Primary Access Procedures in M2M Networks

Abdullah Balci, Radosveta Sokullu
EEE Dept., Faculty of Engineering, Ege University
Izmir, Turkey
email : abdullah.balci@ege.edu.tr, radosveta.sokullu@ege.edu.tr

Abstract — The immensely increased number of networked devices in the last several years is believed to be only the beginning of a new era of machine-to-machine (M2M) communications, where billions of devices will connect over the network and react to events in the environment without human intervention. This emerging technology poses new challenges to the existing connectivity mechanisms. Current network access methods rely on random access (RA) back-off based mechanisms with inherent control of the congestion and the delay. In M2M communications, devices will be sending only very small amounts of data (several bytes so several kilobytes) but their excessive numbers and the possibility that many of them will try to connect simultaneously will cause high collision rate and unacceptable delays. A lot of active research is directed to these subjects and the goal of this paper is to summarize and classify the suggested methods in order to provide a clear picture of the open research issues.

Keywords-M2M communications; random access procedure; congestion; delay

I. INTRODUCTION

A lot of research in recent years is concentrated on new technologies, enabling the communication between “things”, “machines” and “devices” like “the Internet of Things” (IoT), “Machine-to-Machine” (M2M) and Device-to-Device (D2D) communications. The basic idea behind these concepts is that “things” or “objects” will be able to communicate with each other and perform actions without human intervention [1]. In 2015, 604 million machines were connected to the internet and the number is predicted to rise over 3,075 million by 2020. This amazing growth in terms of number of devices and data volume is faster than the growth of human population and consequently has resulted in a new paradigm defined as M2M communications [2][3].

The aim of M2M networks is to connect devices together and enable them to make smart decisions based on the generated and transferred data. The characteristic of these M2M networks are quite different from those of current wired and wireless networks. Human-to-human (H2H) communications focus on high data rates and high QoS, while M2M communications aim generally at low data rates with very strict time constraints. Furthermore, because of the large number of devices, the number of simultaneous attempts to connect to the network will be much greater than those in H2H centered networks. This renders existing access algorithms ineffective resulting in high collision rate and extreme delays and has forced major standardization organizations such as ETSI and 3GPP to concentrate their efforts on these issues [4]. Specifically 3GPP is working on network architectures that allow the integration of M2M communications with cellular networks such as LTE and LTE-A. In LTE, the devices have to perform the RA procedure to connect to the network using the Physical Random Access Channel (PRACH) in the uplink direction. With H2H communications this procedure gives satisfactory results but is not suitable for M2M scenarios. Furthermore, since both H2H and M2M devices will perform this procedure on the same uplink resources, it will also cause significant degradation in the QoS for H2H users and this is quite an active research area [4].

In this paper, we focus on the challenges arising in the RA procedure due to the introduction of an extremely large number of M2M devices. The paper is organized as follows: first a brief overview of the RA procedure in LTE is presented, then in Section III, the RA challenges are defined and the existing solutions are summarized introducing a clear taxonomy of the suggested methods. Section IV concludes the paper by defining the major open research issues.

II. RANDOM ACCESS PROCEDURE IN LTE

The two most important situations when the Random Access (RA) procedure is initiated by a device are: when it is turned on and it has no allocated uplink resources, and when handing over from one eNB to another. The devices can send their access request only on the allocated PRACH, which consists of 6 Resource Blocks (RB). There are 16 different RA resource configurations for different system bandwidth and different number of cells per eNB. LTE allows two types of RA procedure: contention based where the devices compete for the channel access, and contention-free. The first one is used by UEs for initial establishing a connection and synchronization, while the second one is reserved only for new downlink requests or handover, which are very time-sensitive operations [6]. The contention-based RA procedure has four steps:

Step 1: Preamble Transmission: The RA procedure begins with the selection (from a predefined set) of a preamble, which is used as a signature. Each device randomly selects one of the 64-Nf orthogonal pseudo-random preambles, reserved for contention-free requests, without knowing which is already selected or used. When the cell size is large, a longer preamble will improve the reliability of reception at the cell edge. If a preamble is selected by more than one device, a collision can occur in step 3. Otherwise, the
different preambles are easily detected by eNB because of orthogonality. After the transmission is completed, the device waits for the response from eNB.

**Step 2: Random Access Response:** For each successfully decoded preamble, the eNB sends a random access response (RAR) on the PDSCH, and a Random Access Radio Network Temporary Identifier (RA-RNTI), which identifies the time-frequency slot where the preamble was detected. The RAR message includes the identity of the detected preamble, a timing alignment instruction to synchronize the uplink transmission, an initial uplink resource grant for transmission of the “step 3” message, an assignment of a temporary Cell Radio Network Temporary Identifier (C-RNTI), and a back-off indicator, which determines the waiting time before a new RA attempt. Each device expects to receive the RAR within a time window, specified and broadcasted by eNB. The earliest subframe can occur 2 ms after the end of the preamble sub-frame but typical delay is 4 ms. If the device does not receive a RAR within this time window, it selects and transmits another preamble. If multiple devices select the same preamble in the same time-frequency resource, they would both receive the same RAR.

**Step 3: Connection Request:** The “step 3” message is the first scheduled uplink transmission on the Physical Uplink Shared Channel (PUSCH) sent using HARQ. It conveys the C-RNTI and the actual RA message, such as Radio Resource Control (RRC) connection request or scheduling request. If multiple devices select the same preamble in step 1, these devices will be allocated the same time-frequency resource by eNB, and a collision will occur at the eNB. If no acknowledgement is received by the eNB, the devices will retransmit the same message after the timeout expires.

**Step 4: Contention Resolution Message:** The contention resolution message is addressed to the C-RNTI as an answer to the connection request message, which is sent in step 3. Upon reception of the contention resolution message there are three possibilities: 1) the UE correctly decodes the message, detects its own identity and sends back a positive ACK; 2) the UE correctly decodes the message, discovers that it contains another device’s identity, then it sends nothing back; 3) the UE fails to decode the message or misses the resource allocation.

### III. RACH Challenges and Possible Solutions

#### A. RACH Challenges

As mentioned above, one of the major characteristic of M2M communication is the unprecedented high number of devices which leads to much higher access request rate as compared to H2H communication. Many devices may simultaneously try to connect to the network to send only small amounts of data. So the bottleneck is not high network traffic in general but the burst traffic created during accessing the channel. The high collision probability (CP) and low success rate in network access will cause unexpected access delays, waste of resources, and extra energy consumption. Our simulation results (Table 1) confirm that while for uniform distribution of the access requests (typical for H2H – light grey) the burst arrival traffic (typical for M2M – dark grey) creates much larger CP and AD for large number of devices.

<table>
<thead>
<tr>
<th>Performance Metrics</th>
<th>Number of Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5000</td>
</tr>
<tr>
<td></td>
<td>15000</td>
</tr>
<tr>
<td>Collision Probability (CP)</td>
<td>0.005%</td>
</tr>
<tr>
<td></td>
<td>0.05%</td>
</tr>
<tr>
<td></td>
<td>0.16%</td>
</tr>
<tr>
<td></td>
<td>0.23%</td>
</tr>
<tr>
<td>Access Success Probability</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Number of Preamble Transmissions</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
</tr>
<tr>
<td>Access Delay (AD) (ms)</td>
<td>26.61</td>
</tr>
<tr>
<td></td>
<td>27.67</td>
</tr>
<tr>
<td></td>
<td>28.63</td>
</tr>
<tr>
<td></td>
<td>28.87</td>
</tr>
</tbody>
</table>

The RA procedure allows some devices to establish connections but does not solve the overload of repeated attempts. The results in Fig.2 show the increased number of unsuccessful attempts for burst arrival traffic.

#### B. Suggested Solutions

Efficient overload mechanisms are required for M2M communication over LTE. In [7], the authors suggest application level schemes to control the congestion by scheduling the M2M devices in less loaded periods (midnights), but these solutions bring inconvenience to the end user and greatly limit the application areas of M2M communication, making them undesirable for providers and costumers. 3GPP [8] defines the following major criteria for access methods to be used with M2M communications:

- M2M integration shouldn’t affect the H2H performance,
- Access delay should be considered predicting the behavior of M2M device in the radio access network
- Access methods should be easy to integrate and minimize the effect M2M have on the existing network. Compiling with these major criteria there are a number of various solutions proposed for the primary access to the RACH. The main goal of our paper is to present a taxonomy that allows comparing these solutions and pointing out the areas where more research is needed. The general structure of the suggested taxonomy is provided in Fig.2.
C. Access Class Barring Schemes

A large group of solutions is based on Access Class Barring (ACB) access algorithm, where the eNB groups incoming requests into several classes according to their service requirements and broadcasts a barring factor (ACB parameter) and a barring time. When a device initiates the RA process, it draws a random number and compares it with the barring factor. If the number is less than the barring factor, it continues with its RA process, otherwise, it waits for a period equal to the barring time before reinitiating the RA process. Since the probability of collision is reduced, the throughput of RACH is improved but barring increases the access delay because devices may be barred for an undesirably long time. A version of ACB, the Extended Access Class Barring (EACB) is proposed in which delay-tolerant applications are not permitted to perform the RA process if there is congestion. Thus the number of channel access attempts as a whole is reduced at the price of increased access delay for delay-tolerant applications.

Lien et al [10] proposed a Cooperative Access Class Barring (CACB) scheme in a heterogeneous multi-tier network with picocells and macrocells. In this scheme, neighboring eNBs select the barring factor jointly based on the network congestion level. CACB achieves around 30% improvement both in the average access delay and the worst case delay performance. A major drawback is that it requires devices to be located in more than one eNBs coverage area. Hsu et al [11] proposed the Enhanced Cooperative ACB (ECACB), where they add the number of M2M devices attached to an eNB to determine the M2M device access probability and obtain the barring factor. The ECACB continues to monitor the M2M resource allocation after the devices have been successfully attached to network. It presents a new RRM method that reserves a fixed amount of Physical RBs (PRBs) for M2M devices. The number of reserved PRBs is based on the access rate of M2M device. Results show that ECACB has a lower access delay than CACB even in the worst case scenario. Duan et al [12] proposed a Dynamic ACB (DACB) scheme to reduce the congestion in RACH. Their goal is to manage the access attempts of the M2M devices instead of dismissing the access requests. They propose a novel algorithm that reduces both congestion and access delay by changing the barring factor adaptively. As expected reducing of access delay is reduced but at the price of increased complexity since continuous monitoring is required. The authors of [13] use ACB mechanism together with timing advance information to reduce the RA overload. A novel algorithm is proposed to estimate the number of devices requiring access to the eNB in a given RA slot. They try to find the optimal ACB using the fact that for stationary devices propagation delay is nearly constant. Each device stores the timing advance value for a successful RA run, and compares it with the timing advance value in the next RA run. If the two are the same, the device continues to send the connection request message in step 3. Results show that using ACB together with time advance information allows around 50% of the RA slots be saved for M2M devices when compared to other schemes, which use only timing advance information, only ACB, or only cooperative ACBs.

The main advantage of ACB based solutions is the decreased probability of collision; however, the access delay continues to monitor the M2M resource allocation after the devices have been successfully attached to network. It presents a new RRM method that reserves a fixed amount of Physical RBs (PRBs) for M2M devices. The number of reserved PRBs is based on the access rate of M2M device. Results show that ECACB has a lower access delay than CACB even in the worst case scenario. Duan et al [12] proposed a Dynamic ACB (DACB) scheme to reduce the congestion in RACH. Their goal is to manage the access attempts of the M2M devices instead of dismissing the access requests. They propose a novel algorithm that reduces both congestion and access delay by changing the barring factor adaptively. As expected reducing of access delay is reduced but at the price of increased complexity since continuous monitoring is required. The authors of [13] use ACB mechanism together with timing advance information to reduce the RA overload. A novel algorithm is proposed to estimate the number of devices requiring access to the eNB in a given RA slot. They try to find the optimal ACB using the fact that for stationary devices propagation delay is nearly constant. Each device stores the timing advance value for a successful RA run, and compares it with the timing advance value in the next RA run. If the two are the same, the device continues to send the connection request message in step 3. Results show that using ACB together with time advance information allows around 50% of the RA slots be saved for M2M devices when compared to other schemes, which use only timing advance information, only ACB, or only cooperative ACBs.

The main advantage of ACB based solutions is the decreased probability of collision; however, the access delay...
may be increased a lot and is difficult to predict. This is unacceptable for delay-intolerant emergency applications and event-driven applications where the congestion can rise considerably in very short periods of time. An interesting point is covered in the [14]. Using simulation, the study investigates the RACH performance based on a combination of different barring factors (0.9, 0.7, 0.5) and different RA attempt periods (10 s, 60 s). While almost all devices access the network successfully when the attempt period is set to 60 s, for the same barring factor the access probability falls to 60-70% when the attempt period is reduced to 10 s, i.e., for short attempt periods the barring factor considerably affects the collision rate and the access probability. The results provide valuable insight into the operation of ACB schemes under light and congested conditions.

D. Slotted Schemes

In telecommunication systems, media access is either contention-based (Aloha, CSMA in IEEE 802.11), or is channelized, where the frequency or time is shared between users based on pre-defined slots (FDMA, TDMA). The main advantages of TDMA-type access methods are that there is no collision and slots can be assigned on demand. However, slotted-type methods require very precise synchronization. 3GPP proposed [5] a slotted access method for RA in which the slots are dedicated to M2M devices and each M2M device only accesses in its dedicated slot. However, such contention-free mechanisms are exclusive to two use cases: when the device is in connected state but needs uplink synchronization information to be able to transmit positive or negative ACK or when the device is performing handover from one cell to another.

Slotted mechanisms, well known from many MAC layer protocols, solve collision issues but for their operation the number of participating devices has to be known ahead. Previous research in some related areas (i.e., wireless sensor networks) has shown that very good results in terms of resource efficiency can be achieved if slotted schemes are combined as two stage solutions with contention based methods. How approaches like these can be utilized in the context of M2M communications is a research area to be exploited in the future.

E. Preamble Splitting Schemes

One of the main criteria for integrating M2M communication over LTE is minimizing the effects of M2M communication on H2H communication performance. 3GPP described 64 different preambles for random access procedure that are used by both M2M and H2H devices. So the probability of selecting the same preamble will increase, degrading H2H devices. For this reason, some researchers investigate the possibilities for separation of RA preambles for M2M and H2H devices as an indirect way of reducing congestion and ensuring QoS. Lee et al [15] compare the throughput of two methods for separating RA preambles. The first method is to completely split the set of available RA preambles into two disjoint subsets one for M2M and one H2H communication. The other method is split the set into two subsets – one individually for H2H devices, the other

common for both H2H and M2M devices. They demonstrate that method 2 is slightly better than method 1, but their proposal does not examine the effects on decreasing the congestion. Another study on splitting preambles is [16], where Kim et al propose the Adaptive RA Preamble Splitting (PS) to evaluate three history-based PS schemes in terms of access success rate by using the ratio of non-contention based RA preambles to the total number of RA preambles. These and some other studies using PS show that access success probability can be increased but they do not discuss the effects these methods have on delay performance.

F. Resource Allocation

Another approach to combat congestion on the RACH channel is to separate the RACH resources for H2H and M2M devices [17]. If RACH resources are shared between M2M and H2H devices, the large amount of M2M devices will negatively affect the performance of H2H devices. The RACH resources may be wasted by collisions created by the enormous amount of M2M devices trying to access the network. 3GPP proposed a general algorithm, Dynamic Separate RACH resources (DSRR) in which M2M devices are categorized by types and only devices of the same type contend for the PUSCH resources. Each M2M device listens to the environment before accessing the channel. If there is any activity from devices of the same type, it starts a contention process with them on the dedicated resources; otherwise it is permitted to contend with H2H devices in the PRACH by following the normal RA procedure. Separating M2M from H2H devices in the uplink and classifying the M2M devices increases the access probability for both M2M and H2H devices as well as ensures the performance and access delay for H2H communication. In [18], the authors focus on RA opportunities (RAOs) for group paging. Group paging is a RAN overload control scheme that uses a single paging message to inform a group of M2M devices for system information changes and emergency notifications. In group paging, the eNB assigns a unique group identity to a group of M2M devices. When the group of devices receives the paging message, the devices simultaneously transmit randomly chosen RAOs in the first RA slot. The RA preamble is determined in terms of RAOs and the number of RAO is equal to the number of frequency bands in RA slot times the number of preambles. When a collision occurs, the collided M2M device will perform the back-off algorithm and perform RA procedure with a new chosen RAO in a new RA slot. Because with group paging the number of devices decreases, it is suggested to dynamically allocate RAOs in each RA slot. The proposed method improves the utilization of resources by at least 65% compared to static RAO allocations. Oh et al [19] propose a Dynamic Access Control and RACH Resource Allocation algorithm, which has two phases; “estimating the number of M2M devices” and “access control and RAOs allocation”. The last two studies are similar in the sense that they adjust the RAOs, however the first one applies it to group paging, while the other one integrates it with an ACB mechanism. Both show that improving the utilization of RAOs maximizes the RA efficiency while guaranteeing the average delay. Lo et al [20]
propose a novel overload control scheme called Self-Optimizing Overload Control Scheme (SOOC). SOOC integrates several multiple control schemes (RACH resource separation scheme, the ACB, the SAS and the p-persistent scheme) in order to provide a more efficient, step-by-step congestion control. A M2M device can easily detect an overload on the PRACH channel when it fails to receive a response in the step 4. It assumes there is a state ofnetwork overload and performs an ACB algorithm as described before. Since the collided devices would collide again in the next RA slot, the p-persistent algorithm is used to minimize the chance of a second collision. In the p-persistent algorithm, a device senses the medium and if found idle, transmits with probability p, or else keeps sensing the medium continuously until it becomes idle and then transmits with probability p. A small p leads to long access delay. In SOOC, each device keeps track of the overload indicator and increases it when an access attempt fails. This situation also means that the congestion level of the PRACH channel is rising. That is why each device includes the overload indicator in the step 3 message and according to that the eNB dynamically increases or decreases the number of RA-Slots for PRACH. Unfortunately, the algorithm proposed in this paper is not supported by any simulation or theoretical results.

Separating RACH resources can reduce the impact on H2H devices by only slightly reducing the M2M communication performance as compared to the non-separate RACH case [14]. However, the improvement is limited at high congestion levels. The study also shows that while the dynamic allocation of RACH resources increases the general performance, under heavy traffic the PUSCH resources are extensively allocated for RA procedure.

G. Back-off Scheme

The methods in this group explore different possibilities of adjusting back-off time after collision to regulate the RA procedure. In [21], it is proposed to prioritize H2H devices by using a separate back-off scheme. If the random back-off time for M2M devices is based on a separate back-off parameter, larger than the one assigned to H2H devices, M2M devices will perform RA after a longer time. This increases the success probability for H2H devices. Jian et al [22] suggest M2M class-dependent back-off prioritization to reduce the RACH overload in RAN. They also propose to combine back-off scheme with ACB to control the number of devices that are allowed to start the RA procedure. The suggested algorithm consists of two stages; the ACB stage and the class-dependent back-off stage. The results show that the probability of collision is reduced and the throughput improved by 2-5%. But the disadvantage of this algorithm is increasing the access delay. Bello et al [23] propose a Q-learning based RA scheme (QL-RACH) where a virtual RA slot frame (M2M-frame) is designed specifically for M2M devices. Each RA slot keeps a value according to the success probability in the virtual frame and this value is used in the future to find the best slot for placing an access request. Here the back-off scheme is implemented on top of QL-RACH. Both H2H and M2M devices can use the same frame for initial access. However, after a collision, H2H and M2M devices use different back-off frames, which are restricted for each group, M2M devices cannot transmit in H2H frames and vice versa. The results show an enormous throughput increase by around 70%, but the incurred access delay is not discussed.

Back-off based schemes have only limited potential to improve overall performance and cannot handle overload if the intensity of arrivals is very high. However, they give good results when combined with other schemes; a research question, which can be pursued further [13].

H. Group&Cluster-Based Scheme

In many cases it is required that applications are grouped into clusters based on a specific criteria like; geographic location, application type, QoS requirements, etc. and it is convenient that M2M devices access the RA slot as a group. In the M2M architecture proposed by 3GPP, the resource restricted M2M devices connect to an M2M gateway, which communicates with the eNB on their behalf. Cluster-based access methods have been exploited in WSN quite a lot to provide promising results (e.g. the LEACH protocol, which introduced a revolutionary distributed cluster formation technique enabling self-organization of large number of nodes [24]). Such ideas are quite relevant to M2M communications where clusters can be help increase the efficiency in using the RACH. Kim et al [25] propose to use spatial clustering of devices for preamble reuse during the RA procedure (ERA SGRPA). Preamble reuse for two different devices is suggested when the difference of preamble detection time is larger than the delay spread. All group parameters such as number of groups, a set of group distance, and a set of preambles allocated for each group are broadcasted by eNB. Each device knows its distance from the eNB through the RSS value and determines the affiliated group in a distance based manner. Then, each device selects a preamble set, allocated for its group to start the RA procedure. The proposed method reduces the probability of collision to 1.65% when the number of devices in a cell is 50,000, which is about 9 times less than conventional RA schemes. Another scheme proposed by Lee et al [26] uses location information to form groups. Changes in the RA procedure adapt it to group-based communication. An M2M device periodically transmits the group preamble on behalf of all members. When the eNB receives the group preamble, it sends multiple RARs, which carry information for uplink grant in a RAR-window. Then, each device selects a RAR in the frame and gets the uplink grant. The results show considerable enhancement in the access delay where 95% of the devices successfully complete the RA procedure within 200 ms, which is 40% higher than the devices in legacy RA procedure. A main drawback is that the number of devices in each group must be known before dedicating the resources. Kao et al [27] propose a two-stage group based RA scheme to reduce the collision of M2M devices. The RA procedure is controlled by the M2M gateway for each group. The first M2M device to send a RA request becomes a group leader (M2M gateway) and the eNB broadcasts its identity to all group members. Data transmission is done at two stages: Local Network Stage (LNS) – between M2M device and
M2M gateway - and Global Network Stage (GNS) - between M2M gateway and eNB. Devices contend to send requests to the M2M gateway and the M2M gateway assigns them a RA slot. Each M2M device waits for the arrival of the allocated RA slot and an ACK from the M2M gateway. Finally, the device can perform the ordinary RA procedure on the allocated RA slot. There is no priority between M2M devices and they perform the local contention within the group before GNS. The results show that the access probability is increased nearly 7 times compared to traditional schemes; however the authors do not discuss the delay problem.

IV. CONCLUSION

In this study, we have discussed existing methods for controlling the RA procedures in M2M networks. We have pointed out three major problems arising from the abundance of devices trying to connect to the network simultaneously: increased collisions creating congestion and extreme delays, reduced throughput and reduced QoS for H2H users. We have suggested a taxonomy of the existing solutions and provided their comparative evaluation. ACB based methods alleviate collisions, preamble splitting methods help preserve the H2H QoS but both lead to uncontrollable delays. Resource allocation methods score well on both counts but reduce the general throughput, similar to back-off based ones. Finally, most promising are cluster based and adaptive, two-stage solutions, which require further research.

REFERENCES