CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Electrical Engineering

Department of Telecommunication Engineering

Distributed Queuing-based Random Access Procedure in Mobile Networks

Ing. Programme: Communication, Multimedia and Electronics
Specialisation: Electronic Communication Networks

May 2016

Author: Bc. Yang Ping-Hsun
Supervisor: doc. Ing. Zdeněk Bečvář, Ph.D
Prof. Ray-Guang Cheng

I hereby declare that this master’s thesis is completely my own work and that I used only the cited source in accordance with the instruction about observance of ethical principles of preparation of university final projects.

Prague, May 26, 2016

……………..………

Signature

Acknowledgement

I am strongly appreciated to my supervisor doc. Ing. Zdeněk Bečvář, Ph.D for his invaluable help and instructions on all administrative matters, courses studying and my master thesis during my Double-Degree Study in CVUT. And I am sincerely thankful to Prof. Ray-Guang Cheng for the opportunity to study in CVUT and his continuous encouragement throughout my master study and careful teach on my working attitude, research methodology.

My thanks also belong to BMW lab mates, who supported me when I was confused. Thanks to Ing. Jan Plachý for all kind help to my staying in Prague. Thanks to my parents and brother for their encouragement, support, and concerns. At last, I would like to thank to one person who always pushed me and concern about me.

ABSTRACT

Random access procedure is an access mechanism to radio resources in Long Term Evolution (LTE) mobile networks to handle a competition for radio resources of Machine Type Communication (MTC) devices. In the random access procedure, the MTC devices transmit a request for radio resources in the first available opportunity. The devices which select the same resource fail the access and need to compete again. Thus, in the case that massive number of MTC devices are competing at the same time and transmit the access request simultaneously, the system performance may degrade for the high collision probability. Thus, a challenge is to develop a solution ensuring low collision probability for massive number of MTC devices while guaranteeing fairness. A suitable extension of a conventional random access procedure towards support of massive MTC access is to distribute colliding devices into parallel queues. This approach is known as Distributed Queueing Random Access Procedure (DQRAP). The DQRAP improves the performance regardless of the number of MTC devices and thus enables massive MTC communication. This thesis aims to development of analytical models to estimate the performance of the average maximum access delay and the average number of transmissions in a multi-channel slotted ALOHA system using DQRAP. As the results of a numerical analysis and simulations show, the proposed models accurately approximate the simulations.

Key words: Analytical model, Machine Type Communications, LTE, Distributed Queueing, Random Access Procedure

ANOTACE

Procedura náhodného přístupu je způsob řízení soutěžení velkého množství zařízení typu stroj (Machine Type Communication, MTC) o přístup k radiovým prostředkům v mobilních sítích LTE. Během procedury náhodného přístupu vyšlou MTC zařízení požadavek na přístup ke konkrétním radiovým prostředkům v prvním možném okamžiku. Zařízení, která vybrala pro přístup stejné radiové prostředky jsou v přístupu neúspěšná a musí svůj požadavek opakovat. V případě velkého množství MTC zařízení soutěžících o radiové prostředky ve stejný čas může být efektivita přístupu výrazně snížena z důvodu vysoké pravděpodobnosti kolize žádostí o stejné prostředky. Výzvou tedy je vývoj procedury náhodného přístupu zajištující nízkou pravděpodobnost kolize pro velké množství MTC zařízení a zároveň garantující spravedlivý přístup pro všechna zařízení. Jedno z možných řešení pro velké množství přistupujích MTC zařízení je distribuovat zařízení, která neuspěla v přístupu k radiovým prostředkům do více front. Tento mechanismus je znám jako tzv. Procedura náhodného přístupu s distribuovanou frontou (Distributed Queueing Random Access Procedure, DQRAP). Metoda DQRAP zlepšuje efektivitu přístupu MTC zařízení bez ohledu na jejich počet a tím umožňuje komunikaci velkého množství MTC zařízení. Cílem této diplomové práce je vetvoření analytických modelů pro odhad průměrného maximálního zpoždění přístupu k radiovému kanálu a průměrného počtu přístupů ve vícekanálovém systému ALOHA pomocí DQRAP. Jak ukazují výsledky numerické analýzy a simulací, navržené modely přesně aproximují simulační výsledky.

Klíčová slova: Analytický model, Komunikace strojů, LTE, Náhodný přístup s distribuovanou frontou

Table of Contents

List of Figures

List of Tables

List of Acronyms

MTC	Machine Type Communication
3GPP	Third Generation Partnership Project
RAN	Radio Access Network
H2H	Human-to-human
RACHs	Random Access Channels
LTE	Long Term Evolution
UE	User equipment
UE-ID	UE identity
eNB	Evolved Node B
GID	Group identity
RAP	Random Access Procedure
ACB	Access Class Barring
EAB	Extended Access Barring
DQRAP	Distributed Queueing-based Random Access procedure
RAR	Random Access Response
MAC	Medium Access Control
BI	Backoff Indicator
HARQ	Hybrid Automatic Repeat Request
RRC	Radio Resource Control
ACK	Acknowledgment
NACK	Negative-acknowledgment
CRQ	Contention Resolution Queue
RAS	Random-access slot
PRACH	Physical Random Access Channel

I. INTRODUCTION

Machine Type Communication (MTC), which is also known as Machine to Machine (M2M) communication, is a service defined by standardization organization The 3rd Generation Partnership Project (3GPP) [1] to enable direct communication among electronic devices, machines, and enable communications from MTC devices to a central MTC server or a set of MTC servers [2] over cellular networks [3]. MTC usually involves a large number of devices to support a wide range of applications, such as smart grid, road security, or consumer electronic [4]. However, MTC imposes challenges on cellular networks related to new traffic characteristics and simultaneous accesses of radio resources in radio access network (RAN) by high number of devices [3][4][5]. All these problems may cause serious congestion, peak load and may result in intolerable delays, packet loss, or even service unavailability to conventional human-to-human (H2H) communication services [2][3]. Hence, 3GPP also focuses on finding a proper overload control mechanisms to handle the congestion and guarantee network availability and quality of H2H services under heavy MTC load [3][4].

The overload control of uplink Random Access Channels (RACHs) in radio access network (RAN) is one of the principle working items for 3GPP Long Term Evolution (LTE) and future mobile networks [3][4]. The purpose of RAN overload control is to avoid RAN overloading of simultaneous access of the RACHs by mass MTC devices. Based on MTC traffic generation [6], the RAN overload control schemes can be categorized into push-based and pull-based approaches [7]. In the push-based approach, there is no restriction and the MTC traffic is pushed from MTC devices to the network until RAN overloading is detected. In the pull-based approach, the MTC traffic is pulled by the network, which means the network may properly control the MTC traffic load through a paging to prevent RAN overload [3].

The paging and group paging are potential pull-based RAN overload control schemes [3][8][9]. In LTE, a downlink paging channel is defined to transmit the paging information to a user equipment (UE) and informs the UEs of a change of system information and emergency notifications. In the original paging scheme, a specific UE is activated when the network transmits a paging message at the UE's paging occasion. The paging occasion of each UE is determined according to its UE identity (UE-ID). Current paging mechanism that has been originally designed for H2H services can only page up to 16 devices with a single paging message, and only two paging occasions are available per 10 ms radio frame [3][8]. Therefore, to activate a large number of MTC devices in the original paging scheme, an LTE base station, usually denoted as evolved Node B (eNB), must transmit multiple paging messages over a long period. A group paging mechanism that uses a single group paging message to activate a group of MTC devices solves the multiple paging messages transmission of paging scheme is proposed in [1]. In the group paging, an MTC device is assigned by a unique group identity (GID) after camping on a network and joining a group. All of the MTC devices in the group listen to the same paging channel at the same paging occasion derived from the GID [8]. When the GID appears in a group paging message the corresponding group of MTC devices shall simultaneously perform the standard Random Access Procedure (RAP). The LTE RAP is described in details in Section II.

For access in LTE, every device generates a random preamble, which is sent to the eNB. The preamble is a 6 bit signature that a device uses to attempt an access; there are maximum of 64 (2^6) possible preambles to be selected [11]. In LTE RAP, the device successfully access network if and only if a preamble is chosen solely by it (other devices choose another preamble). Contrary, the device access fails if there is another device choosing the same preamble. When a device fails, it waits for a backoff time and retransmits the attempt according to the backoff parameter value sent by the eNB. The MTC devices, which fail random accesses shall perform the standard LTE random backoff procedure to retransmit their random-access attempts during a paging access interval until the specified retry limitation is exceeded [3]. Note that MTC devices are informed about the paging access interval and the dedicated random-access resources reserved for further communication of group paging by the group paging message.

In the group paging, the number of MTC devices to be paged is known and the MTC devices access the network in a highly synchronized manner once they are paged [3] However, the number of random-access attempts (number of accessing devices) in each random-access slot gradually decreases if any device successfully accesses the RACH because no new arrival is generated after the new paging.

On the particular topic of a contention resolution, i.e., resolution on attempts of more devices trying to access radio resources the same time, Access Class Barring (ACB) and Extended Access Barring (EAB) are proposed in 3GPP Release 8 and Release 11, respectively. Both ACB and EAB are improvements of the access mechanisms of cellular systems to handle massive amount of devices in a single cell [1][10]. However, the limitation of these solutions is that they are based on the backoff periods, which disperse access attempts. This results in a negative impact on the energy consumption and the access delay for the devices [11] Thus, an alternative procedure, denoted as Distributed Queueing-based Random Access procedure (DQRAP) is presented in [11]. DQRAP is the random access procedure based on a tree-splitting algorithm and a distributed queueing mechanism. By the distributed scheduling of the queues, DQRAP provides efficient channel utilization regardless of the number of accessing MTC devices and reduces the average access delay and the energy consumption with a low blocking probability for a massive number of simultaneous access attempts [11]. The standard random access procedure is presented in section II while section III describes the DQRAP mechanism and the integration of DQRAP into the standard random access procedure. In [11], the authors outline the idea of DQRAP and conduct simulations to prove its efficiency; however, analytical analysis is not carried out.

Thus, this thesis focuses on the behavior and the modeling of finite-user random access using Distributed Queueing-based Random Access Procedure (DQRAP) proposed in [11] triggered by the group paging. In the thesis, the group paging is also assumed. It means new arrivals are generated only at the beginning of the first access cycle and the number of contending devices in each access cycle is gradually decreased if any device successfully accesses the channel. In other words, the arrival rate and the successful transmission probability are considerably decayed in each access cycle. In this thesis, we estimate the average number of transmissions and average maximum access delay by adopting analysis model developed in this thesis. We propose a model for each performance metric. Then, the models are adapted in order to reduce their complexity. The numerical results show that the proposed analytical model can accurately estimate the performance metrics and matches the simulation results.

The organization of this paper is as follows. In Section II, the related work is described. Then, LTE Random Access Procedure is presented in Section III. In Section IV, the DQRAP is thoroughly described. The system model is presented in Section V. The developed analytical models are presented and discussed in Section VI. The numerical results of analytical analysis are discussed in Section VII. Finally, conclusions are provided in Section VIIII.

II. RELATED WORK

The radio access network (RAN) overload control is one of the most important issue for 3GPP Long term Evolution (LTE). Plenty of schemes on the topic of contention resolution to improve the overload problem are proposed. In the ACB scheme [18][9], different MTC traffic types are classified into different access classes and each device class is assigned a specific ACB factor. By setting different ACB factors, each MTC access class has different channel access probability, which means the network can control the traffic load by setting the ACB factor. The main disadvantage of ACB mechanism is that some devices may experience the unpredictable increased delay [20]. In the separated RACH resources scheme [21], the network performs overload control by reserving different dedicated Random Access Channel of for the H2H and MTC traffic or allocating different random access slots (RAS) to H2H or MTC devices. In other words, both types of traffic (H2H and MTC) have distinct channel access probability. This solution reduces the negative effect of this method on non-M2M devices, but the performance is still notably decreased if the MTC traffic load is high because the available resources for MTC devices are reduced [20]. In the dynamic allocation of RACH resource scheme [22], the network predicts the MTC traffic load and dynamically allocates RACH resources for MTC devices in the case of congestion. The scheme can solve most cases of congestion, but the allocation occupies the resources originally intended for data transmission [20]. In the backoff adjustment scheme [22][23], different backoff timers are assigned to MTC devices to delay the access attempts. The scheme is not able to cope with peak congestion level because the reduction of an average access delay mainly relies on an improvement of channel access probability [20]. All of the previous proposed schemes aim to deal with the congestion of large number of MTC devices but cannot provide a feasible solution with a considerate balance between the access delay, access probability, and energy consumption A medium access control (MAC) protocol called Distributed-queueing (DQ) mechanism which is first proposed in [24] demonstrates the stability of its performance independently on the number of devices transmitting the access simultaneously. The DQ mechanism utilizes virtual distributed queue to reserve the RASs for collided devices to retransmit the access request. The mechanism can be implemented with simple modification into the standard random access procedure as proposed in [11]. The DQ mechanism reduces the energy consumption of MTC devices and the access delay while maintaining low blocking probability under massive number of devices. Thus, the DQ approach is suitable for the massive number of simultaneous arrivals of MTC devices.

III. LTE RANDOM ACCESS PROCEDURE

This section explains the contention-based Random Access (RA) procedure defined for LTE networks. The RA procedure mainly consists of a four-message handshake between the device (UE) and the eNB. Figure 1illustrates the LTE RA procedure [3][11][12][17]. At first, a device synchronizes to the downlink timing ((1) in Figure 1). Secondly, the device randomly selects a RA preamble from a group of preambles reserved for the RACHs and transmits the RA preamble (Msg1) in a randomly chosen RA slot and a frequency band ((2) in Figure 1). Note that the preamble is transmitted through the RACH shared by multiple devices and the signalling messages are transmitted in a dedicated channel specifically reserved for each device.

If the eNB detects a preamble, it sends back a Random Access Response (RAR) message (Msg2) indicating the identity of detected preamble(s) selected by devices, uplink timing alignment instructions, and the dedicated uplink resource reserved for devices to transmit the Msg2 ((3) in Figure 1) [12]. Each response message carries a medium access control (MAC) header and one or more MAC RARs. The MAC header may carry the backoff parameter values, denoted as Backoff Indicator (BI), for the collided or un-detected UEs [3]. The collided or undetected devices should wait for a specific number of sub-frames before it attempts to access the channel again. Thu number of sub-frames is expressed by the backoff counter. If a device receives the RAR without information that the preamble it selected and transmitted in the Msg1, the device randomly chooses a backoff counter from zero to the BI and retransmits a newly selected RA preamble (Msg1) in the next available RA slot when the backoff counter expires (i.e., decreases to zero). In LTE, the range of BI is from 0 to 960 sub-frames [3][12]. The procedure continues until the maximum number of preamble transmissions is reached. If the maximum number of transmissions is reached, additional attempts are blocked.

After the device receives the RAR message from the eNB, a remaining signaling required for connection setup is transmitted on the assigned dedicated uplink resource in a synchronized manner by using the same procedures as normal data transmission (see [3]). Non-adaptive Hybrid Automatic Repeat Request (HARQ) is subsequently enabled to protect the signaling exchange of the message [3][12][15]. After the device successfully receives the RAR message, it sends the Msg3, a 'Radio Resource Control (RRC) connection request message, carrying the device ID, to the eNB at the radio resource assigned by the eNB ((4) in Figure 1). The eNB responses with an HARQ acknowledgment (ACK) or negative-acknowledgment (NACK) after the time interval required by receiving HARQ ACK (T_{HARQ}) [3].

If the eNB successfully receives the Msg3, the eNB responses with HARQ ACK ((7) in Figure 1). In contrast, if the device receives HARQ NACK, it waits for a gap time for the Msg3 retransmission ($\mathrm{T_{M3}}$) [3] and, then retransmits the Msg3 ((6) in Figure 1). After successful transmission of the Msg3, the eNB waits for a gap time to monitor the Msg4 (T_{A_M4}) [3] and transmits the Msg 4 to inform about the setup of RRC connection ((8) in Figure 1). Like in case of the Msg3, the UE waits for T_{HARQ} and sends another ACK to the eNB if the Msg4 is successfully received ((11) in Figure 1). If the eNB does not receive the ACK for the Msg4, it waits for a gap time for the Msg4 retransmission (T_{M4}) [3] and retransmits the Msg4 ((9) in Figure 1). The number of HARQ retransmission of the Msg3 and the Msg4 is limited to $\mathrm{N_{HARQ}}$ times. The device starts/restarts a contention resolution timer T_{CR} indicating maximum duration of the random access procedure (presented in sub-frames) whenever it transmits the Msg3 [3]. The device declares a random-access failure and reverts to Step (1) to retransmit its RA attempt if the contention resolution timer expires. Note that the Msg 3 and the Msg 4 are used for carrying connection setup signaling messages as well as for contention resolution.

In some cases, the eNB may have a chance to decode the same preamble transmitted by multiple devices and reply a response message [3]. After these devices receive RAR message indicating the same detected preamble, they will transmit their own Msg3 on the same dedicated resource and then realize the random-access failure after the expiry of the contention resolution timer.

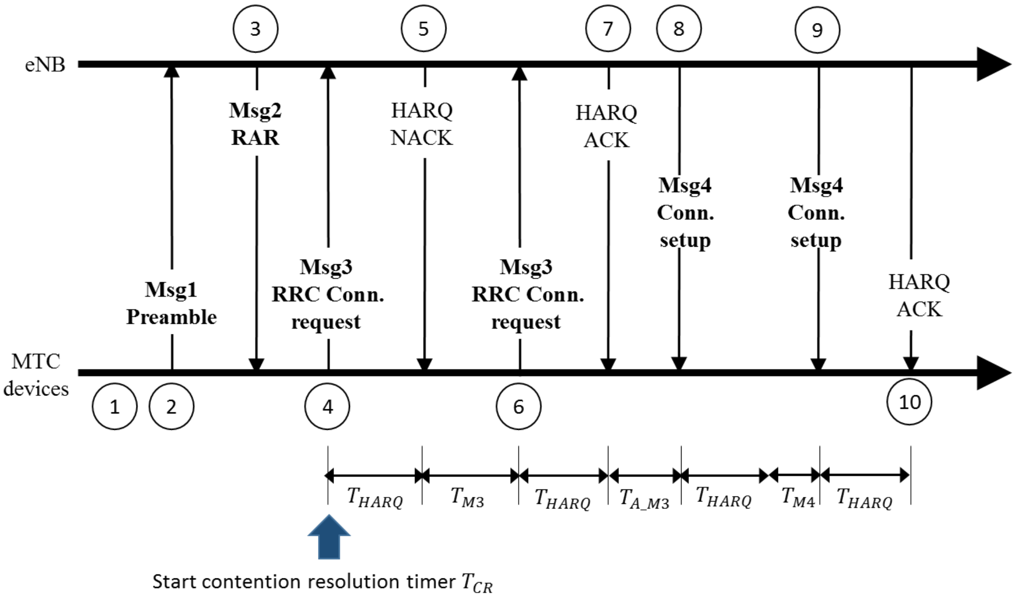


Figure 1. LTE Random Access Procedure [12]

The 3GPP includes the Access Class Barring (ACB) scheme in subsequent amendments to the standard to provide additional control mechanisms [1], which control the overload problem by classified access classes with different access probabilities. Figure 2 shows the application of the ACB scheme to the 3GPP RA procedure [11]. In ACB scheme, the network informs each device about the barring status, which denote if the ACB scheme is active or not. If the ACB is active, each device draws a random number between 0 and 1 with uniform distribution upon a start of initiation of the radio network connection. The random number is compared with the barring rate informed by the network. The device is barred if the random number is smaller than the barring rate. In contrast, the device can continue to initiate the RA procedure.

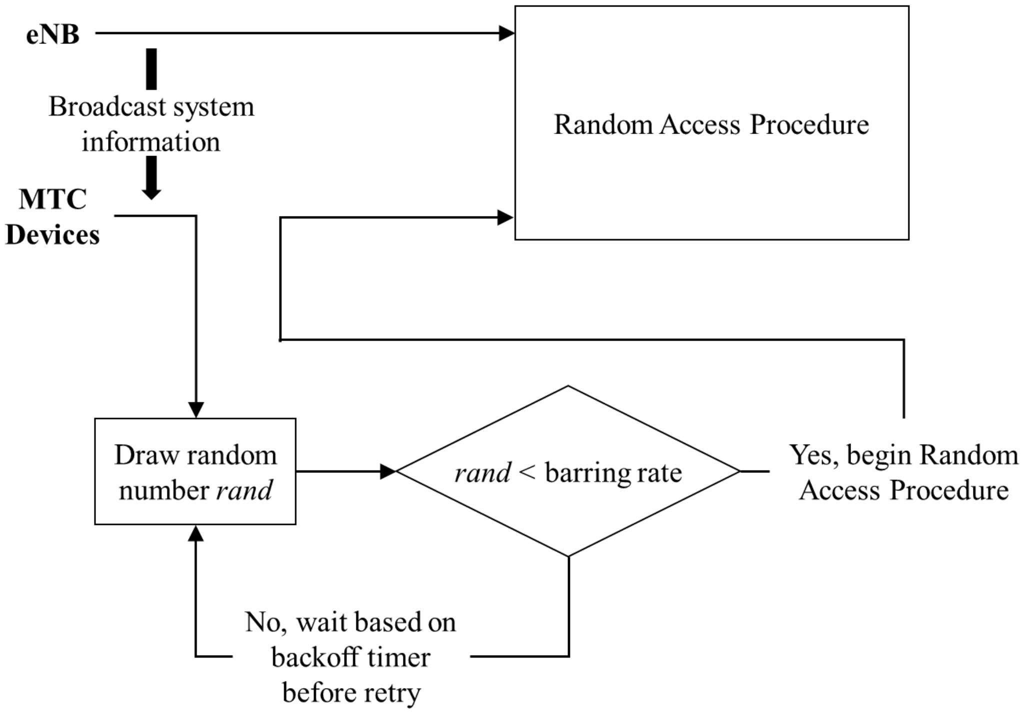

eNB
Broadcast system information
MTC Devices
Random Access Procedure
Draw random number *rand*
rand < barring rate
Yes, begin Random Access Procedure
No, wait based on backoff timer before retry

Figure 2. Application of Access Class Barring (ACB) scheme to Contention-Based Random Access (RA) Procedure for LTE [11].

IV. DQ-BASED RANDOM ACCESS PROCEDURE

This section presents the DQ mechanism and its integration into the standard RA procedure [11] as described in Section III.

A. *Distributed Queueing for Contention Resolution*

The DQ mechanism is based on an m-ary tree splitting algorithm with a simple set of rules to split collided devices into group by utilizing virtual collision resolution queues during an access procedure [11]. When collisions are detected, the devices are split into groups for the subsequent retransmissions. The splitting reduces the probability of collision by decreasing the number of simultaneous attempts. The distributed scheduling of the queues enables almost full channel utilization regardless of its capacity, the number of the transmitting devices, and the traffic pattern [11]. The queues are distributed in the sense that each device uses internal counters to represent the queue length and the position of the device within the queue. The values of each counter are updated based on the network feedback. In this way, the devices handle their transmission turn.

Figure 3 depicts an example of the DQ algorithm execution based on [11]. In the first RA Slot, six devices (*d1 – d6*) request access. The collision happens when more than one device selects the same preamble, (in our example, *d1* collides with *d2*, and *d4* collides with *d5* and *d6*). For each set of colliding devices an RA Slot is exclusively assigned for retransmission of the RA attempt. The colliding devices enter a queue referred to as Collision Resolution Queue (CRQ). Different contention group is set for each collided preamble and the CRQ length is increased by one per each collided preamble. In the example in Figure 3, the devices *d1* and *d2* collide with the preamble 1 in the RA Slot 1, and, thus, *d1* and *d2* enter in the first position in the CRQ; *d3* succeeds with the preamble 2; and *d4, d5* and d6 collide with the preamble 3 and enter in the second position in the CRQ. In the RA Slot 2, *d1* and *d2* contend for access since they are both at the first position in the CRQ while *d4*, *d5* and *d6* wait in the queue until the next RA Slot. The devices *d1* and *d2* collide again in the RA slot 2 with the preamble 1; this group enters at the end of the CRQ. At the RA Slot 3, *d4*, *d5* and *d6* succeed by selecting different preambles and leave the CRQ. At the RA Slot 4, *d1* and *d2* contend again and succeed.

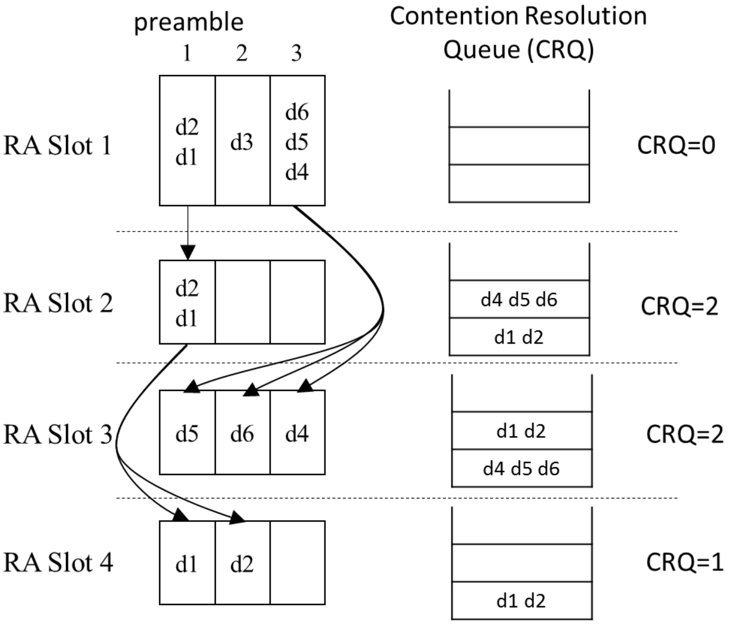


Figure 3. Distributed Queueing algorithm and Contention Resolution Queue (CRQ) behavior in the collision resolutions [11].

B. *Distributed Queueing (DQ)-based Random Access (RA) procedure [11]*

Upon initial access, the device selects a RA Slot and gets the current status of the CRQ from corresponding Msg2 from the eNB. New devices are not allowed to initiate the access if there is an ongoing contention in the selected RA Slot (i.e., the CRQ length is equal or more than one). Therefore, the device will not transmit in the next RA Slot and repeats this procedure until there is no further collisions detected.

If a free RA Slot is found, the device transmits a preamble on the next occurrence of the RA Slot and it waits for the corresponding Msg2. Three states may be provided in the Msg2 and the devices will do as follows:

1. *Empty state*: no preamble was received. The device increases by one the preamble

retransmission counter and reenter the CRQ.

2. *Collision state*: a collision was detected. The device increases by one the preamble retransmission counter and reenter the CRQ.

3. *Success state*: a preamble was received and no collision was detected. The device decodes the RAR message and proceeds to the transmission of Msg3.

V. SYSTEM MODEL

This work considers a fixed number of devices performing a distributed-queueing-based random access in a multichannel slotted ALOHA system [16]. In this system, time is divided into fix-length ‘access cycles’. Each access cycle contains a random-access slot (RAS) specifically reserved for the devices to transmit their attempts for an access of radio channel. Figure 4 shows how the RAS is mapped in the access cycle. The variable that controls the number of RAS per frame is a configuration index in Physical Random Access Channel, denoted as *PRACH Configuration Index*. The *PRACH Configuration index* is broadcasted by the eNB in LTE [5]. The *PRACH Configuration index* value adopted for simulations in this thesis is 3. i.e., one RAS for each access cycle. In each RAS, a device can indicate its attempt by transmitting a preamble randomly chosen from a set of preambles. The number of preambles is equivalent to the number of available channels and the number of available preambles for contention-based access is broadcasted by the eNB in the *Broadcast Downlink Channel* [5]. The attempt for radio channel access is successful if the preamble is chosen by only one user and the attempt is failed (preambles collided) if the same preamble is selected by more than one devices. At the end of each RAS, the eNB broadcasts an *RAS status* containing a list of successful preambles and collided preambles. The device learns the success or failure of its attempt before the end of the access cycle. The successful devices are assigned with dedicated channels. The collided devices follow a distributed queueing algorithm proposed in [11] to determine the next RAS to re-transmit their attempts.

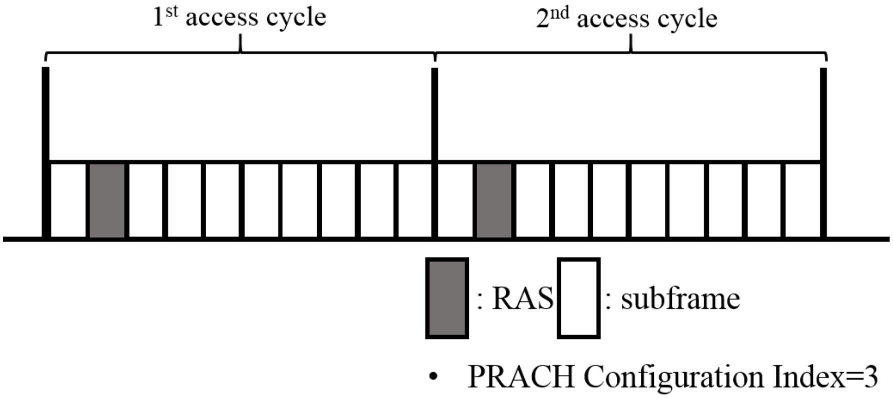


Figure 4. Structure of access cycle

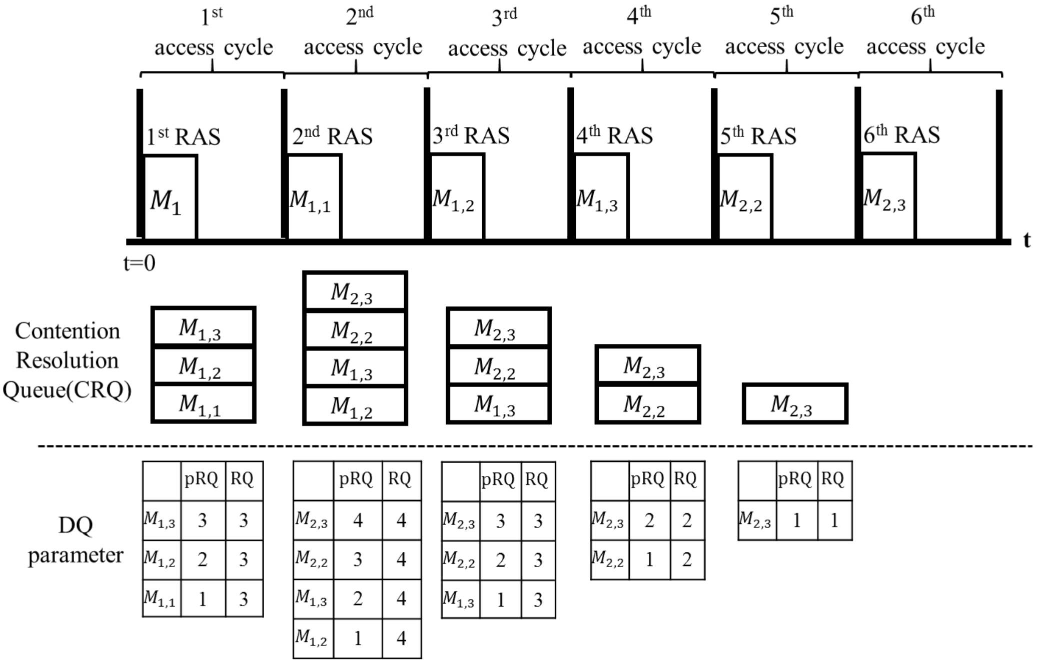
The distributed queueing algorithm utilizes virtual collision resolution queues to split collided devices into groups and thus, reduces the collision of subsequent re-transmissions. Devices, which select the collided preamble forms a contention group and enter a contention resolution queue. The queues are distributed in the sense that each device uses two counters, an RQ counter and a pRQ counter, to represent the queue length and the position of the devices within the queue, respectively [11]. A dedicated RAS is then exclusively assigned to the contention group of devices for retransmitting their attempts. Each device can update the values of both counters and compute its position in the queue based on the feedback of the RAS status. The devices who are in the same position in the queue are seen as one contention group. In distributed queueing, the group of devices located on the bottom of the Contention Resolution Queue (CRQ), i.e., pRQ=1, perform retransmission in the next RAS. The value of RQ is decreased by one when the contention group transmits in the RAS. Contrary, the value of RQ is increased by one when there is a collided preamble in the RAS. The devices, which choose the collided preambles enter the end of the contention resolution queue. The position of the devices in the contention resolution queue is indicated by pRQ.

We consider a one-shot random-access scenario as in [13], which represents access requests transmitted simultaneously by all devices triggered by group paging signal. That is, all of the devices receive a group paging from the eNB and simultaneously transmit their first random-access attempts in the first RAS based on the reception of the same paging signal, e.g., requesting measurement data report from smart meters (electricity, water, and gas). The collided attempts are retransmitted following the Distributed-Queuing Random Access (DQRA) protocol and the maximum number of retransmissions is infinite.

Let $M_{i,j}$ be the number of devices, which choose the j^{th} preamble at the i^{th} RAS. The preamble *j* is collided if $M_{i,j} \geq 2;$ is success if $M_{i,j} = 1;$ and is idle if $M_{i,j} = 0.$ Figure 5 illustrates the operation of the one-shot random access adopting distributed queueing. The lower part of Figure 5 shows the values of RQ and pRQ counters. In this example, M_1 devices simultaneously transmit their attempts in the 1st RAS, $M_1 = \sum_{\forall j} M_{1,j}$. Let's assume 3 preambles are reserved in each RAS. After the 1st transmission, $M_{1,1}$ devices collided by choosing *Preamble 1*; $M_{1,2}$ devices collided by choosing *Preamble 2*; and $M_{1,3}$ devices collided by choosing *Preamble 3*. At the end of the 1st RAS, the $M_{1,1}$, $M_{1,2}$, and $M_{1,3}$ collided devices enter the contention resolution queue in the order, which corresponds to the order of the preamble they choose. Note that the bottom of the queue will be served first. The queue length, i.e., RQ value, is set to be the number of collided preambles, that is, 3. The pRQ values for the three contention groups represent the position of the devices within the queue. That is, pRQ for the $M_{1,1}$, $M_{1,2}$, and $M_{1,3}$ devices is set to be 1, 2, and 3, respectively. The contention group of the $M_{1,1}$ devices is at the bottom of the contention resolution queue and thus, get the chance to transmit at the subsequent (i.e., the 2nd) RAS. We assume that among $M_{1,1}$ devices, $M_{2,2}$ and $M_{2,3}$ devices are collided and enter the contention resolution queue. Note that the total number of collided devices in the 2nd RAS is always lower than or equal to the number of devices which transmit in the 2nd RAS

($i.e.,\ M_{2,2}+M_{2,3}\leq M_{1,1}$). At the end of the 2nd RAS, the collided $M_{2,2}$ and $M_{2,3}$ devices enter the contention resolution queue and the RQ value is updated to be 4, i.e., 3 (collided devices in the 1st RAS) -1 (device sending in the 2nd RAS) +2 (collided devices in the 2nd RAS) = 4. The pRQ values for the existing $M_{1,2}$ and $M_{1,3}$ devices are both decreased by one and are updated to be 1 and 2, respectively. The pRQ values for the newly collided $M_{2,2}$ and $M_{2,3}$ devices are set to be 3 and 4, respectively. The $M_{1,2}$, $M_{1,3}$, $M_{2,2}$, and $M_{2,3}$ devices are scheduled to transmit in the 3rd, 4th, 5th, and 6th RAS, respectively, as shown in Figure 5 (assuming no collisions). Their counters are updated accordingly in similar way as described for the 1st and 2nd RAS above.

As in [11], we consider LTE network where devices are cell-synchronized and they receive all configuration parameters related to the random access procedure. In real environment, the eNB can detect collision of the same preamble sent by more than two devices, e.g., when two devices transmit the same preamble, the eNB will decode and send Random Access Response (RAR) message back to both devices. This may affect the operation of DQRA. Thus, it is also assumed that the eNB is not able to decode simultaneous transmission of the same preamble. Under this assumption, the eNB can detect collision of preambles sent by more than two devices and force these devices with collided preambles to enter the CRQ. We assume there is no transmission error and detection error for signaling messages (i.e., in addition to collision, Msg1 to Msg4 in Figure 1 are received without error), which means there is no condition of the random access failure because of an expiry of the contention resolution timer as described in Section III.



	pRQ	RQ
$M_{1,3}$	3	3
$M_{1,2}$	2	3
$M_{1,1}$	1	3

	pRQ	RQ
$M_{2,3}$	4	4
$M_{2,2}$	3	4
$M_{1,3}$	2	4
$M_{1,2}$	1	4

	pRQ	RQ
$M_{2,3}$	3	3
$M_{2,2}$	2	3
$M_{1,3}$	1	3

	pRQ	RQ
$M_{2,3}$	2	2
$M_{2,2}$	1	2

	pRQ	RQ
$M_{2,3}$	1	1

Figure 5. Distributed queueing mechanism and CRQ behavior in the collision resolution.

One of the advantages of DQRA is that it enables almost full channel utilization regardless of the number of transmitting devices [11]. This advantage results in a low number of retransmissions for MTC devices. The Figure 6 illustrates an example on how DQRA is capable to handle big number of devices. In Figure 6, we assume 3 preambles are available and *M* devices access in the first RAS. In the condition where *M* is considerably high and the CRQ of each RAS is fully utilized, DQRA shows exponential growth of the maximum number of devices, which can successfully access (i.e., capacity, which is equal to the sum of preambles of each RAS in each transmission) with the increasing order of transmissions.

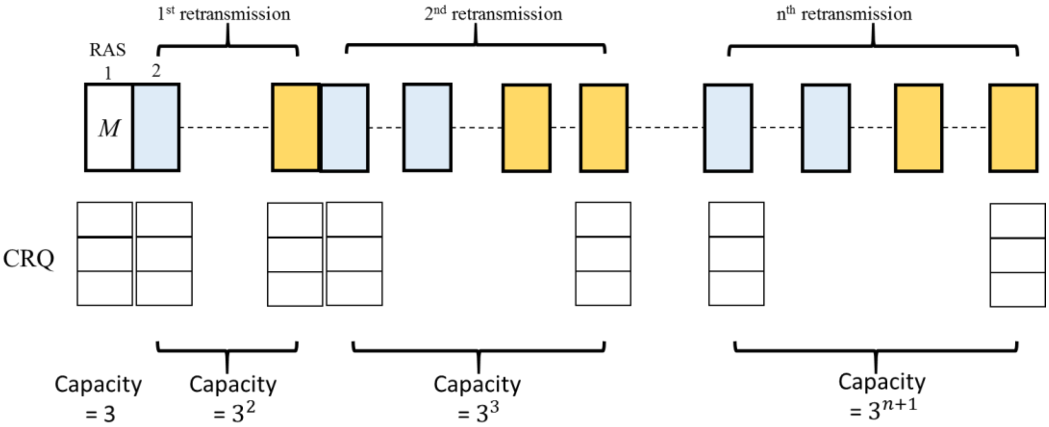


Figure 6. Exponential growth in capacity for each order of retransmission

VI. Developed Analytical Model

Let N be the number of channels in each RAS reserved by the eNB and M be the number of devices, which transmit their first attempts in the 1st RAS using DQRA. The status of the transmission in each RAS is the same as placing M balls (representing devices attempting to access channel) into N bins (representing preambles available for random access). For the bins with two or more balls, the balls in the same bin will be put into another N bins again. The procedure repeats until there is no bin containing two or more balls.

We develop an analytical model for Average Maximum Access Delay and Average number of transmissions. Figure 7 illustrates the difference between both performances metrics. The content of each rectangle is the number of accessing devices in each RAS. The average maximum access delay is the average value of the number of RASs required by all of the M devices to successfully access radio channel. The average number of transmissions is the average value of the number of transmissions required by one device to successfully access radio channel.

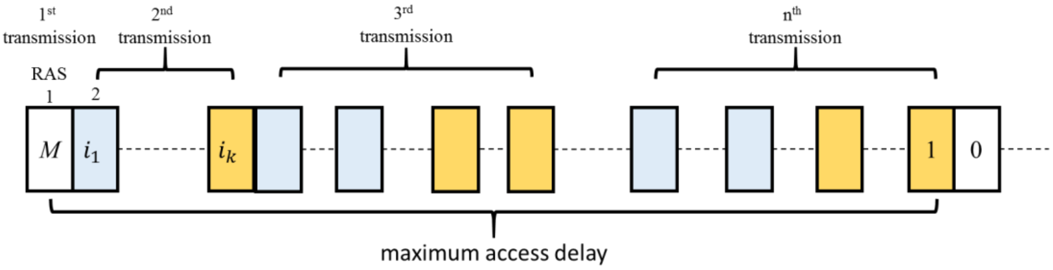


Figure 7. Illustration of Average Maximum Access Delay and number of transmissions

The analytical models for all two metrics are described in following subsections,

A. Average Maximum Access Delay

Let $\bar{K}(M,N)$ be the average value of the maximum number of RASs required by M devices to successfully transmit their attempts for accessing one of N channels. $\bar{K}(M,N)$

is equal to the average number of trials for placing balls into bins. $\bar{K}(M,N)$ can be determined using a combinatory and is given by,

$$\begin{aligned}
&\bar{K}(M,N)\\
&=\sum_{k=0}^{N}\sum_{i_1,\cdots,i_k}\Pr(k\text{ bins fail; each failed bin has }i_1,\ldots,i_k\text{ balls, respectively}) \qquad (1)\\
&\quad\times(1+\sum_{j=1}^{k}\bar{K}(i_j,N))
\end{aligned}$$

where the number "1" represents the first trial of placing balls (i.e., the first RAS) and $\sum_{j=1}^{k}\bar{K}(i_j,N)$ represents the number of trials of balls' placing required by collided balls in each failed bin. We derive the equation with each conditional probability (i.e., probability of each combination of $i_1,\ldots,i_k$) multiplied by the number of trials of balls' placing required by collided balls in each failed bin (represented by the term $1+\sum\bar{K}(i_j,N)$).

Then lets split (1) into a part corresponding to the condition when all balls succeed, which means there is no failed bin (i.e., k = 0), and a part representing collision conditions for situations when k bins fail ($k \geq 0$) and each bin has i_1, …, i_k balls, respectively. Then, $\bar{K}(M,N)$ is given by,

$$\begin{aligned}
&\bar{K}(M,N)=\Pr(0\text{ bin fails; }k=0)\times 1+\\
&\sum_{k=1}^{\min\{\lfloor\frac{M}{2}\rfloor,N\}}\ \sum_{i_1=2}^{M-2(k-1)}\ \sum_{i_2=2}^{M-2(k-2)-i_1}\cdots\sum_{i_k=2}^{M-\sum_{j=1}^{k-1}i_j} p(k;i_1,\ldots,i_k)\times(1+\sum_{j=1}^{k}\bar{K}(i_j,N)) \qquad (2)
\end{aligned}$$

where $p(k;i_1,\ldots,i_k)$ is the probability that k bins fail and each bin has i_1, …, i_k balls, respectively, i.e., bin 1 has i_1 balls, bin 2 has i_2 balls and so on. The remaining $(M-\sum_{j=1}^{k}i_j)$ balls (representing successful devices) are placed in the remaining (N-k) bins. The number of failed bins is equal to a minimum of the number of failed bins when all failed bins have only two balls collided (i.e., $\left\lfloor\frac{M}{2}\right\rfloor$) and the number of total bins (i.e., *N*). The limit of the number of failed bins is reflected in (2) by summation $\sum_{k=1}^{\min\{\left\lfloor\frac{M}{2}\right\rfloor,N\}}$. The number of collided balls in each failed bin is not less than two (i.e., $i_j \geq 2$). The limit of number of balls in each failed bin is reflected in (2) by $\sum_{i_1=2}^{M-2(k-1)} \sum_{i_2=2}^{M-2(k-2)-i_1} \ldots \sum_{i_k=2}^{M-\sum_{j=1}^{k-1} i_j}$. For each combination of $i_1, \ldots, i_k$., the number of balls in each bin is specified in the order of the bin number, i.e., i_1 is specified first, and i_k is specified the last. The limit of the number of balls in each bin is equal to the total number of balls subtracted by the minimum number of balls in the bins where the number of balls is not specified (take i_2 for example, the number of not specified bins is *k*-2, noted that the minimum number of balls in one failed bin is two), and then subtracted by the number of balls in the bins where the number of balls is already specified (take i_2 for example, the bin 1 is already specified, so the number of balls in the bin 1, i.e., i_1 is subtracted).

We then calculate the conditional probability with math operations and the completed analytical model of $\bar{K}(M,N)$ is given by,

$$\bar{K}(M,N) = \frac{C_M^N M!}{N^M} + \sum_{k=1}^{\min\{\left\lfloor\frac{M}{2}\right\rfloor,N\}} \sum_{i_1=2}^{M-2(k-1)} \sum_{i_2=2}^{M-2(k-2)-i_1} \ldots \sum_{i_k=2}^{M-\sum_{j=1}^{k-1} i_j} \frac{C_k^N C_{i_1}^M \cdots C_{i_k}^{M-\sum_{j=1}^{k-1} i_j} C_{M-\sum_{j=1}^{k} i_j}^{N-k} \times (M - \sum_{j=1}^{k} i_j)}{N^M} \times (1 + \sum_{j=1}^{k} \bar{K}(i_j, N)) \tag{3}$$

where C_k^n is the number of *k*-combinations from a given set of *n* elements ($C_k^n = 0$, if *n* < k). The probability that all balls succeed (*k*=0) is calculated by $C_k^N M!$ (the number of permutations that *M* balls are placed into *N* bins and each bin has exactly one ball) divided by N^M (the total permutations that *M* balls are placed into *N* bins). $p(k;i_1,\dots,i_k)$ is calculated as the multiplication of C_k^N (the number of combinations that *k* bins are chosen from *N* bins), $C_{i_1}^M \cdots C_{i_k}^{M-\sum_{j=1}^{k-1} i_j}$ (the number of combinations that the balls in each failed bin are chosen from *M* balls) and $C_{M-\sum_{j=1}^{k} i_j}^{N-k}\left(M-\sum_{j=1}^{k} i_j\right)!$ (the number of permutations that the successful balls, i.e., $M-\sum_{j=1}^{k} i_j$, are placed into the remaining bins, i.e., *N*-*k*. The multiplication is divided by N^M.

The composition of RASs required by balls in each failed bin is illustrated in Figure 8. The number of RASs required by each failed bin contains the first RAS where all of the *M* devices transmit their first attempt and the RASs required by the group of balls in each failed bin for retransmission. The RASs required by each group of balls can be seen as a situation when each group of balls is placed in the first RAS using another (parallel) DQRA that is showed by the colored part of Figure 8. For smaller values of *M* with a given $\bar{K}$, in can be enumerated recursively.

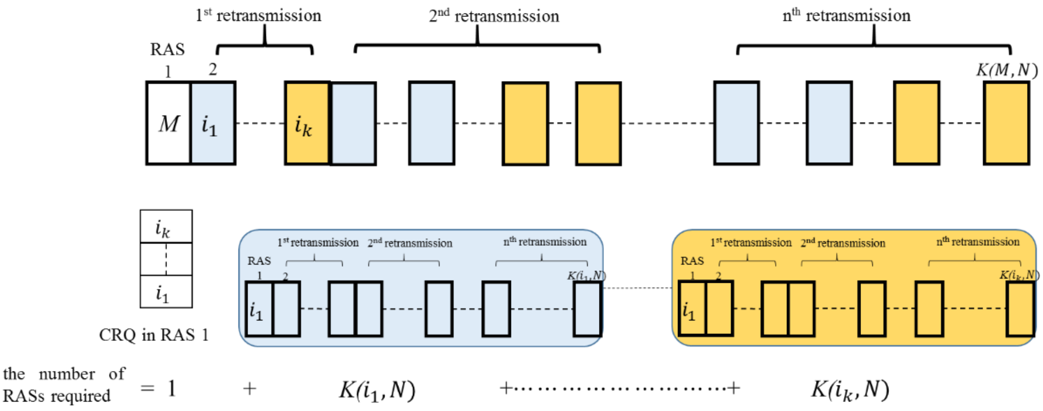


Figure 8. Composition of the number of RASs required in One-Shot DQRA

However, the computational complexity of (3) is high for large *M* since the number of possible combinations quickly increases. To reduce the computational complexity, we propose a new analytical model with low computational complexity, which does not calculate each combination.

We first consider the average maximum access delay $\bar{K}(\mathrm{M},\mathrm{N})$ with low complexity model. Figure 9 illustrates an example showing how the low complexity model reflects original model defined in (3) but with lower computational complexity. We replaced the conditional probability adopted in (3) to simplify the computation, the new adopted conditional probability considers the condition where specific numbers of bins (represented by *s* bins) collided with the same number of balls (represented by *i* balls). The simplification by new adopted probability is given by,

$$\bar{K}(M,N) = \sum_{i=2}^{M}\sum_{s=0}^{N}\sum_{k=0}^{N}\sum_{\substack{i_1,\cdots,i_k \\ s\text{ bins have } i\text{ balls}}} \Pr(k\text{ bins fail; each failed bin has } i_1,\ldots,i_k\text{ balls, respectively}) \times (1+s\times\bar{K}(i,N)) \tag{4}$$

We simplify the probability in (1) by considering some bins (represented by *s*) have exactly specific number of balls (represented by *i*) among failed bins (represented by *k*). The simplification is reflected by the sum $\sum_{k=0}^{N}\sum_{\substack{i_1,\cdots,i_k \\ s\text{ bins have } i\text{ balls}}}$. Because we only consider some of the failed bins with same number of balls, the number of trials of balls' placing required by considered failed bin is reflected by $s\times\bar{K}(i,N)$. Note that the limit of the number of the failed bins is *N* and the limit of the number of collided balls is *M*. We sum up the simplified part and derive the new adopted probability. The low complexity model adopting the new adopted probability is given by,

$$\bar{K}(M,N) =$$
$$\sum_{i=2}^{M}\sum_{s=0}^{N}\Pr(\ s \text{ bins among failed bins collided with exactly } i \text{ balls}) \times (1+s\times\bar{K}(i,N))$$

(5)

The simplification part is reflected by

$$\sum_{k=0}^{N}\sum_{\substack{i_1,\cdots,i_k \\ s \text{ bins have } i \text{ balls}}}\Pr(k \text{ bins fail; each failed bin has } i_1,\ldots,i_k \text{ balls, respectively})$$
$$= \Pr(\ s \text{ bins among failed bins collided with exactly } i \text{ balls})$$

The accuracy of (5) will be shown in the section with results.

The simplification towards low complexity is depicted in Figure 9. The upper part of the figure represents all the combinations when collision happens (i.e., one of the bins has two balls or more). The bin number is marked above the bin, and the number of balls in each bin is marked inside the bin. *p(s;i)* is the probability that *k* bins have exactly *i* balls (e.g., *p*(1;2) represents the probability that one bin has exactly two balls) and the colored part represents the combinations corresponding to the conditional probability calculated in the new proposed model. In Figure 8, we already know that (3) calculates the probability of each combination of i_1, …, i_k one by one. Figure 9 shows the calculation concept of the low complexity model by the example where we assume 5 balls and 3 available bins in the beginning. Equation (3) calculates the probability of each combination when collision happens as illustrated in the upper part of Figure 9, and, thus, the probability calculation is done 21 times (see the number of combinations in Figure 9). The lower part of the figure illustrates the conditional probability calculated by the new proposed model. According to the figure, the new proposed model only calculates the probability 5 times because each conditional probability (i.e., $p_l(k;i)$) includes more than one combination.

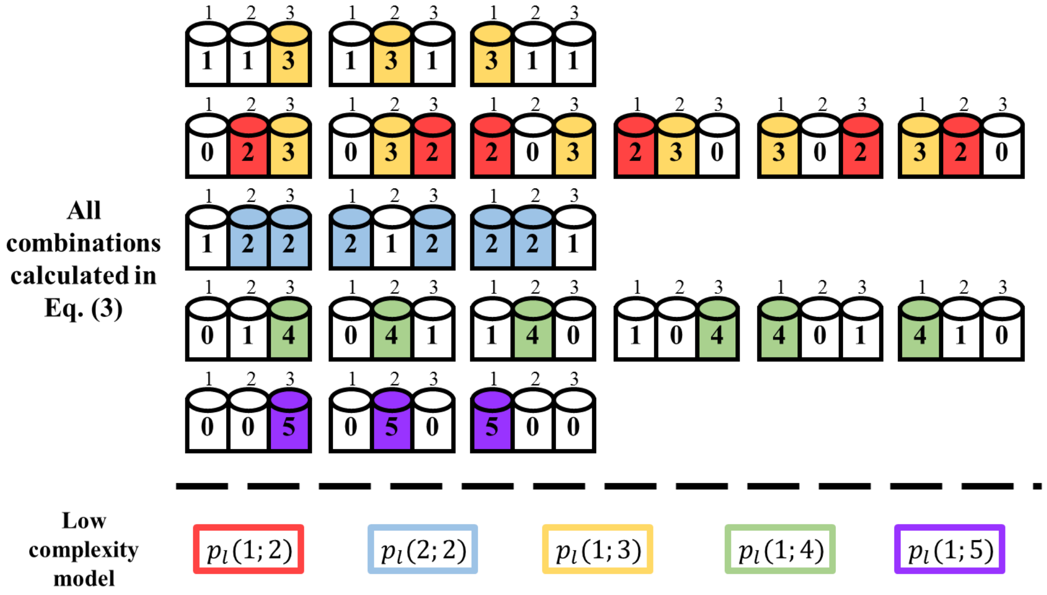


Figure 9. Calculation concept of the proposed model

Based on a part of the example shown in Figure 9, Figure 10 illustrates the difference in calculation between the proposed analytical model (3) and its extension towards low complexity model. In Figure 10, the probability of each combination (e.g., bin 1 and bin 2 have one ball inside and bin 3 has three balls inside) is calculated one by one by (3). The new low complexity model simplifies the computation by calculating the conditional probability that *k* failed bins has exactly *i* balls collided. Note that each conditional probability is multiplied by the number of RASs required by each failed group with specified number of balls (i.e., $p_l(1;2)$ represents one bin with exactly two balls, so $p_l(1;2)$ is multiplied by $\bar{K}(2,3)$; $p(0,2,3)$ specifies that there are zero, two, three balls in the bins, respectively, so $p(0,2,3)$ is multiplied by $\bar{K}(2,3)$ and $\bar{K}(3,3)$). Note that the probability $p(0,2,3)$ is not multiplied by $\bar{K}(0,3)$ as it does not corresponds to failed group. From the lower part of Figure 10, the equalization of the calculation into low complexity model is illustrated. The average number of required trials obtained by the low complexity model (i.e., $p_l(1;2) \times \bar{K}(2,3) + p_l(1;3) \times \bar{K}(3,3)$) is equal to every calculation obtained by (3) (i.e., blue colored part). Thus, the low complexity model reaches the same computation result as (3) but with lower number of calculations. This simplification solves the problem of high computational complexity for large M .

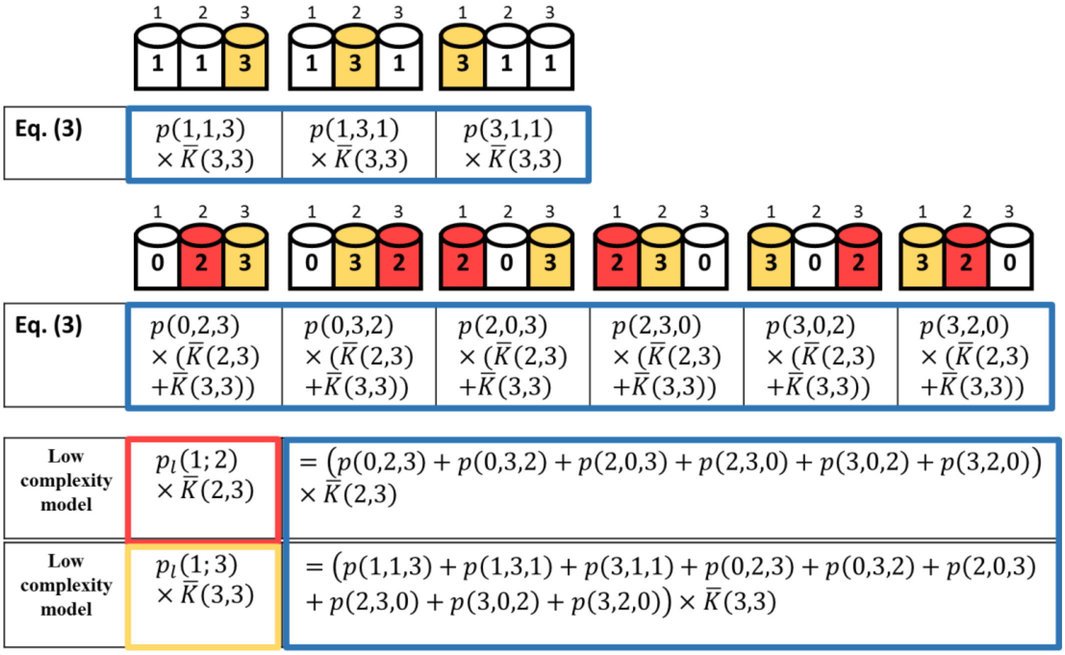

1 2 3
1 1 3
1 3 1
3 1 1
Eq. (3)
$p(1,1,3) \times \bar{K}(3,3)$
$p(1,3,1) \times \bar{K}(3,3)$
$p(3,1,1) \times \bar{K}(3,3)$
0 2 3
0 3 2
2 0 3
2 3 0
3 0 2
3 2 0
Eq. (3)
$p(0,2,3) \times (\bar{K}(2,3) + \bar{K}(3,3))$
$p(0,3,2) \times (\bar{K}(2,3) + \bar{K}(3,3))$
$p(2,0,3) \times (\bar{K}(2,3) + \bar{K}(3,3)$
$p(2,3,0) \times (\bar{K}(2,3) + \bar{K}(3,3))$
$p(3,0,2) \times (\bar{K}(2,3) + \bar{K}(3,3))$
$p(3,2,0) \times (\bar{K}(2,3) + \bar{K}(3,3))$
Low complexity model
$p_l(1;2) \times \bar{K}(2,3)$
$= \big(p(0,2,3) + p(0,3,2) + p(2,0,3) + p(2,3,0) + p(3,0,2) + p(3,2,0)\big) \times \bar{K}(2,3)$
Low complexity model
$p_l(1;3) \times \bar{K}(3,3)$
$= \big(p(1,1,3) + p(1,3,1) + p(3,1,1) + p(0,2,3) + p(0,3,2) + p(2,0,3) + p(2,3,0) + p(3,0,2) + p(3,2,0)\big) \times \bar{K}(3,3)$

Figure 10. Comparison between the original model and the proposed model

Based on above-shown simplification, the low complexity model of $\bar{K}(M,N)$ is given by,

$$\bar{K}(M,N) = \sum_{i=2}^{M} \sum_{s=0}^{N} \Pr(\ s \text{ bins with exactly } i \text{ balls}) \times (1 + s \times \bar{K}(i,N)) \tag{6}$$

where the number "1" represents the first trial (i.e., RAS) and $s \times \bar{K}(i,N)$ represents the average number of trials of balls' placing required by *s* bins with exactly *i* balls. We derive (6) with the probability that *s* bins collide with exactly *i* balls instead of the conditional probability (i.e., probability of each combination of $i_{1},\ldots,i_{k}$) calculated in (3). The probability is multiplied by the number of trials of balls' placing required by collided balls only in the failed bins with *i* balls collided (represented by term 1+$s \times \sum \overline{\mathrm{K}}(i, N)$). We then split (6) into two parts, the first part is the calculation for no bins collided with i balls (i.e., $s = 0$) and the second part is for case when there are one or more than one bins with *i* balls collided (i.e., $s \geq 1$). $\overline{K}(M, N)$ is given by,

$$\overline{K}(M,N) = 1 + \sum_{i=2}^{M} \sum_{s=1}^{\min\{\left\lfloor \frac{M}{i} \right\rfloor, N\}} p_l(s;i) \times s \times \overline{K}(i,N) \tag{7}$$

where the number "1" represents the sum of probabilities that no bins with *i* balls for $1 \leq i \leq M$ (i.e., $\sum_{i=2}^{M} p_l(0;i)$). $p_l(s;i)$ is the probability that *s* bins collided with exactly *i* balls ($2 \leq i$) (i.e., remaining $(M - s \times i)$ balls are randomly placed in the remaining (*N*-*s)* bins) The number of collided balls is from two to M (i.e., $\sum_{i=2}^{M}$) because there are at least two balls choosing one same preamble for collision. The maximum number of bins (i.e., *s*), which have exactly *i* balls is equal to a minimum of the number of failed bins when all failed bins have *i* balls collided (i.e., $\left\lfloor \frac{M}{i} \right\rfloor$) and the number of total bins (i.e., *N*). We then further simplify (7) and $\overline{K}(M, N)$ is given by,

$$\overline{K}(M,N) = 1 + \sum_{i=2}^{M} p_{sum}(i) \times \overline{K}(i,N) \tag{8}$$

where $p_{sum}(i)$ is the sum of probabilities $s \times p(s;i)$ for $1 \leq s \leq \min\{\left\lfloor \frac{M}{i} \right\rfloor, N\}$, i.e., $p_{sum}(i) = \sum s \times p(s; i)$. The enumeration of $p_{sum}(i)$ is shown in the later description. The final form of the low complexity model of $\overline{K}(M, N)$ is given by,

$$\bar{K}(M,N)=1+\sum_{i=2}^{M}\frac{C_1^N C_i^M (N-1)^{M-i}}{N^M}\bar{K}(i,N) \tag{9}$$

where C_k^n is the number of *k*-combinations from a given set of *n* elements ($C_k^n = 0$, if $n <$ k).

The enumeration of $p_{sum}(i)$ is described as follow,

$$\begin{aligned}
&p_{sum}(i)\\
&= \sum_{s=1}^{\min\{\left\lfloor \frac{M}{i} \right\rfloor,N\}} s \times p(s;i)\\
&= \sum_{s=1}^{\min\{\left\lfloor \frac{M}{i} \right\rfloor,N\}} p_{AL}(s;i)\\
&= \sum_{s=1}^{\min\{\left\lfloor \frac{M}{i} \right\rfloor,N\}} \frac{1}{N^M}\{C_s^N \sum_{p=0}^{s-1} C_i^{M-pi}(N-s)^{M-si} - \sum_{q=k+1}^{\min\{\left\lfloor \frac{M}{i} \right\rfloor,N\}} q \times C_q^N \sum_{r=0}^{q-1} C_i^{M-ri}(N-j)^{M-ji}\}\\
&= \frac{C_1^N C_i^M (N-1)^{M-i}}{N^M}
\end{aligned} \tag{10}$$

where $p_{AL}(s;i)$ is the probability that at least *s* bins have *i* balls collided, i.e., $p_{AL}(s;i) = \sum_{k=s}^{\min\{\left\lfloor \frac{M}{i} \right\rfloor,N\}} p(k;i)$. The $p_{AL}(s;i)$ is composed of two parts. The first part of $p_{AL}(s;i)$ is the conditional probability that *s* bins with *i* balls are chosen and remaining balls are randomly placed into the remaining bins. The first part of $p_{AL}(s;i)$ expresses the amount where redundant conditions are included as explained in Figure 11. Figure 11 shows the redundant conditions when 1 bin has 2 balls collided. The redundant conditions represent the repeated computation of the same condition. The second part, which is subtracted from the first part of $p_{AL}(s;i)$, is the sum of probabilities of all redundant conditions. The summation of $p_{AL}(s;i)$ is equal to the sum of probabilities $s \times p(s;i)$ for $1 \le s \le \min\{\left\lfloor \frac{M}{i} \right\rfloor, N\}$, i.e., $\sum p_{AL}(s;i) = \sum s \times p(s;i)$.

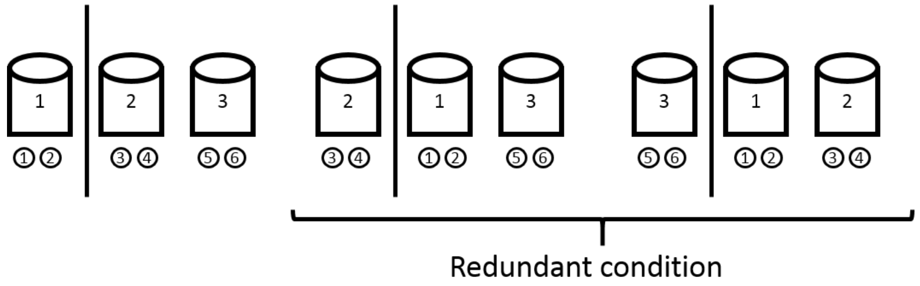


Figure 11. Example of over counted condition for derivation of $p_{AL}(s;i)$ (M=6, N=3, s=1 ,i=2)

B. Average number of transmissions

Now, let $\bar{T}(M,N)$ be the average number of transmissions required by one device to successfully attempt channel access. $\bar{T}(M,N)$ is equal to the average number of trials for one ball to be placed into bins before the ball succeed (i.e., the number of times when there is a preamble selected with two or more devices plus the first shot when all devices attempt to access channel simultaneously). $\bar{T}(M,N)$ can be determined using a combinatory and is given by,

$$\begin{aligned}&\bar{T}(M,N)\\&=\sum_{k=0}^{N}\sum_{i_1,\cdots,i_k}\Pr(k\text{ bins fail; each failed bin has }i_1,\ldots,i_k\text{ balls, respectively})\\&\quad\times(1+\sum_{j=1}^{k}\bar{T}(i_j,N)\times\frac{i_j}{M})\end{aligned} \tag{11}$$

where the number "1" represents the first trial (i.e., the first transmission) and $\sum_{j=1}^{k}\bar{T}(i_j,N)$ represents the average number of trials of placing one ball required by each failed bin. For the calculation of the average number, the number of trials for each failed bin (i.e., $\bar{T}(i_j,N)$) is multiplied by the ratio (i.e., $\frac{i_j}{M}$) where the numerator is the number of balls in each failed bin and the denominator is the number of total devices. We derive the equation with each conditional probability (i.e., probability of each combination of $i_1, ..., i_k$) multiplied by the number of trials of one ball's placing required by collided balls in each failed bin (represented by term $1+\sum \overline{T}(i_j, N) \times \frac{i_j}{M}$). We split (11) into two parts in the same way as (2), the first part is the success condition and the second part is for the collision conditions. Then, $\overline{T}(i_j, N)$ is given by,

$$\overline{T}(M,N) = \Pr(0 \text{ bin fails}; k=0) \times 1 + \sum_{k=1}^{\min\{\lfloor \frac{M}{2} \rfloor, N\}} \sum_{i_1=2}^{M-2(k-1)} \sum_{i_2=2}^{M-2(k-2)-i_1} \cdots \sum_{i_k=2}^{M-\sum_{j=1}^{k-1} i_j} p(k; i_1, ..., i_k) \times (1+\sum_{j=1}^{k} \overline{T}(i_j, N) \times \frac{i_j}{M}) \tag{12}$$

The limit of number of failed bins and the limit of number of balls in each failed bins are represented in the same way as (3). We also represent $\overline{T}(M,N)$ with the same math operations of success probability (k=0) and $p(k; i_1, ..., i_k)$. The completed analytical model of $\overline{T}(M,N)$ is given by,

$$\overline{T}(M,N) = \frac{C_M^N M!}{N^M} + \sum_{k=1}^{\min\{\lfloor \frac{M}{2} \rfloor, N\}} \sum_{i_1=2}^{M-2(k-1)} \sum_{i_2=2}^{M-2(k-2)-i_1} \cdots \sum_{i_k=2}^{M-\sum_{j=1}^{k-1} i_j} \frac{C_k^N C_{i_1}^M \cdots C_{i_k}^{M-\sum_{j=1}^{k-1} i_j} C_{M-\sum_{j=1}^{k} i_j}^{N-k} \times (M-\sum_{j=1}^{k} i_j)}{N^M} \times (1+\sum_{j=1}^{k} \overline{T}(i_j, N) \times \frac{i_j}{M}) \tag{13}$$

Now, let's consider the average number transmissions $\overline{T}(M,N)$ with low complexity model. $\overline{T}(M,N)$ can be determined using a combinatory and is given by,

$$\begin{aligned}&\bar{T}(M,N)\\&=\sum_{i=2}^{M}\sum_{s=0}^{N}\Pr(k\text{ preambles selected by }i\text{ devices})\times(1+s\times\bar{T}(i,N)\times\frac{i}{M})\end{aligned}\qquad(14)$$

where the number "1" represents the first transmission and $s\times\bar{T}(i,N)\times\frac{i}{M}$ represents the average number of transmissions required by one device in the groups with exactly *i* devices collided. We derive (14) with the probability that *s* preambles collided with exactly *i* balls and the probability is multiplied by average number of transmissions required by one device in the groups with exactly *i* devices collided. Note that for the calculation of the average number for one device, the number of transmissions for each failed preamble (i.e., $\bar{T}(i,N)$) is multiplied by the ratio $\frac{i}{M}$ where the numerator is the number of balls in each failed bin and the denominator is the number of total devices. We then split (14) into two parts in the same way as it is done for (7). Then, $\bar{T}(M,N)$ is given by,

$$\bar{T}(M,N)=1+\sum_{i=2}^{M}\sum_{s=1}^{\min\{\lfloor\frac{M}{i}\rfloor,N\}}p_l(s;i)\times s\times\bar{T}(i,N)\times\frac{i}{M}\qquad(15)$$

where $p_l(s;i)$ is the probability that *s* failed preambles are selected by exactly *i* devices, Noted that the remaining successful $(M-s\times i)$ devices randomly select the remaining $(N-s)$ preambles. We further simplify (15) in the same way as (8) so the $\bar{T}(M,N)$ is given by,

$$\bar{T}(M,N)=1+\sum_{i=2}^{M}p_{sum}(i)\times\bar{T}(i,N)\times\frac{i}{M}\qquad(16)$$

where $p_{sum}(i)$ is the sum of probabilities $s\times p(s;i)$ for $1\le s\le\min\{\lfloor\frac{M}{i}\rfloor,N\}$, i.e.,

$p_{sum}(i) = \sum s \times p(s;i)$. The final low complexity model of $\bar{T}(M,N)$ is given by,

$$\bar{T}(M,N) = 1 + \sum_{i=2}^{M} \frac{C_1^N C_i^M (N-1)^{M-i}}{N^M} \bar{T}(i,N) \times \frac{i}{M} \tag{17}$$

VII. NUMERICAL RESULTS

In this section, the performance metrics are introduced, simulations set-up and system parameters are described and then, the numerical results of analytical analysis are presented and compared with simulation results.

A. Performance Metrics

The average maximum access delay presented in RASs and seconds and average number of transmissions are chosen as the performance metrics in this thesis to evaluate the performance of group paging using DQRP.

The average maximum access delay presented in RASs, labelled as $\bar{K}(M,N)$, is defined as the average value of the maximum number of RASs required by M devices to successfully transmit their attempts for accessing one of N channels. $\bar{K}(M,N)$, is obtained based on (9).

Then, the average maximum access delay defined as the time (in seconds) elapsed between the first attempt and the Msg. 4 (see Figure 1) reception of the last successfully accessed device is denoted as $\overline{D_{max}}$. The average maximum access delay presented in seconds ($\overline{D_{max}}$) is calculated in a similar way as $\bar{K}(M,N)$, i.e., according to (9). However, as the $\overline{D_{max}}$ is in seconds, it is multiplied by the interval between two successive RASs, T_{RA_REP}, plus by the time between sending the Msg1 and receiving the Msg4 (see Figure 1). Note that the duration of one sub-frame is 1ms. The time between the end of Msg1 and the beginning of the RAR window, T_{RAR}, is 2ms and the RAR message arrives at the beginning of the RAR window [11]. The time between the Msg2 and the Msg3, T_{RRC}, is 5ms. This value follows the minimum suggested time as defined in [14]. The *macContentionResolutionTimer* indicating maximum duration for receiving the Msg4 is set to 15 sub-frames [11], that is, the Msg4 arrives randomly within range of the 1st and 15th sub-frames after sending the Msg3. Then, the average maximum access delay presented in seconds is calculated as follow,

$$\begin{aligned}&\bar{D}_{\max}(M,N)=\\&\bar{K}(M,N)\times T_{RA\_REP}+T_{RAR}+T_{RRC}+rand[1, macContention\mathrm{Re}\,solutionTimer]\end{aligned}\tag{18}$$

The average number of transmissions for successfully accessed devices, $\bar{T}(M,N)$, is defined by (17) and it expresses the number of transmissions needed for one device to successfully access radio channel using a DQ-based RA procedure. The average number of transmissions is calculated based on (17).

B. Simulations set-up and system parameters

Computer simulations are conducted on top of a C-based simulator to verify the effectiveness of the proposed analytical model. The results are averaged out over 1000 simulation drops. Each sample is obtained by performing One-Shot DQRA with *M* devices and *N* available preambles.

In the simulations, *M* MTC devices are assumed to initiate all random-access attempts and *N* preambles are reserved per RAS. Markers and lines in Figure 13, Figure 14 and Figure 15 are used to present simulation and analytical results, respectively. The analytical results of $\bar{K}$, $\overline{D_{max}}$, $\bar{T}$ are obtained based on (9), (18) and (17), respectively. We duplicate the same simulation environment as in [11]. The setting of the parameters used in the simulation is summarized in Table 1. We consider four different numbers of preambles, i.e., the eNB reserves 6, 18, 36, 56 preambles in each RAS (*N* = 6, 18, 36, 56) to page a group of a size from 10 to 1500 MTC devices (*M* = 10 to 1500).

Table 1. Random-access related system parameters.

Notations	Definitions	Settings in simulation
M	Average number of MTC devices in a cell	10~1500
T_{RA_REP}	Interval between two successive random-access slots (unit: sub-frame)	T_{RA_REP} =10 (PRACH Configuration Index = 3)
N	Total number of preambles in a RAS	6,18,36,56
T_{RAR}	Time between Msg2 and Msg3 (unit: sub-frame) [11]	2
T_{RRC}	Time between Msg3 and Msg 4 (unit: sub-frame) [11]	5
macContention ResolutionTimer	maximum duration for receiving Msg4 (unit: sub-frame) [11]	15

C. *Analytical and simulation results*

The comparison between the high complexity model, low complexity model and the simulation is shown in Figure 12. The analytical and simulation results for $\overline{K},\ \overline{D_{max}}$ and $\overline{T}$ are shown in Figure 13, Figure 14, and Figure 15, respectively.

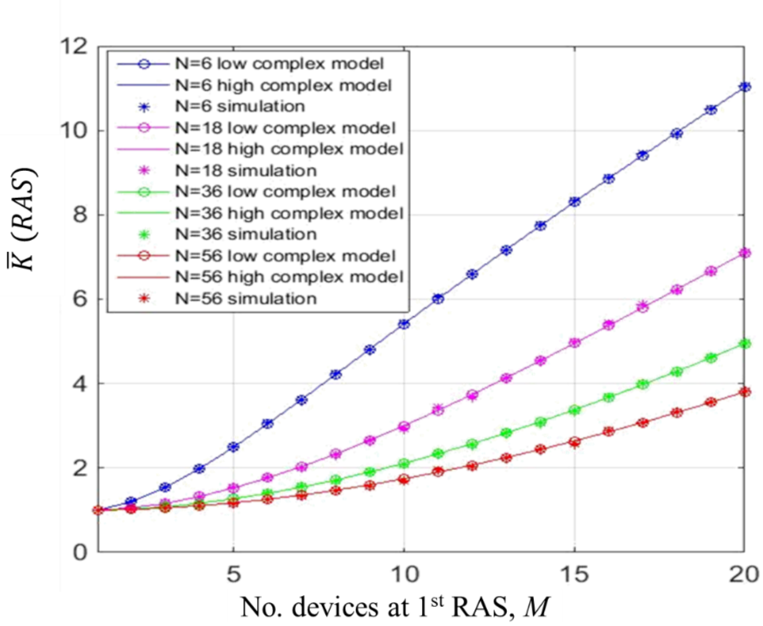


Figure 12. Comparison between high complexity model and low complexity model.

Figure **12** shows that the proposed low complexity model based on (9) perfectly match the result of high complexity model based on (3).

In Figure 13 and Figure 14, the analytical analysis of average maximum access delay presented in RASs and seconds, respectively, is presented and compared with simulation results. The simulation results are all coincided with the analytical results. We can see the similar trends in both figures due to the fact that $\overline{D}$ is based on $\overline{K}$ and there is linear relation between these two performance metrics. In both figures, the average maximum access delay increases with the decreasing of the number of available preambles because the collision probability increases as the number of preambles decrease. For N equals to 18 and 36, the average maximum access delay converges to each other around M equals to 1700. Then, the average maximum access delay reverses. It is because, in DQ mechanism, devices retransmit in the order of their positions in the CRQ, the length of CRQ affects the delay time of devices. Thus, in DQ mechanism, when the number of preambles increases, the chance of successful access increase, but the waiting time for retransmission may also increase because of the longer CRQ.

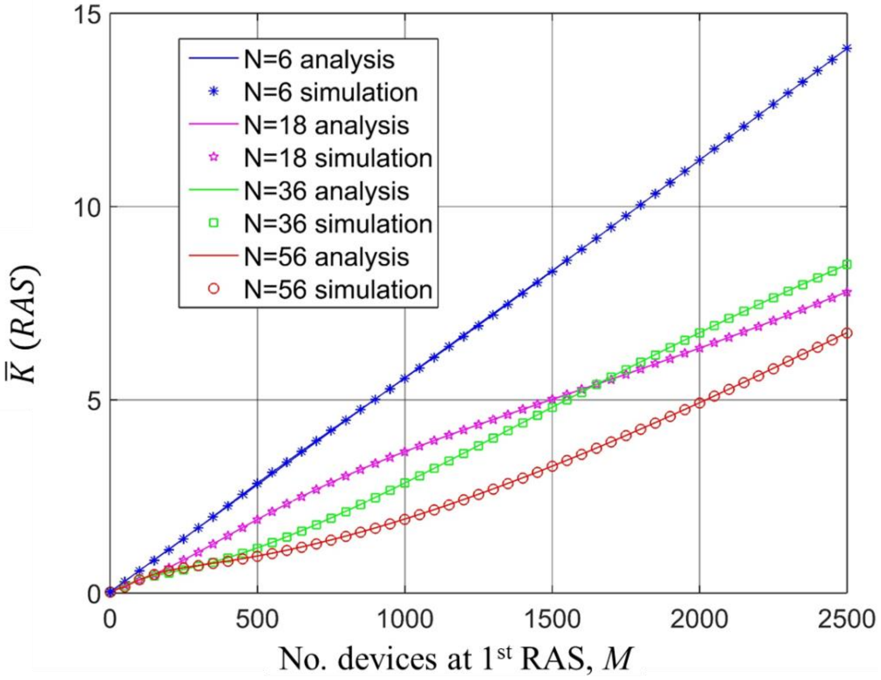


Figure 13. Average maximum access delay (presented in RASs)

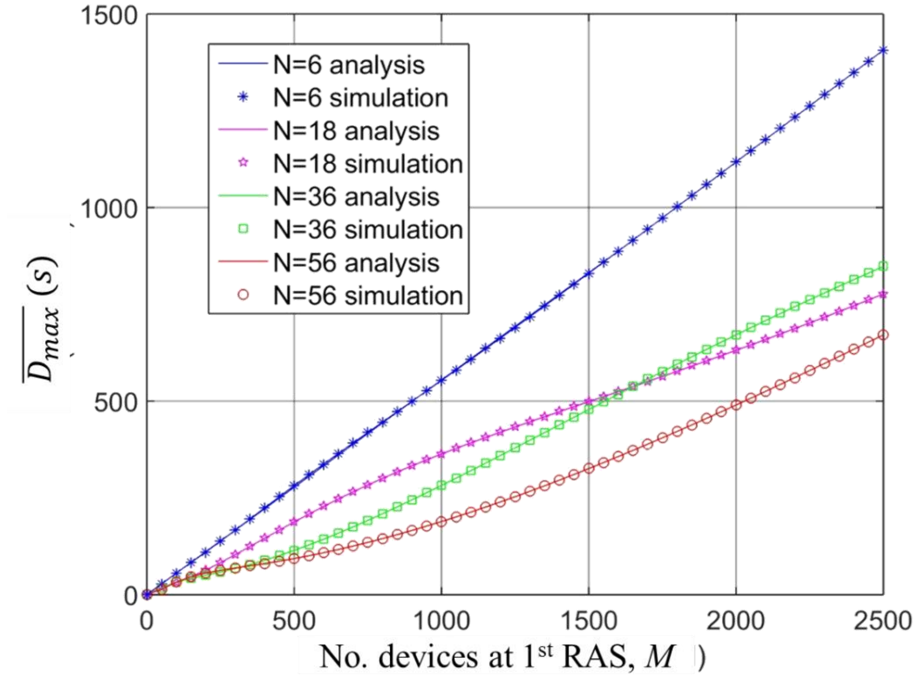


Figure 14. Average maximum access delay (presented in seconds)

In Figure 15, the average number of transmissions, $\bar{T}$, is depicted. It can be seen that $\bar{T}$ is larger when the number of preambles is smaller because of a higher collision probability. $\bar{T}$ increases rapidly when the number of paged devices is small. Then, the increase becomes slow when *M* becomes large. Figure 15 shows that DQRA keeps the number of transmissions low even for high number of devices because DQRA splits devices into groups for the subsequent retransmissions in an exponential manner.

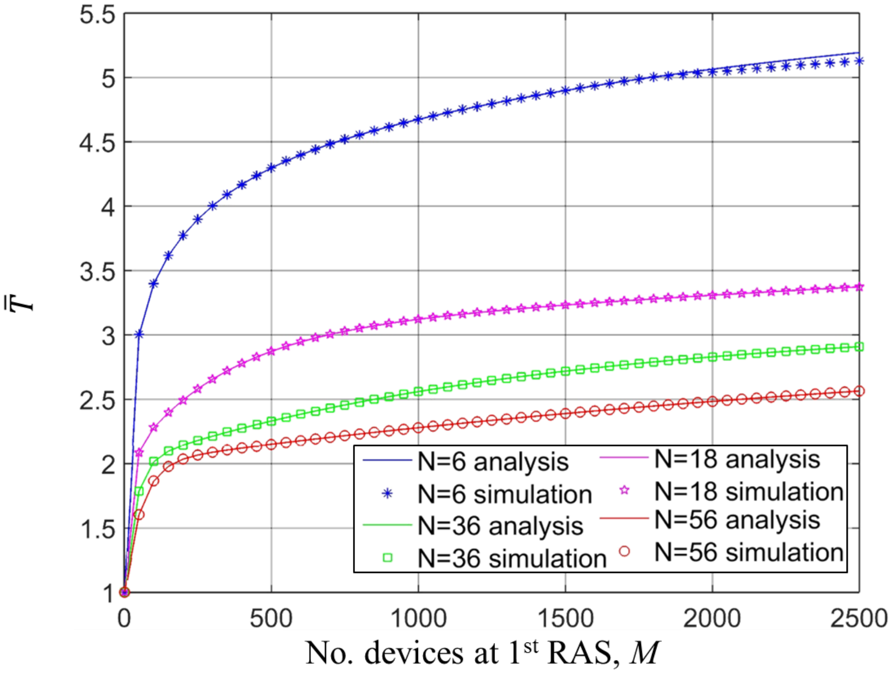


Figure 15. Average number of transmissions

VIII. CONCLUSION

In the thesis, we present the overview of RA procedure in LTE in order to explain that the current standard RA procedure can only deal with massive number of MTC devices by increasing the backoff parameters for retransmitting the preamble. However, this results in increased access delay. We also describe a promising extension of the conventional random access towards the DQ mechanism and the DQRAP to understand how the DQ mechanism is implemented into a common RA procedure and how the DQRAP organizes the devices and split them into virtual queues to reduce the collision probability. We assume the one-shot scenario, which means that all devices are triggered simultaneously to access the network by a group paging signal.

The major contribution in thesis is a development of analytical models for estimation of DQRAP performance. The analytical models are developed for estimation of the Average Maximum Access Delay and Average number of transmissions. The numerical results derived based on the developed analytical models demonstrate that the model accurately matches the simulation result and confirms the behavior of the DQRAP in a sense that a low number of transmissions as expected even for a high number of simultaneously accessing MTC devices. The number of transmissions is getting saturated with a high number of MTC devices as the colliding devices are distributed to parallel queues.

As the DQRAP is a potential solution to handle massive number of devices in LTE. in the future, the more performance metrics, such as an average access delay, average number of devices in queue, blocking probability, and energy consumption should be investigated and models for them should be derived.

REFERENCES

[1] 3GPP TR 37.868, “RAN improvements for machine-type communications,” v. 1.0.0, Aug. 2011.

[2] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: potential, challenges, and solutions," IEEE Communications Magazine, vol. 50, no. 3, pp. 178-184, March 2012.

[3] C. H. Wei, R. G. Cheng and S. L. Tsao, "Performance Analysis of Group Paging for Machine-Type Communications in LTE Networks," IEEE Transactions on Vehicular Technology, vol. 62, no. 7, pp. 3371-3382, Sept. 2013.

[4] 3GPP TR 22.868 V8.0.0, “Study on Facilitating Machine to Machine Communication in 3GPP Systems,” March 2007.

[5] A. Laya, L. Alonso, P. Chatzimisios and J. Alonso-Zarate, "Massive access in the Random Access Channel of LTE for M2M communications: An energy perspective," IEEE International Conference on Communication Workshop (ICCW), pp. 1452-1457, London, 2015.

[6] M. Y. Cheng et al., “Performance evaluation of radio access network overloading from machine type communications in LTE-A networks,” IEEE Wireless Communications and Networking Conference Workshops (WCNCW), pp. 248-252, Apr. 2012.

[7] 3GPP R2-104873, “Comparing push and pull based approaches for MTC,” Institute for Information Industry (III), Coiler Corporation, RAN2#71, Aug 2010.

[8] 3GPP R2-104870, “Pull based RAN overload control,” Huawei and China Unicom, RAN2#71, Aug. 2010.

[9] 3GPP R2-104004, “Group paging for MTC devices,” LG Electronics Inc., RAN2#70bis ,July 2010.

[10] 3GPP TSG RAN WG2 #69bis R2-102296, “RACH intensity of Time Controlled

Devices," Beijing, China, April 2010.

[11] A. Laya, L. Alonso, and J. Alonso-Zarate, "Contention Resolution Queues for Massive Machine Type Communications in LTE," IEEE Personal, Indoor, and Mobile Radio Communications (PIMRC) Workshop on Machine-to-Machine Communications, pp. 2314-2318, Hong Kong, 2015,.

[12] 3GPP TS 36.321 "Evolved universal terrestrial radio access (E-UTRA): Medium access control (MAC) protocol specification," Third-Gen. Partnership Proj, Sophia-Antipolis Cedex, France, ver. 9.3.0, Jun. 2010.

[13] C. H. Wei, R. G. Cheng, and S. L. Tsao, "Modeling and estimation of one-shot random access for finite-user multichannel slotted ALOHA systems," IEEE Communication. Letters, vol. 16, no. 8, pp. 1196–1199, Aug. 2012.

[14] Sesia, S. and Baker, M. and Toufik, I., LTE - The UMTS Long Term Evolution: From Theory to Practice. Wiley, 2011, section 19.3.1.3.

[15] IEEE 802.16m-08/413, "Synchronous Non-adaptive HARQ in IEEE 802.16m Uplink".

[16] L., G. Roberts, "Aloha packet system with and without slots and capture", ACM SIGCOMM Computer Communication Review, vol. 5 no. 2, April 1975.

[17] 3GPP R1-060584, "E-UTRA random access," Ericsson, RAN1#44, Feb 2006.

[18] 3GPP R2-100182, "Access control of MTC devices," CATT, RAN2#68bis, Valencia, Spain, Jan. 2010.

[19] J. P. Cheng, C. H. Lee and T. M. Lin, "Prioritized random access with dynamic access barring for RAN Overload in 3GPP LTE-A networks," *IEEE GLOBECOM*, pp. 368-372, 2011.

[20] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," IEEE Communications Surveys Tutorials, vol. 16, no. 1, pp. 4–16, First Quarter 2014

[21] 3GPP R2-102780, "Discussion on RACH overload for MTC," CATT, RAN2#70, May.

2010.

[22] 3GPP TSG RAN WG2 #71 R2-104662, “MTC simulation results with specific solutions,” ZTE, Madrid, Spain, Aug. 2010.

[23] X. Yang, A. Fapojuwo, and E. Egbogah, “Performance Analysis and Parameter Optimization of Random Access Backoff Algorithm in LTE,” IEEE Vehicular Technology Conference (VTC Fall), pp. 1–5, 2012.

[24] W. Xu and G. Campbell, “A Near Perfect Stable Random Access Protocol for a Broadcast Channel,” IEEE ICC, 1992.