Cellular M2M Network Access Congestion: Performance Analysis and Solutions

Fengming Cao and Zhong Fan
Toshiba Research Europe, Telecommunications Research Laboratory
32 Queen Square, Bristol, BS1 4ND, UK
{fengming.cao, zhong.fan}@toshiba-trel.com

Abstract — Machine to machine (M2M) communication is the direct communication between devices without human involvements. Recently M2M services in cellular networks have been promoted by cellular operators for sustainable profits for the next decade. However, M2M will also bring serious access congestion problems into cellular networks. This paper studies random access (RA) protocols for cellular M2M communication systems and in particular investigates two RA methods including transmission probability control and two-hop Aloha as well as their dynamic extensions. Our analysis illustrates that the proposed solutions can provide much better performance in terms of random access throughput and delay.

Keywords — M2M, random access, Aloha

I. INTRODUCTION

Machine to machine (M2M) communications are the direct communications between devices without human involvements. As an enabling technology for a wide range of Internet of things (IoT) applications such as telematics, smart metering, healthcare, security, and fleet management, M2M has drawn huge attentions in the past few years [1]. Driven by the consumers’ demand for new services and enterprises’ need to reduce costs, significant growth in the M2M market during the next few years has been predicted. It is forecasted that the number of M2M devices will expand to billions by 2020 and cumulative M2M cellular connections are expected to rise to 364.5 million globally by 2016 [2].

With the rapid development of the M2M landscape, it is inevitable to use the ubiquitous cellular infrastructure to facilitate M2M communications. Recently, M2M has been intensively discussed in the cellular standardization bodies such as 3GPP LTE and WiMAX, as mobile operators and vendors are seeking new services and applications for sustainable profits for the next decade [3, 4, 5]. M2M over cellular is one of the best options to connect diverse devices over great distances by using established, ubiquitous and robust networks with proven technologies. For example, British Gas is installing smart meters for their customers, where smart meters take accurate readings of energy usage in small time granularity and automatically send the data via a cellular network [6]. In 3GPP, M2M is also referred to as MTC (machine type communications).

Basically, to facilitate M2M communications, a cellular network has to meet the following requirements:

- support of extremely high number of devices, with different traffic characteristics,
- support of various latencies, high reliability and low power consumption,
- support of different mobility profiles, and
- no negative impact on human to human (H2H) communications.

One of the most critical challenges in cellular M2M is the network access congestion problem, where massive M2M accesses happen in a very short time and consequently cause very low access success probability and long access delays [7]. Fig. 1 shows the population density in London based on 2007 statistics [5]. Based on this information, the average number of households per square kilometer and consequently the number of smart electricity meters per cell for different cell radii (assuming each household has an electricity meter) can be estimated. Table I lists the expected number of households per cell for typical cell radii [5]. Compared to the number of active conventional H2H users in a cell (usually less than 100), the huge amount of M2M devices in a cell could cause severe congestion in network access.

Fig. 1: The population density in London

For a given wireless communication system, multiple access protocols have huge impact on the system performance. Basically there are two kinds of contention-based access protocols. One is the carrier sense multiple access (CSMA) protocol based on the principle of listen before talk. The other is the Aloha-based access protocol, in which a user simply transmits whenever it wants to. Normally, CSMA-based access
is more efficient than Aloha-based access and is widely used in short range networks such as WiFi [13]. However, due to the hidden node problem, it is not suitable for networks with large cell sizes such as cellular and satellite systems. To improve system efficiency, access requests in cellular systems (from 2G to 4G) are typically based on Aloha random access. In LTE, the RA procedure consists of four steps: random access preamble transmission, random access response, RRC (radio resource control) connection request, and RRC connection setup [11]. In LTE, a random access opportunity (RAO) is defined as a three-dimensional radio resource block in a time-frequency domain with a fixed number of preambles. Thus a RAO is the combination of random access slot, frequency band, and preamble signature [12].

The integration with vast M2M devices brings a unique challenge in designing an efficient and reliable random access protocol for cellular networks, and various proposals have been under active discussion in standardization bodies (e.g. [4][5]). In 3GPP, several radio access network contention resolution schemes have been developed: random backoff, access class barring, dedicated resource allocation, etc. Each of these methods has its strengths and weaknesses and none of them is widely acknowledged as the best solution. In the literature, a few papers have also investigated radio access network (RAN) overload issues of M2M communications in 3GPP LTE networks. For example, in [10], the authors propose a radio resource management method based on grouping according to the quality of service (QoS) requirements and packet arrival rate of each MTC group. However, when there is a massive number of MTC access attempts, this scheme does not scale well.

<table>
<thead>
<tr>
<th>Area</th>
<th>Population Density (Sq Km)</th>
<th>Typical Cell Size (Km)</th>
<th>No. Households/cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central London</td>
<td>10000</td>
<td>0.5</td>
<td>4968</td>
</tr>
<tr>
<td>Urban London</td>
<td>7500</td>
<td>2</td>
<td>35670</td>
</tr>
</tbody>
</table>

This paper is motivated to analyze the cellular random access protocol and propose some new solutions. The contributions of the paper are as follows.

1) The impact on the slotted Aloha performance due to M2M applications has been thoroughly studied.
2) Two methods including transmission probability control (TPC) and two-hop Aloha have been analyzed.
3) Dynamic TPC and two-hop solutions have been proposed and their advantages are demonstrated.

The rest of the paper is organized as follows. Section II describes the widely-used slotted Aloha random access protocol. The impact on non-M2M (i.e., H2H) applications is studied in Section III. The new solutions including transmission probability control and two-hop Aloha are proposed in Section IV and Section V concludes the paper.

II. ALOHA RANDOM ACCESS

In the following, we briefly go through the slotted Aloha protocol, which has been widely used in most of cellular systems. The slotted Aloha protocol is an improved version of the pure Aloha protocol, where a user can send the request only at the beginning of a time slot. If the request message collides with another transmission, it will be re-sent at a later time slot.

Suppose there are $M$ users and each user sends the request message with an arrival rate $\lambda'$, and the total arrival rate (transmission attempts) will be $\lambda = \lambda'M$. Assume there are $N$ opportunities for the users to send the request, where an opportunity can be a narrow band, an access code in a given time slot, or a combination of both, depending on the cellular system in place. The arrival of independent transmissions can be modeled as a Poisson distribution, i.e., the probability that $k$ transmissions occur is given as

$$P(k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

(1)

Now the throughput of successful transmission in one opportunity is defined as the probability of having just one transmission, i.e., $P(1)$:

$$T_c = \frac{\lambda}{N} e^{-\lambda/N}$$

(2)

Then the throughput of all the opportunities can be calculated as

$$T = NT_c = \lambda e^{-\lambda/N}$$

(3)

For any given user, the user collision probability in an opportunity is defined as the ratio of number of collisions to the number of total transmissions, i.e.:

$$P_c = \frac{\sum_{i=0}^{\infty} (1 - P(1)) - P(1)}{\sum_{i=0}^{\infty} P(1)}$$

$$= \frac{\lambda}{N} \frac{1}{e^{-\lambda/N}}$$

$$= 1 - e^{-\lambda/N}$$

The success probability at $k^{th}$ transmission after the first $k-1$ attempts failed is given as

$$S_k = P_c^{k-1} P(0)$$

$$= (1 - e^{-\lambda/N})^{k-1} e^{-\lambda/N}$$

(4)

Normally, the access delay distribution depends very much on the backoff policies. Instead of focusing on a specific backoff policy, here we assume there are no retramsmissions and the access delay is defined as a generic term as the average number of trials to send over an access request and can be obtained as
\[ D = \sum_{k=1}^{\infty} k \cdot S_k \]
\[ = \sum_{k=1}^{\infty} k(1 - e^{-\lambda/N})^{k-1}e^{-\lambda/N} \]
\[ = e^{\lambda/N} \]  \hspace{1cm} (5)

Fig. 2 illustrates the throughput versus total arrival rates with different number of opportunities (4, 8, 16, and 32). It is well known that the maximum throughput per opportunity can be achieved as \( T_{c,\text{max}} = e^{-1} \) when \( \lambda = N \), which means the average delay \( D = e \). The figure clearly shows that the throughput will dramatically decline with the increase of the total arrival rate when \( \lambda > N \). Fig. 3 depicts the average delay with different total arrival rates and we can see that the average delay increases with the total arrival rate exponentially. Similarly, Fig. 4 shows the collision probability versus total arrival rate, and it can be seen that with four opportunities, the collision probability will reach 1 when the arrival rate is just about 20. On the other hand, the collision probability drops rapidly as the number of random access opportunity increases.

From Fig. 2 and Fig. 3, we can observe that the throughput and access delay are highly related to the number of access opportunities \( N \) and the total arrival rate \( \lambda \), which depends on the number of users \( M \) and the individual arrival rate \( \lambda' \). Usually, the total number of opportunities for a given system is fixed. When the total arrival rate is larger than the number of opportunities, increasing the arrival rate will reduce the throughput and increase the access delay significantly. To this end, M2M applications with a vast number of M2M devices may result in severe congestion of random access channels. In the following section we will further analyze the impact of massive M2M access on the RA performance of non-M2M applications.

### III. THE IMPACT ON NON-M2M TRAFFIC

When M2M users are integrated into cellular networks, they have to compete with conventional non-M2M users during the random access procedure. Given the total number of users \( M \), all the users are divided into 2 sets for M2M users and non-M2M users as \( C_i \) (1 \( \leq i \leq 2 \)), and \( M = \sum_{i=1}^{2} M_i \). where \( M_i = |C_i| \). Denote \( \lambda'_i \) as the transmission rate of each user in \( C_i \), then the arrival rate for the \( i^{th} \) set will be \( \lambda_i = \lambda'_i/M_i \) and the total arrival rate will be \( \lambda = \sum_{i=1}^{2} \lambda_i \). Now we will analyze the impact on the non-M2M users when the M2M users share the opportunities with the non-M2M users. Based on (3), we can get the access throughput as

\[ T = \lambda e^{\lambda/N} \]
\[ = e^{\lambda/N} \sum_{i=1}^{2} \lambda_i \]
\[ = \sum_{i=1}^{2} T_i \]  \hspace{1cm} (6)

where \( T_i = \lambda_i e^{\lambda/N} \) is the access throughput of the \( i^{th} \) class set. Fig. 5 shows the throughput of non-M2M users vs. the number of M2M users with different number of opportunities, assuming each user’s arrival rate is 0.2 and there are 20 non-M2M users.
It clearly demonstrates that the throughput of non-M2M users will degrade significantly with the increasing number of M2M users.

![Graph showing non-M2M throughput vs. number of M2M users](image)

**IV. M2M RANDOM ACCESS SOLUTIONS**

The above analysis has demonstrated the negative impact resulted from the integration of massive M2M devices. If there are enough number of RA opportunities, the system may be able to cope with the access congestion resulted from M2M users, but this is usually not the case due to the huge number of M2M devices and practical system constraints. Therefore random access schemes have to be enhanced to tackle the access congestion problem. In [5][8], dedicated opportunities have been proposed, where the M2M users are allocated some dedicated opportunities to access the network instead of sharing the same opportunities with non-M2M users. Dedicated opportunities can surely mitigate the impact on the non-M2M users, but network access congestion still remains for the M2M users. Backoff has also been suggested to avoid access collision, whereas a user will retransmit the request at a later time slot through a given backoff policy if a collision happens [9]. Backoff is regarded as an effective means to improve the efficiency, but this advantage will diminish with high collision probability due to a large number of users. In the following, two methods are investigated to specifically address the M2M network access congestion problem. One is the transmission probability control strategy and the other is the two-hop slot Aloha algorithm.

**A. Transmission probability control (TPC)**

As mentioned before, M2M enables a wide range of applications from smart metering to fleet management, all having very different traffic characteristics. Some applications may need to access the network almost in real time while others are delay-tolerant. This actually implies that priorities can be applied to different users during the random access process to reduce congestions. According to different applications’ quality of service (QoS) requirements, the base station (BS) can prioritize the classes of users and process random access based on the priorities. This can be implemented by imposing a transmission probability. This is also known as access class barring in the 3GPP literature.

As analyzed in Section II, the access throughput and delay are directly related to the total arrival rate, which is composed of the individual arrival rate and the number of users. The individual arrival rate can be controlled by the BS through a transmission probability as shown in Fig. 6, where a predefined value \( \alpha \) is assigned to all the users or a given set of users. Whenever a user wants to send a request, it will generate a random value \( \beta \) with uniform distribution in the range of \((0, 1)\), and compare it with \( \alpha \). If \( \beta < \alpha \), the request will be sent straightaway; otherwise, the user has to wait for the next time slot and repeat the aforementioned steps.

![Diagram of random access with transmission probability control](image)

**B. Two-hop M2M random access**

The other solution to the network access congestion problem is to divide the users into clusters or groups, and the users in the same cluster have to go through a gateway or aggregator to access the BS, which makes the access procedure become a two-hop Aloha protocol. Fig. 7 shows a two-hop M2M network, where we assume only the users generate the requests and the gateways are only responsible for collecting and forwarding the request to the base station. Assume there are \( I \)
clusters each having \( M_i \) users with arrival rate of \( \lambda_i \), then the
total number of users is \( M = \sum_{i=1}^{\gamma} M_i \) and the total arrival rate
at BS will be

\[
\lambda = \sum_{i=1}^{\gamma} \lambda_i
\]

where \( \lambda_i \) is the arrival rate from the gateway of cluster \( i \), which
is equal to the access throughput in cluster \( i \):

\[
\lambda_i = M_i \lambda_i e^{-\frac{M_i \lambda_i}{N}}
\]

(10)

Hence the access throughput at the BS is

\[
T_{\text{two-hop}} = \lambda e^{-\frac{\lambda}{N}}
\]

\[
= \sum_{i=1}^{\gamma} M_i \lambda_i e^{-\frac{M_i \lambda_i}{N}} e^{-\frac{\sum_{i=1}^{\gamma} M_i \lambda_i}{N}}
\]

(11)

Then the average access delay from a user to the BS is

\[
D_{\text{two-hop}} = D_{\text{gw-bs}} + D_{\text{us-gw}}
\]

\[
e^{\frac{\lambda}{N}} + \frac{1}{N} \sum_{i=1}^{\gamma} e^{M_i \lambda_i / N} - \frac{\sum_{i=1}^{\gamma} M_i \lambda_i}{N}
\]

(12)

By setting the derivative of \( T_{\text{tpc}} \) with respect to \( \alpha \) to zero, \( \alpha^* \)
is obtained as

\[
\alpha^* = \min\left(\frac{N}{\lambda}, 1\right)
\]

(14)

With \( \alpha^* \), the access throughput and delay become

\[
T_{\text{tpc}}^* = Ne^{-1}
\]

(15)

\[
D_{\text{tpc}}^* = \frac{e}{\alpha}
\]

(16)

**Dynamic two-hop access**

We need to find \( I^* \) such that

\[
I^* = \arg \max_{I \in \mathbb{N}^+} T_{\text{two-hop}}(I)
\]

(17)

where \( \mathbb{N}^+ \) is the set of natural numbers excluding zero.
Similarly, by setting \( \frac{\partial T_{\text{two-hop}}}{\partial \alpha} = 0 \), we can get

\[
\sum_{i=1}^{\gamma} M_i \lambda_i e^{-\frac{M_i \lambda_i}{N}} = N
\]

(18)

With the assumption that each group has the same number of
users and the same arrival rate of \( \lambda_i \), denote \( \mu = \sum_{i=1}^{\gamma} M_i \lambda_i \) as
the sum of all the user’s arrival rate, then (18) will become

\[
\mu e^{-\frac{\mu}{N}} = N
\]

Then \( I^* \) is given as

\[
I^* = \left\lfloor \frac{\mu}{\log\left(\frac{\mu}{N}\right)} \right\rfloor
\]

(19)

where \( \left\lfloor x \right\rfloor \) rounds \( x \) to the nearest integer. It should be
note that \( I^* \) will become negative when \( \mu < N \) or infinity when
\( \mu = N \), and in this case \( I^* \) will be set to 1.

**D. Performance comparison**

The system performance is studied in terms of access delay
and access throughput. It is assumed that there are 16 RA
opportunities and varying number of users, and each user has
the same arrival rate of 0.8. There are five cases to compare:
“Normal” means conventional slotted Aloha, “TPC” is the TPC
method using fixed transmission probability, i.e., \( \alpha = 0.4 \).
“Two-hop” is the two-hop access method using fixed number
of groups, i.e., \( I = 5 \), and “TPC-Op” and “Two-hop-Op” are
dynamic TPC and dynamic two-hop access respectively. In the
following, we are mainly interested in the performance with
larger number of users, i.e., \( M > 30 \).

Fig. 8 illustrates the access delay with different number of
users. Compared with the “Normal” case, access delay can be
significantly reduced even with fixed transmission probability.
From the figure, we can see that dynamic TPC can further
improve the delay, while the two-hop solutions with either fixed
grouping or dynamic grouping provide the lowest delays. The
throughputs vs. number of users are depicted in Fig. 9, where
the dynamic TPC or two-hop solutions outperform the others
with a large margin. As expected, the dynamic solutions are

![Two-hop M2M network diagram](image-url)
adaptive to the number of users and always provide the best throughputs. Even the fixed two-hop access method can give a much better throughput than the “Normal” method.

Fig. 10 shows the delay vs. throughput when the number of users is larger than 30 and it well summarizes the comparisons. In terms of delay, two-hop solutions achieve the best, dynamic TPC comes second, and “Normal” is the worst. In terms of throughput, dynamic TPC is slightly better than dynamic two-hop while the fixed two-hop solution has more gain over the fixed TPC and “Normal”. It should be noted that the throughput per opportunity is used here, and the graph would have the same trend in comparison if the total access throughput was used.

Fig. 9: Throughput per opportunity vs. number of users

Fig. 10: Comparison of delay vs. throughput per opportunity (number of users ranges from 30 to 150)

VI. ACKNOWLEDGMENTS

The authors would like to acknowledge the fruitful discussions with their colleagues at Toshiba and the support of its directors.

REFERENCES