Free Access Distributed Queue Protocol for Massive Cellular-based M2M Communications with Bursty Traffic

Anh-Tuan H. Bui[1], Chuyen T. Nguyen[2], Truong C. Thang[1], and Anh T. Pham[1]
[1]Computer Communications Lab., University of Aizu, Aizuwakamatsu, Japan
[2]School of Electronics and Telecommunications, Hanoi University of Science and Technology, Vietnam
Emails: {m5202101, thang, pham}@u-aizu.ac.jp; chuyen.nguyenthanh@hust.edu.vn

***Abstract*—Long Term Evolution (LTE) networks are expected to play a key role in providing connections to billions of Machine-Type Devices (MTDs) in the 5G big picture. Since the ALOHA-based access framework of LTE-A alone cannot support bursty massive access scenarios caused by the MTDs, the 3GPP has additionally employed the Access Class Barring (ACB) scheme as the baseline traffic control method. While the scheme certainly improves access success probability of the devices, corresponding delay degradation may be unacceptable. In this paper, we renounce the ALOHA-based framework of LTE to propose a new protocol, namely the Free Access Distributed Queue (FADQ), accompanied by an estimation method to resolve the massive synchronized access issue in a more efficient manner. Simulations under the 3GPP's reference setup show that FADQ protocol significantly reduces access delay while maintaining an access success probability on par with the ACB.**

I. Introduction

Machine-to-Machine (M2M) communications refers to autonomous communication between automated Machine-Type Devices (MTDs). It is easy to see that M2M will be an integral part of future's networks thanks to enormous benefits brought about by its realization. In fact, supporting massive connection density for M2M has been recognized as one of the main design objectives of the imminent fifth-generation (5G) networks [1]. Nevertheless, the 5G New Radio (NR) is not yet ready for such task due to its limited coverage and hardware immaturity. Thus, M2M communications in early stages will inevitably rely on the well-standardized, worldwide available Long Term Evolution (LTE) networks which are also considered a major segment of the 5G infrastructure.

Direct integration of MTDs into LTE networks, however, poses significant challenges. Many studies have pointed out that in the scenario where tens of thousands of MTDs try to connect to the network in a bursty manner [2], the LTE's Random Access CHannel (RACH) may be severely congested [3]. As a countermeasure, the 3GPP has implemented the Access Class Barring (ACB) mechanism which statistically spreads the concentrated access attempts to avoid overload. Although the ACB sharply improves the access success probability of the MTDs, the access delay may be severely prolonged at the same time [4].

Clearly, it is due to the incapability to handle bursty access scenario of the ALOHA-based contention resolution mechanism used in the RACH that the ACB scheme must be applied in advanced to improve access success probability [4]. Motivated by such fact, we employ a different approach in this paper to propose a new access protocol, namely Free Access Distributed Queue (FADQ). Our protocol uses a logical access queue to resolve contentions, and is aided by an estimation method to efficiently handle the massive synchronized access issue. The advantages of FADQ will be verified via means of computer simulation under the 3GPP's reference setup [2].

II. ACB Scheme & Motivations of the Work

This section provides a brief overview of the LTE Random Access (RA) procedure (with the ACB scheme implemented) which consists of a four-message handshake between MTDs and the Base Station (BS). For more details, the readers are referred to [4] and references therein.

An MTD starts by sending a randomly chosen preamble sequence (out of K orthogonal ones) to the BS in the nearest Random Access Opportunity (RAO), then listens during a time window to receive a *Random Access Response* (RAR) message that signifies the identities of successfully decoded preambles without collision in that RAO. If the MTD does not find its preamble's identity in the RAR, it performs a random backoff before attempting a new preamble transmission. Otherwise, the MTD schedules an uplink *Connection Request* message i.e., Msg. 3, based on corresponding resource grant in the RAR and wait for a responsive *Contention Resolution* message i.e., Msg. 4, from the BS. Devices who successfully receive Msg. 4 is considered to complete the RA procedure, while those who exceed the maximum number of preamble transmissions give up i.e., "blocked" from the network.

As seen, the LTE RACH uses a random backoff principle to resolve preamble collisions when multiple MTDs select the same preambles in an RAO. Although such simple mechanism is adequate when the MTDs come randomly, it cannot handle the bursty access scenario envisioned by the 3GPP [2], in which the devices arrive (and initiate access) following the time-limited *Beta* distribution. In that case, a large portion of the MTDs is blocked due to short-term congestion [4].

The Access Class Barring (ACB) scheme, thus, is employed to defer RA procedure initiations at the MTDs and in effect, reshape the bursty access pattern. When ACB is in use, an

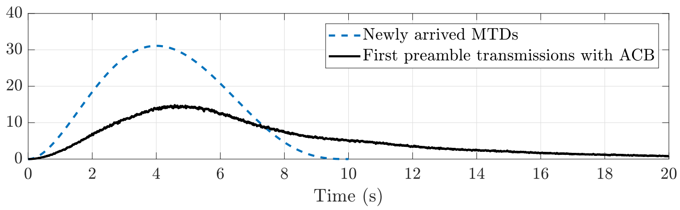

Fig. 1. Spreading effect of the ACB scheme, $p_{ACB} = 0.5$, $T_{ACB} = 4$s

MTD first performs a barring check whose probability of success is p_{ACB}. If the MTD passes the check, it may initiate the RA procedure. Otherwise it must wait for a random period of $[0.7, 1.3) \times T_{ACB}$ before redoing the check [4]. The effect of ACB with $\{p_{ACB}; T_{ACB}\} = \{0.5; 4\text{s}\}$ on 30,000 MTDs whose arrivals follow $Beta(3{,}4)$ distribution over 10s is illustrated in Fig. 1. It is seen that while the ACB greatly reduces the number of new access attempts per RAO during congestion, it excessively defers devices outside of the congested period as well. As a result, overall access delay is severely degraded.

Evidently, the ACB scheme is employed to increase access success probability at the cost of access delay because of the inability to cope with bursty traffic of the RACH's random backoff mechanism. To eliminate the need for ACB, more effective contention resolution mechanisms must be considered. Among them, DQ-based protocols that resolve contentions by breaking contending MTDs in an RAO into smaller groups and organize the groups into a logical access "queue" are one promising option. Thanks to the use of queuing discipline, this protocol family inherently attains very high access success probability. Nevertheless, the conventional DQ-based protocol for LTE i.e., the CRQ [5], is still inefficient delay-wise as it automatically forms a new group for MTDs involving in a preamble collision without considering their actual number. That is, CRQ may create many tiny groups (each contains only a few colliding MTDs) and drastically prolong delay because a group, tiny or not, always occupies a full RAO to transmit [6]. Furthermore, although the previously proposed DQ-based protocol [6] has addressed such issue with the aid of load estimation, it 1) only does so for the case where all MTDs arrive simultaneously, and 2) employs a physical-layer estimation method, both of which are highly impractical.

Thus, we are motivated to propose a new DQ-based protocol, namely the "Free Access Distributed Queue" (FADQ), to practically and effectively supports a massive number of MTDs accessing in a bursty manner over a limited period. To achieve good delay performance, FADQ is assisted by an existing feasible estimation method that will also be introduced.

III. Proposed Random Access Protocol

In this section, we aim to describe our newly proposed FADQ protocol and the relevant estimation technique.

A. Estimation method

Load estimation is an indispensable component in the design of optimal access protocols and has been intensively studied, albeit mainly in RFID context [7], [8]. Here, we employ an approach similar to the one in [7] to estimate the number of colliding MTDs in an RAO as follows.

Let us respectively denote by C, S and E the observed number of preambles that are selected by more than one (Collision), only one (Singleton) and none (Empty) of the MTDs in an RAO. Note that $C + S + E = K$. The averages of C, S and E can be respectively derived as

$$\bar{C} = K - \bar{S} - \bar{E}, \quad \bar{S} = n_t \left(1 - \frac{1}{K}\right)^{n_t - 1}, \quad \bar{E} = K\left(1 - \frac{1}{K}\right)^{n_t}. \qquad (1)$$

The number of transmitting MTDs i.e., n_t, can then be estimated as the $\hat{n}_t$ that minimizes the distance between the observed triplets (C, S, E) and the theoretical average i.e.,

$$\hat{n}_t = \arg \min_{n_t \in \mathbb{N}} \left\{ \left(C - \bar{C}\right)^2 + \left(S - \bar{S}\right)^2 + \left(E - \bar{E}\right)^2 \right\}. \qquad (2)$$

In the unlikely event of $C = K$, $\hat{n}_t$ cannot be found and is interpolated as the integer multiple of K that is greater and closest to $\hat{n}_t$ at $C = K - 1$. Finally, the estimated number of colliding MTDs is expressed as $\hat{n}_c = \hat{n}_t - S$. Note that $\hat{n}_c$ for different (C, S, E) can be calculated once and stored in a matrix for continuous use.

B. Free Access Distributed Queue (FADQ) protocol

The first key idea of FADQ is that whenever preamble collisions occur in an RAO, all n_c colliding MTDs are randomly split into G groups. More importantly, the value of G is calculated based on the estimate of n_c i.e., $\hat{n}_c$, to maintain system's efficiency. For convenience, let us first denote by r the maximum expected number of MTDs successfully obtaining uplink grants in an RAO, and by $d(r)$[1] the number of MTDs in an RAO such that r can be attained. The calculation of G, which is our first main contribution, is then detailed as follows.

- If $\hat{n}_c > d(r)$, then $G = [\hat{n}_c / d(r)]$ where $[\cdot]$ denotes the "round" operator. The rationale is that after the division, the average number of MTDs per group will be around $d(r)$ and thus, r may be achieved when these groups retransmit in the future.
- If $\hat{n}_c \leq d(r)$, then $G = 1$ as at this point, further divisions will cause the average number of MTDs per groups to drop too low, which leads to under-utilization.

These groups are then "pushed" to the end of a logical access queue and in an RAO, only the group at the queue's head may exit to perform preamble retransmissions. The queue does not exist physically, but is realized using two counters named DQ (at the BS) and pDQ (at each MTD). The former is current "length" of the queue while the latter is current "position" of the MTD inside the queue e.g., $pDQ = 0$ devices are at the queue's head and may transmit. Both counters are updated after each RAO as follows

For DQ (at the BS):

- If $DQ > 0$ i.e., a contention session is going on, then $DQ = DQ - 1$ due to removal of the head entry.

[1] $d(r)$ is derived and justified in Appendix A.

- If preamble collisions occur, $DQ = DQ + G$ to reflect the addition of G groups of MTDs to the queue's end.

For pDQ (at individual MTD):

- If the MTD is waiting in the queue i.e., $pDQ > 0$, then $pDQ = pDQ - 1$ due to removal of the head entry.
- If the MTD collides with others of same preamble, it randomly selects an integer g between $[0, G-1]$ and set its $pDQ = DQ - G + g$ to indicate that it has chosen the g-th group and re-entered the queue from the end. Note that G and DQ are assumed to be included in the RAR.

To finalize our design, we now specify a rule following which FADQ treats new MTDs. It can either hold them until current contention session is over i.e., until $DQ = 0$ again, or let them join the group at the queue's head to instantly transmit preambles in next RAO i.e., "free access". The former strategy is reasonable for completely random access pattern, as the number of new MTDs accumulated during a session tends to be stable. In bursty scenario where arrivals of MTDs are temporally clustered, however, that number explodes as time progresses toward concentration point, thus results in excessively long resolution sessions that degrade access delay.

Therefore, our protocol employs the "free access" strategy which is much more beneficial in terms of delay. However, the actual number of MTDs in an RAO would then be higher than what we expected while designing G, as new devices of unknown quantity are continuously added to the head group. In other words, given the "free access" rule, our calculation of G is sub-optimal and represents a *best-effort solution* in keeping the number of MTDs per RAO at the optimal level. The adoption of "free access" to cope with bursty access scenario is our second main contribution that signifies another fundamental difference between FADQ and existing DQ-based protocols for LTE [5], [6].

Fig. 2 is an example of our protocol with $K = 4$ preambles. For demonstration purpose, we assume in this example that $W_{RAR} \times N_{RAR}$ is big enough so $d(r) = K$ (see Appendix A) and that the BS knows the exact number of colliding MTDs i.e., n_c, in an RAO. The left hand side is the visualization of FADQ's operation as a splitting tree where each rectangle represents an RAO. The upper and lower number inside a rectangle represent the number of transmitting *"old"* MTDs and n_c, respectively, while the number with the addition sign next to it denotes the number of newly arrived MTDs in that RAO. In the 1st RAO, 15 new MTDs arrive and transmit their preambles. None of them succeeds due to collisions, and the BS randomly divides these MTDs into $G = \lceil 15/d(r) \rceil = 4$ groups in a best-effort attempt to keep the number of MTDs per group close to $d(r)$. The first group of 4 MTDs retransmits in the 2nd RAO where there is also an unexpected new MTD who transmits immediately using "free access" rule. As a result, there are a total of $4+1 = 5$ transmitting MTDs in this RAO, and two of them are involved in collision ($n_c = 2$). Since $n_c < d(r)$, the BS sets $G = 1$ to stop further division, and the two remaining MTDs thus simply rejoin the queue from the end to retry later in the 6th RAO. This process continues until all MTDs finish

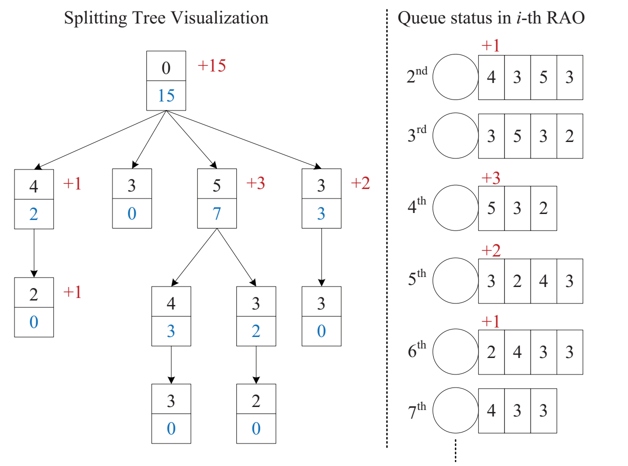

Fig. 2. Example of the operation of FADQ

their RA procedure.

Lastly, when RAO periodicity is short, multiple parallel queues are used as in our previous work [6] to extend the interval between consecutive RAOs seen by a queue to ensure that from a queue's perspective, RARs for an RAO are received before next RAO's occurrence. In 3GPP's setup where *PRACHConfIndex* = 6 i.e., one RAO per 5ms, and $W_{RAR} =$ 5ms, this translates into two parallel queues, each sees an RAO every 10ms. Upon entering the network, an MTD choose a queue to associate with and all following preamble transmissions must be in the RAOs seen by that queue.

C. Preamble Detection Probability & Limited Grants

According to the 3GPP's simulation setup, a *singleton preamble* is assumed to be detected with probability $\left(1 - 1/e^i\right)$ where i denotes the i-th preamble transmission by the corresponding MTD [2]. Furthermore, even if the singleton preamble is correctly detected, the limit in the number of grants that can be sent during the RAR window may result in the MTD not being granted resources for its Msg. 3 transmission. Here, we assume that all MTDs whose Msg. 1 fails due to these reasons are aware of such fact and perform "free access" in the very following RAO alongside newly arrived ones.

IV. Simulation Setup, Results & Discussion

A. Simulation Parameters and KPIs

Our simulation parameters are displayed in Table I, which agree with the reference setup for massive bursty access scenario [2]. The processing delays follow the diagram in [3] while barring factor and barring time are set to $p_{ACB} = 0.5$ and $T_{ACB} =$ 4s [3], respectively.

To assess the protocols, we use four Key Performance Indicators (KPIs) [3] described as follows. Note that the MTDs who successfully finish the RA procedure before exceeding *preambleTransMax* are referred to as *successful* MTDs

1) Collision probability, P_c: The ratio between total number of preambles transmitted by more than one MTDs and total number of preambles available during simulation period.

TABLE I
SIMULATION PARAMETERS

Parameters	**Values**
Number of MTDs	$N = 30000$
Arrival distribution	Beta(3,4) over 10 s
PRACH configuration index	$PRACHConfIndex = 6$
Subframe length	1 ms
Available preambles for contention-based random access	$K = 54$
Maximum number of preamble transmissions	$preambleTransMax = 10$
RAR window size	$W_{RAR} = 5$ subframes
Maximum number of uplink grants per subframe	$N_{RAR} = 3$
Preamble detection probability for the i-th preamble transmission	$P_d = 1 - \frac{1}{e^i}$
Backoff Indicator	$BI = 20$ ms
Retransmission probability for Msg 3 and Msg 4	0.1
Maximum number of Msg 3 and Msg 4 HARQ transmissions	5
Round-trip time of Msg 3 (Msg 4)	8 (5) subframes

TABLE II
SIMULATION RESULTS

Key Performance Indicator		*FADQ protocol*	*ACB scheme*
Collision probability, P_c		12.72%	2.3%
Access success probability, P_s		99.83%	97.45%
Access delay, D	$\mathbb{E}[D]$	2249.6 ms	4143.53 ms
	D_{10}	21.63 ms	21.52 ms
	D_{50}	1609.48 ms	2951.67 ms
	D_{90}	6164.82 ms	11837.22 ms
	D_{95}	6998.24 ms	15819.29 ms
Average number of preamble transmissions, $\mathbb{E}[k]$		2.8024	2.4507

2) Access Success Probability, P_s: The ratio between the number of successful MTDs and total number of MTDs.

3) Statistics of the access delay of successful MTDs: We define the access delay of an MTD as the duration from when the MTD arrives to when it correctly receives Msg. 4 from the BS. Then, this KPI is assessed via the mean access delay, the 10th, 50th, and 95th percentiles ($\mathbb{E}[D]$, D_{10}, D_{50} and D_{95} respectively) of the CDF of the access delay.

4) Statistics of the number of preamble transmissions for successful MTDs: We assess this KPI in terms of its mean value $\mathbb{E}[k]$.

B. Results & Discussions

In this section, computer simulations using MATLAB under the settings in Table I are performed to validate the effectiveness of the proposed protocol in comparison with the ACB scheme. Corresponding results for the four KPIs are summarized in Table II.

1) Let us first examine P_c. It is seen from Table II that there are significantly more preamble collisions occurring in our proposed protocol than the ACB scheme. A lower P_c, however, does not necessarily imply a more efficient contention resolution mechanism and may hint that the RACH is under-utilized for most of the time. Later analysis will reveal that this is indeed the case for ACB. Our FADQ, on the other hand, posts a high P_c because it keeps the RACH constantly busy, albeit at a sub-optimal level.

2) The access success probability i.e., P_s, is an indicator of how suitable (but not necessarily efficient) a protocol is in providing the massive MTD population with connections. As seen, both protocols are able to maintain very high P_s which are above the 95% threshold. Also, compared to the ACB scheme, our protocol achieves a slightly higher P_s despite having a much higher P_c.

3) Access delay is our most important KPI and is the KPI at which the advantages of FADQ become pronounced. Under the same settings, FADQ shows a remarkable 45.7% reduction in average access delay compared to the ACB, and half of the population of the former finish their RA procedure within 1.609 seconds, which is 45.6% quicker than those of the latter. This delay gap widens further as one progresses toward the tail of the CDFs. At the 90th and 95th percentile, our protocol outperforms the ACB by a margin of 47.9% and 55.1%, respectively. D_{10} of FADQ, meanwhile, is as good as that of the ACB thanks to the former's adoption of "free access" rule that allows quick access for newly arrived MTDs.

To further demonstrate the effectiveness of FADQ over the ACB, we plot the access delay CDFs of the two in Fig. 3. It is observable that the ACB exhibits a long trail starting from around the 75th percentile. Such undesirable behavior is clearly the toll of consecutive failed barring checks [3]. Plus, since there is no limit to the maximum number of barring check failures, there is also no bound to the maximum delay of the ACB scheme i.e., its $D_{100} \rightarrow \infty$ as evidently seen on the figure. On the other hand, our proposed protocol displays a much steeper CDF, and all of its successful MTDs are granted access within a limited period of less than 11.82 seconds from their arriving time during the simulation period.

This delay analysis has verified our earlier arguments about P_c. That is, the high P_c of FADQ comes from good use of random access resources while the low P_c of the ACB is due to the fact that it generates many under-utilized RAOs (which prolong access delay) with its barring mechanism.

4) The final KPI i.e., $\mathbb{E}[k]$, is also of interest as it is closely related to the energy consumption of the MTDs. Here, the benefit of having a lower P_c applies directly and thus, the MTDs of the ACB scheme need undeniably less contending attempts on average to succeed than those of ours. In practice, however, the bursty access scenario is likely to arise from urgent situations under which the tradeoff between access delay and $\mathbb{E}[k]$ should be carefully considered.

To provide a look at the system's evolution, we have plotted in Fig. 4 the average numbers of transmitting MTDs and MTDs who receive uplink grants in each RAO of the ACB scheme and FADQ protocol. Looking at the upper part, one can see the that the ACB actually performs very well during the congested period (from the 343th to 1329th RAO) where

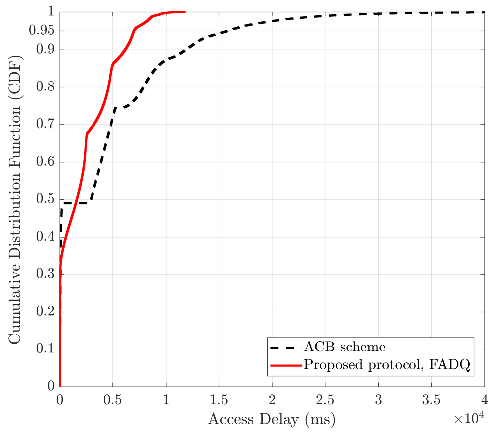

Fig. 3. CDF of access delay of the ACB and FADQ protocol

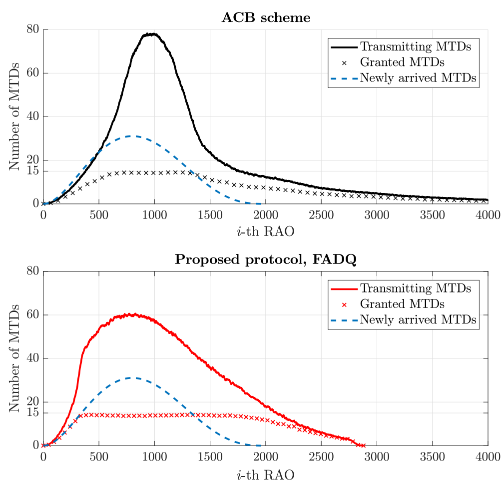

Fig. 4. Temporal behaviors of ACB scheme and FADQ

the number of uplink grants allocated in each RAO is close to the system's maximum of $W_{RAR} \times N_{RAR} = 15$ for most of the time. After that, however, ACB quickly shows its drawback as the static barring parameters $\{p_{ACB}; T_{ACB}\} = \{0.5; 4\text{s}\}$ become excessive for the relaxed post-congestion traffic.

Our proposed protocol, on the other hand, shows a consistently good performance over time and exhibits an adaptive behavior. That is, the number of granted MTDs rises in harmony with the number of newly arrived ones until the system hits its limit of $W_{RAR} \times N_{RAR}$ at the 343th RAO, then stays near that limit for a long duration before finally settling down at around the 2000th RAO i.e., the end of activation period of $T = 10$s. The drop at the end is justified by the fact that without any new arrivals, the population of each group inside the queue gradually decreases below $d(r)$ as the old MTDs succeed and leave. The number of uplink grants obtained in each RAO thus decreases accordingly.

V. Conclusions

This paper has proposed a new DQ-based random access protocol i.e., the FADQ, assisted by a load estimation method to support the massive bursty access issue of cellular-based M2M communications. To accomplish the goal, the proposed protocol makes use of two main ideas of free access and load estimation. As a result, our proposal is able to achieve greatly reduced access delay compared to the ACB scheme while still maintains a very high access success probability, as proven by computer simulations under the 3GPP's reference setup. More importantly, being designed with practicality in mind, FADQ fully complies with LTE specifications and arises as a suitable access solution for 5G's massive M2M use case.

Appendix A: Derivation of $d(r)$

We define r as the maximum expected number of MTDs that can be provided with uplink grants in an RAO. When the system is not bounded by the maximum number of grants per RAR window, r is achieved when the number of MTDs in an RAO is equal to the number of preambles K. Otherwise r is limited by $N_{RAR} \times W_{RAR}$. Thus, we can write

$$r = \min\left\{K\left(1-\frac{1}{K}\right)^{K-1},\ N_{RAR} \times W_{RAR}\right\}. \tag{3}$$

We are interested in keeping the number of MTDs in an RAO at a certain level $d(r)$ such that the corresponding expected number of singleton preambles is approximately r to ensure the RACH's efficiency. The expected number of singleton preambles in an RAO given n_t transmitting devices, however, is expressed as $\bar{S}$ in (1) and thus, $d(r)$ can be found as the $n_t \leq K$ that minimizes the distance between $\bar{S}$ and r i.e.,

$$d(r) = \arg \min_{n_t \in \mathbb{N}, n_t \leq K} \left|\bar{S} - r\right|. \tag{4}$$

For the parameters in Table I i.e., $W_{RAR} \times N_{RAR} = 15$ and $K = 54$, we get $r = 15$ and $d(r) = 22$.

References

[1] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5g: physical and mac-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, September 2016.

[2] 3GPP, "Study on ran improvements for machine-type communications," 3rd Generation Partnership Project (3GPP), Tech. Rep. 37.868, Sep. 2011, version 11.0.0.

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

[4] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in lte-a networks with massive m2m traffic," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, 2017.

[5] A. Laya, L. Alonso, and J. Alonso-Zarate, "Contention resolution queues for massive machine type communications in lte," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Aug 2015, pp. 2314–2318.

[6] A. T. H. Bui, C. T. Nguyen, T. C. Thang, and A. T. Pham, "Design and performance analysis of a novel distributed queue access protocol for cellular-based massive m2m communications," *IEEE Access*, vol. 6, pp. 3008–3019, 2018.

[7] H. Vogt, "Efficient object identification with passive rfid tags," in *Pervasive Computing*, F. Mattern and M. Naghshineh, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 98–113.

[8] M. Shahzad and A. X. Liu, "Fast and accurate estimation of rfid tags," *IEEE/ACM Trans. Netw.*, vol. 23, no. 1, pp. 241–254, Feb 2015.