Modeling of Distributed Queueing-Based Random Access for Machine Type Communications in Mobile Networks

Ray-Guang Cheng, *Senior Member, IEEE*, Zdenek Becvar, *Member, IEEE*, and Ping-Hsun Yang

***Abstract*— Machine type communications (MTC) devices stay in idle mode to save energy and should perform random access (RA) procedure to obtain radio resources for data transmission. The RA procedure introduces access delay and extra power consumption for the MTC devices. Thus, RA needs to be optimized. In this letter, we develop low complexity analytical models to rapidly estimate maximum access delay and average number of preamble transmissions for distributed queueing-based random access (DQRA) protocol, which improves the performance of standard RA for MTC in LTE. The proposed model can be used to analyze the performance of group paging using DQRA. The performance analysis shows that the proposed analytical models accurately match the simulation results.**

***Index Terms*— Random access, machine type communications, analytical model, distributed queuing.**

I. Introduction

MACHINE type communications (MTC) is a new way enabling communication between electronic devices and machines through cellular networks. A key characteristic of MTC is that it involves massive amount of simultaneous attempts of devices to access radio resources. In LTE-based cellular networks, a slotted ALOHA protocol is adopted as the access control protocol. However, this protocol may lead to a high collision probability and a huge access delay due to serious congestion in random access channels (RACHs) [1]. There are many schemes developed to solve the collision problem in the ALOHA-based system (see, e.g., [2], [3]). Another drawback of ALOHA, reported in [4], is a low throughput. The throughput limitations can be overcome by a Distributed Queueing (DQ) protocol [5]. The DQ protocol demonstrates a stability of its performance and near optimum behavior in terms of throughput and access delay. Such features makes the DQ a suitable protocol for the RA of massive MTC devices in future mobile networks.

In the DQ protocol, time is divided into slots and each slot consists of two parts, access part and data part. The first part is further divided into minislots representing access opportunities. The DQ protocol is based on a tree splitting algorithm with simple rules to organize each device into two virtual queues: i) contention resolution queue (CRQ) for solving previous collisions and ii) data transmission queue for the data transmission.

There are two implementations of the DQ protocol into LTE RA structure. One is denoted as Distributed Queuing Access Protocol for LTE (DQAL) [6] and the other one is known as Distributed Queueing-based Random Access [7]. Both implementations apply the CRQ into LTE RA and the minislots in DQ protocol are virtually mapped to the preambles in LTE. The devices, which select the same preamble form a collided group and transmit a new preamble in later random access slots (RASs). The order of the collided groups in the CRQ is based on the selected preamble (lower preamble number served first). In DQAL, several collided groups can retransmit different preambles in one RAS, while in DQRA only one collided group retransmits the attempt in each RAS.

Simulation results in [7] show that the DQRA can reduce access delay and energy consumption while maintaining a low blocking probability for a high amount of devices simultaneously attempting to access radio resources. Comparing to DQAL, DQRA conducts a simpler mechanism and outperforms the conventional LTE RA even for low number of simultaneous arrival. Nevertheless, [7] does not provide analytical modeling, which is needed to estimate the RA delay bound for optimization of the RACH's resource allocation. The contribution of this paper consists in the analytical modeling of DQRA to estimate the maximum access delay and the average number of access attempts before devices obtain the radio resources. We also address complexity of the models and develop low complexity approaches for both metrics. The developed models match simulation results, thus, can be exploited for extensive evaluation of the impact of key parameters on the DQRA performance and its optimization.

The rest of the paper is organized as follows. Section II summarizes the principle of DQRA. Section III presents the developed analytical model for both performance metrics. Numerical results and conclusions are provided in Section IV and Section V, respectively.

II. System Model for DQRA Modelling

In this section, a system model is outlined and a principle of DQRA is described in order to provide background for the analytical modelling.

We consider a fixed number of devices performing DQRA in multichannel slotted ALOHA in LTE with the time divided into fix-length access cycles. Each access cycle is composed of ten sub-frames and contains a RAS reserved for the devices to transmit their attempts to access the radio channel. As in [8], we consider a one-shot RA scenario where all the

Manuscript received June 28, 2017; revised August 10, 2017; accepted September 13, 2017. Date of publication September 21, 2017; date of current version January 8, 2018. This work was supported in part by the Ministry of Science and Technology, Taiwan, under contract No. MOST 105-2221-E-011-033-MY3 and in part by the project No. SGS17/184/OHK3/3T/13 funded by CTU in Prague. The associate editor coordinating the review of this paper and approving it for publication was Y. Wu. *(Corresponding author: Ray-Guang Cheng.)*

R.-G. Cheng and P.-H. Yang are with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: crg@mail.ntust.edu.tw).

Z. Becvar is with the Department of Telecommunication Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, 166 27 Prague, Czech Republic (e-mail: zdenek.becvar@fel.cvut.cz).

Digital Object Identifier 10.1109/LCOMM.2017.2755020

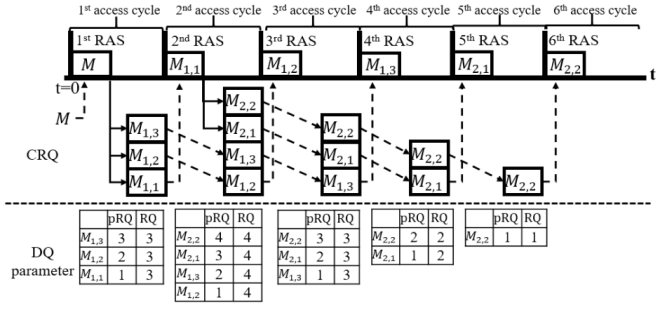

Fig. 1. Example of DQRA operation and principle of queue management.

devices receive a group paging message from the eNB and simultaneously transmit their first RA attempts in the first RAS. We also assume an infinite number of retransmissions. In each RAS, any device can indicate its willingness to obtain communication resources by transmitting a preamble randomly chosen from a predefined set of preambles.

As in [7], it is assumed that the eNB is not able to decode simultaneous transmission of the same preamble by multiple devices. Therefore, the eNB considers the preamble transmitted by multiple devices as collided preamble and devices do not receive radio resources for communication. After receiving the preamble, the eNB broadcasts a Random Access Response (RAR) containing the collision status of each preamble. The device learns the success or failure of its attempt according to the information in RAR. The successful devices are assigned with dedicated channels for further communication based on the Connection Request message and the access procedure is finished after reception of the Connection Setup message by the device. The collided devices follow DQRA proposed in [7] to determine the next RAS for new RA attempts. New devices are not allowed to enter the process if there is an ongoing contention in the selected RAS.

The DQRA protocol utilizes virtual CRQ to split collided devices into groups to reduce the collision probability of subsequent retransmissions. The queues and devices' positions in the queue are identified by two counters: RQ counter and pRQ counter. The RQ counter represents the queue length (i.e., the number of the groups of devices in the queue) and the pRQ indicates the position of the group of devices within the queue [7]. The devices colliding with the same preamble form a contention group and enter the CRQ at the same position. Each device computes its position in the queue (defined by the pRQ) based on the information received from the eNB in RAR message. In DQRA, the devices located at the head of the CRQ (i.e., $pRQ = 1$) perform retransmission in the upcoming RAS. After a group of devices retransmits the new preambles in the new RAS, the RQ counter is decremented. Contrary, the RQ is incremented if there is a new collision of the devices with the same preamble. The counter pRQ is decremented by all devices after each RAS until $pRQ = 1$. The devices can transmit preamble in the next RAS when the devices' pRQ become equal to 1.

Fig. 1 illustrates an example of the DQRA operation. Let's assume three preambles are available in each RAS and $M_{i,j}$ represents the number of devices belonging to the j^{th} group of devices colliding in the i^{th} RAS. The lower part of Fig. 1 shows the values of RQ and pRQ counters in individual RASes. In this example, M devices simultaneously transmit randomly chosen preambles in the 1^{st} RAS. There is a collision if two or more devices select the same preamble. Let's assume $M_{1,1}$, $M_{1,2}$, and $M_{1,3}$ devices collide by choosing preamble 1, 2, and 3, respectively, in the 1^{st} RAS. The RQ value is set to the amount of collided preambles, i.e., 3 in our example. The amount of collided preambles is learned by the devices from the RAR message sent by the eNB.

The pRQ value represents the position of the group of devices within the CRQ and thus pRQ differs for all three groups of collided devices. The collided devices enter the CRQ in the order of the preambles they choose, i.e., pRQ is set to 1 for the devices selecting the lowest preamble (in our case $M_{1,1}$). Hence, the pRQs for the groups of $M_{1,1}$, $M_{1,2}$, and $M_{1,3}$ devices are set to 1, 2, and 3, respectively. The group of the $M_{1,1}$ devices is at the head of the CRQ (i.e., $pRQ = 1$), thus, these devices transmit the new preambles in the 2^{nd} RAS.

Now, let's assume that out of $M_{1,1}$ devices colliding in the 1^{st} RAS, $M_{2,1}$ and $M_{2,2}$ devices collide again in the 2^{nd} RAS. Both groups of $M_{2,1}$ and $M_{2,2}$ devices enter the CRQ again. Thus the RQ counter is incremented by two and the new RQ value is set to 4 (3 groups colliding in the $1^{st}RAS$ minus one group served in the 2^{nd} RAS plus two groups collided in the 2^{nd} RAS). The pRQ values for the groups of $M_{1,2}$ and $M_{1,3}$ devices, which are already in the CRQ from the 1^{st} RAS, are decremented, i.e., the new pRQ values are set to 1 and 2, respectively. The pRQ values for the newly collided $M_{2,1}$ and $M_{2,2}$ devices are set to 3 and 4, respectively. In the 3^{rd} RAS, the group of $M_{1,2}$ devices is allowed to retransmit their preambles as their pRQ is equal to 1. In similar way, the groups of $M_{1,3}$, $M_{2,1}$ and $M_{2,2}$ devices are scheduled to transmit and succeed in the 4^{th}, 5^{th}, 6^{th} RAS, respectively. Their counters are updated accordingly in each RAS as shown in Fig. 1.

III. Analytical Model for DQRA

In this section, we describe analytical models for maximum access delay and average number of preamble transmissions as these are key performance indicators for the RA.

A. Maximum Access Delay

Let $\overline{T_{max}}(M, N)$ be the average value of the maximum access delay (unit: RAS) required by M devices to successfully access N channels. After the first transmission in the first RAS, K preambles collide ($0 \leq K \leq N$) and remaining $(N - K)$ preambles are either idle or successful. In other words, DQRA splits the devices into K collided groups. Let $M_{1,k}$ ($M_{1,k} \geq 2$) be the number of devices in the k-th collided group ($0 \leq k \leq K$). From the second RAS, the K collided groups operate independently. Hence, the average value of the maximum access delay of the K collided group is the sum of the average value of the maximum access delays for these K groups (i.e., $\sum_{k=1}^{K} \overline{T_{max}}(M_{1,k}, N)$. Hence, we can define $\overline{T_{max}}(M, N)$

as follows:

$$\overline{T_{max}}(M,N) = 1+\sum_{K=0}^{N}\sum_{M_{1,1},M_{1,2},\ldots,M_{1,K}} P(M_{1,1},M_{1,2},\ldots,M_{1,K}|K) \times\left(\sum_{k=1}^{K}\overline{T_{max}}(M_{1,k},N)\right) \quad (1)$$

where the number $''1''$ represents the first RAS, $P(M_{1,1},M_{1,2},\ldots,M_{1,K}|K)$ is the conditional probability that K preambles collided in total, each preamble colliding with $M_{1,1},M_{1,2},\ldots,M_{1,K}$ devices, respectively (i.e., probability of the possible combinations for all K collided groups). $P(M_{1,1},M_{1,2},\ldots,M_{1,K}|K)$ is calculated as a multiplication of the number of combinations of the collided devices chosen out of M devices and the number of combinations of the successful devices chosen out of the remaining devices.

The concept of (1) is straightforward, however, it is a time-consuming process to calculate $\overline{T_{max}}(M_{1,k},N)$ since the total number of terms in (1) becomes intractable for large values of M and N as the computational complexity of (1) is $O(N\times M^N+1)$. The computational complexity of (1) can be reduced by reordering the collided groups from a variety of the combinations and by merging together the collided groups, which have an identical number of devices. The reordering and merging process can greatly reduce the number of individual terms to be calculated (the computational complexity is $O(N\times M^2)$) while the results of computation are identical. Hence, $\overline{T_{max}}(M,N)$ can be rewritten as:

$$\overline{T_{max}}(M,N)=1+\sum_{i=2}^{M}\sum_{K=1}^{\min\{\lfloor \frac{M}{i},N\rfloor\}} P(i,K)\times K\overline{T_{max}}(i,N) \quad (2)$$

where $P(i,K)$ is the probability that M devices access N channels, and among them, there are K collided groups and each collided group has exactly i devices. The second summation covers all combinations of the numbers of devices in collided groups$P(i,K)$ is given by:

$$P(i,K)=\frac{C_K^N C_i^M\cdots C_i^{M-(K-1)i}}{N^M}\times\frac{N!}{(N-(M-Ki))!} \quad (3)$$

where C_x^y is the number of x-combinations from a given set of y elements ($C_x^y=0,$ if $y<x$).

B. Average Number of Preamble Transmissions

Let $\overline{R}(M,N)$ be the average number of the preamble transmissions for M devices to access N channels. Analogically to (1), K preambles collide ($0\le K\le N$) and remaining $(N-K)$ preambles are either idle or success after the first transmission in the first RAS. From the second RAS, the K collided groups operate independently. Hence, the average number of the preamble transmissions is the sum of the average number of the preamble transmissions for these K groups. That is, $\sum_{k=1}^{K}\overline{R}(M_{1,k},N)\times\frac{M_{i,k}}{M}$. Hence, we can obtain $\overline{R}(M,N)$ as:

$$\overline{R}(M,N) = 1+\sum_{K=0}^{N}\sum_{M_{1,1},M_{1,2},\ldots,M_{1,K}} P(M_{1,1},M_{1,2},\ldots,M_{1,K}|K) \times\left(\sum_{k=1}^{K}\overline{R}(M_{1,k},N)\times\frac{M_{i,k}}{M}\right) \quad (4)$$

Applying $P(i,k)$ as in (2), $\overline{R}(M,N)$ can be written as:

$$\overline{R}(M,N)=1+\sum_{i=2}^{M}\sum_{K=1}^{\min\{\lfloor \frac{M}{i},N\rfloor\}} P(i,k)\times k\overline{R}(i,N)\times\frac{i}{M} \quad (5)$$

IV. Evaluation and Discussion

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

A. Performance Metrics

The maximum access delay (in seconds) and the average number of preamble transmissions are chosen as the performance metrics.

The maximum access delay is the time elapsed between the first attempt to access the radio resources (1^{st} RAS) and reception of the *Connection Setup* message by the last device out of all M devices. The maximum access delay (in seconds) is modeled based on (2) and transformed to time units as:

$$\overline{T_{max}}(M,N)\times T_{RA\_REP}+T_{RAR}+T_{RRC}+rand[1,T_{CR}] \quad (6)$$

where T_{RA_REP} is the time interval between two successive RASs, T_{RAR} is the time interval between RA attempt (preamble) and RAR, T_{RRC} is the time interval between the RAR and *Connection Request*, and T_{CR} indicates the maximum duration for receiving *Connection Setup*. Note that T_{CR} is set for a device to consider that *Connection Setup* is not received, therefore, we assume the *Connection Setup* arrives in random time between 1 and T_{CR} after sending *Connection Request*.

The average number of transmissions is obtained directly by enumeration of (5).

B. Simulation Set-Up and System Parameters

Simulations are conducted in C-based simulator to verify the accuracy of the proposed analytical model. The results are averaged out over 1 000 simulation drops. Each sample is obtained by performing one-shot DQRA with M devices and N preambles available for each RAS. According to LTE, the duration of one sub-frame is 1ms. The simulations environment is duplicated from [7], i.e., $M=10\sim 2500$; $N=6$, 18, 36, 56; $T_{RA\_REP}=10$ subframes; $T_{RAR}=2$ subframes; $T_{RRC}=5$ subframes; $T_{CR}=15$ subframes.

C. Analytical and Simulation Results

The analytical and simulation results for both performance metrics are shown in Fig. 2 and Fig. 3, respectively.

As depicted in Fig. 2, the maximum access delay increases with decreasing number of available preambles since the

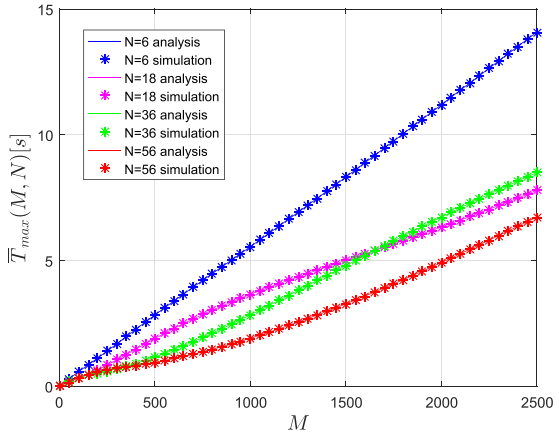

Fig. 2. Maximum access delay (in seconds).

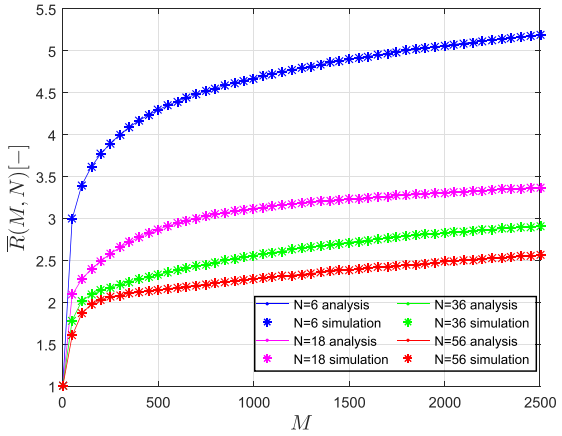

Fig. 3. Average number of preamble transmissions.

collision probability is inverse function of the number of preambles. For N equals to 18 and 36, the maximum access delay converges and reverse around M equals to 1700 devices. It is because, in DQ mechanism, the devices retransmit in the order of their positions in the CRQ and the length of CRQ affects the access delay of devices. Thus, when the number of preambles increases, the chance of successful access increases as well, but the waiting time for retransmission can also become high due to the longer CRQ. Comparison of the analytical models and the simulations shows the analytical model perfectly matches the simulation results.

Fig. 3 shows that the average number of preamble transmissions rises with the number of devices and then saturates for large M. This confirms that DQRA is able to serve access of massive amount of MTC devices. Like in the previous figure, the analytical model perfectly fits the simulations.

Fig. 4 shows that for smaller values of M, increasing N effectively reduces the maximum access delay because the success probability is significantly increased. However, for a larger M, the retransmission delay first drops rapidly with rising N to a local minimum (about $N = 16$ and $N = 18$ for $M = 1500$ and $M = 2500$, respectively). Then, the access delay slightly increases to a local maximum and starts slowly decreasing again. This behavior is due to a combination

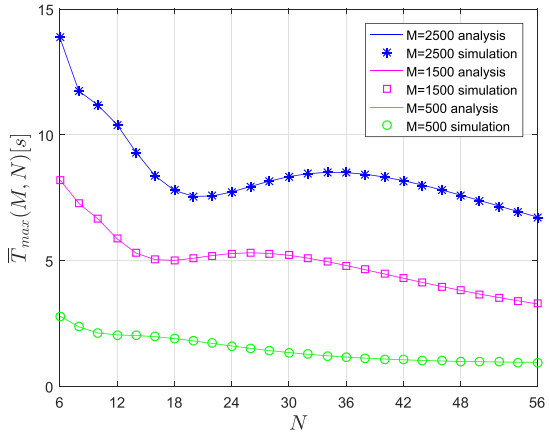

Fig. 4. Maximum access delay (in seconds) as a function of the number of preambles.

of two effects. First, a lower collision probability due to a higher number of preambles leads to a lower access delay (as for the low N). Second, for a higher N, a larger time is required for retransmissions as the distributed queues are served sequentially. A combination of both effects leads to a non-monotonic behavior for the larger M, as shown in Fig. 4.

V. Conclusion

In this paper, we have developed the analytical model for fast estimation of DQRA performance. The analytical models of the average number of transmissions and the maximum access delay can be used for an optimization of the RACH's resource allocation. The numerical results derived by the analytical models demonstrate that the models accurately match the simulation results and confirm the behavior of the DQRA in the sense that a low number of transmissions is needed to obtain access to radio resources even for a high number of simultaneously accessing MTC devices.

As the DQRA is a potential solution to handle massive number of devices in LTE. In the future, more performance metrics, such as energy consumption should be modeled.

References

[1] F. M. Awuor and C.-Y. Wang, "Massive machine type communication in cellular system: A distributed queue approach," in *Proc. IEEE ICC*, May 2016, pp. 1–7.

[2] *Discussion on RACH Overload for MTC*, document 3GPP R2-102780, CATT, RAN#270, May 2010.

[3] X. Yang, A. Fapojuwo, and E. Egbogah, "Performance analysis and parameter optimization of random access backoff algorithm in LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2012, pp. 1–5.

[4] B. Yang, G. Zhu, W. Wu, and Y. Gao, "M2M access performance in LTE-A system," *Trans. Emer. Telecommun. Technol.*, vol. 25, no. 1, pp. 3–10, 2014.

[5] W. Xu and G. Campbell, "A near perfect stable random access protocol for a broadcast channel," in *Proc. IEEE ICC*, vol. 1. Jun. 1992, pp. 370–374.

[6] A. Samir, M. M. Elmesalawy, A. S. Ali, and I. Ali, "An improved LTE RACH protocol for M2M applications," *Mobile Inf. Syst.*, vol. 2016, Jul. 2016, Art. no. 3758507.

[7] A. Laya, L. Alonso, and J. Alonso-Zarate, "Contention resolution queues for massive machine type communications in LTE," in *Proc. IEEE PIMRC*, Jul. 2015, pp. 2314–2318.

[8] 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 Commun. Lett.*, vol. 16, no. 8, pp. 1196–1199, Aug. 2012.