward for SR transmissions, and the traffic input rate. While the optimal inactivity timer and the corresponding maximum network utility decrease with the resource penalty factor. With a large resource penalty factor, i.e., a high level of sensitivity that the system has toward the SR resource utilization ratio, the utility of the CBRA scheme would be small and the system should choose the PBRA scheme for mMTC services.

Note that in this article, we do not consider the uplink resource constraint on SR transmission. In practice, the BS may release the connections with MTDs due to resource constraints rather than no traffic exchange during the inactivity timer. It is, therefore, of great importance to further extend the analysis to incorporate the resource constraint. Moreover, we do not consider the discontinuous reception operation of each device in RRC states, when we characterize the behavior of each device. How the discontinuous reception operation affects the setting of inactivity timer is an interesting topic that deserves much attention in the future study.

Appendix
Derivation of (12)

To derive the PMF of $\gamma_{i,T_{\text{in}}}$ (i.e., the number of busy slots in the period from slot 2 to slot $i-T_{\text{in}}-1$ given no streak of T_{in} empty slots), let us first simplify the notation for easing the understanding by denoting

$$L = i - T_{\text{in}} - 2 \text{ and } x = \gamma_{i,T_{\text{in}}}. \tag{30}$$

Equivalently, we derive the PMF of the number of busy slots within L slots given no streak of T_{in} empty slots. It can be easily obtained that $x \in \{\lfloor (L/T_{\text{in}})\rfloor, \lfloor (L/T_{\text{in}})\rfloor + 1, \ldots, Y\}$ and the PMF of x as

$$p(x) = \frac{\lambda^x(1-\lambda)^{L-x}\cdot M_L^x}{\sum_{x=\lfloor L/T_{\text{in}}\rfloor}^{i-T_{\text{in}}-2}\lambda^x(1-\lambda)^{L-x}\cdot M_L^x} \tag{31}$$

where M_L^x represents the total number of possibilities for x busy slots within L slots given no streak of T_{in} empty slots. In the following, we focus on deriving M_L^x and divide the discussion into two parts depending on whether $L-x \le T_{\text{in}}-1$ or not.

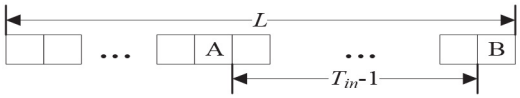


Fig. 12. Illustration of the special slot A and slot B.

First, for $L - x \le T_{\text{in}} - 1$, in this case, there will be no streak of T_{in} empty slots no matter how x busy slots are distributed throughout L slots. Accordingly, the total number of possibilities is given by

$$M_L^x = \binom{L}{x}, \text{ for } L - x \le T_{\text{in}} - 1. \tag{32}$$

On the other hand, for $L - x > T_{\text{in}} - 1$, there will be a streak of T_{in} empty slots if x busy slots are improperly placed within L slots. To derive M_L^x in this case, we adopt a recursive approach by focusing on two cases depending on whether the last slot is empty or not. For simplicity, we refer the last slot as slot B and the slot that is $T_{\text{in}} - 1$ slots ahead of slot B, as slot A, as show in Fig. 12.

1) If the slot B is a busy slot, then we can exclude it from the period of interest and focus only on the remaining part, which contains $x-1$ busy slots and is $L-1$ slots long. For the remaining part, the total number of possibilities for $x-1$ busy slots within $L-1$ slots given no streak of T_{in} empty slots is given by M_{L-1}^{x-1}.
2) If the slot B is an empty slot, then x busy slots are distributed throughout $L-1$ slots and the corresponding number of possibilities is M_{L-1}^x if the slot B is excluded. Yet, with slot B being an empty slot, there should be no streak of $T_{\text{in}} - 1$ empty slots ahead of slot B, (which indicates that slot A should be a busty slot). It is equivalent to calculate the total number of possibilities for $x-1$ busy slots within $L-T_{\text{in}}-1$ slots given no streak of T_{in} empty slots which can be denoted by $M_{L-T_{\text{in}}-1}^{x-1}$. Therefore, the total number of possibilities with slot B being empty is $M_{L-1}^x - M_{L-T_{\text{in}}-1}^{x-1}$.

By combining the number of possibilities in both cases, i.e., slot B is either a busy slot or empty slot, we have

$$M_L^x = M_{L-1}^{x-1} + M_{L-1}^x - M_{L-T_{\text{in}}-1}^{x-1}, \text{ for } L - x > T_{\text{in}} - 1. \tag{33}$$

To this end, we have gotten the recursive expression of M_L^x in (32) and (33). Next, we will calculate the initial condition, i.e., only one busy slot within l slots given no streak of T_{in} empty slots by focusing on the following three cases.

1) For $l < T_{\text{in}}$, there will be no streak of T_{in} empty slots no matter how the busy slot is placed within l slots. Accordingly
$$M_l^1 = l. \tag{34}$$
2) For $T_{\text{in}} < l \le 2T_{\text{in}}-1$, the busy slot should be put around the middle of the period to avoid a streak of T_{in} empty slots on both sides. The possible locations for proper placement is $[l-(T_{\text{in}}-1), T_{\text{in}}-1]$ (suppose that slots in the period are indexed increasingly from 1 to T_{in}). Accordingly, we have
$$M_l^l = 2T_{\text{in}} - l. \tag{35}$$

3) For $l > 2T_{\text{in}} - 1$, there will be a streak of T_{in} empty slots no matter how the busy slot is placed within l slots, i.e.,

$$M_l^1 = 0. \tag{36}$$

Combining (34)–(36) yields the initial condition in (14). Finally, (12) can be obtained by combining (30)–(36).

REFERENCES

[1] V. B. Mišić and J. Mišić, *Machine-to-Machine Communications: Architectures, Standards and Applications*. Boca Raton, FL, USA: CRC Press, 2014.

[2] "Cisco visual networking index: global mobile data traffic forecast update, 2017–2022," Cisco, San Jose, CA, USA, White Paper, Feb. 2019.

[3] V. W. S. Wong, R. Schober, D. W. K. Ng, and L. C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[4] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.

[5] C.-H. Wei, G. Bianchi, and R.-G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, Apr. 2015.

[6] J. J. Nielsen, D.-M. Kim, G. C. Madueño, N. K. Pratas, and P. Popovski, "A tractable model of the LTE access reservation procedure for machine-type communications," in *Proc. IEEE Global Commun. Conf. (Globecom)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[7] R. R. Tyagi, F. Aurzada, K.-D. Lee, and M. Reisslein, "Connection establishment in LTE-A networks: Justification of poisson process modeling," *IEEE Syst. J.*, vol. 11, no. 4, pp. 2383–2394, Dec. 2017.

[8] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, Dec. 2017.

[9] J. Choi, "On the adaptive determination of the number of preambles in RACH for MTC," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1385–1388, Jul. 2016.

[10] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec. 2016.

[11] C. Di, B. Zhang, Q. Liang, S. Li, and Y. Guo, "Learning automata-based access class barring scheme for massive random access in machine-to-machine communications," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6007–6017, Aug. 2019.

[12] Y. Sim and D.-H. Cho, "Performance analysis of priority-based access class barring scheme for massive MTC random access," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5245–5252, Dec. 2020.

[13] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in LTE networks: Modeling and throughput optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771–2785, Apr. 2018.

[14] W. Zhan and L. Dai, "Access delay optimization of M2M communications in LTE networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1675–1678, Dec. 2019.

[15] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) Protocol Specification, V13.3.0*, 3GPP Standard TS 36.331, Jan. 2017.

[16] E. Dahlman, S. Parlvall, and J. Sköld, *4G: LTE/LTE-Advanced for Mobile Broadband*. Amsterdam, The Netherlands: Elsvier, 2014.

[17] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures, V14.2.0*, 3GPP Standard TS 36.213, Apr. 2017.

[18] "4G–5G interworking RAN-level and CN-level interworking," Samsung, Seoul, South Korea, White Paper, Jun. 2017.

[19] G. Foddis, R. G. Garroppo, S. Giordano, G. Procissi, S. Roma, and S. Topazzi, "LTE traffic analysis for signalling load and energy consumption trade-off in mobile networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 6005–6010.

[20] S. Lin, J. Yu, and X. Jiang, "Signalling overhead analysis of small data transmission for machine type communication," in *Proc. IEEE 18th Int. Conf. Commun. Technol. (ICCT)*, Chongqing, China, Oct. 2018, pp. 664–668.

[21] Y. Yamada and Y. Miyao, "Impact of inactivity timer on performance of control plane in LTE core network under burst traffic," in *Proc. IEEE Int. Conf. Commu. (ICC)*, London, U.K., Jun. 2015, pp. 3131–3136.

[22] J. Puttonen, E. Virtej, I. Keskitalo, and E. Malkamäki, "On LTE performance trade-off between connected and idle states with always-on type applications," in *Proc. IEEE 23rd Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sydney, NSW, Australia, Sep. 2012, pp. 981–985.

[23] X. Wang, M.-J. Sheng, Y.-Y. Lou, Y.-Y. Shih, and M. Chiang, "Internet of Things session management over LTE—Balancing signal load, power, and delay," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 339–353, Jun. 2016.

[24] Y. Qi, A. U. Quddus, M. A. Imran, and R. Tafazolli, "Semi-persistent RRC protocol for machine-type communication devices in LTE networks," *IEEE Access*, vol. 3, pp. 864–874, 2015.

[25] S. Huang, G. Feng, L. Liang, and S. Qin, "Power-saving coercive sleep mode for machine type communications," in *Proc. 23rd Asia–Pac. Conf. Commun. (APCC)*, Perth, WA, Australia, Dec. 2017, pp. 1–6.

[26] Y. Gao and L. Dai, "Random access: Packet-based or connection-based?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2664–2678, May 2019.

[27] C. Kahn and H. Viswanathan, "Connectionless access for mobile cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 26–31, Sep. 2015.

[28] A. Höglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3GPP release 15 early data transmission," *IEEE Commun. Stand. Mag.*, vol. 2, no. 2, pp. 90–96, Jun. 2018.

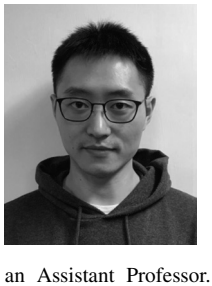
[29] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification, V12.5.0*, 3GPP Standard TS 36.321, Apr. 2015.

[30] Y. Li, X. Cheng, Y. Cao, D. Wang, and L. Yang, "Smart choice for the smart grid: Narrowband Internet of Things (NB-IoT)," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1505–1515, Jun. 2018.

[31] *Service Accessibility, V11.3.0*, 3GPP Standard TS 22.011, Apr. 2013.

[32] *5G; NR; Radio Resource Control (RRC); Protocol Specification, V15.6.0*, 3GPP Standard TS 38.331, Jul. 2019.

[33] S. Hailu, M. Säily, and O. Tirkkonen, "RRC state handling for 5G," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 106–113, Jan. 2019.



Wen Zhan (Member, IEEE) received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2012 and 2015, respectively, and the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2019.

He was a Research Assistant and a Postdoctoral with the City University of Hong Kong. Since 2020, he has been with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China, where he is currently an Assistant Professor. His research interests include Internet of Things, modeling, and performance optimization of next-generation mobile communication systems.

Chen Xu received the B.S. degree from Nanjing University of Science and Technology, Nanjing, China, in 2018, where he is currently pursuing the M.S. degree with the School of Electronic and Optical Engineering.

His research interests include wireless communications, signal processing, and Internet of Things.

Xinghua Sun (Member, IEEE) received the Ph.D. degree from the City University of Hong Kong (CityU), Hong Kong, in 2013.

In 2010, he was a visiting student with INRIA, Rennes, France. In 2013, he was a Postdoctoral Fellow with CityU. From 2015 to 2016, he was a Visiting Scholar with the University of British Columbia, Vancouver, BC, Canada. From 2014 to 2018, he was an Associate Professor with Nanjing University of Posts and Telecommunications, Nanjing, China. Since 2018, he has been an Associate Professor with Sun Yat-sen University, Shenzhen, China. His research interests are in the area of wireless networking and Internet of Things.

Jun Zou (Member, IEEE) received the B.Eng. and Ph.D. degrees in communication and information system from Nanjing University of Science and Technology, Nanjing, China, in 2011 and 2016, respectively.

He is currently an Associate Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology. In 2019, he was a Research Associate with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. His research interests are in the areas of wireless communications, signal processing, and Internet of Things.