Impact of Backoff Algorithm on IoT over Multichannel Slotted Aloha System

Ruki Harwahyu*1, Chu-Chun Huang2, Ray-Guang Cheng2, Riri Fitri Sari1
1Universitas Indonesia, Depok, West Jawa, 16424 Indonesia
2National Taiwan University of Science and Technology

*ruki.h@ui.ac.id

Abstract. The increasing trend of Internet of Things brings new challenges to efficiently manage the resource in the access network. As a prominent solution to serve IoT services, LTE incorporates a small data transmission scheme for this reason. In this scheme, the overall system performance is greatly influenced by its random access procedure. Hence, studying the behavior of backoff during random access procedure is important. This work implements an iterative contending-user estimation model to analyze the performance of the contention-based random access procedure supporting finite-population. For the sake of generality, the system is modeled as a multichannel slotted Aloha. This allows our model to be used in wider specifications, including LTE and WiMAX systems. The behavior and performance of the system supporting multiple load scenarios is studied under different network loads. The simulation result demonstrates the accuracy of our proposed method to predict the normalized throughput, the packet-dropping probability, and the average access delay of each access-class. The performance of different backoff algorithm to resolve the collision in the system is compared and evaluated.

Keywords: Internet of Things (IoT); LTE; WiMAX; Multi-channel slotted ALOHA; Random access; backoff algorithm

1. Introduction

Cellular networks have been used to provide city-wide coverage for broadband services. In recent decade, the demand for data communication is nowadays increasingly high. Since the adoption of connectionless communication for data services, the method to handle non-voice communication over cellular network has been improving. Expanding from its main purpose to serve human-to-human communication, nowadays the widely-deployed cellular network is also a potential solution to serve machine-to-machine (M2M) or Internet of Things (IoT) services. The fact that cellular services are ready in most of the cities and villages has become main motivation to utilize it for serving IoT devices. In fact, in many places, this readiness is far surpassing the readiness of other wireless technologies such as ZigBee, and LoRa. Although these competing technologies are designed from the beginning to mainly serve data communication and IoT, deploying them widely (i.e. to really rivals cellular network) needs many efforts and investments. On top of that, cellular has strong history for being pretty stable business. Meanwhile ZigBee and LoRa are new technologies which possess risk to start the business with, must overcome public scepticisms, needs more advertisements to endorse adoption, etc.
Following the evolution of cellular technologies, it is clear that IoT services will be one of its important markets. A significant amount of research and development focuses have been devoted to make cellular technologies more capable in handling IoT scenarios.

In the recent improvement/evolution of cellular networks to serve IoT, LTE-Cat0 and NB-IoT were introduced. In these ‘derivatives’ of LTE system, most of LTE procedures are still used, while some others are simplified to accommodate IoT requirements such as low power consumption, long range transmission (i.e. better link budget), and simple/light algorithm processing (permitting cheaper hardware module). Both LTE-Cat0 and NB-IoT can be deployed and coexist well with existing cellular networks.

Even before LTE-Cat0 and NB-IoT were introduced, several schemes aiming to improve the service for M2M devices were proposed. Starting from LTE-Advanced Pro, a connectionless uplink transmission, small data transmission, and several other schemes have been proposed and discussed. These schemes are proposed to increase the efficiency in serving small data packets which were presumed (at that time) to be the case for machine-to-machine communication. Notice that in cellular system, there are numerous procedures needed to be completed before an uplink data transmission can be conducted. These procedures comprise synchronization, connection establishment, security and subscription/billing checking. For normal human-to-human communications, this would be normal either from timing and packet sizing perspective. However, for the envisioned (which has become reality today) machine-to-machine communications, these procedures take significant amount of time and frequency resources. These procedures require multiple message handshakes which all together are too much overhead in time and data size for the small data transmitted by the M2M devices. Additionally, it also consumes more energy, which is a precious resource for M2M/IoT devices. The aforementioned schemes enable IoT devices to transmit its small data (smaller than a certain threshold) without the hassle of some preliminary procedures, decreasing the size of the overhead, and in turn increasing overall system efficiency.

LTE and its ‘derivatives’ use random access (RA) procedure to decide resource allocation for data transmission. RA procedure is one of the preliminary procedures mentioned earlier. RA procedure consists of 4 messages that are being exchanged by end device and base station. There are contention-free and contention-based RA procedures. Contention-free RA procedure has its message exchange coordinated by the base station in more detail, making collision between devices impossible to happen. Meanwhile, contention-based RA procedure allows collision between the messages transmitted by end devices. The contention-based RA procedure is the one that has major impact on overall system throughput, especially in IoT services with large number of end devices within one cell of base station.

Since collision is normal to occur in this procedure, collision resolution is incorporated. To resolve collision, the contention-based RA procedure in LTE uses limited number of retransmission with uniform random backoff before each retransmission. The contention-based RA procedure needs to be done by device with backlogged uplink data which no longer have an active data bearer to be used. Typically, this happens when the device is just turned on and attached to the network or when it has been idling for a duration more than the data bearer timer.

Among the important characteristics of IoT services are its small data size and intermittent data transmissions. The intermittent data transmission is mostly caused by the periodic nature of its application to generate and send uplink data. This typically happens in sensor devices with no actuators. In such case, the IoT devices may sleep (the whole device or just its radio module) for considerably long period of time to conserve its energy. In this case, the devices need to conduct contention-based RA procedure to obtain resources to send its data.

With a stripped-down signaling to favor the small-sized data, the overall system performance greatly depends on the RA procedure [1]. Hence, studying and estimating the behavior of RA procedure is important. Additionally, as the system is subject to heavy contention during overload condition, backoff algorithms play important role in resolving the contention. This work studied the impact of uniform random backoff (UB) and binary exponential backoff (BEB) algorithm. We used
our model to estimate the normalized throughput, average access delay, packet dropping probability and variance of access delay of the system.

2. Random Access Procedure
In LTE networks, all end devices/nodes receive synchronization signals and broadcasting messages from base station upon their attachment. A node performs RA through physical RA channel (PRACH) [2] to acquire uplink resource for data transmission. PRACH is a set of consecutive sub-carriers in a time interval which are solely provided for preamble transmission. Preamble is the first out of the four messages exchanged during RA procedure which takes form as a physical signal stream/symbol. Up to 64 orthogonal symbols are provided in each PRACH and can be randomly chosen by transmitting device.

A device starts an RA procedure by transmitting its randomly chosen preamble in the immediate PRACH following its transmission time. There is a possibility that there are more than one devices transmitting preamble at the same PRACH. Subsequently, base station detects the preambles that are received with enough power at its side. Notice that a transmitted preamble can be undetected due to various fading or interferences. Additionally, a preamble which is detected by base station can also be the resultant of more than one transmission, i.e. more than one devices transmitting the same preamble (causing constructive interference since the timing differences are tolerated by preamble’s symbol design).

The second message of RA procedure is RA response (RAR) or Msg2. RAR is broadcasted by base station to indicate the detected preambles. However, there is a limitation of the number of preambles which can be acknowledged in a RAR window. When the number of detected preambles is more than the number of preambles which can be acknowledged, base station randomly choose maximum number of preambles to be acknowledged in the RAR. Each acknowledged preambles are paired with a dedicated resource to be used for transmitting the third and fourth messages of RA procedure.

The third message of RA procedure is called Msg3. In this message, the device whose transmitted preamble is acknowledged in RAR transmits its identity for connection request in the corresponding air interface resource. Hence, two devices which transmitted different preambles and both of them are acknowledged in RAR will not collide. However, if the two devices transmitted the same preamble and they are acknowledged in RAR, both of them eventually transmit their Msg3 at the same resource, causing a collision. I.e., collision occurs if two or more devices choose the same preamble. Collided devices shall perform random backoff and retransmit a new preamble in the next available PRACH. The process repeats until the maximum number of attempts is reached. The device will declare a RA failure if RA procedures still fail in its last attempt.

Concurrent channel access request from massive number of devices in IoT scenarios may severely congest the RACH. The congestion becomes even more costly since IoT should conserve its power instead of wasting them in collisions. In this paper, we demonstrate an analytical model to estimate the performance of RA procedure. For the ease of this study, the RA procedure is generalized and is modeled as a multichannel slotted ALOHA system. Subsequently, the study and evaluation is focused to the backoff mechanism, which is the main and the only mechanism used to resolve the collision.

Several works have been devoted to the single-channel [3-4] or multi-channel [5-12] slotted ALOHA systems. In [3], the throughput of single-channel slotted ALOHA systems as a function of a constant offered load was presented. In [4], the relationship between the throughput and the average access delay of a finite-user single-channel slotted ALOHA system was investigated. For multi-channel slotted ALOHA systems, the performance metrics of throughput [5], [6], [9], average access delay [7-11], and collision and success probabilities [12] have been discussed. A Poisson approximation model to estimate the number of success and collided UEs in the first RA slot was presented in [13]. Several works on modeling of RA channel in LTE with resource sharing schemes was summarized in [14] while also presenting a novel the concept of generalized resource sharing
scheme.

3. System Model

This paper considers $M$ IoT devices in an LTE access network or its derivatives. The process of packet arrival, which triggers the devices to conduct RA procedure, happens in an arrival interval of $[t_F, t_L]$ which follows a general distribution with probability density function of $A(t)$. The uplink data packets of all devices are assumed to have uniform size, and one device has only one data packet to be transmitted. For the RA procedure, the base station reserves $R$ preambles in each RACH. The parameters used in this paper are summarized in Table 1. No background traffics are assumed in this system. Hence, contention is limited only among the IoT devices.

| Notation | Meaning |
|----------|---------|
| $R$ | Total number of reserved preamble |
| $M$ | Total number of mobile user in the network |
| $A(t)$ | PDF of the arrival process of mobile users |
| $t_F$ | Time of the first arrival |
| $t_L$ | Time of the last arrival |

**QoS Metrics**

| Notation | Meaning |
|----------|---------|
| $P_s$ | Success transmission probability |
| $\bar{D}$ | Average transmission delay |

**RACH-specific Parameters**

| Notation | Meaning |
|----------|---------|
| $T_{RAR}$ | Mobile user’s max. processing time to decode transmitted preambles |
| $T_{RA\_REP}$ | Interval between 2 successive RA slots |
| $W_{RAR}$ | Duration of RAR window |
| $N_{UL}$ | Max. total number of acknowledged mobile user in one RAR window |
| $W_{BO}$ | Duration of backoff window |
| $N_{PT\_max}$ | Maximum number of preamble transmission |
| $N_{HARQ}$ | Maximum number of HARQ transmission |

To minimize the signalling before data transmissions, RACH-based small data transmission scheme is assumed to be applied in this system. Figure 1 shows the message sequence chart of the RACH-based small data transmission scheme. It mainly consists of two parts: RA procedure part (Steps 1-4) and data transmission part (Steps 5-6). For RAR transmission, base station reserve a duration of $W_{RAR}$ ms as the RAR window. The maximum number of mobile users that can be acknowledged in one RAR window is $N_{UL}$. Non-adaptive hybrid automatic repeat request (HARQ) is used to protect the transmission of MSG-3 and MSG-4 and the maximum number of HARQ transmissions is $N_{HARQ}$.

The data transmission part starts immediately after receiving the MSG-4. The mobile user transmits the first uplink (UL) small data packet carrying a System Architecture Evolution (SAE) Temporary Mobile Subscriber Identity (S-TMSI) as its identifier and a gateway identifier as the destination of the packets. The UL small data packets are then transmitted on a pre-configured default data radio bearer (DRB) and the transmission continues until the expiry of a DRB timer. A separate connectionless security context is used to protect the connectionless transmission.
Figure 1: RACH-based small-data-transmission scheme [16].

4. Analytical Model
In this section, we apply the iterative contending-user estimation model presented in [17] to estimate the number of success mobile users in each RA slot. Let us consider a regularly slotted RACH, where the random access slots are referred to with the index $i \geq 1$. Let $T_{Ri,\text{max}}$ be the ending time of the observation. $T_{\text{max}}$ equals to the time at which the last mobile user generates its last access request plus its waiting time until the immediate RA slot, $\Delta T_A$, plus its maximum duration to complete a RA procedure [17], which is given in Eq. (1).

$$T_{Ri,\text{max}} = t_i + \Delta T_A + \left(1 + (N_{\text{prev}} - 1) \frac{T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{BO}}}{T_{\text{RAR,REP}}}ight). \quad (1)$$

$\Delta T_A$ comprises the time difference between the actual arrival time of the last mobile user among mobile users and the immediate RA slot at which it can transmit its preamble, since preamble can only be sent at RA slot. Note that $\Delta T_A$ is 0 if this mobile user arrives at RA slot.

$$\Delta T_A = \begin{cases} 0, & \text{if } t_i \mod T_{\text{RAR,REP}} = 0, \\ T_{\text{RAR,REP}} - (t_i \mod T_{\text{RAR,REP}}), & \text{otherwise}. \end{cases} \quad (2)$$

Hence, the total number of RA slot during observation interval, $I_{\text{RA}}$, can be defined as

$$I_{\text{RA}} = \left\lceil \frac{T_{\text{max}}}{T_{\text{RAR,REP}}} \right\rceil - \left\lceil \frac{T_{\text{min}}}{T_{\text{RAR,REP}}} \right\rceil + 1. \quad (3)$$
A. Estimation of total number of contending mobile users

During the contention at RA slot $i$, a total of $M_i$ mobile users from all ACs transmit their preamble. Let $M_i$ be the total number of mobile user that transmit preamble at RA slot $i$. Likewise, let $M_i[n]$ be the total number of mobile user that conduct its $n$th preamble transmission in RA slot $i$, for $n$ being the index of preamble transmission, where $1 \leq n \leq N_{PT_{\text{max}}}$, such that $M_{i,k} = \sum_{n=1}^{N_{PT_{\text{max}}}} M_{i,k}[n]$. From the contention, $M_{i,F}[n]$ out of $M_{i}[n]$ mobile users are not acknowledged. In this work, $M_{i,F}[n]$ does not include mobile users that experience failure in MSG-3 and MSG-4 transmission since it is theoretically proven to be very small and negligible amount [17].

B. Definition of successful and failed mobile users

A mobile user is considered to be successful in its RACH procedure (and hence named successful mobile user) when it is acknowledged, since after being acknowledged in RAR message, it can proceed to MSG-3 and MSG-4 transmission and eventually obtain the resource for data transmission. A mobile user is considered to be failed in its RACH procedure (and hence named failed mobile user) only when it exceeds the preamble transmission limit without being acknowledged in RAR message.

Total number of new arrivals (i.e., mobile users that conduct its first preamble transmission, i.e. $n = 1$) in RA slot $i$, $M_{i}[1]$, can be calculated based on the arrival PDF as

$$M_{i}[1] = \begin{cases} 
M \sum_{t = \text{max}(0,t_{i-1})}^{\text{min}(t_{i},t_{i-1}+T_{\text{REP}})} A(t) & \text{if } i = 1 \text{ and } t_{f} \leq t_{i} \\
M \sum_{t = \text{max}(t_{i-1}+T_{\text{REP}},t_{f})}^{\text{min}(t_{i},t_{i-1}+T_{\text{REP}})} A(t) & \text{if } i > 1 \text{ and } t_{f} \leq t_{i}, \text{ and } t_{i-1} + T_{\text{REP}} \leq t_{i} \\
0 & \text{otherwise}
\end{cases}$$

where $t_{i}$ be the time (unit: sub-frame) of $i$th RA slot ($1 \leq i \leq I_{RA}$), which can be calculated as $t_{i} = t_{i-1} + (i-1)T_{\text{REP}}$. Note that $M_{i}[n]$ for $n > 1$ should be calculated separately from $M_{i}[1]$ since it also comprises preamble retransmission. $M_{i}[n]$ for $n > 1$ is given in Eq. (5).

$$M_{i}[n] \approx \sum_{P_{\text{min}}}^{P_{\text{max}}} \alpha_{p,i} \times M_{p,F}[n-1], \quad \text{for } 2 \leq n \leq N_{PT_{\text{max}}}. \quad (5)$$

with $\alpha_{p,i}$ indicates total number of mobile users that conduct preamble retransmission in $i$th RA slot from being failed at the range of $P_{\text{min}}$th to $P_{\text{max}}$th RA slot. $\alpha_{p,i}$, $P_{\text{min}}$ and $P_{\text{max}}$ are defined in [17] as follows
\[
\alpha_{p,i} = \begin{cases} 
\frac{(p-i+3)T_{\text{RA,REP}} + T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{BO}}}{W_{\text{BO}}} & \text{if } P_{\text{safe}} \leq p \leq i - \frac{T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{BO}}}{T_{\text{RA,REP}}} \\
\frac{T_{\text{RA,REP}}}{W_{\text{BO}}} & \text{if } i - \frac{T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{BO}}}{T_{\text{RA,REP}}} < p < i - 1 - \frac{T_{\text{RAR}} + W_{\text{BAR}}}{T_{\text{RA,REP}}} \\
\frac{(i-p)T_{\text{RA,REP}} - T_{\text{RAR}} + W_{\text{RAR}}}{W_{\text{BO}}} & \text{if } i - 1 - \frac{T_{\text{RAR}} + W_{\text{RAR}}}{T_{\text{RA,REP}}} \leq p \leq P_{\text{max}} \\
0 & \text{otherwise.}
\end{cases}
\]

\begin{align}
P_{\text{min}} &= \left[i - 1 + \frac{1-T_{\text{RAR}} - W_{\text{RAR}} - W_{\text{BO}}}{T_{\text{RA,REP}}} \right]. \\
\end{align}

and

\begin{align}
P_{\text{max}} &= \left[i - \frac{1+T_{\text{RAR}} + W_{\text{RAR}}}{T_{\text{RA,REP}}} \right].
\end{align}

After preamble transmission and the reception of RAR message, we can determine total number of mobile users that are acknowledged and unacknowledged. Out of unacknowledged mobile users, some of their preamble transmission are possibly detected and not collided (they are unacknowledged due to \(N_{\text{UL}}\) limitation). Total number of detected-and-not-collided mobile users in its \(n\)th preamble transmission at RA slot \(i\) is denoted as \(M_{i,\text{safe}}[n]\) herein. \(M_{i,\text{safe}}[n]\) can be estimated as

\begin{align}
M_{i,\text{safe}}[n] &= M_i[n] \times p_n \times e^{-\frac{M_i}{\bar{r}}}. 
\end{align}

The first part of Eq. (9), \(M_i[n]\), comprises the total number of mobile users that transmit its \(n\)th preamble at RA slot \(i\). The second part, \(p_n\), denotes the power ramping effect. With power ramping mechanism, the transmission power is gradually increased along with the retransmission counter, which increases the transmission success transmission probability in the later transmission round.

According to [12], \(p_n = 1 - e^{-n}\) is used in this paper. The third part, \(e^{-\frac{M_i}{\bar{r}}}\), denotes the estimated success transmission probability of one mobile user in a contention for \(R\) preambles, where the total number of contending mobile users are \(M_i\).

Detected-and-not-collided mobile users should receive acknowledgment before it can be considered as successful mobile users. Let \(M_i[n]\) be the total number of mobile users that is acknowledged after transmitting its \(n\)th preamble at RA slot \(i\). When total number of detected-and-non-collided mobile users in all ACs from \(i\)th RA slot, \(M_i\), is lower or equal to \(N_{\text{UL}}\), base station acknowledges all of them. However, due to resource limitation in RAR message, the base station will randomly choose \(N_{\text{UL}}\) mobile users out of \(M_{i,\text{safe}}\) mobile users to acknowledge when \(M_{i,\text{safe}}\) is more than \(N_{\text{UL}}\). Hence,

\begin{align}
M_{i,\text{s}}[n] &= \begin{cases} 
M_{i,\text{safe}}[n] & \text{if } M_{i,\text{safe}} \leq N_{\text{UL}} \\
\frac{M_{i,\text{safe}}[n]}{N_{\text{UL}}} & \text{otherwise.}
\end{cases}
\end{align}

where obviously,
\[
M_{i,\text{safe}} = \sum_{n=1}^{N_{\text{tx}}} M_{i,\text{safe}}[n].
\] (11)

By knowing the total number of successful mobile users that conduct \(n\)th preamble transmission in RA slot \(i\), \(M_{i,S}[n]\), total number of failed mobile users that conduct \(n\)th preamble transmission in RA slot \(i\) in Eq. (5) can be obtained as a simple subtraction as follow:

\[
M_{i,F}[n] = M_i[n] - M_{i,S}[n].
\] (12)

C. Performance Metrics

In this work, success transmission probability and average transmission delay are chosen as performance metrics. From the mobile users that transmit its \(n\)th preamble at RA slot \(i\), \(M_i[n]\), \(M_{i,S}[n]\) of them are acknowledged and \(M_{i,F}[n]\) of them are unacknowledged, such that \(M_i[n] = M_{i,S}[n] + M_{i,F}[n]\). The success transmission probability, \(P_S\), can then be derived as total number of successful (i.e., acknowledged) mobile users during the observation period divided by total number of mobile users. \(P_S\) can be expressed as

\[
P_S = \frac{\sum_{i=1}^{N_{\text{tx}}} \sum_{n=1}^{N_{\text{seq}}} M_{i,S}[n]}{M}.
\] (13)

Notice that some references refer this success probability as throughput, which is usually denoted as \(\tau\). Transmission delay for a mobile user is calculated as the time difference between when the data packet is ready to be transmitted until it is completely transmitted to the base station. The transmission delay comprises the waiting time between actual arrival time and the time of immediate RA slot, time for preamble transmission and retransmissions, the time for finishing MSG-3 and MSG-4 transmission, and the time for finishing the small data transmission. Obviously, the averaged value of transmission delay is calculated only for mobile users that successfully receive the resource grant from the RACH procedure. The average value of transmission delay is denoted as \(\overline{D}\), and is expressed as

\[
\overline{D} = \frac{\sum_{i=1}^{I_{\text{tx}}} \sum_{n=1}^{N_{\text{seq}}} M_{i,S}[n](\overline{T}_\tau + T_d)}{\sum_{i=1}^{I_{\text{tx}}} \sum_{n=1}^{N_{\text{seq}}} M_{i,S}[n]},
\] (14)

where \(\overline{T}_\tau\), is the average duration required by a mobile user that successfully finish its RACH procedure after transmitting preambles exactly \(n\) times. In other word, \(\overline{T}_\tau\) is the average time counted since the mobile user transmits its first preamble transmission until it receives MSG-4. The calculation of \(\overline{T}_\tau\) is given in Appendix B in [13]. Let \(T_d\) denote the duration for completing small data transmission. \(T_d\) is counted from the transmission time the first bit of the data until the transmission time the last bit of the data. In this work, the length of the data, \(L\), must not exceed 1KB as the specification of small data [1]. \(T_d\) depends on the transmission rate, \(R_c\), which is specified for each modulation and coding scheme (MCS) level. \(T_d\) can be obtained as

\[
T_d = \begin{cases} 
\frac{L}{R_c}, & L \leq 1\text{KB}, \\
1\text{KB}, & 1\text{KB} < L < 1\text{KB}, \\
\frac{L}{R_c}, & L \geq 1\text{KB}.
\end{cases}
\] (15)

Since the data is transmitted at the time which was originally used for signaling messages, the lowest MCS level with \(R_c = 16\) kbps is assumed.

5. Experimental Results

Computer simulation is used to verify the result from our analytical model. In the experiment, different load conditions are represented by different packet generation probability (\(e\)), while BEB algorithm is used by users in access class 1 and UB algorithm is used by users in access class 2. The
parameter of each backoff algorithm is listed in Table I. Fig. 1, 2, 3 and 4 shows the normalized throughput, packet dropping probability, average access delay and variance of access delay, respectively. In the four figures, all results obtained by our analytical model closely coincide with the results obtained by simulation. This shows that our analytical model can accurately estimate the system’s behavior.

The system is also evaluated under different number of users in each access class. Firstly, it is evaluated when access class 1 (AC1) contains 10 users and access class 2 (AC2) contains 90 users. Secondly, it is evaluated when access class 1 contains 10 users and access class 2 contains 50 users.

In the following, simulation (S) and mathematical analysis using the derived analytical model (A) are conducted to evaluate the systems. In this evaluation, we calculate and observe the result of success probability, $P_S$; normalized throughout, $\tau$; packet-dropping probability, $P_d$; average access delay, $D$; and variance of the access delay ($Var(D)$).

| Table II. Backoff Parameters |
|--------------------------------|
| BEB | Initial backoff window of AC 1 | 1 |
|     | Maximum backoff window of AC 1 | 16 |
| UB  | Initial backoff window of AC 2 | 1 |
|     | Maximum backoff window of AC 1 | 15 |

Figure 1 Normalized throughput

Figure 2 Packet dropping probability
The effect of backoff different backoff algorithms used by the two access classes can be shown in the four figures. In Figure 1, it is observable that uniform backoff (UB) obtains higher normalized throughput in both cases compared to binary exponential backoff (BEB). Meanwhile in Figure 2, it can be observed that binary exponential backoff (BEB) has higher packet-dropping probability in the two provided cases compared to uniform backoff (UB).

The results in Figure 3 show that uniform backoff (UB) yields higher average access delay in both cases compared to binary exponential backoff (BEB). This is because the uniform backoff provides enough time separation or spreading factor in time domain to let more devices to be successful. As the consequence, the average delay is higher. For most of IoT scenarios where battery energy consumed by transmission activity is far higher than its idle counterpart, having longer waiting time while ensuring higher chance to be successful in each transmission attempt can be beneficial, keeping in mind that most of such IoT applications are delay tolerant and less mission critical.

The results in Figure 4 demonstrates that the access delay in binary exponential backoff (BEB) is more stable, i.e. less differences among occurrences, for the two included cases when compared to the uniform backoff (UB)’s results. This can be traced back to its root of the generation of random backoff counter in each device. The results in Figure 4 are important for those who designed IoT application which needs timely report or uplink data delivery. Smaller variance of the access delay means that it is more likely for the accessing devices to have similar delay performance, which is more preferred in some IoT application, to simplify the timing procedure in the generated log or reports to be sent to the centralized pool.
Overall, we can observe that the uniform backoff yields higher throughput, but also comes with higher delay as the tradeoff. Meanwhile the binary exponential backoff yields lower delay which comes with lower throughput.

Experiment is then continued by considering an IoT application assumed as follow. There are hundreds (assumed as $M = 100$ to $1000$) of IoT devices reside in the service coverage of one access point or base station. Being a simple, cheap, and autonomous end device, these devices have to send a small measurement data which is generated periodically. The timing of data generation/arrival in each device is synchronized with the others. The periodicity of the data arrival is $T_{\text{arrival}}$ ms, which is assumed to be enough for finite number of transmission attempts to accommodate the contention and various types of fading. Following the design assumed earlier, the maximum number of transmission attempts for each generated packet is $N_{\text{PTmax}}$.

To optimally utilize the two types of backoff algorithm, the application designer should be aware of the tradeoff between normalized system throughput and the periodicity of data generation. In general, longer backoff window may resolve the contention which in this case may be caused by high number of devices in the cell. However, longer backoff window also restrict the application to generate its data too frequently.

Fig 5 depicts the maximum delay in respect to the maximum backoff window and maximum number of transmission attempt for each generated packet. $N_{\text{PTmax}}$ of 1 to 6 and maximum $W_{BO}$ of 1 to 32 are examined. In this figure, the results are similar in both backoff algorithms, which is due to obvious reason represented in the above equations. From the result in this figure, we can observe that having more retransmission attempt and longer maximum backoff window spends more time which eventually prolong the minimum interval between two successive data generation by the application layer. When the next data is generated after a period shorter than this minimum interval, with the assumed simplicity of IoT devices, the data may be dropped since the previous data is still in the buffer waiting for retransmission. This is of course by assuming that the channel is used by other devices too. In practice, the more device shares the channel (i.e. the cell), the more effective the investment (of the access point/base station) is.

![Figure 5 Minimum interval of data packet generation](image)

Collision is expected in normal multichannel slotted aloha system. The severity level of the collision is affected by number of channels and number of contending devices. Depending on the severity of collision, the maximum delay which become the lower bound of data generation interval depicted in Fig. 5 may be relaxed. The severity of collision is roughly the reverse of the system’s success probability in each interval of data generation. Fig. 6 shows the success probability for different duration of maximum backoff window, maximum number of transmission attempts, and
number of contending devices. For the ease of visual observation, only $N_{PT_{\text{max}}}$ of 2, 4, and 6 are displayed and 20 channels are assumed. By combining the results in both figures 5 and 6, application designer can decide its data generation interval based on his chosen $N_{PT_{\text{max}}}$, max $W_{BO}$, target number of served IoT devices per cell, and QoS factor of success probability.

![Figure 6 Minimum interval of data packet generation](image)

Let us consider an exemplary case where a delay-tolerant IoT application requires up to 400 devices deployed in each cell (the cell coverage area can be known for example by consulting its cell plan map) and 90% of successful transmission rate. The client wants to gather data as frequent as possible but keeping the cost low by having simpler/cheaper IoT devices. The application designer wants to know the minimum data generation interval. Let us further assume that the radio system can only support 20 channels. Hence, looking at Fig. 6 can be used to choose the smallest maximum backoff window and the smallest $N_{PT_{\text{max}}}$ for $M=400$ which can achieve success probability of at least 0.9. For this case, we can choose $N_{PT_{\text{max}}}=4$ with maximum backoff window of 11 slots, or $N_{PT_{\text{max}}}=6$ with maximum backoff window of 6 slots. Subsequently, by referring to Fig. 5, we can obtain the information for comparison summarized in Table 2.

| Table 2. Two feasible options for the example case |
|-----------------------------------------------|
| $N_{PT_{\text{max}}}$ | Max. $W_{BO}$ (slots) | Min. data generation interval (slots) |
|-----------------------|----------------------|-------------------------------------|
| Option 1              | 4                    | 42                                  |
| Option 2              | 6                    | 43                                  |

By observing the comparison in Table 2, although only 1 slot difference, applying option 1 to the system is better since the devices can generate more frequent data and send it to the centralized system, e.g. for more accurate analysis.

6. Conclusion
We propose an analytical model to estimate the performance of small data transmission. Our analytical model can be used to estimate a generalized system in consider with hybrid traffic which is included two different backoff algorithms. Each access-class adopts either UB or BEB algorithm independently. This differential-equation-based analytical model can derive the normalized throughput, the packet-dropping probability, the average access delay, and the variance of the access.
delay for random access channels in generalized multi-channel slotted ALOHA systems. Comparison with simulation results shows that the analytical model is extremely accurate in predicting the performance in this system.

7. References
[1] 3GPP TR 23.877, “Machine-type and other mobile data applications communications enhancements,” v. 0.9.0, Apr. 2013.
[2] 3GPP TS 36.321 V13.1.0, “Medium Access Control (MAC) protocol specification”, 2016.
[3] J. C. Arnbak and W. Van Blitterswijk, “Capacity of slotted ALOHA in Rayleigh-fading channels,” IEEE J. Sel. Areas Commun., vol. 5, no. 2, pp. 261–269, Feb. 1987.
[4] L. Kleinrock and S. S. Lam, “Packet switching in a multi-access broadcast channel: Performance evaluation,” IEEE Trans. Commun., vol. COM-23, no. 4, pp. 410–423, Apr. 1975.
[5] I. Rubin, “Group random-access discipline for multi-access broadcast channels,” IEEE Trans. Inf. Theory, vol. IT-24, no. 5, pp. 578–592, Sep. 1978.
[6] H. H. Tan and H. Wang, “Performance of multiple parallel slotted ALOHA channels,” in Proc. INFOCOM, Mar. 1987, pp. 931–940.
[7] V. Naware, G. Mergen, and L. Tong, “Stability and delay of finite user slotted ALOHA with multipacket reception,” IEEE Trans. Inf. Theory, vol. 5, no. 7, pp. 2636–2656, Jul. 2005.
[8] P. R. Jelenkovic and J. Tan, “Stability of finite population Aloha with variable packets,” Tech. Rep. EE2009-02-20, 2009, arXiv:0902.4481v2.
[9] Y.J. Choi, S. Park, and S. Bahk, “Multichannel random access in OFDMA wireless network,” IEEE J. Sel. Areas Commun., vol. 24, no. 3, pp. 603–613, Mar. 2006.
[10] P. Zhou, H. Hu, H. Wang, and H. H. Chen, “An efficient random access scheme for OFDMA systems with implicit message transmission,” IEEE Trans. Wireless Commun., vol. 7, no. 7, pp. 2790–2797, Jul. 2008.
[11] L. Kleinrock and F. Tobagi, “Packet switching in radio channels, Part I— Carrier sense multiple-access modes and their throughput—Delay characteristics,” IEEE Trans. Commun., vol. COM-23, no. 12, pp. 1400–1416, Dec. 1975.
[12] R. G. Cheng, C. H. Wei, S. L. Tsao, and F. C. Ren, “RACH collision probability for machine-type communications,” in Proc. IEEE 75th Veh. Technol. Conf., May 2012, pp. 1–5.
[13] IEEE, “Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems”, IEEE Std. 802.16TM, May 2009.
[14] R. Harwahyu, J. Loo, R. G. Cheng, Y.Y. Chu, “Smoothing the drive, Critical Access fo Intelligent Transportation Systems,” IEEE Veh. Technol. Magz., June 2017.
[15] Y. Y. Chu, R. Harwahyu, R. G. Cheng, C.H. Wei, “Study of Generalized Resource Allocation Scheme for Multichannel Slotted ALOHA Systems,” IEEE 26th PIMRC, Aug. 2015.
[16] 3GPP TR 37.869, “Study on enhancements to machine-type-communication (MTC) and other mobile data applications enhancements,” v1.0.0, Aug. 2013.
[17] C. H Wei, G. Bianchi, R. G. Cheng, “Modeling and Analysis of Random-access Channels with Bursty Arrivals in OFDMA Wireless Networks,” IEEE Trans. Wireless Comm., vol. 14, no. 4, pp. 1940-1953, 2015.

Acknowledgements
This article is presented at the International Conference on Smart City Innovation 2018 that supported by the United States Agency for International Development (USAID) through the Sustainable Higher Education Research Alliance (SHERA) Program for Universitas Indonesia’s Scientific Modeling, Application, Research and Training for City-centered Innovation and Technology (SMART CITY) Project, Grant #AID-497-A-1600004, Sub Grant #IIE-00000078-UI-1.