Load Estimation and Connection Request Barring for Random Access in Massive C-IoT

Riri Fitri Sari, Senior Member, IEEE, Ruki Harwahyu, Member, IEEE, Ray-Guang Cheng, Senior Member, IEEE

Abstract—Cellular Internet of Things (C-IoT) is a new term referring to access technology introduced by 3GPP to support IoT applications. C-IoT has several advantages compared to unlicensed Low Power Wide Area (LPWA) technology. However, C-IoT’s random access (RA) procedure is easily overloaded by massive access from IoT applications. This work proposes a novel mechanism which combines a simple algorithm to assess the instantaneous load and a dynamic tuning of the contention level on the third message of RA. The technical consideration for integrating our mechanism to the existing C-IoT system is also presented. Performance evaluation is conducted under several practical scenarios. The results show that our proposed mechanism brings a significant increase of normalized throughput with cost of just slight increase of the access delay compared to the original system. Additional performance comparisons with the other existing approaches which target similar problem are also included. The results show that our mechanism performs well, especially when considering how simple it is to be implemented.

Index Terms—C-IoT, LPWA, mMTC, M2M

I. INTRODUCTION

Cellular Internet of things (C-IoT) is a new term referring to access technology introduced by 3GPP to support IoT applications. It includes the extended coverage GSM (EC-GSM), LTE for machines (LTE-M), and narrowband IoT (NB-IoT). Compared to regular LTE networks, C-IoT is more suitable for serving cheaper battery-powered IoT devices with small intermittent data transmission. C-IoT enables operators to serve wider area and more IoT devices within one cell.

It is envisioned that there will be 20.4 billion connected devices in 2020 [1]. With this large market potential, it is necessary to provide an effective solution for connecting IoT devices. Different to human-to-human (H2H) communications, IoT communications generally have small packets with intermittent arrival. Although this seems to be trivial at first glance, the biggest challenge raised by IoT for the access network is its massive number of devices which potentially exhaust the available resources in the cell. Additionally, IoT device also needs to be affordable or even cheap to accelerate its adoption in the market, able to operate with small battery for a long time, and fits in various scenarios and applications with different requirements in term of communication capability, security, and reliability. These characteristics limit the usage of conventional wireless technology in the IoT domain.

The small data packets of IoT application requires a slimmer-down data encapsulation and signaling before it is sent to the network. A study in [2] assumes this small packet is around 1 KB. In addition to that, it usually has intermittent transmission, allowing it to save the power by going into the sleep mode. This is possible since most IoT sensor devices only needs to communicate a small data every several minutes or hours. In this case, keeping the connection bearer alive is inefficient. Thus, a connection setup procedure (e.g. handshake, CSMA/CA, etc.) is needed for each transmission. In respect to IoT’s short data and energy constraints, this connection setup procedure needs to be fast.

Massive number of IoT devices per cell is expected in many IoT scenarios. According to the study document considered in 3GPP [3], there can be thousands of devices needs to be served by an access point/base station. A Massive Autonomous Report (MAR) scenario is introduced, in which a lot of devices attempt to send their data in a short period of time. Such massive sporadic access request may congest the radio access network.

With this congestion, many of the precious resources are wasted in a collision, with only a few devices can eventually (after many transmission attempts) transmit their data with higher delay and power consumption. Additionally, it may also disrupt or completely block the access request by other higher-priority services. On top of that, to be affordable and power-efficient, IoT devices need simpler protocols, relaxed crystals oscillator timing accuracy, and simplified baseband signal processing. All of these features of IoT require reengineered connectivity mechanism.

C-IoT emerges as the answer for unlicensed Low Power Wide Area Network (LPWAN) technologies targeting to serve IoT applications with the aforementioned characteristics. Unlicensed LPWAN uses unlicensed band in its service, which comprises technologies such as IEEE 802.11ah, LoRa, SIGFOX, Ingenu RPMA, etc. Meanwhile, C-IoT uses licensed cellular spectrum and thus, C-IoT is often referred to as licensed LPWAN.

Compared to traditional H2H technologies such as LTE, WiFi, WiMAX, Bluetooth, LPWAN has a better link budget...
(can have further coverage beyond obstacles), consumes less energy and have smaller overhead during connection setup procedure. Compared to unlicensed LPWAN technologies, C-IoT inherits the potential advantages from cellular technology: less interference (since it uses unlicensed frequency bands); potentially larger service coverage since it can use the existing cellular infrastructure readily deployed city/nation-wide; and the adoption and operation is usually supported and organized by more mature business model and companies (i.e., telecommunication operators).

Similar to LTE, each device in C-IoT needs to be in connected state to obtain uplink time-frequency resource allocation (often called ‘data bearer’) for data transmission. To decide which devices, among hundreds or even thousands device in the cell, can obtain this limited resources, C-IoT adopts random access (RA) procedure. While the detail of RA procedure is elaborated in the next section, normally, it takes form of a contention-based mechanism, whose throughput and efficiency decrease significantly under simultaneous massive load [4].

Massive access by IoT devices is likely to happen in various IoT applications due to synchronized periodic reporting [3], synchronized paging [5], or simultaneous emergency reporting [6]. Combined with a limited number of sub-carriers in NB-IoT, the devices need to retransmit more attempts in order to be successful (causing longer access delay), more energy is wasted in the collisions and collision resolutions, and the actual number of devices that can utilize the resource for data transmissions is very limited. Additionally, to cover a wider area, different RA configuration may be applied in each coverage enhancement (CE) level, which is made to ensure good signal quality at different Reference Signals Received Power (RSRP). This increases the complexity of solving the overload problem in RA.

This paper aims to improve the access success probability and decrease access delay of RA under massive access from power and computationally constrained IoT devices, by carefully taking into account the support for multiple CE level in C-IoT. The characteristic of RA under high load is well studied, and many improvements have been proposed in the literature, which are briefly discussed in Section II.B. However, most of them require a considerable amount of modification on RA procedure, the method cannot quickly adapt to the load condition, and the model is less robust to deal with various network conditions and multiple CE levels. This work, on the other hand, tries to effectively solve the problem with minimal modification to the existing RA mechanism. Overall, this work has three contributions:

- An agile mechanism to approximate the instantaneous load without the need for historical record.
- A mechanism to increase success probability by dynamically tune the contention level on the third message of RA.
- A model to assess the performance of RA serving multiple CE levels, either it is equipped with the proposed mechanism or not.

The rest of this work is organized as follows. Section II delivers the brief of RA in C-IoT followed by literature studies and state-of-the-art of RA improvements. The system model is then developed in Section III. The two proposed mechanisms are elaborated back-to-back in Section IV. A formal model to assess the proposed mechanisms when implemented in RA is constructed in Section V. Section VI describes the investigation scenarios and discusses the obtained results highlighting the performance of interest. Finally, the conclusion is delivered in Section VII.

II. RANDOM ACCESS, STUDIES, AND IMPROVEMENTS

A. Random Access in C-IoT

RA procedure is a 4-message handshake, which is illustrated in Fig. 1. It starts when UE transmits preamble (number 1 and 3 in Fig. 1) to enhanced node-B (eNB, or base station). Preamble is a physical signal and is transmitted in the shared channel of Physical Random Access CHannel (PRACH)1. It is randomly chosen by UE which has an urge for data transmission from a pool containing several orthogonal preambles (informed to all UEs residing in the cell). While it does not convey any specific information about the transmitting UE, preamble is used to denote that there is (or ‘are’) UE(s) wanting to contend for the available resource to transmit their data. Preamble in C-IoT has two different implementations [7]. In LTE-M and Cat-0, different preambles are represented with different orthogonal symbols. Meanwhile in NB-IoT, only one symbol is available, and different preambles are represented with different sub-carriers where the symbol is transmitted.

![Fig. 1. Random access procedure](image)

Next, eNB replies the detected preamble(s) with random access responses (RARs, number 2 and 4 in Fig. 1). This reply is not directed to certain UEs since one detected preamble may be a resultant from several UEs choosing the same preamble and transmitting it at the same time. It is also due to the fact that the preamble does not convey UE’s identity.

The number of RARs that is sent can be equal to or less than the number of detected preambles. In this case, detected preamble may not be included in RAR. When UE’s transmitted preamble is not acknowledged with RAR, UE needs to conduct backoff (i.e. randomizing retransmission time) and retransmit a new preamble in the next available PRACH. The size of backoff window is informed in RAR. Note that the number of transmissions (including retransmission) per RA attempt (i.e. per one data bearer request) is limited.

Subsequently, UE sends its unique connection request (Msg3, number 5 and 7 in Fig. 1) in a dedicated resource indicated by the RAR associated with the acknowledged preamble. By doing

\[1\] Also called Narrowband PRACH (NPRACH) for narrowband systems such as NB-IoT.
this, one acknowledged preamble is mapped to one resource to transmit Msg3. Hence, when 2 or more UEs transmit the same preamble, and the preambles are acknowledged with RAR, these UEs transmit their Msg3 in the same resource, resulting in a collision. Msg3 may employ hybrid automatic repeat request (HARQ, number 6 in Fig. 1), such that when NACK is received, the UE retransmits Msg3. This is helpful when Msg3 failure is due to fading. When the failure is due to collision, it is resolved via backoff and preamble retransmission.

In the next step, eNB replies the correctly received Msg3 with ACK (8) and Msg4 (9) for connection setup. This informs the UE about the subsequent process until bearer assignment for data transmission. If Msg4 is not received, it can be retransmitted upon a timer (10). It is worth to notice that when a UE already receive Msg3’s ACK, the probability for it to be failed (and cannot proceed to transmit the data packet) is almost zero [8].

C-IoT is designed for battery-constrained and simple (i.e., low computational power) UEs. The technology is also expected to have wider (or ‘deeper’ for basement/indoor usage) coverage. To meet these requirements of LPWA, preamble (and other transmissions) in C-IoT may be repeated several times, depending on UE’s RSRP. eNB may define several RSRP thresholds, each of which corresponds to a CE level. In this case, the lower RSRP, e.g., due to greater distance to eNB, is mapped to the higher CE level. In each CE level, eNB defines how many repetitions each transmission should be conducted, which sub-carriers can be used for preamble transmission, how many retransmissions a UE can perform, size of the backoff window, and several other timing parameters. These are to ensure good reception for UEs with various RSRPs.

Each UE initially transmit their preamble in the sub-carriers which are allocated to its CE level. In this case, contention happens locally in each CE levels. When its preamble or Msg3 is not acknowledged (due to fading, not selected, or collided), retransmission with backoff is executed. When it meets the retransmission limit in the current CE level, it uses the parameters of the next higher CE level and retransmits. Notice that in this case, this UE contends with other UEs from this higher CE level. The process repeats until it reaches the highest CE level, or the absolute limit of retransmissions counts.

**B. Existing Studies and Improvements**

The studies in [9] [10] [11] [12] have investigated the contention in RA from MAC layer perspective. The work in [9] analyzes the normalized throughput and access delay under transient condition. In this work, three CE levels are evaluated without backoff. The work in [10] extended the model and simplified assumptions in [9] by considering implementation details of narrowband PRACH in NB-IoT. This work also proposes an approach to find the optimal configuration which maximizes overall system’s success probability for a given maximum access delay upper bound. The work in [11] enhances NB-IoT to support multicast services. The enhancements were evaluated under the presence of background traffic. The work in [12] investigated a narrowband cognitive radio system which operates with slotted Aloha mechanism to serve IoT services. In this mechanism, each backlogged device can transmit when they found an empty channel. In this work, different channel sensing mechanisms are evaluated for their throughput. To sum up, all of these studies consider perfect preamble detection, which is not always the case for C-IoT which is designed to be able to operate below the noise floor.

The cross-layer perspective of PHY and MAC for contention in narrowband RA is studied in [13] [14]. The work in [13] analyzes success probability under time-correlating interference due to low SNR and collision. This work reveals that repeated transmission is effective in improving success probability under lower access load. However, under heavier load, repeated transmission brings only a slight improvement with very inefficient resource utilization. The work in [14] presents reconfiguration of physical preamble structure by considering the effect of collisions at the MAC layer. In this work, the long preamble sequence is proposed to be chopped into several groups of short preambles. This increases the number of orthogonal opportunities and reduces collision probability. Nevertheless, this work overlooked the fact that the receiver should process all the data in each of the repeated preamble sequence.

Coded RA [15] was introduced to decrease the contention. In this scheme, each transmission attempt consists of multiple preambles in series. Thus, it virtually expands the number of RA opportunities and decreases the chance of collision. Nevertheless, it increases energy and time spending. Queueing approach was introduced in [16] to divide colliding devices into several queueing groups. Each device maintains a local counter to decide the queued transmission time. However, under higher load, this method suffers from high access delay since the queue grows extensively. In [17], the devices are grouped according to their transmission delay requirements and assigned to different time slots accordingly. An added overhead of feedback mechanisms is employed to guarantee the delay requirements. In such setting, when there are too many devices having tight delay requirements, more resource are required to satisfy the delay requirements.

The performance of RA can also be improved by reducing simultaneous high load by spreading device’s preamble transmission time [18] or reduce their probability to transmit the preamble with barring mechanisms [19] [20] [21]. The barring can be implemented to completely prevent certain UEs or class of UEs from transmitting preamble in a certain interval, or reduce their probability to transmit using a barring factor broadcasted in SIB2, which can be adjusted based on the load. The work in [22] enhances the standard Access Class Barring (ACB) scheme by providing mechanism to control RA failure probability. It adaptively adjusts the barring factor and the number of preamble transmissions. The study implies that number of supported devices are decreased as a tradeoff to increase reliability. In [23], ACB is enhanced by a binary countdown contention resolution to allow more successful devices in RA. The configuration of such mechanisms is made adaptive based on the load to maximize success probability and minimize resource consumption. However, since the real collision happens at Msg3 transmission, barring the preamble is less effective. Additionally, SIB2 periodicity is too long, hence less robust for adapting to the fast-changing load.

The fact that preamble is transmitted with repetition in NB-IoT is exploited in [24], which argued that the repetition could be relaxed by time diversity introduced via retransmission, yielding higher success probability under higher RA load. This
is because in principle, reducing repetition decreases the number of detected transmissions, which in turn decreases collusion. However, this yields inefficient power consumption. A study in [25] presented a preamble allocation matrix for prioritization in RA. However, it was presented only to relieve the collision in higher priority classes and move it to lower priority classes without essentially solve the overload problem of RA. In respect to the existence of multiple CE levels, the matrix can be applied for preamble allocation toward multiple CE level.

III. SYSTEM MODEL

The underlining mechanism of RA in C-IoT is similar to Multi-channel Slotted Aloha. The presence of several CE levels can be exhibited as several parallel Multi-channel Slotted Aloha systems that interact in a certain manner for accommodating retransmissions. This work considers a C-IoT cell serving M stationary IoT devices, that is, no handover due to relatively short data transmission duration and low mobility. The devices are spread in an extended coverage of a base station where the base station may configure 1 to C CE levels. Let us then represent different parameters in each CE level with an index c, with 0≤c≤C-1, following CE level numbering used by 3GPP [7].

The number of IoT devices which resides in CE level c is denoted as \( M_c \), such that \( M = \sum_{c=0}^{C-1} M_c \). In practice, \( M_c \) depends on devices’ location and RSRP thresholds of each CE level. Packet generation/arrival per unit time for these devices is represented with a general probability density function \( A_c(t) \). Each packet generation is followed by preamble transmission to initiate RA procedure. Since a device can send only 1 preamble at a time and its transmission time is restricted only at the beginning of a slot (i.e., PRACH), the expected number of arrival (i.e., the first preamble transmission attempt) at the \( i \)-th slot of CE level \( c \), \( m_{i,c}^c \left( 1 \right) \), can simply be calculated as

\[
m_{i,c}^c \left( 1 \right) = M_c \int_{t=t_{i-1}}^{t_{i}} A_c(t) \, dt,
\]

where \( c' \) denotes the CE level where this attempt are conducted (in this case, since it is the first preamble transmission attempt, \( c' = c \)); the upper bound of integral, \( t_i \), implies the starting time of slot \( i \); and the lower bound, \( t \rightarrow t_{i-1} \), implies the time right after the starting time of slot \( i-1 \). This general form of arrival allows it to be expanded when necessary, e.g., for different devices types and QoS classes.

Base station reserves \( R_c \) distinct preambles in CE level \( c \). Since different CE levels may dictate a different number of repetitions per transmission, the duration of a preamble transmission in each CE level may vary. For CE level \( c \), let \( T_c \) be the PRACH periodicity and \( d_c \) be the offset of starting time of the PRACH relative to the beginning of the radio frame. This model assumes that \( T_c \) is enough for a preamble transmission and its corresponding RAR, such that retransmission attempt with backoff counter of 0 can be conducted in the succeeding PRACH. In CE level \( c \), backoff counter is upper bounded by a backoff window, \( B_c \). Each device may transmit up to \( P \) preamble transmission attempts (referred to as “attempts” herein) with up to \( P \) attempts can be conducted at CE level \( c \).

It is assumed that the device whose transmitted preamble is not detected may still transmit Msg3. However, the transmitted Msg3 is also undetected in such case, and by considering the capture effect, it will not cause any collision. The collision in CE level \( c \) is realized by the involved devices \( d_{i,c} \) ms after the beginning of preamble transmission.

IV. PROPOSED MECHANISMS

Based on the above system model, the proposed mechanisms are elaborated in this section. Firstly, the load approximation is explained. The resulted approximation is used to adjust the parameter of the proposed barring mechanism, which is explained subsequently.

A. Load Approximation

As suggested in [26], the number of preambles that is detected by the base station can be useful to estimate the actual number of transmitting devices, i.e., the instantaneous load at the current PRACH. We exploit this to further decrease the collision probability during Msg3 transmission by adjusting the Msg3 baring factor. This is conducted in order to optimally adjust the number of devices which finally transmit Msg3.

Let us denote by \( R_{3,c} \) the number of dedicated resources at PUSCH to transmit Msg3 in CE level \( c \). \( R_{3,c} \) can be equal or less than \( R_c \). Obviously, the optimal number of devices which transmit Msg3 in CE level \( c \), \( m_{3,c}^* \), would be \( R_{3,c} \), as it maximizes the probability of successful Msg3 transmission, \( \Pr[S_{3,c}] \). After a device’s preamble is acknowledged in RAR, its Msg3 transmission will be successful if its preceding preamble does not collide, i.e., no other devices use that preamble. When there are \( R_{3,c} \) acknowledged preambles and \( m_{3,c} \), acknowledged devices, the probability of a device to be successful in its Msg3 transmission can be obtained by multiplying the probability of a preamble is chosen by the device, which is \( \frac{1}{R_{3,c}} \), and the probability that the chosen preamble is not chosen by the other \( m_{3,c} - 1 \) devices, which is \( \left( 1 - \frac{1}{R_{3,c}} \right)^{m_{3,c} - 1} \). Hence, \( \Pr[S_{3,c}] \) can be calculated as

\[
\Pr[S_{3,c}] = \frac{1}{R_{3,c}} \left( 1 - \frac{1}{R_{3,c}} \right)^{m_{3,c} - 1} \approx \frac{1}{R_{3,c}} e^{-\frac{m_{3,c}}{R_{3,c}}}.
\]

From [24], \( \Pr[S_{3,c}] \) can be further approximated as \( \frac{1}{R_{3,c}} e^{-\frac{m_{3,c}}{R_{3,c}}} \), which is useful for our recursive calculation explained later in Section V. Note that Msg3 transmission is successful if the dedicated resource where it is transmitted at is not chosen by other devices.

During preamble reception, base station only knows the number of detected preambles. In fact, base station knows nothing about how many devices transmit each of those detected preambles. Hence, for calculating the estimate of number of devices that transmit preambles, \( \hat{m}_c \), we can only use the number of detected preambles and of course the number of available preambles as our input for CE level \( c \).

Assuming that there are \( m \) (unknown) number of devices which transmitted preambles and \( r_c \) detected preambles at a PRACH in CE level \( c \), the expected number of unused preambles, \( E[R_c-r_c] \), in this PRACH can be calculated by
multiplying the number of ways of choosing 1 out of $R_c$ preambles, which is exactly $R_c$, multiplied by its probability for not being chosen by all $m_c$ devices, which is $(1 - \frac{1}{R_c})^{m_c}$. Hence,

$$E[R_c - r_c] = R_c \left(1 - \frac{1}{R_c}\right)^{m_c} = R_c \left(\frac{R_c - 1}{R_c}\right)^{m_c}. \quad (3)$$

When it is known that there are $r_c$ preambles being used, i.e. the other $R_c - r_c$ preambles are unused, for $r_c < R_c$, $\tilde{m}_c$ can be estimated by replacing $E[R_c - r_c]$ with $R_c - r_c$ in (3) and converting it into logarithmic form. Hence,

$$\tilde{m}_c = \begin{cases} \log_{R_c - r_c} \frac{R_c - r_c}{R_c}, & \text{for } r_c < R_c; \\ \text{otherwise.} & \end{cases} \quad (4)$$

The second clause is provided since the first clause would grow exponentially to infinity when $r_c \to R_c$. Hence, $\tilde{m}_c$ is upper bounded by $m_{max}$, which is discussed in the later subsection.

Once the estimate $\tilde{m}_c$ of the total number of devices competing over the PRACH is computed, the optimized average number of devices allowed to transmit Msg3 is provided by our Msg3 barring mechanism.

Table I showcases numerical values for the estimate $\tilde{m}_c$, for the case $R_c = 12$ (the minimum number of sub-carriers in a single NB-IoT CE level) and $r_c < R_c$. As expected, when $r_c$ is small, the number of estimated devices is close to $r_c$. Conversely, when the $r_c$ grows closer to $R_c$, the estimate rapidly increases.

In real use, this equation can only be used for $r_c \in \{0, 1, 2, ..., R_c - 1\}$. Meanwhile for $r_c = R_c$, base station can only assume that $m$ equals $m_{max}$. However, this limitation is negligible since the number of preambles that can be acknowledged in RAR and the number of available PUSCH resources may not exceed $R_c$ (otherwise RAD capacity and PUSCH resources are considered overprovisioned and will be unused).

<table>
<thead>
<tr>
<th>$r_c$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[\tilde{m}_c]$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>13</td>
<td>16</td>
<td>21</td>
</tr>
</tbody>
</table>

| $m_{max}$ | 0 | 1 | 2 | 3 | 4 | 5 | 7 | 8 | 11 | 13 | 16 | 21 |

B. Dynamic Connection Request Barring

Our proposed mechanism reduces the collision in Msg3 transmission step by limiting the number of devices which transmit Msg3. In the RA procedure, Msg3 contains device’s identity and other information for connection request. Originally, each device whose transmitted preamble is indicated in RAR (i.e. the transmitted preamble is detected by base station) transmits its Msg3 at the time-frequency resources that are specified for its preamble.

In the 4-step RA procedure, the actual collisions are mostly detected in Msg3 step instead of in preamble transmission step. If more than one device transmits the same preamble, and at least one of the transmitted preambles is detected by the base station, base station may indicate this detected preamble in the subsequent RAR. Those devices, which have transmitted the same preamble, then transmit their Msg3 at the same resource, causing them to collide.

Up until this point, we know that when devices are collided in their preamble transmission, most of the time, their Msg3 will collide too. However, not all of the detected preambles result in Msg3 collision. Hence, instead of limiting the devices from transmitting their preamble altogether (presumably to decrease the collision, which is exactly the approach taken by ACB and Extended Access Barring (EAB) [21]), our proposed method do not limit the preamble transmission and exploit it for guessing the instantaneous load, i.e. number of devices currently requesting connection to send their data.

The estimated number of devices which has transmitted their preamble is then used to cleverly decrease the contention in Msg3 step. This is done by allowing only some of those devices to actually transmit their Msg3. We introduced a single parameter called “Msg3 barring factor”, $Z_c$, at CE level $c$, whose valid value is between $[0, 1]$. In this case, $Z=1$ allows every device in CE level (which is indicated in RAR) to transmit its Msg3, and $Z_c=0$ allows no device at CE level $c$ to transmit their Msg3. Since base station can always estimate the load in the current RA slot, $Z_c$ can be updated to match the estimated load in every PRACH and broadcasted via RAR message. Notice that from implementation point of view, this provides us an advantage since updating $Z_c$ via RAR is closer to real-time than updating other parameters via System Information Blocks (SIBs) or even Master Information Block (MIB). Finally, upon receiving RAR message, each device has to pick a random number from $[0, 1]$. A device in CE level $c$ can transmit its Msg3 if the chosen number is less or equal to $Z_c$. Otherwise, the device acts as if its preamble is not acknowledged in RAR, i.e., retransmit a new preamble.

To keep the efficiency at its peak, we need to maintain the ratio between the number of available resources and the number of devices which actually transmit Msg3 to be no less than 1. Hence, to approach this ideal condition, a simple rule is configured in the base station to set the Msg3 barring factor as follow

$$Z_c = \begin{cases} 1, & \text{for } \tilde{m}_c \leq R_c; \\ \frac{r_c}{R_c / \tilde{m}_c}, & \text{otherwise.}\end{cases} \quad (5)$$

We admit that our estimation is rough and we are unable to estimate the load when all preambles are occupied. Hence, we propose the following formula for calculating $m_{max}$ as

$$m_{max} = kR_c + \log_{\frac{R_c - 1}{R_c}} \frac{R_c - 1}{R_c} \quad (6)$$

In this equation, the predefined load is denoted as $kR_c$. $k$ is added for scaling the $R_c$ (or $R_c$ since this method can be applied either locally in a specific CE level or globally for all CE levels.), since the load condition is always relative to $R_c$ (for example, it is common to say ‘high load’ for the load which is larger than $R_c$ and ‘low load’ for the load which is smaller or equal to $R_c$).

As the result, our estimation cannot achieve the ideal ratio between the number of available resources and the number of devices which actually transmit Msg3. I.e., when $m$ is between $\frac{R_c - 1}{R_c}$ and $kR_c + \log_{\frac{R_c - 1}{R_c}} \frac{R_c - 1}{R_c}$, the number of devices which actually transmitting Msg3 is less than $R_c$. On the contrary, when $m$ is larger than $kR_c + \log_{\frac{R_c - 1}{R_c}} \frac{R_c - 1}{R_c}$, the number of devices which actually transmit Msg3 is greater than $R_c$. The effect of this less ideal setting is explored in the next section.

Since the purpose of the proposed mechanisms of load approximation and dynamic Msg3 barring are to increase...
number of successful devices in each PRACH, they are implemented in each PRACH as illustrated in Pseudo-code 1.

Pseudo-code 1. Implementation of the proposed mechanism in each PRACH

```c
base station receives preambles;
if number of detected preambles == number of preambles {  
   assume number of transmitting devices to be maximum;
} else {  
   approximate number of transmitting device with (4);
}
base station adjusts Msg3 barring factor with (5);
base station informs Msg3 barring factor via RAR message;
transmitting devices select a random number;
if the number c = Msg3 barring factor {  
   the device transmits Msg3;
} else {  
   the device conducts backoff and retransmit;
}
if Msg3 collide {  
   the device conducts backoff and retransmit;
} else {  
   the device transmits Msg4 (connection setup) and succeed;
}
```

V. RANDOM ACCESS MODELING

An analytical model is used to estimate the performance metrics of the conventional/standard and our improved RA procedure. The measured performance metrics comprises success probability and average access delay, whose calculation can be specified for each CE level. The construction of this model follows iterative contending user estimation framework [8], in which the RA in a cell is modeled as multi-band multi-channel slotted Aloha system [10], with different CE level, preamble and PRACH altogether representing different band, channel, and slot, respectively. The produced results, including those shown in Section VI, are verified by computer simulation to have less than 2% error. The source of estimation error in this model is discussed later in this section.

Following the model that has been developed in Section III, it is expected that there are \( m_{c,i} \) devices from CE level \( c \) to conduct their first preamble transmission attempt in slot \( i \) of CE level \( c' \). In addition to these first attempts, retransmissions (i.e. the second attempt, third, and so on) may also occur in this slot. These retransmitting devices can be originally from CE level \( c \) or lower. To represent all of these transmission attempts, let \( m_{c,i}^{r}(n) \) denotes the number of devices residing in CE level \( c \) which transmit its \( n \)-th attempt at slot \( i \) in CE level \( c' \).

Let \( N \) denote the maximum number of attempts that a device can perform. When multiple CE levels are defined, \( N \) takes form of a matrix where \( N_{c,c'} \) be the element in \( c \)-th row and \( c' \)-th column of \( N \), denoting the maximum number of attempts that a device from CE level \( c \) can perform in CE level \( c' \). \( N_{c,c'} \) for \( 0 \leq c' \leq c \leq C-1 \) can be obtained from \( P_{c} \) and \( P_{c'} \) as

\[
N_{c,c'} = \{ \min(P_{c}, P - \sum_{c''=c+1}^{C-1} P_{c''}), \text{for } c \leq c' < C - 1; \max(0, P - \sum_{c''=c+1}^{C-1} P_{c''}), \text{for } c \leq c' = C - 1. \}
\]  

With this, we can introduce \( m_{c,i}^{s} \) as the summation of \( m_{c,i}^{r}(n) \) over \( 0 \leq c' \leq c \) and \( 1 \leq n \leq N_{c,c'} \), denoting the total number of devices which transmit at slot \( i \) of CE level \( c' \). I.e.,

\[
m_{c,i}^{s} = \sum_{c'=0}^{c} \sum_{n=1}^{N_{c,c'}} m_{c,i}^{r}(n). \tag{8}
\]

These devices contend for \( R_{c} \) preambles. Preamble collision yields Msg3 collision since devices transmitting the same preamble will transmit Msg3 using the same resource. In the presence of our dynamic connection request barring mechanism, some devices which transmit preamble may be barred and will not transmit Msg3 when the Msg3 barring factor of CE level \( c' \), \( Z_{c'} \), is greater than 0. In this case, the legacy system without our dynamic connection request barring mechanism can easily be represented with \( Z_{c'} = 1 \).

Successful transmission of Msg3 yields successful RA since the error probability of Msg4 is negligible [8]. For slot \( i \) in CE level \( c \), let \( s_{c,i}^{r}(n) \) be the expected number of devices which transmit its \( n \)-th attempt with a chosen preamble that is not chosen by the other \( m_{c'}^{r} - 1 \) devices. In short, these \( s_{c,i}^{r}(n) \) devices are successful devices. \( s_{c,i}^{r}(n) \) can be calculated as

\[
s_{c,i}^{r}(n) = Z_{c'} \cdot m_{c,i}^{s}(n) \frac{1}{R_{c}} (1 - \frac{1}{R_{c}})^{Z_{c'} m_{c,i}^{s} - 1}. \tag{9}
\]

For convenience in the later iterative calculation, by exploiting the well-defined Bins and balls problem [27], \( s_{c,i}^{r}(n) \) can be approached by

\[
s_{c,i}^{r}(n) \approx Z_{c'} \cdot m_{c,i}^{s}(n) e^{-\frac{Z_{c'} m_{c,i}^{s}}{R_{c}}}. \tag{10}
\]

The accuracy of this approximation grows as \( Z_{c'} m_{c,i}^{s} \) grows [27], which is most likely the case for C-IoT. Hence, it is practical for the following iterative calculation.

For realistic modeling of the RA, power ramping during preamble transmission and limitation of PUSCH should also be considered, demanding \( s_{c,i}^{r}(n) \) to be refined. Power ramping increases the detection probability of transmitted preamble. Let us adopt \( 1 - e^{-a} \) as the detection probability of the \( n \)-th preamble transmission attempt [28]. Meanwhile, PUSCH limitation acts as the upper bound of the number of acknowledged preambles in a RAR messages. Let \( U \) denote the limitation of PUSCH in each PRACH (i.e., each slot). When number of detected preambles exceeds \( U \), base station randomly selects \( U \) detected preambles to be acknowledged in RAR. Hence, \( s_{c,i}^{r}(n) \) is refined as

\[
s_{c,i}^{r}(n) \begin{cases}  
(1 - e^{-n})Z_{c'} m_{c,i}^{s}(n) e^{-\frac{Z_{c'} m_{c,i}^{s}}{R_{c}}} & \text{if } \sum_{c'=0}^{c} \sum_{n=1}^{N_{c,c'}} (1 - e^{-n})Z_{c'} m_{c,i}^{r}(n) e^{-\frac{Z_{c'} m_{c,i}^{r}}{R_{c}}} \leq U; \\
U(1 - e^{-n})Z_{c'} m_{c,i}^{s}(n) e^{-\frac{Z_{c'} m_{c,i}^{s}}{R_{c}}} & \text{otherwise.} 
\end{cases} \tag{11}
\]

As the compliment to \( s_{c,i}^{r}(n) \), let \( c_{c,i}^{r}(n) \) be the number of devices from CE level \( c \) whose \( n \)-th attempt in slot \( i \) of CE level \( c' \) is not successful. It can be easily obtained as

\[
c_{c,i}^{r}(n) = m_{c,i}^{s}(n) - s_{c,i}^{r}(n). \tag{12}
\]

Subsequently, the devices which have not successful and still have chance(s) to retry will retransmit. Such retransmission may occur at the same CE level or at the higher CE level. In addition, the devices which have not successful yet has no remaining chance are failed.

2327-4662 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information. Authorized licensed use limited to: Universitas Indonesia. Downloaded on March 23, 2020 at 05:00:33 UTC from IEEE Xplore. Restrictions apply.
Retransmission at the same CE level is conducted after backoff. In this case, their retransmission time is spread into different slots. A device may end up transmitting the n-th attempt at slot i when it was collided in its (n-1)-th attempt at slot j<i, still has chance to retransmit at the same CE level, and its backoff counter ends at the interval of [l_i+1,l_i]. Notice that j depends on i and backoff window, such that total number of retransmission which occurs at slot i of CE level c whose previous attempt is also conducted at CE level c, \( m^c_{E,i}(2 \leq n \leq N_{c,E}) \), is

\[
m^c_{E,i}(n) = \sum_{j=\text{max}(0,i-\frac{d_{col,i}}{\tau_c})}^{\text{min}(n,l_i)} w_{j,i} m^c_{E,j}(n-1) \quad (13)
\]

with \( w_{j,i} \) denotes the likelihood that a device retransmits at slot i after collided at slot j, \( w_{j,i} \) for CE level c is calculated as

\[
w_{j,i} = \begin{cases} \frac{(j-1)\tau_c + d_{E,i} - (i-2)\tau_c}{\tau_c} & \text{for } 1 \leq j \leq i \text{ and } \frac{d_{E,i}}{\tau_c} \leq j \leq i - \frac{d_{E,i}}{\tau_c}; \\ \frac{2\tau_c + d_{E,i} - (i-2)\tau_c}{\tau_c} & \text{for } i - \frac{d_{E,i}}{\tau_c} \leq j \leq \text{max}(0,i-\frac{d_{E,i}}{\tau_c}) \text{; otherwise.} 
\end{cases} \quad (14)
\]

Retransmission at the next CE level is conducted without backoff. This type of retransmission is considered to be the first transmission in the new CE level. For example, with \( N_{c,E}=3 \) and \( N_{c,E+1}=2 \), when a device residing in CE level c collides at its 3rd attempt in CE level c, it retransmits at CE level c+1, and of course this retransmission is considered to be its first attempt in CE level c+1. Note from this example that if \( N_{c,E+1}=0 \) this device cannot retransmit at CE level c+1. Thus, \( m^c_{E,c}(1) \) for \( c < c' \leq C-1 \) can be calculated as

\[
m^c_{E,c}(1) = \frac{c^{c'} - 1}{c^{c'-1}} \left( (l_i-1)\tau_c + d_{E,i} - d_{E,i-1} - d_{E,i-2} - \tau_c \right) (N_{c,c'-1})
\]

\[
\sum_{j=\text{max}(0,i-\frac{d_{col,i}}{\tau_c})}^{\text{min}(n,l_i)} w_{j,i} m^c_{E,j}(n-1) \quad (15)
\]

up to now, we can calculate \( m^c_{E,1} \) in (8) from (1), (13) and (15). Hence, (11) and (12) can be calculated iteratively for all slots in all CE levels to obtain the performance metrics.

Success probability of the system, \( P_S \), measures ratio between the number of successful devices to the number of arrivals. For this, we need recursive summation of \( s^c_{E,1}(n) \) from (11) for the working ranges of \( c, c', i, \) and \( n \). Meanwhile, to measure the success probability of devices residing in CE level c, \( P_{S,c} \), the summation over \( 0 \leq c' \leq C-1 \) can be removed. Thus, for CE level c,

\[
P_S = \frac{\sum_{c=0}^{C-1} \sum_{c'=0}^{C-1} \sum_{i=1}^{\text{max}(0,i-\frac{d_{col,i}}{\tau_c})} \sum_{n=1}^{N_{c,c'-1}} s^c_{E,i}(n)}{\sum_{c=0}^{C-1} \sum_{c'=0}^{C-1} \sum_{i=1}^{\text{max}(0,i-\frac{d_{col,i}}{\tau_c})} m^c_{E,i}(1)}. \quad (16)
\]

VI. EVALUATION

Computer simulations were conducted to verify the accuracy of the proposed analytical model and the proposed load estimation. The verified analytical model is then used for assessing the performance of our proposed method applied in NB-IoT system. The performance of the original random access in NB-IoT is included as a baseline for comparison. For this evaluation, all IoT devices are assumed to transmit its data at the same time and thus initiates RA procedure by transmitting preamble simultaneously. This demonstrates the extreme cases of RAN overload which may be experienced by safety IoT applications during emergency. The RSRP of the devices to the base station are distributed uniformly.

The accuracy of the proposed model is examined for number of devices, \( m, \) of \( \{1, 2, \ldots, 120\} \) with number of preambles, \( R, \) of 12, which represents very low to very high load condition. The estimated number of devices which transmit preamble is compared with the actual one. 10^6 samples from the simulation are taken for each value of \( M \). Fig. 2 shows the result of our estimation with \( k=1 \) and 2, compared with an ideal/perfect estimation. Subsequently, Fig. 3 shows the relative error of the estimations in Fig. 2. From both figures, it is observed that the
estimation cannot always predict the load precisely. Especially, the estimation error is increasing along with \( m \) for \( m > m_{\text{max}} \). With higher \( m \), most likely all preambles are occupied and detected. In such situation, our estimator always gives \( \hat{m} = m_{\text{max}} \). However, the next result shows that despite of its high error, this estimation is still useful for our dynamic connection request barring scheme.

We compare our proposed dynamic connection request barring scheme with the original system. For this comparison, we consider an NB-IoT system serving an IoT application whose transmission delay cannot exceed 3 s. This is to simulate MAR applications which require its ACK packet to be received no later than 10s after data packet’s transmission [3]. We consider one-shot arrival with 12 preambles in each CE level. The observation is conducted with number of devices in each CE level of 1, 2, ..., 120.

![Figure 2: Estimated load for \( R=12 \)](image)

![Figure 3: Estimation error for \( R=12 \)](image)

| TABLE II. PARAMETER’S VALUE IN SETTINGS 1, 2 AND 3 |
|-----------------|-----------------|-----------------|-----------------|
| Parameters      | Settings        |                |                |
| \( C \)         | 1               | 2              | 3              |
| \( P \) and \( P_c \) | 5 and 5        | 4 and 4        | 4 and 3        |
| Repetitions     | 2              | 2, 8, 32       | 2, 8, 32       |
| \( B_c \) (ms)  | 512            | 512, 512, 256  | 512, 512, 256  |

For simpler observation in this study, it is assumed that the number of configured PRACH repetition is enough for perfect detection of the preamble, and number of uplink resources for Msg3 matches the number of preambles. Three settings are considered as shown in Table II. Setting 1 represents an NB-IoT cell with 1 CE level, while Settings 2 and 3 each represent an NB-IoT cell with 3 CE levels. Cross-CE-level retransmission is enabled in Setting 3. They are provided to reveal the effectiveness of our proposed improvement scheme under various settings. In this table, the values are chosen such that any successful devices in the original system (i.e. without our proposed improvement) will not exceed application’s maximum tolerable RA delay of 3s.

Investigation is conducted with number of IoT devices in the cell, \( M \), varying from 12 to 120. Each device poses one backlogged data with the same size and the same transmission time as the others, representing a synchronized MAR scenario. Figs. 4 and 5 show the success probability and average access delay for Setting 1, respectively. In these figures, we compared our improvement with \( k=1 \) and 2, our improvement with perfect load estimation, and the original system without our improvement. Fig. 4 shows that our proposed improvement increases the success probability under higher load. Fig. 5 shows that our proposed improvement obtains not only a higher success probability, but it also a lower average access delay. It is also observed from both figures that our improvements with \( k=1 \) and 2, recall equation (6), obtain almost similar performance to the one obtained by our improvement with perfect load estimation. This signifies that our estimator, despite of its error which is shown in Fig. 3, can still be used to yield a very close performance obtained with a perfect estimator. This is because the estimation error emerges for higher load (see Fig. 3). Meanwhile, under higher load, the base station sets \( Z \) to \( \bar{R}_c / \bar{m}_c \), where \( \bar{m}_c = m_c \) for perfect estimation. Note that the absolute difference between \( \bar{R}_c / \bar{m}_c \) and \( R_c / m_c \) is smaller than the absolute difference between \( \bar{m}_c \) and \( m_c \). Thus, the large discrepancy between \( \bar{m}_c \) and \( m_c \) is translated as small error in \( P_s \).
Let us observe the number of preamble transmissions and successful preamble transmissions over time, starting from the arrival time, which is displayed in Fig. 6. This figure shows the situation for Setting 1 with \( M=120 \). Our proposed improvement scheme with \( k=1 \) is compared with the original system. Additionally, number of devices which eventually transmit Msg3 is also displayed. From this figure, it can be observed that our barring mechanism successfully adjust Msg3 contention load to be around 12 (i.e. ideal for \( R=12 \)), except for the first arrival only. The maximal load (i.e. number of devices which transmit preamble) other than the first arrival is always fallen within low error area (see Figs. 2 and 3). Hence, the high error displayed in Figs. 2 and 3, which are found in relatively higher load, does not hinder our improvement scheme to deliver superior performance compared to the original system.

Figures 7 and 8 show the success probability and average access delay for Setting 2, respectively. In this setting, three CE levels are considered, representing a usage scenario with extended cell coverage. Number of PRACH repetition for each of the three CE levels follows [29]. Fig. 7 shows that our improvement increases the success probability in each CE level when there are more devices. This yields higher total success probability, which is simply calculated by dividing number of successful devices from all CE levels with the number of devices in all CE levels.

Figure 8 shows that for CE levels 0 and 1, the average access delays of our proposed improvement are lower than those without one. This is because more devices can be successful with less retransmission since our barring mechanism limits the number of transmitted Msg3, which then decrease the collision. However, our proposed improvement has slightly higher \( D \) in CE level 2 for \( M > 120 \). This is because the original system yields very few (almost none) successful devices. Additionally, it is also observed from this figure that the overall \( D \) is almost similar for both improved and original systems. This is because the number of barred devices in the improved system is lower than the number of collided devices in the original system. Eventually, on average, devices need fewer attempts to be successful. Fig. 8 also shows the average access delay for all CE levels, which is calculated by simply totaling the time taken by all successful devices in all CE levels and divide it by the number of successful devices in all CE levels.

Notice that higher \( M \) does not always yield higher \( D \). Lower \( D \) happens as \( M \) increases when \( P_s \) is very low, i.e. \( \leq 0.2 \) (e.g. \( M > 72 \) in CE level 2 and \( M > 216 \) in CE level 1 of the original system). Under higher \( M \), heavy collision at the arrival slot is prolonged until the later slots. Thus, fewer devices succeed in the later slots. Compared to lower \( M \), higher \( M \) has faster decrease in number of successful devices over time (observing successively from the first until the last slots). Since \( D \) is simply the total access duration spend by all successful devices divided by number of successful devices, this decrease is sharper when the decerase of \( P_s \) between the lower and higher \( M \) is very small. Hence, when number of successful devices from the later slots is multiplied by its slot time, following (20), higher \( M \) yields lower value for the numerator, which finally returns lower \( D \). To be more intuitive, let us consider 1 CE level with 12 preambles \((R=12)\), up to 3 attempts, no backoff, and simultaneous arrival at slot 1. The result for \( M=24 \) and 36 are sampled in Table III. Following (10), number of successful devices in each slot can be calculated based on the number of transmitting devices (e.g. transmitting devices at slot 2 are those who failed at slot 1). Finally we can calculate total number of successful devices from the three slots and their average access delay, \( D \). In this example, \( D \) of \( M=36 \) is lower than that of \( M=24 \).

<table>
<thead>
<tr>
<th>Table III: Example of Lower ( D ) at Higher ( M )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of devices</td>
</tr>
<tr>
<td>Average number of successful devices in slot 1 (1st attempt)</td>
</tr>
<tr>
<td>Average number of successful devices in slot 2 (2nd attempt)</td>
</tr>
<tr>
<td>Average number of successful devices in slot 3 (3rd attempt)</td>
</tr>
<tr>
<td>Total number of successful devices</td>
</tr>
</tbody>
</table>
Figures 9 and 10 show the success probability and average access delay for Setting 3, respectively. This setting considers three CE levels with cross CE-level retransmission, representing a usage scenario with more extended coverage. Similar to Setting 2, number of PRACH repetition for each of the three CE levels follows [29]. Fig. 9 shows that our improvement increases the success probability in each CE level when there are more devices, which eventually yields higher total success probability.

Figure 10 shows that our proposed improvement always yield lower $D$ except in CE level 2 when the original system yields very low $P_S$, which is also happening in Setting 2. Hence, overall average access delay of our improved system is close to that of original system and is slightly higher when the original system yields very low $P_S$, especially when $M > 150$.

Finally, from Figs. 7-10, it is observed that the effectiveness of our proposed improvement scheme is not affected by the presence of multiple CE level with and without cross-CE-level retransmission.

Now let us compare the proposed method with the original RA and RA with preamble barring mechanism [21], in which 50% of the preamble transmission attempts at each RA opportunity are prevented to avoid contention overload. For this comparison, Setting 3 is assumed. Figs. 11 and 12 exhibit the overall (includes all CE levels) success probability and overall average access delay, respectively. In Fig. 11, the proposed Msg3 barring method demonstrates its superiority compared to the other two by achieving the highest success probability under all examined $M$. Reflecting this result to the one in Fig. 12, it is shown that under higher load of about $M \geq 144$, both barring methods (i.e. Msg3 barring and preamble barring) should pay a longer delay than the system with original RA. This is inevitable since the barring decreases contention by preventing the attempts and push it to the later RA opportunity, causing the increased delay. Although in this case the proposed mechanism yields the highest delay, this tradeoff can be a good choice for delay-tolerant IoT application since it provides the highest success probability.

VII. CONCLUSION

Computer simulations were conducted to verify the accuracy of the proposed analytical model and the proposed load estimation. The verified analytical model is then used for assessing the performance of our proposed method applied in NB-IoT system. The performance of the original random access in NB-IoT is included as a baseline for comparison. Fig. 4 shows that our proposed improvement increases the success probability under higher load. Fig. 5 shows that our proposed improvement does not only obtain higher success probability, but it also obtains lower average access delay. It is also observed from both figures that our improvements with $k=1$ and $2$ obtain almost similar performance to the one obtained by our improvement with perfect load estimation. This signifies that our estimator, despite its error which is shown in Fig. 3, can still be used to yield a very close performance obtained with a perfect estimator.
ACKNOWLEDGMENT

This research and article’s publication is supported by the World Class Professor (WCP) program from Ministry of Research, Technology and Higher Education (RISTEKDIKTI) of Republic of Indonesia under contract number 123.52/D2.3/KP/2018.

REFERENCE