Reliable Radio Access for Massive Machine-to-Machine (M2M) Communication

Madueño, Germán Corrales

Publication date:
2015

Document Version
Accepted author manuscript, peer reviewed version

Department of Electronic Systems
Aalborg University

Reliable Radio Access for Massive Machine-to-Machine (M2M) Communication

PhD Thesis

Germán Corrales Madueño (gco@es.aau.dk)

Reliable Radio Access for Massive Machine-to-Machine (M2M) Communication.
Germán Corrales Madueño, Ph.D. thesis.

ISBN: 978-87-7152-060-6 J
February 23, 2015

Academic advisors:

Prof. Petar Popovski. Aalborg University, Denmark.

Dr. Cedomir Stefanović. Aalborg University, Denmark.

Assessment Committee:

Prof. Jens Zander. KTH Royal Institute of Technology, Sweden.

Prof. Mikael Gidlund. Mid Sweden University, Sweden.

Assoc. Prof. Rasmus Løvenstein Olsen (Chairman). Aalborg University, Denmark.

Publication List:

Paper 1: Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "How Many Smart Meters can be Deployed in a GSM cell?" (Published) IEEE International Conference on Communications 2013 (ICC). Budapest, Hungary, 2013. Pages: 338 - 343.

Paper 2: Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering" (Published) IEEE International Conference on Smart Grid Communications 2014 (SmartGridCom). Venice, Italy, 2014.

Paper 3: Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling" (Published) IEEE Wireless Communications Letters (WCL), Volumen 3. August 2014. Pages: 429 - 432.

Paper 4: Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Efficient LTE Access with Collision Resolution for Massive M2M Communications" (Published) IEEE Global Communications Conference 2014 (GLOBECOM) Workshops - Ultra-Low Latency and Ultra-High Reliability in Wireless Systems. Austin, USA, 2014.

Paper 5: Germán Corrales Madueño, Nuno K. Pratas, Čedomir Stefanović, and Petar Popovski. "Massive M2M Access with Reliability Guarantees in LTE Systems" (Accepted) IEEE International Conference on Communications (ICC), 2015.

Paper 6: Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Efficient Coexistence of Regular M2M Traffic and Alarm Reporting in IEEE 802.1ah Networks" (Submitted) IEEE Internet of Things Journal (IoT-J), 2015.

Paper 7: Germán Corrales Madueño, Nuno K. Pratas, Čedomir Stefanović, René B. Sørensen and Petar Popovski. "What Can Wireless Cellular Technologies Do about the Upcoming Smart Metering Traffic?" (Submitted) IEEE Communications Magazine (Special issue on Internet of Things), 2015.

Abstract

English Abstract

Machine-to-Machine (M2M) communication is a term that identifies the emerging paradigm of interconnected systems, machines, and things that communicate and collaborate without human intervention. The characteristics of M2M communications are small payloads and sporadic transmissions, while the service requirements can range from massive number of devices to ultra-reliable. This PhD thesis focuses on novel mechanisms to meet these requirements in a variety of wireless systems, from well-established technologies such as cellular networks, to emerging technologies like IEEE 802.11ah.

Today an overwhelming 89% of the deployed M2M modules are GPRS-based. This motivates us to investigate the potential of GPRS as a dedicated M2M network. We show that by introducing minimal modifications to GPRS operation, a large number of devices can be reliably supported. Surprisingly, even though LTE is seen as the preferable solution for M2M, no mechanisms are in place to guarantee reliable M2M access. Contrary to mainstream solutions that focus on preventing overload, we introduce mechanisms to provide reliable M2M service. We also investigate what cellular networks can do about upcoming smart metering traffic. Finally, we consider the 802.11ah protocol, which represents an interesting alternative to cellular networks. We propose an allocation mechanism that matches the dominant operation regime in the network (i.e., periodic

vs. alarm reporting). With the proposed improvements, 802.11ah exhibits outstanding results in terms of efficiency and reliability for massive M2M. The research in this dissertation shows that reengineering of protocols has a great potential to adapt the existing system to the new M2M requirements. However, it also has limitations, that are visible when there is a requirement for a significantly higher data rate per device or requirement for latency that is lower than the fundamental limit of the existing system.

Dansk resume

Machine-to-Machine (M2M) kommunikation er en betegnelse som identificerer den fremkommende paradigme af forbundne systemer, maskiner og enheder som kommunikerer og samarbejder uden menneskelig indblanding. M2M kommunikation er karakteriseret af transmissioner som er sporadiske og indeholdende smådatamængder, samtidig med at servicekravene for disse transmissioner spænder fra at betjene et stort antal enheder til at muliggøre meget pålidelig kommunikation. Denne Ph.D. afhandling fokuserer påinnovative mekanismer til at opfylde disse servicekrav i diverse systemer med trådløs kommunikation, fra veletablerende teknologier såsom mobiltelefonnetværk, til nye teknologier såsom IEEE 802.11ah.

Over 89% af M2M modulerne som anvendes i dag er baserede påGPRS. Dette motiverer os at undersøge potentialet af GPRS som et dedikeret M2M netværk. Vi viser at ved at indføre minimale ændringer til den nuværende virkemåde, kan et stort antal enheder understøttes pålideligt. Selv om LTE opfattes som den foretrukne løsning til M2M, er der overraskende ingen mekanismer i brug som garanterer pålidelig adgang til dette netværk for M2M enheder. I modsætning til nuværende løsninger som fokuserer påat forebygge overbelastning, introducerer vi mekanismer til at levere pålidelig tjeneste af M2M. Vi undersøger ogsåhvad mobiltelefonnetværk kan gøre ved den forestående datatrafik fra intelligente måleenheder (såkaldt smart metering). Desuden betragter vi 802.11ah kommunikationsprotokollen, som repræsenterer et interessant alternativ til mobiletelefonnetværk. Vi foreslår en allokeringsmekanisme som er tilpasset den primære virkemåde af netværket (dvs. periodiske transmissioner eller rapportering af alarmer). Med disse foreslåede forbedringer giver 802.11ah protokollen fremragende resultater, når der ses påeffektivitet og pålidelighed for M2M kommunikation med et højt antal enheder. Forskningen fra den dissertation viser at reengineering af protokollerne har et stort potentiale for at tilpasse de eksisterende systemer til nye M2M krav. Men det har sine begrænsninger der

er synlige når der er krav til signifikant højere data hastighed per enhed eller krav til latenstid der er lavere end den fundamentale grænse for det eksisterende system.

Preface

The thesis was under the supervision of Professor Petar Popovski and PostDoc Čedomir Stefanović, and was supported by the Danish Council for Independent Research (Det Frie Forskningsråd) within the Sapere Aude Research Leader program, Grant No. 11-105159 “Dependable Wireless bits for Machine-to-Machine (M2M) Communications”.

The thesis is comprised by five publications plus one submitted journal and one submitted magazine (i.e. seven contributions in total), where two are on reliability in GPRS, three are on reliability and delay constraints in LTE, one on efficiency and alarm detection in 802.11ah, and one on cellular networks performance for Smart Grid traffic.

All publications included in this thesis follows the template required for their respective publication channel. Hence the page layout might differ between the contributions.

Acknowledgements

First of all, I would like to thank my supervisors Petar and Ceda, who guided me through the ups and downs of a Ph.D. project. To be honest, I only began to consider a Ph.D after I met Petar and worked with him as student helper, where I saw the wonders of applied academic research. Knowing how busy Petar is, makes me value even more that he has been always there to listen and discuss my research, when I needed it. Ceda, who is also well known for spotting variations of 0.25 points in the font size between two figures in a paper, passion for the details that I now share (for the better or the worst). I cannot not thank him enough for all the help and suggestions during my PhD.

I would also like to thank Nuno Pratas for the many discussions on the white board for the past three years about many topics, from M2M research to hijacking quadcopters. This discussions made the day-a-day much more interesting. Special thanks to Rasmus Krigslund, for his support during the first part of my PhD and many other things. Of course, I have to thank the 'Spanish Mafia', Lucas, Dereje, Alex, Carmen, Isa, Marta, Pablo, Nacho, who understand the need of longer lunch breaks. Also my colleagues at Aalborg University, Henning, Jimmy, Kasper, Jesper.

Also, I would like to thank my parents, who made possible the Aalborg experience. Finally, thanks to Laura with whom I have shared the joys, successes and inevitable frustrations during the PhD in the past three years and many before!

Contents

CHAPTER 1

Introduction

Telemetry is the process through which is possible to perform remote measurements. The collected information is then sent to the system operator via radio or wired connection. In the same sense as telemetry, machine-to-machine (M2M) communications is used in a wide range of applications: monitoring of critical infrastructures, smart grid, e-health, sensors, traffic control, etc. M2M takes the telemetry concept to the next level by using sophisticated wireless networking, rather than simple point-to-point links. The usage of wireless techniques for M2M communication has been made possible due to the level of maturity attained by wireless technologies. The majority of these applications are based on the sporadic reception and transmission of small payloads. Specifically, the service requirements for M2M can be divided into *massive* and *ultra-reliable* [1]:

- Massive Device Transmission: handling of simultaneous or near simultaneous transmission attempts to access the network from a large number of devices. In fact, the traffic model considered by 3rd Generation Partnership Project (3GPP) foresees up to 30k devices [2] per cell.

- Ultra-Reliable: connections relative to critical control data require high reliability, such that the information shall be delivered more than 99.XX % of the time.

On the other hand, cellular networks are mainly designed for high-rate human-type communications, such as voice calls, web browsing or video streaming. The improvements required to deal with the M2M requirements, as mentioned above, came only recently into the attention of the standardization bodies. As a general observation, the key technical problem in enabling M2M communications is not on how to increase the overall system data rate, but on how to distribute it efficiently to many terminals with reliability guarantees. Therefore, the underlying question is how well these networks, such as GSM or LTE, handle M2M communications. This thesis investigates the potential of cellular networks to serve as a M2M network and how its performance can be improved by reengineering parts of the protocol. In addition, this thesis considers the new IEEE 802.11ah protocol, which has been targeted for the Internet of things (IoT).

1.1 State-of-the-Art

M2M traffic characteristics and requirements play a fundamental role in the performance analysis of any system. The three main sources providing this information are: 3GPP, 802.16p task group and OpenSG for utilities communication architecture user group. 3GPP has defined two traffic models for M2M [2]: one for uncorrelated report arrivals and one for correlated report arrivals. The foreseen number of devices per cell for both models range between 1k to 30k. In addition, the requirements in terms of tolerable delay and payload for smart grid are described in [3]. In [4] and [5] 802.16p task group provides an exhaustive description of M2M traffic characteristics and deployment density for a multitude of devices, such as smart meters, credit card machines, traffic sensors, traffic lights, elderly sensors, home security devices, etc. OpenSG smart grid networks system requirements specification [6] is the most coherent and exhaustive description of smart grid characteristics and requirements, which includes a list of all the messages exchanged in the network, indicating the arrival rate, latency and reliability requirements.

Latest report on active M2M cellular devices indicates that 64 percent of these devices are GPRS-only with an additional 25 percent 3G/GPRS compatible, 1 percent is LTE capable and the remaining 10 percent corresponds to other technologies such as WCDMA [7]. Surprisingly, even though the market is clearly dominated by GPRS and appears to continue to be so for several years, little effort has been made to investigate and improve the potential of GPRS. It is not until as recent as June of 2014, that notable efforts from the industry and the standardization bodies has been made to adapt GPRS for M2M, where the main goal is to achieve 160 bps[8]. Despite these recent activities, GPRS studies on M2M are limited to a simplified capacity analysis in 2010. Specifically, in [9] 3GPP investigates the maximum number of M2M devices that the random access channel (RACH) of GPRS can support. Moreover, in [10] 3GPP analyzes the capacity in terms of maximum supported devices in the access grant channel (AGCH). Furthermore, the limitation of the identifier (UFS) used in GPRS to coordinate uplink transmissions is considered in [11]. By comparing the results obtained in the different stages, 3GPP provides a maximum number of supported M2M devices in a GSM cell with one and multiple carriers. However, as we have shown in Paper 2, this procedure is flawed as the different stages cannot be considered separately in the case of M2M.

On the other hand, LTE as a M2M network, has gained a lot of focus from both the industry and the research community. One main challenge is how to handle the massive number of expected devices per cell. LTE RACH provides 10.8k random access opportunities per second (typical configuration), which are several orders of magnitude above the average number of human-type requests per second [2]. However, the number of requesting M2M devices can easily exceed the capacity of the RACH in LTE [12]. To overcome this, four main solutions have been proposed: backoff, extended access class barring (EAB), separation of RACH resources for M2M and human-type and dynamic RACH allocation. The main idea of backoff-based schemes is to spread the requests from the different devices over time, so the RACH does not become overloaded. For example, an initial backoff prior to any transmission by M2M devices is discussed in [13], where the load of the RACH is decreased by a factor of 20. The EAB technique limits the number of M2M devices that access the RACH with the aim of preventing overload [14]. This technique is based on the well-known class barring method, but aimed only for delay tolerant M2M devices. These devices only access the network if a randomly chosen value between 0 and 1 is above the barring factor indicated by the EAB. However, there is considerable delay until the overload is detected and the network can announce the new access parameters. In addition, this technique is not suitable for non-delay devices (e.g., earthquake sensors). Further, in order to avoid M2M traffic degrading human-centric traffic, the authors in [15] proposed to split the contention space of LTE in two groups: one for human-centric communications and one for M2M. The main drawback of this technique is that the RACH congestion in case of overload for M2M worsens as the number of resources have been reduced. Finally, the dynamic allocation is a straight forward approach, where upon detection of RACH overload the number of random access opportunities

(RAOs) per second is increased (i.e., the contention space grows when required). The main drawback of dynamic allocation is the time required to inform the contending devices that the RACH parameters have been modified, which can take up to 5.12 s [14]. In addition, it is not enough to accommodate high loads [12].

Obviously, M2M is not limited to cellular networks. There is an ongoing effort into evolving WiFi (IEEE 802.11), named 802.11ah, to exploit the unlicensed 900 MHz band, where it is expected that a single access point (AP) covers up to 1 km. In [16] the authors investigate the feasibility of 802.11ah radio technology for M2M, determining the maximum data rates achievable based on the link budget. In [17] the authors present a theoretical MAC model, the main conclusion is that the protocol supports a large number of stations with low energy consumption. In [18], the authors investigate the impact of the restricted access window (RAW) in 802.11ah, where throughput variations can be observed just due to the length of the RAW. However, these studies rely mostly on stations having always pending data to transmit (i.e., full-buffer approach), which might not be a valid assumption for many M2M applications. Further, in [19] the authors propose an algorithm to compute the optimal size of the RAW based on the number of estimated active stations.

1.2 Thesis Objectives

The main goal in this thesis is to introduce novel mechanisms for evolving the current wireless communications systems, in order to meet M2M communications requirements. In particular, we consider two important requirements: (1) to provide efficient access to a massive number of M2M devices and (2) to provide ultra-reliable service. The first feature is necessary due to the large number of devices foreseen in each cell, which can go up to 30k devices. It should be noted that if the RACH is congested, no cell service is possible including emergency calls. The second feature is required in order to enable wireless monitoring/control of critical applications, such as earthquake sensors or vehicle-to-vehicle (V2V) communications.

In order to introduce these mechanisms, first it is necessary to identify the limitations of the existing systems. In other words, we aim to identify the limitations imposed by the full access protocol rather than only focusing on data resources limitations. From the communication perspective, it is also relevant to investigate the potential of well established technologies such as GSM, as a dedicated M2M network. Therefore, this thesis also aims to contribute the debate on the future of GSM, where many stakeholders foresees that the GSM spectrum should be re-farmed for LTE.

Finally we note that an important aspect is that the new methods and protocols will be built on top of the existing wireless systems. This does not imply that the research in the thesis will be incremental or comparable to development; but the existing systems will rather be used as a "system model" for investigating novel theoretical and implementation aspects of the communication protocols. The practical implication of such an approach is an easier adoption of the technology by the companies that manufacture the M2M equipment or provide M2M services.

1.3 Structure of the Thesis

The thesis is divided into four parts. The first part investigate the potential of GPRS as a dedicated M2M network. The second part seeks to improve the random access procedure of LTE, providing reliable radio access for massive M2M. The third part investigates the improvements for alarm reporting in 802.11ah. Finally, the fourth part studies the performance of cellular networks for smart metering.

CHAPTER 2

Contributions in This Thesis

The motivation and main results for each contribution in this thesis are summarized in this chapter. The contributions are divided into four sections. Section 2.1 includes the work that investigate the potential of GPRS as a M2M network. Section 2.2 comprises the mechanisms and improvements required in LTE for reliable radio access. Further, Section 2.3 summarizes the work relevant to the new IEEE 802.11ah protocol. Finally, Section 2.4 presents the performance studies of cellular networks for smart grid.

2.1 The Potential of GPRS as a M2M Network

2.1.1 Paper 1

How Many Smart Meters can be Deployed in a GSM cell?
Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.
Presented at the IEEE International Conference on Communications 2013: IEEE ICC'13 - Second IEEE Workshop on Telecommunication Standards: From Research to Standards. Budapest, Hungary, 2013.

Motivation

Smart metering is becoming a key technology for any utility, as these devices are capable of reporting the consumption to a central server without human intervention. The deployment of these devices should be functional for a large period of time (typically a decade). Although mobile operators see LTE as the preferable network for smart metering, its coverage and cost are far away from other well-adopted technologies such as GSM with coverages of almost 99% of the territory in many countries. In this paper, we investigate the potential of GSM as a M2M-dedicated network for smart metering.

Paper Content

In this paper we reengineer GSM to boost its capacity for M2M without degrading the stability of the network. We show that an important limitation for massive M2M in GSM is the lack of uplink identifiers, the so-called uplink state flag (USF). The USF is a three bit field used to coordinate a maximum of seven uplink transmissions in every time slot, where the remaining combination is reserved for other purposes. It should be noted that a typical cell with a single carrier can only support up to 49 active connections (7 per time slot). We propose a logical reinterpretation of the protocol to virtually increase the number of USFs. The modifications are such that no alterations to the physical layer are needed. Moreover, the proposed mechanism allows coexistence with legacy devices.

We also provide with the analytical tools to determine the maximum number of devices that can be allocated in the system for a given report interval and maximum tolerable delay.

Main Results

The main result of this contribution is to determine the maximum number of M2M devices that can be allocated in the system. We consider smart meters with report sizes of 100 bytes and 500 bytes, where the report interval is set in the range of 1 to 15 minutes. We also compare the proposed analytical model with an event-driven simulator developed in MATLAB, where it can be seen that the analytical model provides a lower bound on the maximum number of M2M devices. Specifically, we show that a single GSM cell can provide service to 5k devices reporting every minute with a reliability of 99.99%. By increasing the report interval 15 minutes, the number of supported devices raises to an outstanding 50k devices per cell.

2.1.2 Paper 2

Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering
Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski
Presented at the IEEE International Conference on Smart Grid Communications 2014. Venice, Italy, 2014.

Motivation

The clear dominance of GSM M2M modules in the market motivate us to investigate how GSM can be adapted for M2M traffic. However, little effort has been made in this area with only capacity studies by 3GPP. It should be noted that while the average payload of M2M applications is relatively small, there are other limitations that shall be considered. One important difference with respect to Paper 1, is to consider the access reservation procedure, which plays a fundamental role in the the capacity of the system.

Paper Content

This paper provides an exhaustive description of the different M2M applications characteristics. The devices considered are: smart metering, home security devices, elderly sensors, credit machines, roadway signs, traffic lights, traffic sensors and movie rental machines. Further, the information provided includes the arrival rate, number of devices, average payload size and the traffic distribution in a 1000-m radius cell in suburban area.

In addition, to the traffic characteristics, we illustrate the different GSM access limitations. Specifically, we describe the different stages involved in the establishment of data connections in GSM: random access stage, access granted stage and data stage. We show that, in order to properly assess the capacity of GSM for M2M, the inter-dependencies between the stages must be accounted. Therefore, the simplified studies by 3GPP are not valid. Contrary to the conclusions drawn by 3GPP where the USF is the main bottleneck, we show that the access granted channel (AGCH) with only 32 AGCH messages per second is the main limitation.

We describe possible solutions for the different limitations in GSM. Based on the observation that M2M request are similar, a single access granted message could grant multiple devices, which will boost the number of granting messages per second. In this manner, the access granted stage bottleneck is solved. In addition, we improve the USF solution proposed in paper 1, which is exploited to increase the addressing space of the data stage.

Main Results

The first main result corresponds to a numerical comparison of the arrival rates at each stage. Specifically, we show that the approach by 3GPP assuming the same arrival rate for all the stages is not correct. The arrival rate of the radio access channel (RACH) is underestimated as the additional traffic due to retransmissions is not taken into account. In addition, the AGCH stage acts as bottleneck of the system. Therefore, the arrival rate present at the data stage is no longer given by a Poisson process but rather by a truncated Poisson, where the cut-off rate is the maximum number of AGCH messages per second that the system is able to deliver. One important message is that the AGCH becomes the main limitation when the inter-dependencies between the stages are considered.

The second main result is the maximum number of devices that can be supported in the system considering the aggregated traffic from all the expected M2M devices in the cell. Specifically, we show that with the proposed solutions a GSM cell with a single carrier can support up to 70 arrivals per second with an outage below 2%, which supposes an increment of 133% with respect to the legacy system.[1] For example it could support the equivalent of 13k smart meters with a report interval of 5 minutes in addition to the other M2M applications.

The third main result illustrates the performance of the system when the smart meters are triggered almost simultaneously by an alarm event, for example a power outage. With the proposed solutions, GSM is able to cope with up to 1500 synchronously smart meters with an outage below 0.1%. It should be noted that the legacy GSM presents already 10% outage for very low number of synchronous devices.

[1]Note that outage is defined as the probability of a device not being served before it reaches the maximum number of allowed retransmissions.

2.2 Reliable Massive Access in LTE

2.2.1 Paper 3

Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling
Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.
Published in IEEE Wireless Communications Letters, Volumen 3. August 2014.

Motivation

3GPP M2M scenarios foresees up to 30k devices per single cell, which are expected to periodically report small amount of data to a central server. However, the random access of cellular networks where not originally designed for this type of operation. More importantly, current cellular access protocols can be considered as best-effort, or in other words, its performance depends on the current load of the system. We address the question of how many devices can be supported in LTE for a given reliability requirement.

Paper Content

The main contribution of this letter is the proposed mechanism to provide reliable service for M2M in LTE networks. This mechanism consists on a pool of resources that reoccurs periodically over time. The pool comprises two distinct parts, denoted as preallocated pool and the common pool. Every device in the system has a reserved resource in the preallocated pool, which is used to transmit one report and to indicate if there are more reports waiting to be transmitted (denoted as excess reports). On the other hand, the common pool is shared by all the devices to transmit the excess reports and for retransmissions. We determine the maximum number of devices that can be allocated in LTE for a given amount of resources reserved for M2M and a given reliability requirement. In other words, the data resources are split into two: one for M2M and the remaining for other traffic (e.g., human oriented services). Obviously, we aim to minimize the amount of resources that shall be reserved for M2M.

In addition, it should be noted that the analytical model provided is able to determine the individual reliability of every report generated by each device. In order to do so, we exploit the central limit theorem and the total law of probability.

Main Results

We first validate the analysis provided by using an event-driven simulator developed in MATLAB, where a tight match between the analytical model and the simulation results can be observed. In addition, we compare the performance of the proposed scheme with the legacy LTE using typical configuration parameters. The performance metric is given as the percentage of system capacity used for M2M to serve a given number of M2M devices. It should be noted that M2M devices can only access the data resources that have been reserved for M2M. Results show that up to 30k devices can be served requiring only 9% of system capacity (5 MHz system) when the lowest data modulation is considered. On the other hand, legacy LTE would require to reserve twice the amount of resources.

2.2.2 Paper 4

Efficient LTE Access with Collision Resolution for Massive M2M Communications
Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.
Published in IEEE Global Communications Conference (Globecom) Workshops - Ultra-Low Latency and Ultra-High Reliability in Wireless Systems, 2014.

Motivation

LTE random access channel is well suited for uncorrelated report arrivals. However, its performance is severely affected in case of time correlated arrivals, such as in alarm event with thousands of devices triggered simultaneously. It should be noted that if the random access channel collapses, no cell service is possible (including emergency services). Contrary to the mainstream solutions that tries to spread arrivals over time to avoid collisions, we investigate the potential of collision resolution algorithms.

Paper Content

The main contribution of this paper is a collision resolution algorithm implemented on the top of the existing LTE random access procedure. This algorithm is targeted for delay sensitive M2M devices. The mechanism is activated by the eNodeB upon detection of a large number of collisions in the random access channel. Instead of wasting resources and time trying to avoid collisions, the main idea is to resolve the collisions. In this manner an astonishing number of devices can be resolved in a short interval of time. Since providing data resources for thousands of devices might not be feasible (i.e., lack of resources), the proposed scheme can be used to gather short messages from each device (e.g., alarm message identification). By exploiting this information, the eNodeB is able to provide data resources only to critical devices and avoid redundant information.

Main Results

Obviously a typical random access configuration with one random access opportunity every 5 ms is not enough to support thousands simultaneous arrivals. Therefore, we consider the so-called dynamic allocation technique present in LTE, which increases the number of random access opportunities upon overload detection. In order to illustrate the potential of the proposed mechanism, we compare our solution to an ideal dynamical allocation, i.e., there is no delay since the overload is detected until the number of random access opportunities is increased.

The main performance metrics are: delay, average access attempts and the outage probability (defined as percentage of devices not being served before the maximum number of access attempts is reached). We show that the proposed scheme shows zero outage for up to 30k devices with an average of 4 access attempts and average delay of 1.2 seconds. Meanwhile, the dynamic allocation approach already shows 60% outage with 5k devices. Surprisingly, we also show that the proposed mechanism is more efficient than the dynamic allocation for massive arrivals.

2.2.3 Paper 5

Massive M2M Access with Reliability Guarantees in LTE Systems
Germán Corrales Madueño, Nuno K. Pratas, Čedomir Stefanović, and Petar Popovski.
Accepted in IEEE International Conference on Communications (ICC), 2015.

Motivation

Main strategies in LTE to deal with massive arrivals are based on the well-known access class barring and dynamic allocation schemes. The former scheme blocks access to devices belonging to a certain class, while the latter increases the number of random access opportunities upon overload detection. However, these schemes can be considered as reactive approaches, i.e., they only take place once the stability of the system has been compromised. Instead, we investigate a proactive approach, where the number of contending devices are estimated and then the necessary resources are provided. In this manner, the system stability is always guaranteed.

Paper Content

This paper designs a proactive approach to solve the problem of the reliability guarantee in massive M2M communications. The proposed solution is divided into two phases: estimation and serving. In the first phase, the number of M2M arrivals is estimated and then the amount of resources in the serving phase is tuned. Compared with legacy reactive approaches, the proposed mechanism is instrumental in delivering high access reliability to the M2M devices according to a LTE event-driven simulator implemented in MATLAB.

Main Results

We consider a scenario where the number of devices can be as high as 30k devices. These devices are classified into high priority and low priority classes. Devices belonging to the high priority class reports critical data highly correlated in time (e.g., alarm message). On the other hand, devices who belong to the low priority class can be considered as uncorrelated in time. First we show that under the presence of massive arrivals the legacy LTE random access procedure collapses, and the reliability of the system drops below 10% for 30k devices for both classes. On the other hand, the proposed mechanism is able to guarantee high reliability (above 90%) for the high priority class; while the low priority class can be proactively delayed until the “storm” caused by the high priority class has passed. Finally, we note that the proposed approach guarantees the stability of the network at all times.

2.3 Efficient Coexistence of Regular M2M Traffic and Alarm Reporting in IEEE 802.1ah Networks

2.3.1 Paper 6

Reliable Service for Massive Machine-to-Machine (M2M) Communications in IEEE 802.11ah.
Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.
Submitted to IEEE Internet of Things Journal (IoT-J), 2015.

Motivation

IEEE 802.11ah is a new protocol mainly designed for the Internet of Things in the sub-1GHz band. With distances of up to 1 km, it can be considered as a wide area network. Furthermore, the protocol can support slightly more than 8k stations with a single access point (AP). Therefore it is an interesting alternative to cellular networks, which were not originally designed for such large populations. Contrary to current studies that focus on throughput, we investigate the potential of 802.11ah to provide reliable service for M2M. Specifically we focus on the performance of the system in case of an alarm event.

Paper Content

This journal extends the allocation algorithm proposed in paper 3, which consisted on a pool of resources that reoccurred periodically over time. This algorithm was used by stations that reported periodically to a remote server. In this contribution, we consider the case of alarm reporting, which presents shorter deadlines when compared to periodic reporting (typically differs in more than one order of magnitude).

Similarly to Paper 3, we provide a periodically reoccurring pool of resources that consists on a preallocated pool and a common pool. The preallocated pool consists on a set of reservation slots (RS), where $\Omega \geq 1$ stations are allocated in each RS. Under normal operation conditions (i.e., periodic reporting), the probability of more that one station transmitting within the same RS is kept to a tolerable level. If a collision does take place in the RS and depending on the number of observed collided RSs, the AP determines the dominant operation regime (i.e., periodic reporting vs. alarm reporting). In periodic reporting regime, the AP provides an additional frame with L RSs, where station contend.

If the dominant regime is alarm reporting, the AP selects a contention-free access mechanism, where each station in the collided RSs get an unique additional RS. In addition, we investigate how alarm events can be analytically modeled, and we provide useful insights about the beta distribution, which is typically used by 3GPP to model alarm events.

Main Results

We derive an analytical expression to compute the average cost of the proposed mechanism. This cost is based on the Bayes risk for different parameters that can be configured, such as k, Ω and Δ_C. We consider a fully loaded scenario, where the number of stations is set to 8k, which under normal operation conditions report every 5 minutes (i.e., the report interval is 5 min). The alarm scenario corresponds to the power outage in smart metering due to an earthquake propagating at 4 km per second, where the deadline is set to 2.5 s. Therefore, the described scenario requires that every station is provided with at least 120 transmissions opportunities per report interval. However, we show that with the proposed algorithm with an average of 6 RS will suffice. In addition we note that alarms are detected with very high reliability.

2.4 Smart Grid Traffic in Cellular Networks

2.4.1 Paper 7

What Can Wireless Cellular Technologies Do about the Upcoming Smart Metering Traffic?
Jimmy J. Nielsen, Germán Corrales Madueño, Nuno K. Pratas, Čedomir Stefanović, René B. Sørensen and Petar Popovski.
Submitted to IEEE Communications Magazine (Special issue on Internet of Things), 2015.

Motivation

Currently, the main purpose of smart metering is to report the energy consumption to a remote server few times per day and sporadically indicate the state of the device. In this paper, we provide an exhaustive description of all the messages exchanged between these devices and the remote server. This includes the arrival rate, payload size, message frequency and the maximum tolerable delay. In addition, we consider how the traffic will be shaped in the next generation of smart metering, which is expected to play an active role in the stability of the power grid. The main motivation is to determine how current cellular networks, specifically GPRS and LTE, can handle current and next generation of smart grid traffic. One important difference with respect to Paper 1 and Paper 2 is that we focus on what the legacy GPRS has to offer, i.e., no modifications to the system operation.

Paper Content

The contribution of this paper can be divided into three categories: smart grid traffic model, performance of cellular systems and standardization efforts. The first category includes an exhaustive description of the traffic generated by smart meters (SM) based on the information available in OpenSG smart grid network system requirements specification. The traffic model includes the arrival rate, payload size, message frequency and maximum tolerable delay. In addition, we assume that the next generation of smart meters, denoted as enhanced smart meters (eSM), will be similar to wide area measurement system (WAMS) nodes. This allows us to determine the payload and frequency required by the eSM devices. Secondly, we analyze the performance of current cellular systems to handle the aforementioned smart grid traffic, including both traditional SM and eSM. Finally, we describe the on-going effort and possible solutions for the identified limitations to deal with smart grid traffic in cellular networks.

Main Results

We analyze the performance of cellular systems in a 1000 m radius cell with 4500 SM. First, we show that both GPRS and LTE are capable of providing service to SMs with no outage, where outage is defined as the percentage of devices reaching the maximum number of access attempts without being served. This is a very important result, as it indicates that the GPRS infrastructure can be reused for SM. In regards to eSM we note that an LTE system with 1.4 MHz can barely support a penetration of eSM below 2% requiring up to 10 MHz to reach 30% penetration, while GPRS cannot support support eSM at all. This is due to the aggressive report interval of 1 s and large payloads ($\approx$ 4 kb) of eSMs. We show that by considering lower payloads ($\approx$ 400 b) LTE can support up to 20% of eSMs. However, we notice that the impact of the access reservation procedure must be taken into account, where differences of up to 20% can be observed if only data stage is considered. Finally, we show that by increasing the reporting interval of SM, the observability of the power grid can be increased from hourly intervals to minutes.

CHAPTER 3

Discussion

3.1 Spectrum Debate

Many stakeholders see GSM/GPRS as a dying technology. Shutting down GPRS networks will allow mobile operators to reduce costs, where three different cellular network generations are now coexisting (2G, 3G and 4G). Perhaps, more importantly, it will allow to re-farm its spectrum for LTE. However, the facts contradict this line of reasoning, still today an overwhelming 89% of modules are GPRS-based and only 1% LTE capable [7].[1] GPRS is an interesting option for M2M communications, due to its low cost, worldwide ubiquity and technology maturity. In fact, the cost of a GPRS module is similar to a WiFi module, and several times cheaper than a LTE module [20]. As previously shown, many M2M applications require small payloads (in some cases a single byte will suffice) and sporadic transmissions. These traffic characteristics are a good match for the low rate service provided by GPRS. Moreover, cloud radio access network allows to virtualize the air interface of GPRS into the next generation of cellular systems, which is seen as viable manner to keep GPRS [8]. Therefore, it is not surprising the recent focus from industry to continue GPRS for M2M [8]. We believe that with a suitable reengineering as shown in this

[1] The remaining percentage corresponds to other cellular technologies.

thesis, GPRS has a great potential in the years to come as a dedicated M2M network.

3.2 The Potential of Protocol Reengineering

In this thesis, new methods and protocols have been built on the top the existing systems. This does not imply that the research in the thesis has been incremental or comparable to development; but the existing system has been boosted with novel theoretical methods to meet the M2M requirements. The practical implication of such an approach is an easier adoption of the technology by the companies that manufacture the M2M equipment or provide M2M services. We have denoted this procedure as protocol reengineering and we discuss its potential and limitations in the following text.

One clear example of the potential of protocol reengineering is GPRS. We have shown that the access and data stages limitations for massive M2M can be addressed with protocol reengineering to the MAC layer, leaving the physical layer unaltered. When these limitations are addressed, a surprisingly large number of devices can be supported within a single cell. Specifically, a 3-sector cell can approximately support 13k smart meters reporting every 5 minutes, including the expected incoming traffic from a variety of M2M applications. This supposes an increment of 133% with respect to the legacy system. Furthermore, the proposed improvements also enables GPRS for alarm reporting, supporting up to 1500 synchronously activated smart meters within 120 s window.

LTE is clearly superior to GPRS in almost every aspect: data rate, spectral efficiency, delay, number of simultaneous connections, etc. The access procedure is also superior in term of resources, for example a typical RACH configuration in LTE provides 10.8k access opportunities per second. This is approximately one order of magnitude higher than GPRS cell with 6 carriers. Current LTE operation can be considered as best-effort for M2M, since no mechanisms are in place to guarantee access to the network. Therefore, there is a need to reengineer current LTE protocol to introduce such mechanisms.

In one hand, periodic reporting is typically characterized by a Poisson process. The main challenge of periodic reporting is to make efficient use of the network resources at the same time that a reliability is guaranteed. We proposed a mechanism consisting on a pool of resources that reoccurs periodically on time, where transmissions are guaranteed 99.99% of the time. Higher efficiency is achieved by removing the random access procedure and grouping the transmissions over a a report interval. This approach shows less uncertainty, reducing the need of reserved resources for reliable service. In fact the proposed mechanism requires half of the resources when compared to the legacy system. This mechanism is not limited to LTE and can be adapted to other systems.

On the other hand, alarm reporting is characterized by almost simultaneous requests made from a large number of devices, which could potentially affect the stability of the network. More specifically, the RACH becomes overloaded and many collisions will occur. Note that if the RACH collapses, no cell service is possible (including emergency services). The majority of existing mechanisms focus on preventing the overload rather than providing reliable service for M2M. From our perspective, two main approaches can be considered: embrace collisions or estimate number of contending devices. In the first approach, instead of wasting resources and time trying to avoid collisions, we exploit collision resolutions algorithms. We described how a collision resolution mechanism can be implemented on the top of the existing LTE RACH procedure with minimal modifications. This technique is able to resolve 30k devices with an average of 4 access attempts and an average delay of 1.2 s. It should be noted that dynamic allocation technique used by LTE shows 60% of outage with 5k devices. In the second approach, we estimate the number of contending devices, allowing to determine the number of resources required. If the load is above the system capacity, a barring factor is introduced. Therefore, the number of devices accessing the network is known and the stability guaranteed. In this mechanism, the maximum number of supported devices for a given reliability and maximum delay requirements can be determined.

3.2.1 Limitations of Protocol Reengineering

Nowadays, smart metering is mainly restricted to billing purposes, but as shown in this thesis, this might change in the short term. The surge of distributed energy resources, such as solar panes, wind turbines and electric vehicles introduce variations in the power quality. Therefore, there is a need for more frequent monitoring and control of the grid. This can be achieved by introducing eSMs. The main challenges are larger payloads (≈ 4 kb) and aggressive arrival rate (1 report per second) from a massive number of devices. It should be noted that a penetration of eSM around 20% will require up to 10 MHz of bandwidth in LTE, which is not a reasonable solution for a low cost service such as smart metering.[2] If the message size cannot be reduced, the problem become a data rate issue, which cannot be addressed with protocol reengineering. Hence, it is

[2] Note that GPRS cannot support eSMs at all.

to be expected that the next generation of smart metering will a require a new cellular network generation (i.e., 5G).

IEEE 802.11ah protocol poses an interesting alternative to cellular networks. A single AP can provide service to approximately 8k stations within a 1 km cell, and more importantly uses the unlicensed band (i.e., lower operational costs). The main challenge these system shall face is how to serve massive number of stations efficiently, while supporting alarm reports with high reliability. The RAW is a very convenient feature in this protocol, since it allows to limit the number of contending stations in a given interval of time. However, due to the difference in tolerable delay between periodic and alarm reporting, all stations requires multiple transmissions opportunities within a reporting interval. Therefore, the RAW approach becomes very inefficient. With the proposed allocation mechanism, alarms are always detected while a massive number of stations are served efficiently. Further, this mechanism allows to tune how “massive” is the event we aim to detect; enabling the AP to take the appropriate measurements. With the proposed improvements, 802.11ah presents outstanding results both in terms of efficiency and reliability for massive M2M.

Another fundamental limitation of the reengineering, besides rate, is the minimal latency. There could be time constants used in the protocol (slots, uplink/downlink period sequence, etc.) which puts a lower bound on the achievable latency, such that achieving extremely low latency requires redesign from scratch.

CHAPTER 4

Conclusions & Future Work

The main goal of this thesis is to introduce novel mechanisms for evolving current systems to meet M2M service requirements, focusing on (i) ultra-reliable M2M communications and (ii) massive M2M communications. We show that by introducing minimal modifications to current wireless systems, the performance can be boosted. For example, in GPRS, a surprisingly number of devices can be supported in a single cell (133% increase with respect to the legacy system).

Regarding LTE, current operation for M2M can be considered as best-effort, as there are no mechanism in place that guarantee reliable service. In fact, most of the proposed solutions proposed both by the industry and research focus on preventing system overload by blocking M2M devices. However, we show that reliable mechanisms can be introduced into LTE with minimal modifications to the current system operation. Specifically, we propose a pool of resources that reoccur periodically over time with the aim of providing ultra-reliable service. Furthermore, contrary to main stream solutions that try to avoid collisions by blindly spreading transmissions over time, we consider two main approaches: to embrace collisions and to estimate the number of contending devices. Both approaches are significantly better than legacy solutions, supporting up to 30k devices per cell reporting simultaneously with no outage. On the other hand, legacy solutions present more than 50% of outage already with 5k devices.

IEEE 802.11ah protocol poses an interesting alternative to cellular networks. We propose a mechanism that resolve all alarms while efficiently allocates large number of devices. With the proposed improvements, 802.11ah exhibits outstanding results both in terms of efficiency and reliability for massive M2M.

Finally, we note that there are certain limitations that protocol reengineering cannot overcome. Nowadays, smart metering is mainly restricted to billing purposes but this might change in the short term. The surge of distributed energy resources, such as solar panes or wind turbines, will require faster quality measurements. Thus, we expect smart meters to evolve in order to provide this information (denoted as eSMs), requiring large payloads and more aggressive reporting intervals. We show that too many resources are required in LTE for eSMs. Therefore, if these requirements cannot be relaxed (for example reducing the payload), faster and better systems such as 5G are required.

4.1 Future Work

There are a considerable number of research questions regarding wireless systems and M2M which have not been addressed in this thesis. One line of investigation is the analysis of the latency/performance achieved at the higher layers. Another interesting consideration is the impact of the proposed schemes on the energy consumption, which is a fundamental aspect of M2M. In addition, the usage of sophisticated random access schemes and its potential in GPRS or LTE to deal with massive arrivals, such as coded random access. These schemes use advanced techniques as successive interference cancellation at the physical layer. Finally, the potential of cloud radio access networks (cloud RAN) and its impact on current wireless systems is a very promising and worthy research.

References

[1] Hugo Tullberg, Petar Popovski, David Gozalvez-Serrano, Peter Fert, Zexian Li, Andreas Hoglund, Heinz Droste, Omer Bulakci, Joseph Eichinger, and Krystian Pawlak. "METIS System Concept: The Shape of 5G to Come". In: *Submitted to Communications Magazine, IEEE* (2015).

[2] 3GPP. *Study on RAN Improvements for Machine-type Communications*. TR TR-37868. 3rd Generation Partnership Project (3GPP), 2011.

[3] 3GPP. *Smart Grid Traffic Behavior Discussion*. TR R2-102340. 3rd Generation Partnership Project (3GPP), 2010.

[4] IEEE 802.16p. *IEEE 802.16p Machine to Machine (M2M) Evaluation Methodology Document (EMD)*. EMD 11/0014. IEEE 802.16 Broadband Wireless Access Working Group (802.16p), May 2011.

[5] IEEE 802.16p. *IEEE 802.16p Machine to Machine (M2M) Evaluation Methodology Document (EMD)*. EMD 11/0005. IEEE 802.16 Broadband Wireless Access Working Group (802.16p), 2011.

[6] Open Smart Grid SG-Network task force. *Smart Grid Networks System Requirements Specification, Release Version 5*. 2010.

[7] AB Ericsson. *Interim Ericsson Mobility Report, February 2014*. 2014.

[8] *Overview of 3GPP Release 13*. Tech. rep. 3GPP, 2015.

[9] 3GPP. *RACH Capacity Evaluation for MTC*. TSG GERAN #46 GP-100892. 3rd Generation Partnership Project (3GPP), 2010.

[10] 3GPP. *Downlink CCCH Capacity Evaluation for MTC*. TSG GERAN #46 GP-100893. 3rd Generation Partnership Project (3GPP), 2010.

[11] 3GPP. *USF Capacity Evaluation for MTC*. TSG GERAN #46 GP-100894. 3rd Generation Partnership Project (3GPP), 2010.

[12] A. Laya, L. Alonso, and J. Alonso-Zarate. "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives". In: *IEEE, Communication Surveys Tutorials* 16.1 (2014), pp. 4–16. ISSN: 1553-877X. DOI: 10.1109/SURV.2013.111313.00244.

[13] 3GPP. *Backoff enhancements for RAN overload control*. TR R2-112863. 3rd Generation Partnership Project (3GPP), May 2011.

[14] 3GPP. *Radio Resource Control (RRC) protocol specification*. TR 36.331. 3rd Generation Partnership Project (3GPP).

[15] Ki-Dong Lee, Sang Kim, and Byung Yi. "Throughput Comparison of Random Access Methods for M2M Service over LTE Networks". In: *IEEE GC Workshops*. Dec. 2011, pp. 373–377.

[16] Ali Hazmi, Jukka Rinne, and Mikko Valkama. "Feasibility study of IEEE 802.11 ah radio technology for IoT and M2M use cases". In: *Globecom Workshops (GC Wkshps), 2012 IEEE*. IEEE. 2012, pp. 1687–1692.

[17] Toni Adame, Albert Bel, Boris Bellalta, Jaume Barcelo, Javier Gonzalez, and Miquel Oliver. "Capacity analysis of IEEE 802.11 ah WLANs for M2M communications". In: *Multiple Access Communications*. Springer, 2013, pp. 139–155.

[18] Lei Zheng, Lin Cai, Jianping Pan, and Minming Ni. "Performance analysis of grouping strategy for dense IEEE 802.11 networks". In: *Global Communications Conference (GLOBECOM), 2013 IEEE*. 2013, pp. 219–224. DOI: 10.1109/GLOCOM.2013.6831074.

[19] Chul Wan Park, Duckdong Hwang, and Tae-Jin Lee. "Enhancement of IEEE 802.11ah MAC for M2M Communications". In: *Communications Letters, IEEE* 18.7 (2014), pp. 1151–1154. ISSN: 1089-7798. DOI: 10.1109/LCOMM.2014.2323311.

[20] Geoff Sarnay. "Smart Energy for Smart Cities Webinar". In: *Groupe Speciale Mobile Association (GSMA)* (2012).

[21] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "How Many Smart Meters can be Deployed in a GSM cell?" In: *International Conference on Communications (ICC) 2013 - Workshop Telecom R2S*. Budapest, Hungary, June 2013, pp. 1283–1288.

[22] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering". In: *International Conference on Smart Grid Communications (SmartGridComm) 2014*. 2014, pp. 338–343.

[23] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling". In: *IEEE Wireless Communications Letters (WCL) 2014* 3.4 (2014), pp. 429–432. ISSN: 2162-2337.

[24] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Efficient LTE Access with Collision Resolution for Massive M2M Communications". In: *IEEE Global Communications Conference 2014 (GLOBECOM) - Workshop Ultra2*. 2014.

[25] Germán Corrales Madueño, Nuno Pratas, Čedomir Stefanović, and Petar Popovski. "Massive M2M Access with Reliability Guarantees in LTE Systems". In: *International Conference on Communications (ICC) 2015*. 2015.

[26] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Efficient Coexistence of Regular M2M Traffic and Alarm Reporting in IEEE 802.1ah Networks". In: *Internet of Things Journal, IEEE* (2015).

[27] Jimmy Nielsen, Germán Corrales Madueño, Nuno Pratas, Čedomir Stefanović, René B. Sørensen, and Petar Popovski. "What Can Wireless Cellular Technologies Do about the Upcoming Smart Metering Traffic?" In: *Communications Magazine - Special issue on Internet of Things, IEEE* (2015).

[28] P. Popovski, G.C. Madueño, L.C. Gimenez, L. Luque Sanchez, and N.-C. Gjerrild. "Protocol coding for reliable wireless bits under jamming: Concept and experimental validation". In: *Military Communications Conference, 2011 - MILCOM 2011*. 2011, pp. 113–118.

[29] Germán Corrales Madueño, Nuno K Pratas, Henning Thomsen, Čedomir Stefanović, and Petar Popovski. "Massive Machine-to-Machine Communications". In: *European Telecommunications Standards Institute (ETSI) M2M Workshop*. Sophia Antipolis, France, Oct. 2012.

[30] Germán Corrales Madueño, Nuno K Pratas, Čedomir Stefanović, and Petar Popovski. "2G Cellular Networks Reengineering for massive number of devices with Low-Rate". In: *J.No. 2012-521/10-0291*. 2012.

Contributions In This Thesis

Contents

Paper 1

How Many Smart Meters can be Deployed in a GSM cell?

Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.

Presented at the IEEE International Conference on Communications 2013: IEEE ICC'13 - Second IEEE Workshop on Telecommunication Standards: From Research to Standards. Budapest, Hungary, 2013.

How Many Smart Meters can be Deployed in a GSM cell?

Germán Corrales Madueño, Čedomir Stefanović, Petar Popovski

Department of Electronic Systems, Aalborg University, Denmark
Email: {gco,cs,petarp}@es.aau.dk

***Abstract*—The need to deploy large number of wireless devices, such as electricity or water meters, is becoming a key challenge for any utility. Furthermore, such a deployment should be functional for more than a decade. Many cellular operators consider LTE to be the single long term solution for wide area connectivity serving all types of wireless traffic. On the other hand, GSM is a well-adopted technology and represents a valuable asset to build M2M infrastructure due to the good coverage, device maturity, and low cost. In this paper we assess the potential of GSM/GPRS/EDGE to operate as a dedicated network for M2M communications. In order to enable M2M-dedicated operation in the near future, we reengineer the GSM/GPRS/EDGE protocol in a way that requires only minor software updates of the protocol stack. We propose different schemes to boost the number of M2M devices in the system without affecting the network stability. We show that a single GSM cell can support simultaneous low-data rate connections (e. g. to smart meters) in the order of 10^4 devices.**

I. Introduction

In the emerging area of machine-to-machine (M2M) communications, smart metering is a showcase application: a large number of electricity/water/heat meters use sophisticated wireless networking for two-way communication with a central controller/data collector. The usage of wireless techniques for M2M communication has been made possible due to the level of maturity attained by the wireless technologies: small, inexpensive embedded devices have significant computational power and operate at very low power levels.

M2M communication has significantly different requirements from, e. g. human-to-machine (H2M) services (download, web browsing, video streaming), where large data volumes are sent and high data rate is required. In majority of the scenarios, M2M communication is based on intermittent transmission/reception of small data portions and pose requirements that are different from the ones according to which the common wireless protocols are designed. Some of the most important requirements are the following:

- Transmission from a massive number of devices and maintenance of a large number of active connections;
- Ability to send a small amount of data while decreasing the overhead percentage;
- Real-time communication with low latency;
- Certain connections that carry critical control data require a high degree of reliability, such that a connection should be kept alive more than 99.XX % of the time.

These requirements become more challenging when one considers the forecasts that state that by 2020 there will be 50 billion M2M connected wireless devices [1], spanning a wide application range: smart grid, smart metering, control/monitoring of homes and industry, e-health, etc. While there are many ongoing standardization activities [2], M2M communication solutions have started to be deployed through the existing cellular interfaces, such as GSM and LTE. Specifically, there is an increasing deployment of cellular-based wireless smart meters, such as the ones based on GPRS [3]. Some of those deployments are very large, such as Hydro-Quebec in Canada [4], with about 3.8 million devices that periodically send only a few bytes (KW/h consumption for instance). Another example is happening in Spain and Portugal, where Endesa, the largest Iberian operator, will replace a total of 13 million electric meters with smart meters by 2018 [5]. Since neither GSM nor LTE are originally designed to support massive M2M communication, there are ongoing research and standardization activities to modify those interfaces, notably LTE, in order to support the M2M traffic characteristics [6].

In this paper we explore the potential of GSM to operate as an M2M-dedicated network and support a large number of active connections. The advantages of GSM as a basis for a future M2M infrastructure include ubiquitous coverage, worldwide frequency availability, device maturity, and low cost. In our approach, the GSM[1] protocol is reengineered in a way that the physical layer is kept intact, the frame sent over the air has a backward-compatible structure, while changes are made in the algorithms and the local protocol-related variables that are used by the Base Station and the M2M devices. The key idea is to enable the communication nodes to address/use the resources sent over the air with a finer level of granularity. Such an approach allows smooth evolution of the current 2G networks towards M2M-dedicated networks. We have analyzed and evaluated the capacity of the reengineered GSM network to support the M2M traffic patterns prescribed by 3GPP [6]. The evaluation is made for a traffic with limited latency, such that if a packet is in a queue for excessive time, it is dropped. Our initial results are very promising, confirming that it is possible to scale GSM towards a regime in which a single cell and over a signal

[1]Hereinafter, we use the generic term GSM to refer to data services on GSM, GPRS, and EDGE.

978-1-4673-5753-1/13/$31.00 ©2013 IEEE

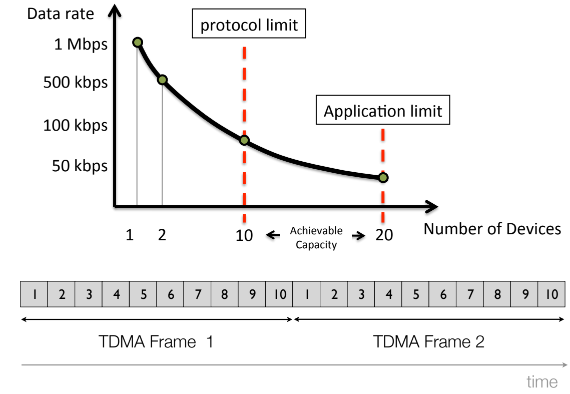

Fig. 1: Ideal system in which the bandwidth is shared among the multiplexed devices. The protocol operation is limiting the number of devices, despite the application requirements.

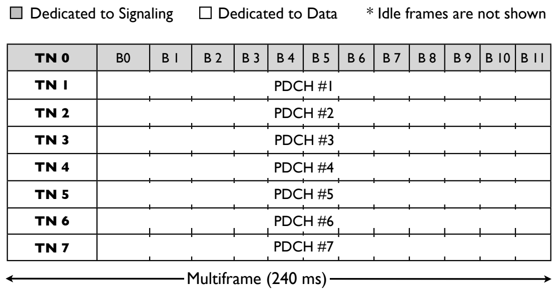

Fig. 2: Multiframe structure composed by 7 time slots (TN0 to TN7) and 7 Packet Data Channels (PDCH).

frequency channel, the number of active low-rate connections is very large ($> 10^4$). Moreover, we show that our system can achieve a good tradeoff, in a large range, between the packet dropping probability and the total number of meters served in a cell.

The rest of the paper is organized as follows. Section II presents the relevant aspects of GSM that have influence on the analysis and the major limitations in regards to M2M. Section III, gives in a compressed manner what is to be modified in the current GSM system. In Section IV, we present the mathematical model used to characterize the system. In Section V the tradeoff between the number of meters and the dropping probability that can be achieved in a GSM system is presented. The last section provides a conclusion.

II. Protocol Limitations for Massive M2M Communication

In this section we first discuss the generic case of TDMA system, and illustrate the mismatch between application requirements and protocol design. Then, we briefly describe the related limitations of the GSM radio access.

A. Protocol Limit in a Generic TDMA System

Ideally, a TDMA system should be able to allocate as many as possible devices as long as the quality of service is guaranteed. The target operation can be described as follow: if a nominal data rate of a system is R bps and there are N potential users, in the ideal case the bandwidth is shared such that each device gets a data rate of R/N bps. At the same time, the limit on N should be posed by the application requirements, rather than protocol limitations. However, in practice, systems are typically not able to operate in this manner.

For example, consider an example TDMA system in which a time frame consists of 10 slots. Each slot has a duration of 1 ms and it carries 10 Kbits; the total available bandwidth is thus 1 Mbps. Further, assume that a device can be allocated from one to all ten slots, while the minimum bandwidth needed by a device to properly operate is 50 Kbps. Fig. 1 illustrates how the system operates: when the system is empty and a new device arrives, the full bandwidth (i.e., 1 Mbps) is allocated to the only user. Each time a new device arrives, the system re-allocates the slots so that all devices get a data rate of $1/N$ Mbps, where N is current number of devices admitted in the system. A data rate of 100 Kbps per device is provided when the system if full, when there are 10 devices being served. Due to *protocol limitation, i.e., the granularity of available resources*, any new arrival is rejected, even though the bandwidth is enough to serve up to 20 devices.

B. GSM Protocol and its Limitations for M2M

The GSM radio-link is TDMA based, where both uplink and downlink are organized in *multiframes*. A multiframe has duration of 240 ms and is composed of 12 radio blocks and 8 time slots. Its structure is depicted in Fig. 2. A radio block is the minimum amount of information that can be sent or received by the devices. Time slots are independent TDMA channels, usually denoted as Packet Data Channels (PDCHs) that can be used for signaling or data-transmission; in this paper we assume a typical configuration, where PDCH #0 is dedicated to signaling, while the remaining seven PDCHs carry data.

The phases of the radio access in GSM are resource request, data transmission/reception and resource release. The resource-request procedure is as follows. First, a device accesses the medium by sending a request message in a so-called Packet Random Access Channel (PRACH), which is a logical channel defined over a chosen PDCH. Upon reception of the PRACH request, the base station grants the resource request assigning the resources in one or more PDCHs. Alternatively, the base station can reject a request if there are no free resources. The grant message is carried in another logical channel - the Packet Access Grant Channel (PAGCH). This message contains the PDCH and the Uplink Status Flag (USF) allocated to the device. Finally, the device is allowed to transmit only in block $k + 1$ of the allocated PDCH in the uplink if its USF was announced in block k of the same PDCH in the downlink.

Through the above stages of a random access procedure, 3GPP recommends that a blocking probability of 2% per stage should not be exceeded [7]. Furthermore, a numerical comparison of the bottlenecks in the respective stages of the procedure is presented in [7], where the USF constraint happens to be the most limiting factor. Particularly, as a USF is 3 bits long and the value 000 is reserved, a maximum of 7 devices can be multiplexed per PDCH, resulting in a maximum of 49 simultaneously allocated devices (i.e, 7 USFs in 7 data PDCHs)[2].

III. Reengineering the GSM system

In this section we propose the solution to overcome the USF limitations by allocating/multiplexing more devices per PDCH. The allocation space is expanded by reinterpretation the usage of USF.

The main idea is that a USF value does not hold for all the blocks of the allocated PDCHs anymore, but is valid only for a subset of blocks, during a predefined set of reoccurring multiframes. Hence, several devices could be identified with the same USF in the same PDCH and block, but the method is collision free as devices are allocated in different multiframe sets, and thus orthogonally in time.

The flow of the procedure is represented in Fig. 3. The allocator for the device now consists of the following parameters: PDCHs, USF, K and M, where K is the number of blocks per PDCH for which the USF is valid and M is the period by which the allocated multiframes reoccur. We assume that M and K are equal for all devices (i.e., we consider a case with homogeneous traffic), these parameters can be broadcast in the Cell Information messages. The device is only allowed to transmit if the USF received in the downlink matches its own in the specific combination of multiframes, PDCHs and blocks; this approach allows for much finer granularity than in the ordinary GSM case.

For a better understanding we provide an example (see Fig. 4) in which 4 devices are multiplexed into a single PDCH with 3 USFs. The parameter M is set to two, therefore, devices have a chance to transmit every second multiframe. Devices #1, #2 have been granted access in multiframe 1 in block 0 with $USF1$, $USF2$ respectively. They all transmit without collisions because they have been assigned to different USF values (this is an example of the ordinary GSM multiplexing strategy). Then, in multiframe 4, device #3 gets the same USF value than device #1 in the same block 0. No collision occurs because #1 transmits in even multiframes, while #3 transmits only in odd multiframes. Finally, in multiframe 7, a new device, device #4, arrives, for which the base station allocates block 1 with USF3, again, no collision occurs as both devices are assigned to independent channels (block 0 and block 1).

[2]We assume that only one radio channel is used. In general case, if there are B radio channels, then a maximum of $49B$ devices can be admitted in the system.

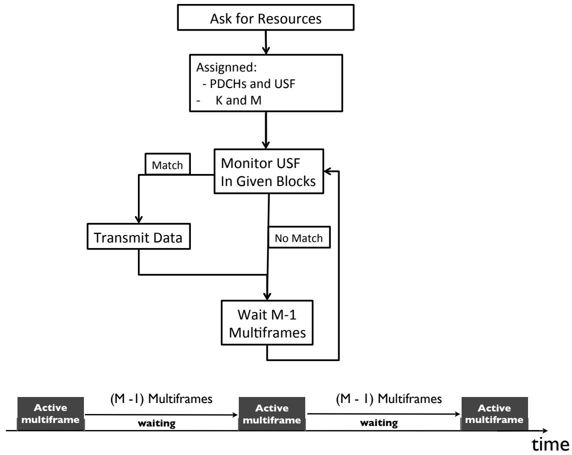

Fig. 3: Flow-chart outlining the proposed method.

IV. Analysis

In this section we first describe a system that can host a large number of smart meters at a target data rate. Moreover, we introduce the requirements of smart meters. Subsequently, we present the maximum achievable capacity in the system for a given amount of resources. Later, we introduce the analytical tools to analyze the performance in terms of reliability.

A. System Model

The system can be described as follows: first the device requests resources, indicating the amount of data needed and the reporting interval RI. Then the base station grants access and schedules the device to transmit in specific time intervals, allowing an efficient sharing of the resources among all the devices in the cell in a coordinated way. The system essentially operates in a circuit-switched manner, where each meter is allocated a portion of the link time. The circuit-switched structure is periodically reestablished (e. g., once a day, week, month, etc.) through the PRACH procedure. The resources used for PRACH are considered negligible.

Smart meters are expected to be installed in fixed locations with high device intensity, where the traffic patterns correspond to device-originated transmissions with small payload (in the range of 100 to 1000 bytes) and periodical reporting in 1 min, 5 mins, 15 mins, 1 hour, 6 hours, 12 hours and 24 hours [8]. In addition, devices tolerate a delay up to the next scheduled transmission opportunity if the message was not successfully delivered. Furthermore, a delay of up to 1 min is allowed in event-triggered alarms [8].

In this paper, we consider a delay tolerance equal to the report interval; any message arriving beyond the delay tolerance (deadline) is dropped and counted as a lost report. The loss probability is defined as the long-run fraction of messages that do not arrive after the deadline.

We model the traffic by a Poisson process as defined by 3GPP [9] [10] and IEEE 802.16p Machine to Machine group

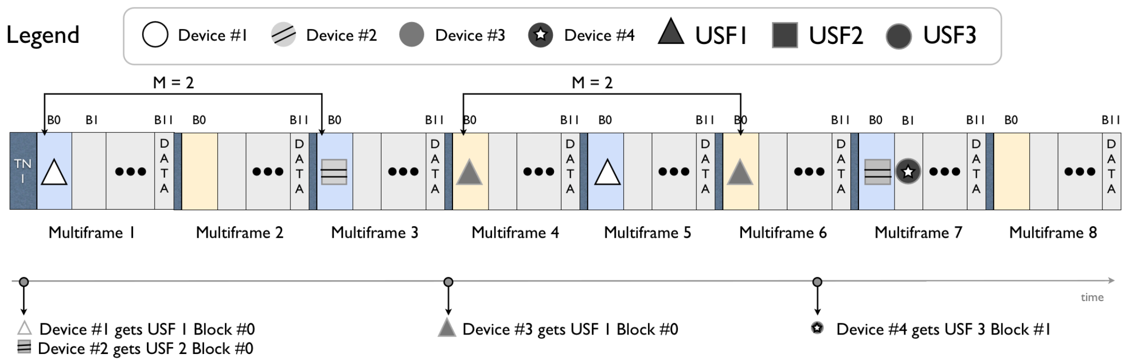

Fig. 4: Example of the proposed expansion of the allocation space.

[11]. The arrival rate is given by $\lambda = 1/RI$, where RI is the report interval. We assume that each device has a packet of constant size RS. Also, each of the devices is allocated a portion of the link-time, as elaborated in Section III. The radio link conditions are assumed to be ideal (i.e., no channel-induced errors are considered).

B. Achievable Capacity in the modified system

The total capacity of the cell (i.e., the number of simultaneous active connections) is determined by the amount of resources needed by each of the devices allocated in the system. The resources are given in terms of blocks needed to transport a report of size RS bytes and the used modulation coding scheme (MCS). Thus, the total capacity is[3]:

$$C = M \cdot \frac{X \text{ Users} \cdot 12 \text{ Blocks} \cdot L \text{ PDCHs}}{\lceil \frac{RS}{MCS} \rceil} \text{ [Devices]}, \quad (1)$$

where L is the number of "data" PDCHs and X the number of USFs available.

In Table I, we show the total cell capacity for different values of M, and the amount of information carried out depending on the coding scheme used and L is the number of PDCHs present in the cell. In addition, the tradeoff between number of devices and the time between transmissions is shown. The larger M is, the more devices are allocated in the system. However, more devices implies larger time between scheduled opportunities T. The duration of this period is given by:

$$T = M \cdot X \cdot 0.24 \text{ [s]} \quad (2)$$

where 0,24 corresponds to the duration of a multiframe expressed in seconds. In addition, the resources given to any device should preserve the reporting interval. Thus, we set a constraint on T:

$$T \leq RI \quad (3)$$

[3]Again, we assume a single frequency.

M Value	Bytes per Transmission MCS1	MCS5	Time Between Scheduled Opportunities (T)	Capacity (devices)
18	22	56	30 s	≈ 10.5 K
35	22	56	≈ 1 min	≈ 20 K
175	22	56	≈ 5 min	≈ 102 K
2100	22	56	≈ 1 h	≈ 1.2 M

TABLE I: Total cell capacity with M configured to example reporting intervals in smart metering when single frequency is used. The assumed values of X and L are 7.

C. Queue Model

The total cell capacity only indicates what is the number of devices provided with a transmission opportunity. However, the actual system capacity depends also on the delay tolerance of the devices, and might significantly differ from the results displayed in Table I. In this subsection we analyze the probability that the report delivery time exceeds the delay tolerance (i.e., loss probability), which can be used to calculate what is the number of devices that can be effectively served.

As elaborated before, we assume that all devices are pre-allocated for service each M multiframes - we assume that the duration of this period is equal to T seconds. Further, we assume that the amount of resources given to a device is limited to a single report, and that the delay tolerance T_d within which the report has to be sent is also equal to reporting interval of the smart meter (RI):

$$T_d = RI \quad (4)$$

any report with delivery time larger than the delay tolerance is dropped and considered as lost.

The above system can be modeled as a multiple vacation queue with limited service and impatient customers (i.e. a queue with reneging). In this system, a customer is only willing to wait up to T_d seconds, after which becomes a lost

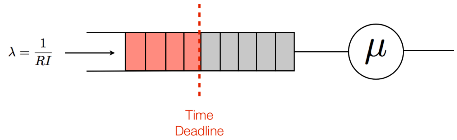

Fig. 5: M/D/1 Queue with impatient customers used to model the system. The service time μ depends on the system parameters M and MCS.

customer. Each time the server visits the queue it serves a single customer if the queue is not empty, and then goes on vacation. If the queue is empty, the server goes to vacation immediately.

The literature on queuing systems with vacations, limited service and reneging is scarce and, to the best of our knowledge, offers no closed form solutions that could be applied to the case in consideration. In order to analytically model the above system, we consider an approximation in which the service time is T seconds. In other words, we approximate a server with vacations with a server with service rate of $\mu = \frac{1}{T}$; as shown later, this approximation will actually yield and upper bound on the loss probability. We note that similar assumptions have been used to model the service rate in GSM in [12].

Without loss of generality we assume that each device sees a dedicated channel, as the service rate does only depend on the parameter M. Thus, the system can be modeled as a queue with a deterministic service time with impatient customers. The number of reports initiated by the device per second is given by a Poisson process $P \sim \text{Poisson}(\lambda)$, where the arrival rate λ is given by the inverse of the reporting interval, i.e., $\lambda = 1/RI$.

The (report) loss probability P_L is the probability that the sum of its waiting time in the queue and the time required for its service exceeds RI, and it can be calculated by the means of the steady-state cumulative distribution function (cdf) of the amount of unprocessed work $U(x)$ as [13]:

$$P_L = 1 - U(\mu\tau) \tag{5}$$

where $\tau = T_d - \frac{1}{\mu}$.

The amount of unprocessed work in the modified queue can be expressed in terms of the amount of unprocessed work in a standard M/D/1 queue U_∞(x) [14]:

$$U(x) = \frac{U_\infty(x)}{1 - \rho + \rho U_\infty(\mu\tau)} \quad \text{for } 0 \leq x \leq \mu\tau \tag{6}$$

where, ρ is the utilization factor equal to $\rho = \frac{\lambda}{\mu}$.

By the PASTA property [15], the amount of unprocessed work $U_\infty(x)$ is identical to the waiting time of a customer when service is first-come, first-served. The waiting time distribution $W_\infty(x)$ in a M/D/1 system can be written as [16]:

$$W_\infty(x) = (1-\rho)\sum_{j=0}^{q}(-1)^j \frac{(\rho x - \rho j)^j}{j!} e^{\rho(x-j)} \tag{7}$$

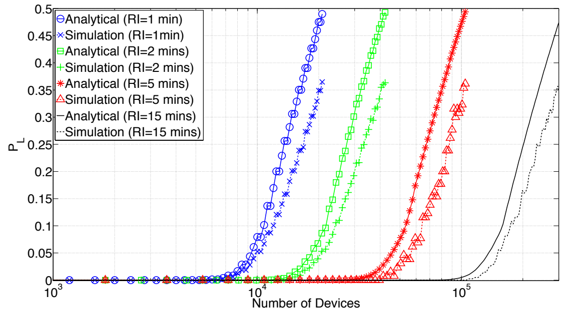

Fig. 6: Comparison between event driven system simulation and the analytical model.

where q is the largest integer less than or equal to x, $q = \lfloor x \rfloor$.

Finally P_L is:

$$P_L = \frac{1}{\rho}\cdot \left[\rho - 1 + \left\{1 + \rho\, e^{(\lambda\tau)} \sum_{j=0}^{q}(-1)^j \frac{(\lambda\tau - \rho j)^j}{j!}\; e^{(-\rho j)}\right\}^{-1}\right] \tag{8}$$

Before we proceed with the performance analysis in Section V, we provide a comparison between a simulation of the proposed system behavior and the presented approximate analysis in Fig. 6. From Fig. 6 it can be observed that the analysis provides an upper bound on P_L.

V. Results

In this section we investigate the inverse relationship between the number of devices and the P_L that can be achieved in a GSM for the example cases of smart meters with report sizes of 100 and 500 bytes and report intervals 1, 2, 5 and 15 minutes. We assume that the delay tolerance is equal to RI and that the devices use modulation coding scheme MCS-5 (56 bytes per block).

Fig. 7 presents the results when report size is 100 bytes. It is noticeable that, for the most demanding case when RI=1min, a single cell could provide service for up to $5{\cdot}10^3$ simultaneous connections with a reliability of 99.99%. This number rises to outstanding value of $5 \cdot 10^4$ simultaneous connections that are served with 99.99%, if the reporting interval is set to 15 min.

Finally, the scenario when the report size is 500 bytes is illustrated in Fig. 8. The effects of carrying larger messages are obvious, as compared to Fig. 7; now the base station has to provide more blocks per device, which decreases the capacity. Nevertheless, the system is able to provide service to 10^4 active devices with at least one transmission opportunity each 15 min.

Finally, we note that the above results represent lower bounds on the number of devices that could be supported in a cell; for the given P_L the actual number could be even higher (see Fig. 6).

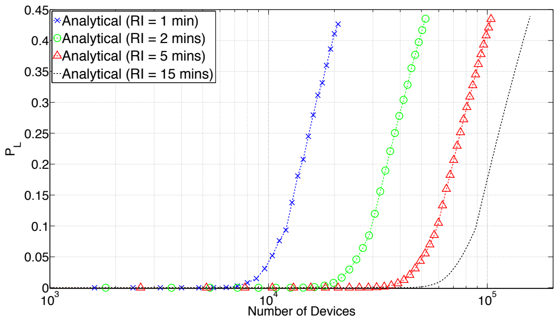

Fig. 7: Loss probability P_L as function of report interval RI, report size 100 bytes.

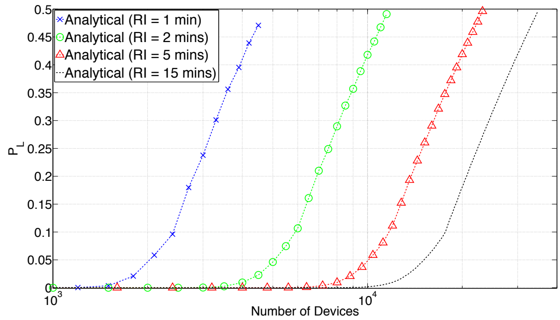

Fig. 8: Loss probability P_L as function of report interval RI, report size 500 bytes.

VI. Conclusions

In this work, we presented a method to reengineer the way devices are allocated resources in GSM, providing for much finer granularity and thus extensively boosting the capacity of the system. Particularly, we have shown that the proposed method for expanding the allocation space allows for support of number of devices that is several orders of magnitude higher than it is achievable in the standard GSM use. The suggested modifications incur only modest changes at the MAC layer, while the physical layer (i.e., air interface) remains unaltered. Moreover, the proposed solution can coexist with the current standard - a subset of the available PDCHs in the cell can be operated in the proposed fashion, while the rest can be operated in the standard way. Further, we demonstrated the performance of the proposed method for the example case of smart metering applications, where the delivery of meter reports is subject to deadlines, by deriving the analytical bounds on the report loss probability and calculating the corresponding call capacities. The obtained results show the considerable potential of GSM to serve as a carrier for smart metering applications.

Acknowledgment

The research presented in this paper was supported by the Danish Council for Independent Research (Det Frie Forskningsråd) within the Sapere Aude Research Leader program, Grant No. 11-105159 Dependable Wireless bits for Machine-to-Machine (M2M) Communications.

References

[1] Q. D. Vo, J. P. Choi, H. M. Chang, and W. C. Lee, "Green perspective cognitive radio-based m2m communications for smart meters," in *Information and Communication Technology Convergence (ICTC), 2010 International Conference on*. IEEE, 2010, pp. 382–383.
[2] L. X. D. Niyato and P. Wang, "Machine-to-machine communications for home energy management system in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 53–59, 2011.
[3] *Sierra Wireless Product Webpage.*, Accessed on October 2012, http://www.sierrawireless.com/en/Solutions/customer_stories/EDMI.aspx.
[4] *Quebec Press Release Smart Metering.*, Accessed on December 2012, https://www.rogers.com/web/link/showNewsDetail?fromWhere=linkInRSSXml&rssBusiUnit=W&NewsID=1906177072.
[5] *Endesa Press Release Smart Metering.*, Accessed on January 2012, http://www.endesasmartgrid.com/index.php/en/smart-homes/smart-metering.
[6] 3GPP, "Service Requirements for Machine-Type Communications (Stage 1)," 3rd Generation Partnership Project (3GPP), TS 22.368, Jun. 2010. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/23368.htm
[7] 3GPP, "Bottleneck Capacity Comparison for MTC," 3rd Generation Partnership Project (3GPP), TSG GERAN #46 GP-100893, 2010.
[8] 3GPP, "RACH intensity of Time Controlled Devices," 3rd Generation Partnership Project (3GPP), TSG RAN WG2 R2-102340, 2010.
[9] 3GPP, "RACH Capacity Evaluation for MTC," 3rd Generation Partnership Project (3GPP), TSG GERAN #46 GP-100892, 2010.
[10] 3GPP, "USF Capacity Evaluation for MTC," 3rd Generation Partnership Project (3GPP), TSG GERAN #46 GP-100894, 2010.
[11] IEEE 802.16p, "IEEE 802.16p Machine to Machine (M2M) Evaluation Methodology Document (EMD)," IEEE 802.16 Broadband Wireless Access Working Group (802.16p), EMD 11/0014, May 2011.
[12] H. Dahmouni, B. Morin, and S. Vaton, "Performance modeling of gsm/gprs cells with different radio resource allocation strategies," in *Wireless Communications and Networking Conference, 2005 IEEE*, vol. 3, march 2005, pp. 1317 – 1322 Vol. 3.
[13] A. De Kok and H. Tijms, "A queueing system with impatient customers," *Journal of applied probability*, pp. 688–696, 1985.
[14] H. Tijms, *A first course in stochastic models*. Wiley, 2003.
[15] W. W. Ronald, "Poisson arrivals see time averages," *Operations Research*, vol. 30, no. 2, pp. 223–231, 1982.
[16] V. Iversen and L. Staalhagen, "Waiting time distribution in M/D/1 queueing systems," *Electronics Letters*, vol. 35, no. 25, pp. 2184–2185, 1999.

Paper 2

Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering

Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski

Presented at the IEEE International Conference on Smart Grid Communications 2014. Venice, Italy, 2014.

Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering

Germán Corrales Madueño, Čedomir Stefanović, Petar Popovski

Department of Electronic Systems, Aalborg University, Denmark
Email: {gco,cs,petarp}@es.aau.dk

***Abstract*—GSM is a synonym for a major success in wireless technology, achieving widespread use and high technology maturity. However, its future is questionable, as many stakeholders indicate that the GSM spectrum should be re-farmed for LTE. On the other hand, the advent of smart grid and the ubiquity of smart meters will require reliable, long-lived wide area connections. This motivates to investigate the potential of GSM to be evolved into a dedicated network for smart metering. We introduce simple mechanisms to reengineer the access control in GSM. The result is a system that offers excellent support for smart metering, as well as the other massive machine-to-machine traffic patterns that are envisioned in 3GPP.**

I. Introduction

Smart metering is a key machine-to-machine (M2M) application, where meters autonomously report usage and alarm information to the grid. It requires sending of low amounts of data from a very high number of meters, differing significantly from the high data rate requirements in human-oriented services. Cellular networks are mainly optimized for the latter, and the amendments required for the effective support of M2M services, including smart grid services, came only recently in focus of the standardization [1]. As a general observation, the key technical problem in enabling M2M communications is not how to increase the overall system data rate, but how to distribute it efficiently to many terminals.

The clear dominance of 2G based solutions in M2M [2] motivates us to investigate if and how GSM networks can be evolved into efficient smart metering networks. Our findings shows that, with a suitable reengineering, GSM networks can support a surprisingly massive M2M devices at even a single frequency channel. This suggest that it is viable to keep one or few GSM channels for M2M operation in the coming years, and thus take advantage of its maturity, low cost and worldwide availability.

The paper is centered on the assessment of the stages of the GSM radio access, which are random access (RACH), access granted (AGCH) and data transmission (DATA). Specifically, we elaborate the operation and the limitations of the GSM access and propose enhancements of the AGCH and DATA stages, with the aim of supporting large number of smart meters per cell. We also present a model of GSM radio access that considers the interstage dependencies, and show that the adopted 3GPP methodology, where the access stages are treated independently [3], [4], leads to unreliable results. Therefore, besides the main message of the paper, which endorses dedicated networks for smart metering based on GSM, the findings presented in the paper constitute an important contribution to the M2M-related 3GPP standardization process.

The rest of the paper is organized as follows. Section II describes the referent M2M traffic scenario. Section III provides an overview of the GSM random access procedure. Section IV presents the proposed approach for the GSM access reengineering. Section V describes the model used to asses the system performance and points out the shortcomings of the methodology used by 3GPP. Section VI presents the results. Section VII concludes the paper.

II. Traffic Models and General Requirements for Machine Type Communications

To assess the potential of GSM as smart metering network, we consider a referent M2M traffic scenario from [5]. The scenario includes the traffic originating both from smart meters and from commercial and home devices. The devices are deployed in a sub-urban GSM cell with a radius of 1000 m and three sectors.[1] The traffic parameters considered are the average message size, the average message arrival rate, and the arrival distribution. Table I summarizes these parameters for the referent scenario, listing also the expected number of devices in the cell. As presented, smart metering differs from other M2M applications, foreseeing two operational modes: *periodic* and *alarm* reporting. Periodic reporting is characterized by variable reporting rates and tolerance of report losses, i.e., if a report is not successfully received, the metering application waits for the next scheduled reception. The alarm reporting is event-triggered, where the allowed reception delay is up to 1 minute and loss of reports is not tolerated [6]. Further, the presented traffic patterns can be divided into two categories - synchronous and asynchronous. For *asynchronous* traffic, the arrivals are not correlated across devices, and traffic patterns with uniform and Poisson distribution fall into this category. On the other hand, the traffic generated by alarm-reporting is *synchronous*, as the alarm event is typically detected by a multitude of smart meters, thus correlating the initiation of their transmission requests.

It can be inferred from Table I that a GSM base station should handle communication scenarios with high density

[1]Compared to the other GSM cell type - urban GSM cells, sub-urban cells have an increased coverage zone, potentially serving more devices.

Appliances/ Devices	Arrival rate [s^{-1}]	Average Message Size [byte]	Number of Devices	Distribution
Smart Meters - Periodic Reporting	1.67e-2, 3.33e-3, 1.11e-3, 2.78e-4, 4.63e-5, 2.32e-5, 1.16e-5	$<$1000	13941	Poisson
Smart Meters - Alarm Reporting	/	$<$1000	13941	Beta(3,4)
Home Security System	1.67e-3	20	3098	Poisson
Elderly Sensor Devices	1.67e-2	128	310	Poisson
Credit Machine in Grocery	8.3e-3	24	72	Poisson
Credit Machine in Shop	5.56e-4	24	1100	Poisson
Roadway Signs	3.33e-2	1	2963	Uniform
Traffic Lights	1.67e-2	1	360	Uniform
Traffic Sensors	1.67e-2	1	360	Poisson
Movie Rental Machines	1.16e-4	152	36	Poisson

TABLE I
TRAFFIC PARAMETERS FOR HOME AND CITY COMMERCIAL M2M DEVICES IN A SUB-URBAN AREA FOR A 1000 M-RADIUS CELL, WHERE 3 SMART METERS PER HOME ARE CONSIDERED [5]–[7].

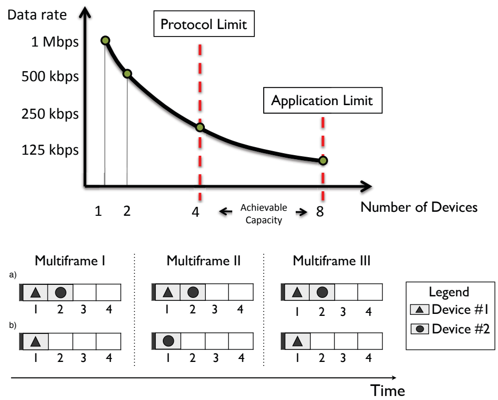

Fig. 1. Ideal system in which the bandwidth is shared among the multiplexed devices. The protocol operation limits the number of devices: a) allocation capacity is limited to 4 devices, b) allocation capacity is improved to 8 devices.

of devices, small payload sizes, and sporadic transmissions. One of the main limitations of GSM access, which prevents its application in such scenarios, is the limited allocation granularity of the access resources. This type of bottleneck is valid for any TDMA-based system, as described next.

A. The Resource Allocation Granularity Problem

Ideally, the granularity of resource allocation in a TDMA system should be dictated by the application requirements and tuned to the application with the lowest demand of resources. However, in practice, TDMA allocation does not work this way, as the resource granularity is fixed by the allocation mechanism. Consider a toy example of a TDMA system with transmission data rate of 1 Mbps, in which transmissions are organized in frames, and each frame consists of 4 slots, Fig. 1. It is assumed that the minimum number of slots that the system can allocate is a single slot per device per frame. Thus, if only one device is present in the system, it is allocated all 4 slots and thus gets a total data rate of 1 Mbps. If there are 4 devices, each gets one slot and a data rate of 250 Kbps, as illustrated in Fig. 1. However, if devices require only 125 Kbps, the system has the capacity to potentially accommodate 8 of them, but, due to the constraints of the allocation mechanism, more than 4 simultaneous connections cannot be supported at the same time. A straightforward approach is to re-allocate 4 devices in every new frame, but this introduces extra control traffic.

On the other hand, the fact that overall allocation is deterministic, i.e., each device should have the opportunity to transmit once in every two frames, could be exploited to design a more efficient allocation method. An approach proposed in this paper is to *logically* extend the allocation space by using frame numbers, known to all devices. Then both device #1 and device #2 can be allocated the same slot, but #1 accesses it only in odd- and #2 only in even-numbered frames, as shown in Fig 1b). Although rather simple, the example illustrates both the limitations that are present in a real system such as GSM, as elaborated in the next section, and the main idea behind the method for their mitigation, elaborated in Section IV.

III. GSM SYSTEM OPERATION AND LIMITATIONS

A. GSM Access Mechanism

The access in GSM is TDMA-based, where both the uplink and the downlink are organized in multiframes with duration of 240 ms. The multiframe structure, in its usual 2D representation, is shown in Fig. 2. A multiframe consists of 12 blocks, where a block is composed of four TDMA frames and a TDMA frame contains 8 time slots. Each time slot is interpreted as a separate TDMA channel, referred to as Packet Data Channel (PDCH). A PDCH can be dedicated either to signaling or data; in a typical configuration, PDCH $\#0$ is for signaling and the remaining 7 PDCHs for data transmissions.

The procedure to establish a connection with the base station (BS) consists of three stages, denoted as RACH, AGCH and DATA stage. The first stage is slotted ALOHA-based random access, in which devices contend for data resources. Specifically, a device with a pending data transmission first waits between 0 and $T-1$ frames, where the actual waiting time is randomly selected and T is a parameter broadcasted by the BS. In the next step, the device transmits its resource request in the random access channel (RACH) that is logically defined in the uplink PDCH #0, and waits for the response

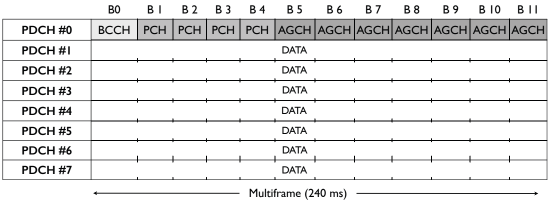

Fig. 2. The structure of the downlink multiframe: the signaling channel PDCH #0 consists of one block dedicated to broadcast channel (BCCH), four blocks for paging channel (PCH) and the remaining blocks are is devoted to access granted channel (AGCH); the remaining PDCHs are devoted to data.

during the next S frames, where S is another parameter broadcasted by the BS. As the devices choose the transmission instants in an independent and uncoordinated manner, their resource requests can potentially collide in the RACH, as in any slotted ALOHA-based scheme.

The BS responds to a resource request only if: (1) the request was successfully received, (2) there are available downlink resources to send the response, and (3) the requested uplink data resources are also available. The response is transmitted in the access granted channel (AGCH) that is logically defined in the downlink PDCH #0, and it has to be delivered before the waiting time of S frames expires at the device side. If no AGCH message is received from the BS during this period, the device repeats the procedure until either a response is received or a maximum of M retries is reached, where M is also broadcasted by the BS.

The response AGCH message assigns to a device an uplink PDCH and an uplink state flag (USF). USF is an identifier that controls the pending uplink data transmission; specifically, a device is allowed to transmit in block $k + 1$ of the assigned uplink PDCH, if its USF was announced in block k in the same downlink PDCH.

B. GSM Access Limitations

As outlined, in order to establish a data connection, a device has to go through three different stages: RACH, AGCH and DATA. The amount of resources (i.e., blocks of the multiframe) devoted to each stage should be scaled to accommodate both the number of connection attempts and the expected traffic volume. Otherwise, if there are no sufficient resources in any of the stages, the ultimate result will be a situation in which a large number of devices are retransmitting their resource requests. This in turn will cause the RACH channel to collapse due to collisions of the retransmissions. Thus, to assure that the operation of the access network is not compromised, 3GPP recommends a blocking probability below 2% in each of the stages [4]. The focus of the further text is on the limitations present in AGCH and DATA stages and the methods how to overcome them by MAC protocol reengineering.[2]

[2]As already outlined, the RACH stage is based on slotted ALOHA, whose limitations and improvements have been in the research focus for a long time.

A typical configuration foresees that 7 out of 11 blocks of the signaling PDCH are dedicated to AGCH, providing capacity to send approximately 30 AGCH messages per second [8]. However, as shown in Section VI, this configuration is inadequate for the referent scenario with a high density of low-rate devices, as a single cell cannot deliver enough AGCH messages within the time required to grant all resource requests. This limitation can be partially solved by dedicating more PDCHs to signaling, but this comes at the expense of the resources available for the DATA stage.

In the DATA stage, the granularity of data resources is limited by the USF allocation mechanism. Specifically, USF is only 3 bits long and the value 000 is reserved to indicate that the upcoming uplink block can be used for RACH contention. Thus, a maximum of 7 devices per PDCH can be multiplexed simultaneously. Therefore, in a single frequency configuration with 7 PDCHs devoted to data, the limit is 49 data connections per uplink multiframe, which is insufficient to support the traffic patterns of the referent scenario.

IV. Reengineering GSM for Massive Smart Metering

The solution for the AGCH bottleneck, proposed by Qualcomm [9], is based on the observation that, in principle, M2M devices share the same capabilities and are likely to request the same type of service. Therefore, most of the AGCH message content can be directed to multiple devices requesting data resources. Specifically, [9] foresees that four consecutive RACH requests could be granted with a single AGCH message, i.e., the capacity of the AGCH stage is increased four times in comparison to the legacy system.

However, once the AGCH bottleneck is removed, the USF limitation becomes even more pronounced, as shown in Section VI. In the following text we present a method to remove the USF bottleneck, which in combination with the AGCH solution results with the improved GSM access scalability.

The modification of the USF allocation mechanism is based on the method partially presented in [10]. The main conception behind it is that the *validity range of USF is reinterpreted*, allowing for accommodation of a substantially increased number of active connections. In the following text we present the extended version of the solution, called *expanded USF (eUSF)*, designed both for the periodic and alarm reporting.

In case of periodic reporting, the allocated USF does not hold anymore for an entire PDCH in all multiframes during which the connection is active, as prescribed by the GSM standard, but it is valid only for a specific set of blocks within a PDCH during X consecutive multiframes, reoccurring periodically every M multiframes. In this way, several active devices reuse the same USF in the same PDCH, as they are multiplexed in the non-overlapping blocks and/or multiframes. The price to pay for this increased multiplexing capability is decreased data rate per active connection; however, the majority of M2M services require low data rates (see Table I), thereby rendering the proposed solution highly relevant.

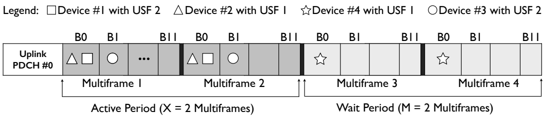

Fig. 3. An example of expanded USF allocation, where the parameters X and M are set to 2: four devices are allocated in a single uplink PDCH with only three USFs.

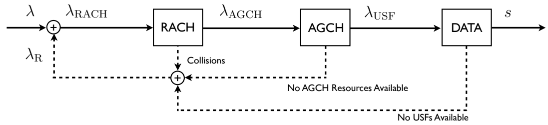

Fig. 4. Illustration of the GSM access stages. The blocking probability for each stage should remain below 2%.

We illustrate the eUSF method through an example depicted in Fig. 3. There are four devices in the example, while the parameters X and M are set to 2, i.e., the allocated USFs' validity pattern consists of periods of 2 "valid" multiframes followed by 2 "non-valid" multiframes. Devices #1, #2 and #3 arrive in multiframe 1; device #1 is allocated USF1 that is valid in block 0, device #2 is allocated USF2 in block 0, and device #3 is allocated USF2 in block 1. Devices #1 and #2 are multiplexed in the same block of the multiframe using different USFs, which is supported by the standard GSM. Devices #1 and #3, share the same USF in the same multiframe, but are multiplexed in different blocks - this way of operation is not supported in the standard GSM. Furthermore, device #4 arrives in multiframe 3 and it is allocated USF1 in block 1. Devices #1 and #4 now share the same USF in the same block of the same PDCH, yet their transmissions are multiplexed as they take place in different multiframes; this way of operation is also not supported in the standard GSM operation. Finally, the standard GSM system has to allocate 4 USFs in order to accommodate 4 users in a single PDCH, whereas in the above example this is done by using only 3 USFs. If the number of devices requiring service increases, a standard GSM system would rapidly run out of available USFs, as it is limited by its inflexible allocation method. On the other hand, the scalability of the eUSF solution is superior and limited only by the required data rates. Another advantage of the proposed method is that, once a device has been allocated a USF, this allocation can in principle last as long as required. Hence, the allocated devices do not have to go through RACH and AGCH stages anymore, relaxing the operation of these stages as well.

The solution for the case of alarm traffic is similar, the main difference is that allocated USF is valid only for the specific set of blocks in the next X consecutive multiframes, rather than being periodic. Once the device sends the report, the data connection is terminated; on the next occasion, the device has to go through RACH and AGCH stages again.

Finally, we briefly outline potential methods to identify the devices compatible with the proposed improvements, while allowing the remaining devices to operate as usual. A simple option is to split the RACH resources into two different groups, one reserved for the standard operation and the other reserved for M2M traffic from the devices compatible with the new solution. However, the drawback of this approach is that it may leave unused RACH resources. In an alternative method, the device informs the BS of its compatibility through values reserved for future use in the RACH request message [11].

V. Capacity Analysis of GSM Access

Fig. 4 depicts the stages of the GSM random access. The "fresh" traffic arrivals are represented by arrival rate λ, which takes into account all the arrivals of newly generated traffic of all the smart meters in the cell. In the ideal case, all the devices with a pending transmission go through all the stages without being blocked, and re-enter the idle mode once the transmission is finished. However, due to collisions in the RACH, lack of AGCH messages and lack of free USFs, the actual arrival rate at each stage is decreased by the success probability of the previous stage:

$$\lambda_{\mathrm{AGCH}} = \lambda_{\mathrm{RACH}} \cdot P_{\mathrm{RACH}}, \tag{1}$$

$$\lambda_{\mathrm{USF}} = \lambda_{\mathrm{AGCH}} \cdot P_{\mathrm{AGCH}}, \tag{2}$$

where P_{RACH} is the probability that RACH request is successfully received and P_{AGCH} is the probability that a successful RACH request is being granted timely access by the BS. In other words, in general it holds:

$$\lambda_{\mathrm{RACH}} \geq \lambda_{\mathrm{AGCH}} \geq \lambda_{\mathrm{USF}}. \tag{3}$$

The devices that are blocked in any of the stages retransmit their requests, generating an additional traffic represented by the arrival rate λ_R. Thus, the total arrival rate present in the RACH λ_{RACH} equals the sum of retransmission λ_{R} and "fresh" traffic arrivals λ.

3GPP has studied the capacity of GSM access to serve M2M traffic in [4], [12], [13]. The adopted methodology assumes that the arrival rates to the AGCH stage λ_{AGCH} and the DATA stage λ_{USF} follow the same Poisson process that is present at the RACH stage, and does not take into account the impact of the retransmissions. I.e., 3GPP methodology assumes that:

$$\lambda = \lambda_{\mathrm{RACH}} = \lambda_{\mathrm{AGCH}} = \lambda_{\mathrm{USF}}. \tag{4}$$

However, this is rather approximate, as demonstrated next.

Fig. 5a) presents the distribution of the traffic arrivals of smart meters for the RACH and AGCH stages, when the newly generated traffic is Poisson distributed with a mean of $\lambda = 40$ arrival/s.[3] It can be observed that the impact of the retransmissions is particularly pronounced, as the mean arrival rate to the RACH stage is $\lambda_{\mathrm{RACH}} \approx 120$ arrival/s, i.e., it is three times larger. Also, the mean AGCH arrival rate in this case is approximately equal to $\lambda_{\mathrm{AGCH}} = 61$ arrival/s, far below λ_{RACH}. The corresponding traffic arrival distribution to

[3]The rest of the parameters used for this study, i.e, the payload size and the GSM coding scheme, are described in Section VI.

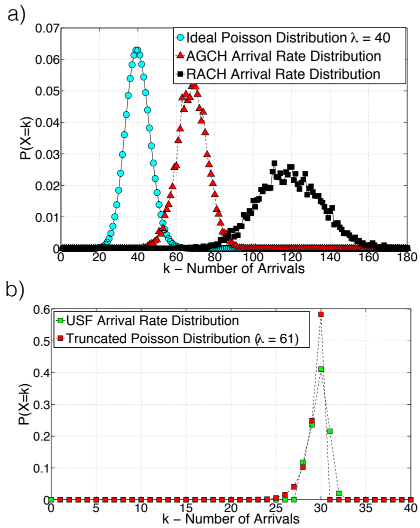

Fig. 5. Arrival rate distributions of the smart meters traffic at a) RACH and AGCH stages and b) DATA stage, when the newly generated traffic is Poisson distributed with $\lambda = 40$.

the DATA stage is depicted in Fig. 5b). Obviously, the DATA stage arrival rate is limited to (approximately) 30 arrival/s - which is a direct observation of the AGCH stage limitation, see Section III-B.[4] Our findings indicate that arrival process to the DATA stage can be described by a truncated Poisson distribution, obtained by truncating the Poisson distribution with mean λ_{AGCH} at the limit established by the AGCH capacity bottleneck; this is also demonstrated in Fig. 5b).[5]

Further, according to the 3GPP study [4], the USF bottleneck is identified to be the most restrictive one with respect to M2M communications. Although this conclusion seems expected, due to the limited number of USF identifiers that can be allocated per multiframe, our findings indicate that the AGCH bottleneck is actually the most restrictive. The main reason is that the AGCH channel cannot grant enough number of devices per second in order to reach the maximum number of active connections supported in the DATA stage. Moreover, we again emphasize that in [4] the USF capacity is characterized using an input Poisson process with mean arrival rate $\lambda_{\mathrm{USF}} = \lambda_{\mathrm{RACH}}$, while the input arrival process should be actually modeled with a truncated Poisson distribution, as demonstrated above.

In the next section, we demonstrate the potential of the proposed bottleneck solutions. Due to the interdependencies among GSM access stages, we adopt a numerical approach when investigating the performance of the GSM access, both for the legacy and the reengineered system.

[4]We note that in [12] a maximum of 38 AGCH/s is assumed; nevertheless, this fact does not impact our conclusions.

[5]The small deviation between analytical and simulation results is due to fact that a second is not multiple of multiframe duration.

VI. Results

The results presented in this section assume that all the asynchronously reporting devices have a uniform payload size of 152 bytes and that the most robust physical coding scheme (CS1) is used. This coding scheme represents the worst case scenario with a payload of 22 bytes per block. As shown in Table I, the selected application payload size is the upper limit for the presented M2M applications, excluding smart metering. In case of smart metering, the M2M-related capacity analysis of 3GPP foresees payloads of 100, 500 and 1000 bytes [13]. Due to space constraints, we show the results only for the synchronously reporting smart meters with a payload size of 100 bytes, and note that similar improvements can be observed for other payload sizes. We also assume that there are no channel-induced errors in the uplink/downlink transmissions, as our study is concentrated on the characterization of the access mechanism. For the simulations we have used a typical GSM configuration, where the RACH parameters are $T = 20$, $S = 105$ frames and maximum $M = 4$ RACH retransmissions [8]. The performance parameters of interest are blocking probabilities of the AGCH and DATA stages, and the outage, defined as the fraction of the accessing devices that have reached the maximum number of connection attempts M without establishing data connection.

Fig. 6a) compares the blocking probabilities P_B of the AGCH and DATA stages as function of the system input arrival rate λ, when only asynchronous traffic is present in the cell (e.g., no alarm events.). The results are presented for (1) the legacy GSM access, (2) the reengineered GSM access where only AGCH solution is applied, denoted as AGCH improvement, and (3) the reengineered GSM access where both the AGCH and eUSF solutions are applied, denoted as eUSF improvement, both for the 3GPP methodology and numerical analysis. Obviously, the results obtained by the 3GPP methodology deviate both qualitatively and quantitatively from the ones obtained by the approach that takes into account interstage dependences. Specifically, the simulations show that the 3GPP blocking probabilities are overestimated and that, in the legacy system, the AGCH stage bottleneck is actually reached before the USF bottleneck. Also, when the AGCH bottleneck is removed, the effects of the USF bottleneck become obvious, and the eUSF solution demonstrates its full potential. The remaining blocking probabilities are not shown, as their values are zero for all the considered arrival rates.

Fig. 6b) compares the outage performance in the same scenario, obtained by numerical analysis. Obviously, superior results are achieved when AGCH and eUSF solutions are combined. E.g., a single-frequency GSM cell can support up to 70 arrivals per second with an outage below 2%, which is an improvement of 133% in comparison to the legacy system. In a 3-sector cell, this translates to approximately 13000 smart meters reporting every 5 minutes in addition to the expected asynchronous M2M traffic shown in Table I.

Fig. 6c) presents the analysis of the proposed AGCH and USF improvements for the traffic scenario presented in

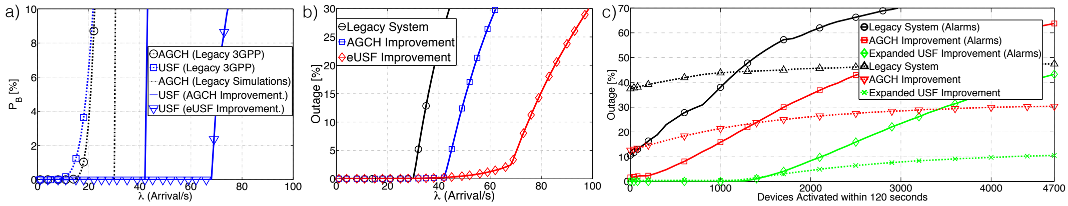

Fig. 6. a) Blocking probabilities for AGCH and DATA stages, obtained by 3GPP model and simulations for asynchronous traffic. b) Outage of the legacy system, the system with AGCH improvement, and the system with AGCH and eUSF improvements for asynchronous traffic. c) Outage for asynchronous and synchronous reporting. The total asynchronous traffic is 42 arrival/s and the synchronously reporting smart meters are activated within 120 seconds.

Section II. The total expected arrival rate of the asynchronous traffic, modeled by uniform and Poisson distributions, is 42 arrival/s. The behavior of the synchronously reporting smart meters is modeled by a Beta distribution; we assume that the activation period, a central parameter of the Beta distribution, is set to 120 seconds.[6] Our goal is to investigate the GSM access performance as the number of synchronously reporting meters is increasing - i.e, we are interested to assess the behavior of the system when it is "stressed" by synchronously initiated resource requests. As pointed out by [8], the RACH performance is severely affected by synchronous arrivals; the RACH limitations are outside the scope of this work and therefore we assume that the resource requests for each traffic type are transmitted in separate RACH channels. The comparison of the results reveals that the combination of the AGCH and eUSF solutions outperforms by far the legacy system and the system where only AGCH solution is implemented. Specifically, the combined AGCH/eUSF solution can roughly support up to 1500 synchronously reporting devices with an outage that is below rather demanding 0.1%, as required for massively deployed sensors [14]. In case of more relaxed upper bounds on outage levels, the number of supported synchronously reporting meters rises, e.g., it is 2300 for the outage below 10%. We note that neither the legacy system nor the AGCH only solution can assure outage level below 0.1%.

VII. Conclusions

In this paper we have presented a concept to transform GSM into a dedicated network for massive smart metering, based on a simple redefinition of the access mechanisms. The demonstrated reengineering principles have been applied to AGCH and DATA stages of the GSM access, significantly boosting the performance in comparison to the legacy system. Results show that a single 3-sectorial cell can provide service up to 13000 smart meters reporting every 5 minutes in addition to the expected M2M traffic per cell. Also, up to 1500 synchronously reporting meters can be supported for rather demanding outage levels of 0.1%. The proposed changes are incurred only at the access control layer, leaving the physical interfaces intact, which a highly desirable feature in practice.

Another important conclusion presented in the paper is that 3GPP modeling methodology of the GSM access is not valid, and that the corresponding results are incorrect.

Finally, we note that 3GPP access protocols from different generations share the same stages: random access, access granted and data transmission. Hence, the proposed approach can be extrapolated to other cellular systems; the investigation how to exploit the analogues principles for the improvement of the resource granularity in 4G is part of our ongoing work.

[6]The similar performance was observed for shorter activation periods.

Acknowledgement

The research presented in this paper was supported by the Danish Council for Independent Research (Det Frie Forskningsråd), grant no. 11-105159 "Dependable Wireless Bits for Machine-to-Machine (M2M) Communications" and grant no. DFF-4005-00281 "Evolving wireless cellular systems for smart grid communications".

References

[1] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2M: From mobile to embedded internet," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 36–43, 2011.
[2] ABI Research, "Cellular M2M connectivity services," Tech. Rep. 2012.
[3] H. Dahmouni, B. Morin, and S. Vaton, "Performance modeling of GSM/GPRS cells with different radio resource allocation strategies," in *IEEE WCNC, 2005*, vol. 3, march 2005, pp. 1317 – 1322 Vol. 3.
[4] 3GPP, "Bottleneck Capacity Comparison for MTC," 3rd Generation Partnership Project (3GPP), TSG GERAN #46 GP-100895, 2010.
[5] IEEE 802.16p, "IEEE 802.16p Machine to Machine (M2M) Evaluation Methodology Document (EMD)," IEEE 802.16 Broadband Wireless Access Working Group (802.16p), EMD 11/0005, 2011.
[6] 3GPP, "RACH intensity of Time Controlled Devices," 3rd Generation Partnership Project (3GPP), TSG RAN WG2 R2-102340, 2010.
[7] 3GPP, "Smart Grid Traffic Behavior Discussion," 3rd Generation Partnership Project (3GPP), TR R2-102340, 2010.
[8] R. Paiva, R. Vieira, and M. Saily, "Random Access Capacity Evaluation with Synchronized MTC Users Over Wireless Networks," in *IEEE VTC Spring, 2011*.
[9] 3GPP, "Packet Channel Assignments to Multiple Mobile Device," 3rd Generation Partnership Project, TSG GERAN #48 GP-101894, 2010.
[10] G. Corrales Madueño, C. Stefanovic, and P. Popovski, "How Many Smart Meters can be Deployed in a GSM cell?" in *IEEE ICC 2013 - Workshop Telecom R2S*, pp. 1283–1288.
[11] 3GPP, "General Packet Radio Service (GPRS); Mobile Station (MS) - Base Station System (BSS) interface; Radio Link Control / Medium Access Control (RLC/MAC) protocol," 3rd Generation Partnership Project (3GPP), TS 44.060, Sep. 2008.
[12] 3GPP, "Downlink CCCH Capacity Evaluation for MTC," 3rd Generation Partnership Project (3GPP), TSG GERAN #46 GP-100893, 2010.
[13] 3GPP, "USF Capacity Evaluation for MTC," 3rd Generation Partnership Project (3GPP), TSG GERAN #46 GP-100894, 2010.
[14] FP-7 METIS, "Requirements and general design principles for new air interface," Deliverable D2.1, 2013.

Paper 3

Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling

Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.

Published in IEEE Wireless Communications Letters, Volumen 3. August 2014.

Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling

Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski

***Abstract*—This letter considers a wireless M2M communication scenario with a massive number of M2M devices. Each device needs to send its reports within a given deadline and with certain reliability, e.g., 99.99%. A pool of resources available to all M2M devices is periodically available for transmission. The number of transmissions required by an M2M device within the pool is random due to two reasons—random number of arrived reports since the last reporting opportunity and requests for retransmission due to random channel errors. We show how to dimension the pool of M2M-dedicated resources in order to guarantee the desired reliability of the report delivery within the deadline. The fact that the pool of resources is used by a massive number of devices allows basing the dimensioning on the central limit theorem. The results are interpreted in the context of LTE, but they are applicable to any M2M communication system.**

***Index Terms*—Wireless cellular access, M2M communications, LTE.**

I. Introduction

In the past two decades, the main focus of the cellular access engineering was on the efficient support of human-oriented services, like voice calls, messaging, web browsing and video streaming services. A common feature of these services is seen in the relatively low number of simultaneous connections that require high data rates. On the other hand, the recent rise of M2M communications introduced a paradigm shift and brought into research focus fundamentally new challenges. Particularly, M2M communications involve a massive number of low-rate connections, which is a rather new operating mode, not originally considered in the cellular radio access.

Smart metering is a showcase M2M application consisting of a massive number of devices, up to 30000 [1], where meters periodically report energy consumption to a remote server for control and billing purposes. On the other hand, the report size is small, of the order of 100 bytes [2]. The current cellular access mechanisms, considering all the associated overhead, cannot support this kind of operation [3]. There are on-going efforts in 3GPP that deal with the cellular access limitations, investigating methods for decreasing the access overhead for small data transmissions [4], access overload control [5] and guaranteed quality of service [6]. Besides LTE, [7] proposes an allocation method for reports with deadlines in GPRS/EDGE, showing that up to 10^4 devices can be effectively supported in a single cell by avoiding random access and using a periodic structure to serve the devices such that the deadlines are met.

In this letter, we consider a system with a periodically occurring pool of resources that are reserved M2M communications and shared for uplink transmission by all M2M devices. The re-occurring period is selected such that if a report is transmitted successfully within the upcoming resource pool, then the reporting deadline is met. We note that, if each device has a deterministic number of packets to transmit in each resource pool and if there are no packet errors, then the problem is trivial, because a fixed number of resources can be preallocated periodically to each device. However, if the number of packets, accumulated between two reporting instances, is random and the probability of packet error is not zero, then the number of transmission resources required per device in each transmission period is random. On the other hand, as the number of transmission resources in each instance of the resource pool is fixed, the following question arises: *How many periodically reporting devices can be supported with a desired reliability of report delivery (i.e., 99.99%), for a given number of resources reserved for M2M communications*? We analyze the proposed approach in LTE context; however, the presented ideas are generic and implementable in other systems where many devices report to a single base station or access point. The results show that, for fixed reliability of access, the proposed method requires less LTE resources compared to the comely used random access.

The rest of the letter is organized as follows. Section II presents the system model. Section III is devoted to the analysis of the proposed access method. Section IV presents the numerical results and Section V concludes the letter.

II. System Model

We focus on the case of periodic reporting, where the length of the reporting interval (RI), denoted by T_{RI}, depends on the application requirements [2]. The M2M resources for uplink transmission are reserved to occur periodically, at the end of each RI. The periodic reporting is typically modeled either as a Poisson process with arrival rate $\lambda = 1/T_{RI}$, where devices can actually send none, one, or multiple reports within RI, or as a uniform distribution, where devices send a single packet per RI [8], [9]. Our analysis will focus on the former case, but we note that the derived results can be readily applied to the latter arrival model, as well. We assume that all report arrivals that occur within the current reporting interval are served in the next reporting interval.

The LTE access combines TDMA and FDMA, such that access resources are represented in 2D, see Fig. 1. As depicted,

Manuscript received March 26, 2014; revised May 20, 2014; accepted May 20, 2014. Date of publication May 23, 2014; date of current version August 20, 2014. This work was supported by the Danish Council for Independent Research (Det Frie Forskningsråd) under Grant 11-105159 "Dependable Wireless bits for Machine-to-Machine (M2M) Communications" and under Grant DFF-4005-00281 "Evolving wireless cellular systems for smart grid communications." The associate editor coordinating the review of this paper and approving it for publication was D. Niyato.

The authors are with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: gco@es.aau.dk; cs@es.aau.dk; petarp@es.aau.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LWC.2014.2326674

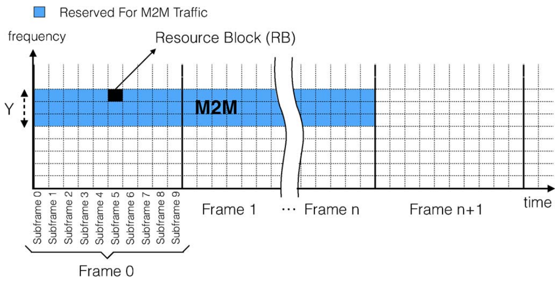

Fig. 1. Representation of the LTE uplink resource structure, where a set of RBs has been reserved for M2M purposes.

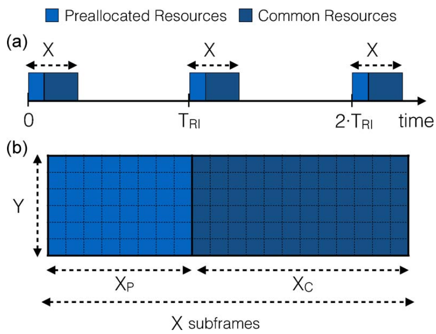

Fig. 2. (a) Periodically occurring M2M resource pool. (b) Division of M2M resource pool in the preallocated and common pool.

time is organized in frames, where each frame is composed of subframes with duration $T_s = 1$ ms. The minimum amount of uplink resources that can be allocated to a device is one resource block (RB), corresponding to a single subframe and 12-subcarriers in frequency. We assume that the uplink resources are split into two pools, one reserved for M2M services (depicted in blue in Fig. 1), and the other used for other services. Note that the approach of splitting the resources for M2M and non-M2M has often been used [10], as their requirements are fundamentally different. Finally, we assume that there is a set of Y RBs reserved for M2M resource pool in each subframe.

The M2M resource pool is divided into two parts, denoted as the preallocated and common pool, which reoccur with period T_{RI}, as depicted in Fig. 2(a). We assume that there are N reporting devices, and each device is preallocated an amount of RBs from the preallocated pool dimensioned to accommodate a single report and an indication if there are more reports, termed excess reports, from the same device to be transmitted within the same interval. The common pool is used to allocate resources for the excess reports, as well as all the retransmissions of the reports/packets that were erroneously received. These resources can only be reactively allocated and in our case we consider the LTE data transmission scheme, where each transmission has an associated feedback that can be used to allocate the resources from the common pool.[1] The length of the M2M resource pool, preallocated pool and common pool, expressed in number of subframes, are denoted by X, X_P, and X_C, respectively, see Fig. 2(b), such that

$$X = X_P + X_C = \alpha N + X_C. \tag{1}$$

where $\alpha \leq 1$ denotes the fraction of RBs per subframe required to accommodate a report transmission and where the value of X_C should be chosen such that a report is served with a required reliability. The analysis how to determine X_C, given the constraints of the required number of RBs per report, number of devices and reliability, is the pivotal contribution of the letter and presented in Section III. Finally, we note that the duration of X has a direct impact on the delay; in the worst case a (successful) report is delivered $T_{RI} + (X \cdot T_s)$ seconds after its arrival, which also defines the delivery deadline.

III. Analysis

Denote by $W_{i,j}$ a random variable that models the total number of transmissions of report j from device i; i.e., $W_{i,j}$ includes the first transmission and any subsequent retransmissions that may occur due to reception failures. Further, assume that the maximum value of $W_{i,j}$ is limited to L, where L is a system parameter, applicable to all reports from all devices. The probability mass function (pmf) of $W_{i,j}$ can be modeled as a geometric distribution truncated at L:

$$P[W_{i,j} = k] = \begin{cases} p_e^{k-1}(1-p_e), & 1 \leq k \leq L-1, \\ p_e^{L-1}, & k = L, \end{cases} \tag{2}$$

where p_e is the probability of a reception failure.

Recall that the reporting of device i is modeled as a Poisson process U_i with arrival rate λ, where the device can send none ($U_i = 0$), one ($U_i = 1$) or multiple reports ($U_i \geq 1$) within RI. As stated above, the first transmission of the first report of a device is sent in the preallocated pool, while all subsequent transmissions take place in the common pool. These include: a) potential retransmissions of the first report, and b) transmissions and potential retransmissions of all excess reports. Denote by R_i the random variable that corresponds to the total number of transmissions from device i requiring resources in the common pool:

$$R_i = \begin{cases} 0, & U_i = 0, \\ \sum_{j=1}^{U_i} W_{i,j} - 1, & U_i \geq 1. \end{cases} \tag{3}$$

The total number of transmissions of all devices requiring resources from the common pool is:

$$R = \sum_{i=1}^{N} R_i. \tag{4}$$

The straightforward way to characterize R is to derive its pmf. However, as the supposed number of reporting devices is very large, R is a sum of a large number of independent identically distributed (IID) random variables. Therefore, we apply the central limit theorem and assume that R is a Gaussian random variable, requiring characterization only in terms of the expectation and variance. We proceed by evaluation of $\mathrm{E}[R]$ and $\sigma^2[R]$, and show in Section IV that this approach provides accurate results.

[1]The minimum latency for the feedback is 6 ms (6 subframes), which includes processing times at the base station and at the device, and which can be assumed negligible taking into account that the RI that we are considering is of the order of thousands subframes.

A. Expectation and Variance of R

The expectation of R is:

$$\mathrm{E}[R] = \mathrm{E}\left[\sum_{i=1}^{N} R_i\right] = N \cdot \mathrm{E}[R_i]. \tag{5}$$

Taking into account (3), it could be shown that:

$$\begin{aligned}\mathrm{E}[R_i] &= \mathrm{E}[R_i|U_i = 0]\mathrm{P}[U_i = 0] + \mathrm{E}[R_i|U_i \geq 1]\mathrm{P}[U_i \geq 1],\\ &= \mathrm{E}[R_i|U_i \geq 1]\mathrm{P}[U_i \geq 1],\\ &= \mathrm{E}\left[\sum_{j=0}^{U_i} W_{i,j} - 1|U_i \geq 1\right](1 - e^{-\lambda T_{RI}}).\end{aligned} \tag{6}$$

By putting $\lambda T_{RI} = 1$ and applying Wald's equation [11], the (6) becomes:

$$\begin{aligned}\mathrm{E}[R_i] &= (\mathrm{E}[U_i|U_i \geq 1]\mathrm{E}[W_{i,j}] - 1)(1 - e^{-1}),\\ &= \frac{1 - p_e^L}{1 - p_e} - (1 - e^{-1}).\end{aligned} \tag{7}$$

where we used the fact that $\mathrm{E}[U_i|U_i \geq 1] = 1/(1 - e^{-1})$. Substituting (7) into (5) yields expectation of R.

The variance of R can be derived in a similar fashion, using the identities related to the variance of the random sum of random variables [11]. Due to the space limitations, we omit the derivation and present only the final result:

$$\begin{aligned}\sigma^2[R] = N \Bigg[&\frac{(2L-1)p_e^{L+1} - (2L+1)p_e^L + p_e + 1}{(1-p_e)^2}\\ &+ e^{-1} \cdot \left(1 - 2 \cdot \frac{1 - p_e^L}{1 - p_e} - e^{-1}\right)\Bigg].\end{aligned} \tag{8}$$

B. Probability of Report Failure

In this subsection, we assess the probability of report failure, i.e., the probability that the report has not been successfully delivered after all attempted (re)transmissions. In general, this probability depends both on the number of resources available and the scheduling policy applied in the common pool. In order to avoid the particularities related to scheduling, we derive an upper bound on the probability of failure which is valid for any scheduling policy.

Denote by Φ the event that a report experiences a failure. Further, denote by l the number of required report transmissions, which includes the first transmission and all the required retransmissions. If we assume for a moment that the number of available resources in the common pool is infinite, then the report fails to be delivered only when the required number of transmissions exceeds L:

$$\mathrm{P}_\infty[\Phi] = \mathrm{P}_\infty[\Phi, l > L] = \mathrm{P}[l > L] = p_e^L. \tag{9}$$

If the common pool that consists of X_C subframes can accommodate C transmissions (i.e., C is the capacity of the common pool in number of transmissions), then the probability of report failure can be written as:

$$\begin{aligned}\mathrm{P}[\Phi] &= \sum_{k=1}^{L} \mathrm{P}[\Phi, l = k] + \mathrm{P}[\Phi, l > L],\\ &= \sum_{k=1}^{L} \mathrm{P}[\Phi|l = k]P[l = k] + p_e^L.\end{aligned} \tag{10}$$

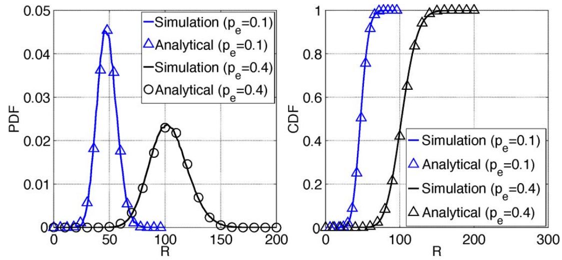

Fig. 3. Comparison of simulated and analytical pdf and cdf of R when $N = 100$ and $L = 10$.

Further, for $1 \leq k \leq L$:

$$\begin{aligned}\mathrm{P}[\Phi|l = k] &= \mathrm{P}[\Phi, R > C|l = k] + \mathrm{P}[\Phi, R \leq C|l = k],\\ &= \mathrm{P}[R > C]\mathrm{P}[\Phi|l = k, R > C]\end{aligned} \tag{11}$$

where we used the fact that $\mathrm{P}[\Phi, R \leq C|l = k] = 0$, i.e., there is no report delivery failure when the total number of required transmissions R is not greater then the capacity of the common pool C, for $l \leq k \leq L$. Regardless of the scheduling policy applied in the common pool, it is always $\mathrm{P}[\Phi|l = k, R > C] \leq 1$, which leads to the following upper bound:

$$\mathrm{P}[\Phi|l = k] \leq \mathrm{P}[R > C]. \tag{12}$$

Finally, substituting (12) into (10) yields:

$$P[\Phi] \leq Q\left(\frac{C - \mu}{\sigma}\right)(1 - p_e^L) + p_e^L \tag{13}$$

where $\mu = \mathrm{E}[R]$, $\sigma = \sqrt{\sigma^2[R]}$, and $Q(\cdot)$ is Q-function, standardly used for Gaussian random variables.

IV. Results

We first validate the assumptions used in the analysis by comparing the probability density function (pdf) and cumulative density function (cdf) of the Gaussian model with the simulated ones. Fig. 3 presents a tight match between the model and simulations, when number of reporting devices is $N = 100$, the maximum number of transmissions per report is $L = 10$, the probability of report error p_e takes values 0.1 and 0.4, respectively, and the number of the simulation runs is set to 10^5 for each parameter combination.

Further, using the bound derived in (13), we determine the fraction of LTE system resources required for reliable M2M services, defined as the ratio of RBs required for reliable M2M services and the total amount of RBs available. The amount of required RBs depends on the modulation[2] and the report size (RS). We assume a typical 5 MHz LTE system [12], typical individual LTE transmission error of $p_e = 0.1$ [13], and the maximum number of report transmissions is again set to $L = 10$. The maximal number of devices is set to 30 000, as indicated by 3GPP in [1]. Finally, the probability of report failure is set to $P[\Phi] \leq 10^{-3}$, i.e., the desired reliability to at least 99.99%. To validate the analytical upper bound, we performed simulations with a random scheduler with 10^5 repeats for each parameter combination. Fig. 4 shows the performance of the proposed

[2]In this paper, we consider the lowest-order (QPSK) and the highest-order (64-QAM) LTE modulation schemes.

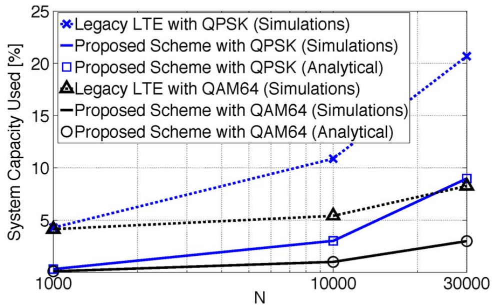

Fig. 4. Fraction of system capacity used for M2M services, when $P[\Phi] \leq 10^{-3}$, RI of 1 minute, RS of 100 bytes, bandwidth of 5 MHz and $p_e = 10^{-1}$.

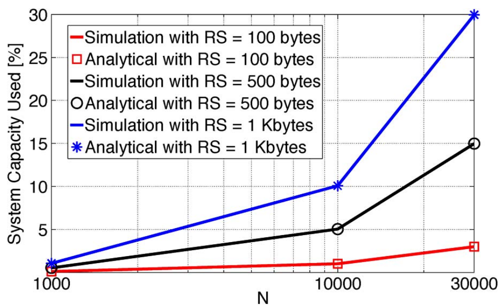

Fig. 5. Fraction of system capacity used for M2M services, when $P[\Phi] \leq 10^{-3}$, RI of 1 minute, bandwidth of 5 MHz, 64-QAM and $p_e = 10^{-1}$.

scheme, when report size RS is 100 bytes and reporting interval RI is 1 minute, corresponding to the most demanding RI according to [2]. It can be observed that for the lowest-order modulation (QPSK), up to 30000 devices can be served with only 9% of the available system resources. If 64-QAM is used, then only 3% of the available resources are required to achieve the target reliability.

Fig. 4 also compares the performance of the proposed scheme with the performance of the legacy LTE, obtained using a LTE simulator with 2 random access opportunities per frame, which is a typical configuration [12]. The devices perform random access for every report, and the reports are also sent in the RBs reserved for M2M traffic. Obviously, the legacy LTE requires about two times more system resources than the proposed scheme; this is due to the uncertainty of the individual report arrivals and retransmissions, demanding a high amount of reserved RBs. In the proposed scheme, the individual reports are grouped (over a RI), and this aggregation exhibits far less uncertainty, requiring less reserved RBs for a reliable service. Furthermore, it can be shown that if the resource reservation is not used and there are only M2M devices present in the cell, the fraction of the system resources required for the target reliability in the legacy LTE is similar as in the proposed scheme. However, in presence of additional traffic in the cell, the reliability of M2M services without reservation cannot be guaranteed anymore.

Fig. 5 depicts the required fraction of system capacity for M2M service, when the RS varies between 100 bytes and 1 kbytes [2], the system bandwidth is set to 5 MHz, modulation scheme is 64-QAM, and $p_e = 0.1$. Obviously, the report size has a large impact in the results, demanding up to 30% of the system capacity in the worst case.

Finally, we note that Figs. 4 and 5 also demonstrate a tight match between the analytical and simulated results.

V. Conclusion

We have introduced a contention-free allocation method for M2M that relies on a pool of resources reoccurring periodically in time. Within each occurrence, feedback is used to reactively allocate resources to each individual device. The number of transmissions required by an M2M device within the pool is random due to two reasons: (1) random number of arrived reports since the last reporting opportunity and (2) requests for retransmission due to random channel errors. The objective is to dimension the pool of M2M-dedicated resources in order to guarantee certain reliability in the delivery of a report within the deadline. The fact that the pool of resources is used by a massive number of devices allows basing the dimensioning on the central limit theorem. Promising results have been shown in the context of LTE, where even with the lowest-order modulation only 9% of the system resources are required to serve 30K M2M devices with a reliability of 99.99% for a report size of 100 bytes. The proposed method can be applied to other systems, such as 802.11ah.

References

[1] "RACH intensity of time controlled devices," Sophia-Antipolis, France, TSG RAN WG2 R2-102340, 2010.
[2] "Smart grid traffic behaviour discussion," Sophia-Antipolis, France, TR R2-102340, 2010.
[3] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, Dec. 2014.
[4] "Study on Machine-Type Communications (MTC) and other mobile data applications communications enhancements," Sophia-Antipolis, France, TR 22.368, Dec. 2013.
[5] "System improvements for Machine-Type Communications (MTC)," Sophia-Antipolis, France, TR 23.888, 2013.
[6] S.-Y. Lien and K.-C. Chen, "Massive access management for QoS guarantees in 3GPP machine-to-machine communications," *IEEE Comm. Lett.*, vol. 15, no. 3, pp. 311–313, Mar. 2011.
[7] G. Corrales Madueño, C. Stefanovic, and P. Popovski, "How many smart meters can be deployed in a GSM cell?" in *Proc. IEEE ICC Workshop*, Budapest, Hungary, Jun. 2013, pp. 1263–1268.
[8] *IEEE 802.16p Machine to Machine (M2M) Evaluation Methodology Document (EMD)*, EMD 11/0014, May 2011.
[9] "USF capacity evaluation for MTC," Sophia-Antipolis, France, TSG GERAN #46 GP-100894, 2010.
[10] K.-D. Lee, S. Kim, and B. Yi, "Throughput comparison of random access methods for M2M service over LTE networks," in *Proc. IEEE GC Workshops*, Dec. 2011, pp. 373–377.
[11] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*. San Diego, CA, USA: Academic, 2004.
[12] B. Yang, G. Zhu, W. Wu, and Y. Gao, "M2M access performance in LTE-A system," *Trans. Emerging Telecommun. Technol.*, vol. 25, no. 1, pp. 3–10, Jan. 2014.
[13] S. Ahmadi, *LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*. Amsterdam, The Netherlands: Elsevier, 2013.

Paper 4

Efficient LTE Access with Collision Resolution for Massive M2M Communications

Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.

Published in IEEE Global Communications Conference 2014 (GLOBECOM) Workshops - Ultra-Low Latency and Ultra-High Reliability in Wireless Systems, 2014.

Efficient LTE Access with Collision Resolution for Massive M2M Communications

German Corrales Madueño, Čedomir Stefanović, Petar Popovski

Department of Electronic Systems, Aalborg University, Denmark
Email: {gco,cs,petarp}@es.aau.dk

***Abstract*—LTE random access procedure performs satisfactorily in case of asynchronous, uncorrelated traffic arrivals. However, when the arrivals are correlated and arrive synchronously, the performance of the random access channel (RACH) is drastically reduced, causing a large number of devices to experience outage. In this work we propose a LTE RACH scheme tailored for delay-sensitive M2M services with synchronous traffic arrivals. The key idea is, upon detection of a RACH overload, to apply a collision resolution algorithm based on splitting trees. The solution is implemented on top of the existing LTE RACH mechanism, requiring only minor modifications of the protocol operation and not incurring any changes to the physical layer. The results are very promising, outperforming the related solutions by a wide margin. As an illustration, the proposed scheme can resolve 30k devices with an average of 5 preamble transmissions and delay of 1.2 seconds, under a realistic probability of transmissions error both in the downlink and in the uplink.**

I. Introduction

Machine-to-Machine (M2M) services span a wide range, including services like car-to-car, smart grid, smart metering, control/monitoring of homes and industry, e-health, traffic control, surveillance, etc. Opposed to the typical human-oriented services, M2M services are not driven by data rates, but by the features of availability and reliability. However, attaining required availability and reliability of M2M services is not a trivial issue, due to a potentially massive number of devices involved. An astonishing 300k devices per cell are foreseen in future M2M scenarios [1], with potentially thousands of them simultaneously trying to access the network. Consider the example of smart grid monitoring - in case of a power outage, thousands of smart meters will try to report the failure. These messages should be delivered before the battery dies (i.e., last-gasp reporting), setting the reporting deadline to 500 ms [2]. In such cases, the LTE random access channel (RACH) becomes overloaded by thousands of simultaneous access attempts [3].

Recently, there has been a large amount of work devoted to investigation of the approaches how to avoid overloading the RACH to protect both network and users against such events. One of the initial approaches is to split the preambles used in the RACH for human and M2M communications [4]. This way human services are not affected, but the major drawback is that the overload problem for M2M services is aggravated, as the number of available preambles is reduced. Another approach is to control the RACH load via backoff adjustments, spreading the preamble retransmissions over time and thus attempting to limit the number of collisions. However, due to the different nature of human and M2M communications, a valid backoff for former might not be suitable for the latter. In [5] specific M2M backoff and class barring parameters are discussed for delay tolerant devices, where the load in the RACH channel is decreased by a factor of 20. However, the delay can raise up to 100 s.

On another hand, only a few solutions for delay-sensitive M2M services have been presented so far. One of these is the dynamic allocation, where additional RACH resources are allocated when an overload is detected [6]. The drawback of this approach is the notification delay of the additional resources availability. In LTE, the number of random access opportunities (RAOs) per frame is broadcasted in the system information block 2 (SIB2); it can take up to 512 radio frames, i.e., 5.12 s, before this broadcast is sent [7]. In [8] a coordinated random access scheme is proposed, where only one or few representatives of every group report the critical information. This is based on the observation that during the congestion period the correlation of messages across devices within a group is very high. The drawbacks in this case are the required coordination among devices within the group and the compromised reliability of relying on a few devices per group to successfully report the delay sensitive information.

In this work we propose a novel approach to deal with massive synchronous access attempts, tailored for delay-sensitive M2M services. Contrary to the mainstream solutions that try to avoid collisions by modifying the parameters of the LTE RACH access procedure, we propose use of a *collision resolution* algorithm to resolve synchronous RACH attempts. The motivation lies in the observation that when RACH is overloaded by synchronous access attempts, the massive number collisions inevitably occurs and it is more efficient to resolve these collisions instead to waste time and LTE resources by trying to avoid them. The basis of the proposed solution is a q-ary tree splitting technique [9], implemented on the top of the existing LTE RACH procedure and activated when RACH overload is detected. Apart from the novel idea of using collision resolution in LTE RACH, the paper contributions are also in presentation of the implementation details and demonstration of the efficiency of the proposed approach to achieve a reliable and timely massive synchronous access.

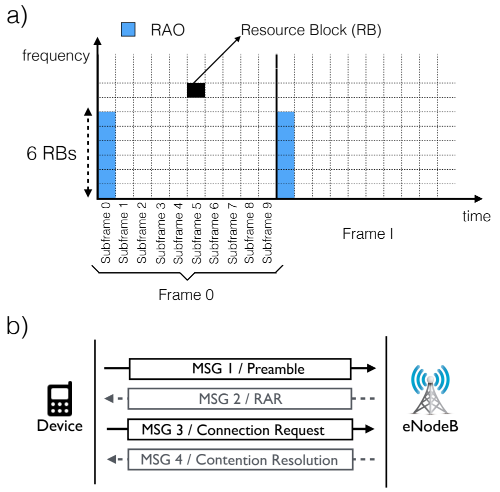

Fig. 1. a) LTE uplink resources with one RAO per frame. b) Message exchange between a device and the eNodeB during the LTE random access procedure.

The rest of the paper is organized as follows. Section II presents a brief overview of the standard LTE random access. Section III describes the proposed solution in details. Section IV demonstrates the performance results. Section V concludes the paper.

II. LTE RACH Overview

The uplink resources in LTE for frequency division duplexing (FDD) can be expressed using a 2D grid, see Fig. 1a), where the x-axis represents time and the y-axis resource blocks (RBs). Time is divided in frames, where every frame is composed of ten subframes, and each subframe is of duration $t_s = 1$ ms. The amount of RBs per subframe is determined by the available bandwidth in the system, which ranges between 6 RBs and 100 RBs. The number of subframes between two consecutive RAOs varies between 1 and 20, where 5 is the most typical value [10], providing one RAO every 5 ms. Finally, every RAO is composed of 6 RBs, as depicted in Fig. 1a), and a maximum of one RAO per subframe is allowed.

The standard LTE random access procedure is of access reservation type, where the devices are contending to reserve resources for their uplink data transmissions using a slotted ALOHA based mechanism. The access procedure comprises exchange of four different messages between a device and the eNodeB, see Fig. 1b). The first message (MSG 1) consists of a randomly selected preamble sent in the next available RAO. There are 64 orthogonal preambles in LTE; some of them are reserved for special purposes and the actual number of available preambles for contention is lower and typically set to 54. A typical premise is that the eNodeB can only detect if a preamble has been activated or not, *but not how many devices have actually activated it* [11]. In other words, if two or more devices send the same preamble in the same RAO, this collision remains undetected. In the next step, the eNodeB replies with the random access response RAR, denoted as MSG 2, to all detected preambles. The contending devices monitor the downlink channel, expecting MSG 2 within the next t_{RAR} seconds. If no MSG 2 is received and the maximum of M MSG 1 transmissions is not reached, the random access procedure restarts after a randomly selected time within the interval $t_r \in [0, B]$, where B is a backoff parameter. If MSG 2 is received, it includes uplink grant information, pointing to the RB where the connection request (MSG 3) should be sent. The connection request indicates the desired operation by the device, such as call/data transmission/measurement report, etc. In case when two or more devices activated the same preamble and received the same MSG 2, their MSGs 3 collide in the RB. In contrast to the collisions of MSGs 1, collisions of MSGs 3 are detected by the eNodeB. The eNodeB replies only to MSGs 3 that did not experience collision, by sending message MSG 4, which allocates the required RBs or denies the request if no resources are available. If no MSG 4 is received after t_{CRT} seconds since MSG 1, the random access procedure is restarted. Finally, if after M MSG 1 transmissions a device does not successfully finish all the steps of the random access procedure, an outage is declared.

The random access in LTE is well suited for asynchronous arrivals, as a typical RACH configuration offers one RAO with 54 available preambles every 5 ms [10], i.e., there are 10.8 k available preambles per second. However, as shown in Section IV, in case of synchronous traffic arrivals, e.g., alarm events with thousands of devices activated simultaneously, the system cannot cope with the excessive collisions of MSGs 3, and the RACH collapses.

III. The Proposed Solution

We start by a high level description of the proposed solution. Assume that an event takes place that causes synchronous RACH access attempts by a massive number of devices. As the number of contention preambles is limited, the ultimate result is a high number of collided MSGs 3 observed by the eNodeB.[1] This could serve as a trigger for eNodeB to modify the LTE RACH operation, by switching from the slotted ALOHA-based collision avoidance to a collision resolution mechanism. Specifically, we propose to use a q-ary tree splitting algorithm [9], leveraging on the LTE orthogonal preambles. The notification to the contending devices about the change of RACH operation, as well as direction of the contending devices through the tree splitting, is performed through the feedback messages sent by eNodeB. These messages could be implemented by modifying the existing eNodeB messages, as outlined further. We proceed by presentation of the details.

A. LTE RACH Modifications

Tree splitting algorithms rely on the use of feedback after every contention attempt; to this end, we propose to use

[1] Note that the eNodeB has only to detect if there is a collision, which could be done in a simple manner, e.g., using an energy detector.

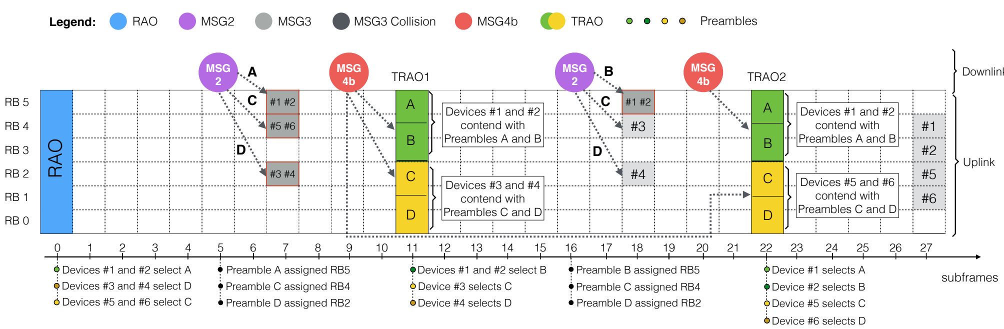

Fig. 2. Illustration of the proposed tree-splitting algorithm.

a new type of MSG 4, denoted as MSG 4b. Contrary to the standard MSG 4, this message is sent to the devices whose MSGs 3 *collided*, notifying them about the collision and specifying the details of the next contention attempt. Specifically, MSG 4b indicates a set of q preambles to be used for the next contention attempt and the RAO where this contention should take place, denoted as tree-splitting RAO (TRAO).[2] The recipients of MSG 4b send new MSGs 1, by transmitting a random preamble from the set of q preambles in the designated TRAO, as directed by the eNodeB. The eNodeB replies with standard MSG 2 to all detected preambles, and the recipients of MSG 2 send standard MSG 3. The eNodeB replies with standard MSG 4 to the non-collided MSGs 3 (i.e., these messages are resolved), and with a new MSG 4b to collided MSGs 3, whose senders continue to participate in the tree-splitting. The above procedure repeats until all MSGs 3 are either resolved or the maximum number of preamble transmissions M is reached, when the affected devices declare outage.

For a better understanding we provide an example in Fig. 2, where there are 6 devices and 4 available preambles, denoted as A, B, C, and D. In subframe 0, devices #1 and #2 send preamble A, devices #3 and #4 send preamble D and devices #5 and #6 send preamble C. The eNodeB detects these three preambles and responds with MSG 2, indicating that MSGs 3 should be sent in subframe 7. When MSGs 3 are transmitted in subframe 7, the collisions are detected and the eNodeB replies with MSGs 4b, indicating that: (i) the devices that sent preamble A should now contend in TRAO in subframe 11 (TRAO1) using preambles A and B, (ii) the devices that sent D should also contend in TRAO1 using preambles C and D, and (iii) the devices that sent C should contend in TRAO in subframe 22 (TRAO2) with preambles C and D. Devices #1 and #2 again choose the same preamble, their MSGs 3 collide in subframe 18, and they are directed to contend again in TRAO2, using preambles A and B. This time #1 and #2 choose different preambles in TRAO2, so MSGs 3

[2]We assume that TRAOs are allocated in subframes that are orthogonal to the subframes containing RAOs; thus, the access performance of other services (e.g., human-oriented services) remains unaffected.

SFN Offset (6 bits)	Preamble ID (6 bits)	TRAO Offset (10 bits)	Group Index (3 bits)

Total Length = 25 bits

Fig. 3. Proposed MSG 4b format.

are allocated different RBs in subframe 27 and do not collide again. Devices #3 and #4 choose different preambles already in TRAO1, so their MSGs 3 are resolved in subframe 18. Finally, devices #5 and #6 choose different preambles in TRAO2, and their MSGs 3 are resolved in subframe 27. We also note that for the sake of clarity MSGs 4 are not shown in Fig. 2.

A possible format for MSG 4b is depicted in Fig. 3. The first two fields are used to indicate the devices affected by the message; specifically they indicate the offset in subframe numbers (SFN) between the current SNF and the subframe in which the devices with preamble ID transmitted their MSG 3. The last two fields are used for the collision resolution, where TRAO offset and Group Index are used to indicate the SFN in which the TRAO takes place and the group of preambles to be used.

Further, we note that the performance of the random access procedure is also affected by the capacity of the control channel (PDCCH) through which the messages MSG 2, MSG 4 and MSG 4b are sent. A straightforward solution is to increase the bandwidth of the system, which indirectly increases the capacity of the PDCCH. In this work we consider another approach, proposed in [12], where one of the reserved radio network temporal identifiers (RNTI) is dedicated for M2M and defined as M2M-RNTI. M2M-RNTI is used by every device to determine who is the recipient of the data or control information. If there are not enough resources in the PDCCH, MSG 2, MSGs 4 and MSG 4b for several devices are bundled into one packet data unit and masked with the M2M-RNTI. This information is transmitted in the packet data shared-channel (PDSCH), allowing to virtually increase the capacity of the PDCCH. Therefore, we assume unlimited downlink capacity, but take into account the amount of required resources when assessing the performance of the proposed solution in

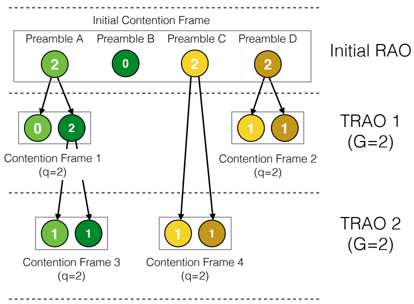

Fig. 4. Illustration of contention resolution with four devices and four preambles.

Section IV.

B. Analysis

In this section we determine the number of transmissions per device, the number of TRAOs required, and the probability of a device exceeding the maximum number of preamble transmissions (outage probability) for the proposed scheme. The presented analysis is the adaptation of the one from [9].

The basic structure of the q-ary tree-splitting algorithm is a contention frame, which is composed of q slots. Users contend by transmitting in a randomly selected slot; if two or more devices transmit in the same slot, a collision occurs and the slot expands into a new contention frame, again with q slots. Every expansion corresponds to a level of the contention tree. This procedure repeats until all slots with collisions are resolved.

We translate the above algorithm into LTE RACH terms in the following way. The root of the tree is the initial RAO where the original collisions happen, and it constitutes a single contention frame. This frame is a exception from all the other frames, as it consists of N_P slots, where the N_P is the total number of available preambles. Also, we assume that the set of available preambles is divided in G non-overlapping sets with q preambles in each, i.e., the total number of available preambles is $N_P = G \cdot q$. The slots of the initial contention frame that contain collisions are expanded in new contention frames containing q slots each. These contention frames take place in TRAOs following the initial RAO; as the available preambles are divided into G sets of q preambles, every TRAO is logically partitioned into G contention frames with q slots in each frame. Starting from the slots of the initial contention frame, every subsequent expansion corresponds to a level of the splitting tree; thus, if every slot splits into q new slots, the number of slots in level m is Gq^m.

Fig. 4 depicts the same example as in Fig. 2, but in the standard tree-splitting representation. There are $N_P = 4$ slots in the root contention frame, and $q = 2$ slots in all other contention frames; numbers in slots denote how many devices contended in them. Note that the contention frames 1, 2 and 4 correspond to the level 2, although they are in different TRAOs, whereas the contention frame 3 corresponds to the level 3, although it is in the same TRAO as contention frame 4. This is due to the fact that every TRAO contains just $G = N_P/q = 2$ contention frames.

To determine the number of levels, which is equal to the number of preamble transmissions required until MSG 3 is received at eNodeB without collision, we recall the approach from [9]. We assume that devices in level m are independently and identically randomly distributed over Gq^m slots. Thus, the probability of only one device transmitting in a slot of level m, when there are total of $N \geq 2$ devices at the start of the tree splitting procedure, is:

$$P_S(m) = \left(1 - \frac{1}{Gq^m}\right)^{N-1}. \tag{1}$$

The probability that m levels are required to resolve the transmission of the device, denoted by $P_L(m)$, is equal to the probability that the transmission is resolved in level m and it was not resolved in level $m-1$:

$$P_L(m) = P_S(m) - P_S(m-1). \tag{2}$$

The outage probability of a device, i.e., the probability that more than maximum of M transmissions are required, and the average number of transmissions T are given by:

$$P_O = 1 - \sum_{m=1}^{M} P_L(m), \tag{3}$$

$$T = \sum_{k=1}^{\infty} i \cdot P_L(m). \tag{4}$$

An approximation to the number of transmissions T can be derived as [9]:

$$\hat{T} = \log_m\left(\frac{N-1}{G}\right) - \left(\frac{1}{2} + \frac{\gamma}{\log_m}\right) + \frac{1}{2N\log_m}, \tag{5}$$

where $\gamma \approx 0.5772$ is Euler's constant. Further, the number of slots with collisions in level m and therefore the number of contention frames in the next level is given by:

$$C(m) = Gq^m\left(1 - \left(1 - \frac{1}{Gq^m}\right)^N\right) - N\left(1 - \frac{1}{Gq^m}\right)^{N-1}. \tag{6}$$

Finally, the expected number of TRAOs required to resolve N devices, denoted as R, can be determined from the number of contention frames as:

$$R = 1 + \sum_{m=1}^{\infty}\left\lceil \frac{C(m)}{G} \right\rceil, \tag{7}$$

where $\lceil\cdot\rceil$ denotes the ceiling function.

IV. Results

In this section we present the performance of the proposed access mechanism, obtained both through the analytical approach and simulations. We also make a comparison with standard LTE RACH procedure [13] and dynamic allocation scheme [6], whose performances are obtained by simulations. For the standard LTE RACH procedure, we use a typical

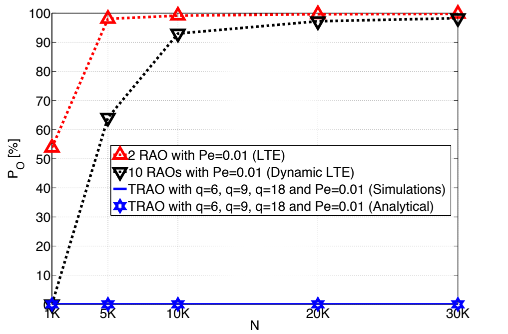

Fig. 5. Outage performance of standard LTE RACH, dynamic allocation and the proposed splitting-tree.

Parameter	Value	Parameter	Value
Total Number of Preambles (N_P)	54	MSG 2	56 bits
MSG 2 Window (t_{RAR})	5 ms	MSG 4	20 bits
MSG 4 Timer	24 ms	MSG 4b	25 bits
Maximum Transmissions (M)	10	System Bandwidth	20 MHz
Contention Timer (t_{CRT})	48 ms	Backoff (B_i)	20 ms
eNodeB and UE Processing Time	3 ms	Modulation	QPSK

TABLE I
SYSTEM PARAMETERS.

configuration of 2 RAOs per frame [10]. For the dynamic allocation scheme, we assume the maximum of 10 RAOs per frame and that there is no delay to activate the additional RAOs, i.e., we compare our method with the best case of dynamic allocation. All the simulations are performed in an event-driven MATLAB simulator, which models the LTE RACH procedure with a probability of error both downlink and uplink of $p_e = 0.01$, which is a typical target error rate in LTE control channel [14], [15]. The number of simulation repeats is set to 100 for every combination of parameters. Since our aim is to compare the performance of the different RACH procedures, we assume that the critical information fits in MSG 3 and no further actions are required; i.e., a device is resolved if MSG 3 is received with no collisions or errors. The rest of the parameters of the random access procedure are listed in Table I; we use a system bandwidth of 20 MHz and note that the similar improvements are observed when less bandwidth is used.

Fig. 5 shows the outage probability P_O, defined as the percentage of devices not completing the RACH procedure before the maximum number of preamble transmissions M is reached, as function of the number of devices N that synchronously start the random access procedure (i.e., in the same subframe). Obviously, a system with 2 RAOs per frame cannot cope with the massive synchronous arrivals and a large percentage of the devices are in outage. The dynamic allocation performs better; nevertheless, its performance is worse by a large margin in comparison to the performance of the proposed scheme. Specifically, the proposed scheme is able to resolve 30K synchronous attempts for any choice of number of preambles q per contention frame within TRAO with insignificant P_O. We also note the negligible differences among the results obtained by the analysis and simulations, where the latter include a realistic error probability. The same holds for the rest of the presented results.

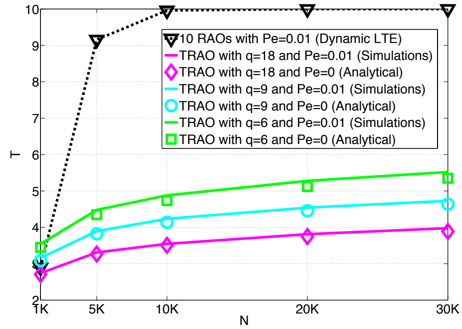

Fig. 6. Average preamble transmissions per device required.

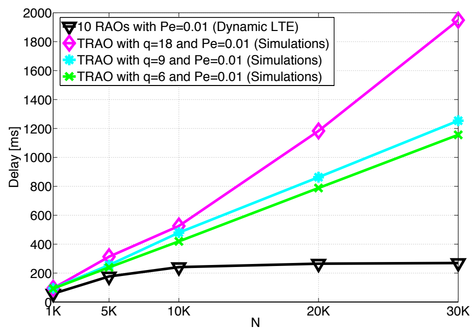

Fig. 7. Average delay experienced by resolved devices.

Fig. 6 shows the average number of preamble transmissions per device T as function of N. It is clear that 10 preamble transmissions (the allowed maximum M) is reached soon by the dynamic procedure, while the proposed scheme requires significantly less preamble transmissions per device. Also, the results show that when more preambles q are available to resolve a collision, less preamble transmissions are required on average. This could be expected from (5), when $G = N_P/q$ is substituted.

The average access delay of devices not in outage is shown in Fig. 7. Obviously, this delay is larger for higher q, even though the number of required transmissions is lower, see Fig. 6. This is due to the fact that for higher q less contention frames G fit in a TRAO, and therefore more TRAOs are needed on average to provide contention frames for the collision resolution. We emphasize that the average delay shown for the dynamic allocation applies only to a small percentage of the devices that are not in outage, c.f. Fig. 5.

The average number of TRAOs required to resolve *all* the

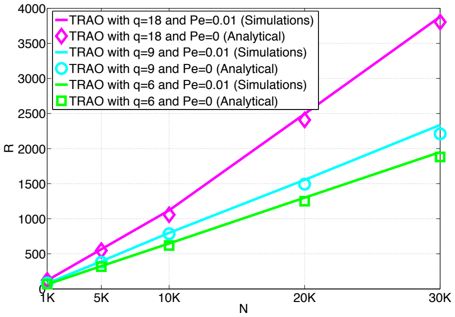

Fig. 8. Average number of TRAOs required.

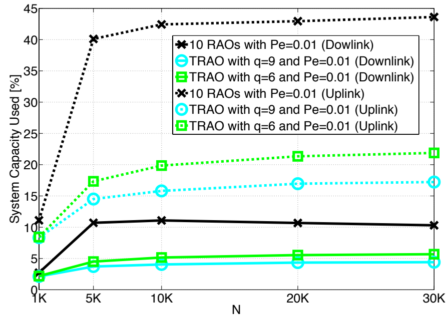

Fig. 9. Percentage of the system capacity for the downlink and uplink used to resolve the devices.

devices R using the proposed scheme is depicted in Fig. 8. Obviously, increasing q increases R; this can be also inferred from the combination of (6) and (7).

Finally, the fraction of the resources used for uplink and downlink for the random access procedure is depicted in Fig. 9. For the downlink, we consider the amount of RBs used to transmit all the required MSG 2, MSG 4 and MSG 4b. For the uplink, we consider the amount of RAOs and TRAOs (6 RBs) together with MSG3 (1 RB). Obviously, the proposed scheme is significantly less demanding than the dynamic allocation, requiring roughly half of the resources both in the downlink and in the uplink. Moreover, we note that these resources are also much more efficiently used, as only an insignificant portion of devices ends in outage, see Fig. 5.

V. Conclusion

In this paper we demonstrated that the LTE RACH becomes easily overloaded with excessive collisions in case of massive synchronous arrivals. We also proposed a scheme to deal with such arrivals, which actively pursues collision resolution instead of trying to avoid them. The scheme is tailored for LTE RACH and requires only modest modifications of the standard protocol, above the physical layer. We demonstrated that the proposed scheme provides reliable and timely service for high numbers of synchronously accessing devices, while requiring less amount of resources than competing schemes. Particularly, an astounding 30k devices can be resolved with negligible outage with an average of 5 preamble transmissions and delay of 1.2 seconds, under realistic probability of transmissions error both in the downlink and in the uplink.

Finally, we note that the proposed scheme allows for efficient and fast delivery of the devices' connection requests, enabling their processing and inspection by the eNodeB. In turn, this could provide an extensive basis for the eNodeB to gain insight in the event(s) that caused the massive synchronous arrivals, filter the redundant connection requests during the critical period, and thus alleviate the requirements for the subsequent data stage.

Acknowledgement

The research presented in this paper was supported by the Danish Council for Independent Research (Det Frie Forskningsråd), grant no. 11-105159 “Dependable Wireless Bits for Machine-to-Machine (M2M) Communications” and grant no. DFF-4005-00281 “Evolving wireless cellular systems for smart grid communications”.

References

[1] FP-7 METIS, “Requirements and General Design Principles for New Air Interface,” Deliverable D2.1, Aug. 2013.
[2] IEEE 802.16p, “IEEE 802.16p Machine to Machine (M2M) Evaluation Methodology Document (EMD),” IEEE 802.16 Broadband Wireless Access Working Group (802.16p), EMD 11/0005, Oct. 2011.
[3] A. Laya, L. Alonso, and J. Alonso-Zarate, “Is the Random Access Channel of LTE and LTE-A suitable for M2M Communications? A Survey of Alternatives,” *IEEE, Communication Surveys Tutorials*, vol. 16, no. 1, pp. 4–16, First 2014.
[4] K.-D. Lee, S. Kim, and B. Yi, “Throughput Comparison of Random Access Methods for M2M Service over LTE Networks,” in *IEEE GC Workshops*, Dec. 2011, pp. 373–377.
[5] 3GPP, “Access barring for delay tolerant access in LTE,” 3rd Generation Partnership Project (3GPP), TR R2-113013, May 2011.
[6] 3GPP, “MTC simulation results with specific solutions,” 3rd Generation Partnership Project (3GPP), TR R2-104662, Aug. 2010.
[7] 3GPP, “Radio Resource Control (RRC),” 3rd Generation Partnership Project (3GPP), TS TS-36331, Aug. 2010.
[8] Y. Chang, C. Zhou, and O. Bulakci, “Coordinated Random Access Management for Network Overload Avoidance in Cellular Machine-to-Machine Communications,” in *20th European Wireless Conference*, May 2014.
[9] A. J. Janssen and M. de Jong, “Analysis of contention tree algorithms,” *IEEE Trans. Info. Theory*, vol. 46, no. 6, pp. 2163–2172, Sep. 2000.
[10] 3GPP, “MTC simulation assumptions for RACH performance evaluation,” 3rd Generation Partnership Project (3GPP), TR R2-105212, Aug. 2010.
[11] N. K. Pratas, H. Thomsen, C. Stefanovic, and P. Popovski, “Code-Expanded Random Access for Machine-Type Communications,” in *IEEE GC Workshops*. IEEE, Dec. 2012, pp. 1681–1686.
[12] B. Yang, G. Zhu, W. Wu, and Y. Gao, “M2M Access Performance in LTE-A System,” *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 1, pp. 3–10, Jan. 2014.
[13] 3GPP, “Medium Access Control (MAC) protocol specification,” 3rd Generation Partnership Project (3GPP), TR 36.321.
[14] S. Ahmadi, *LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*. Elsevier Science, 2013.
[15] A. Ghosh, R. Ratasuk, W. Xiao, B. Classon, V. Nangia, R. Love, D. Schwent, and D. Wilson, “Uplink Control Channel Design for 3GPP LTE,” in *IEEE PIMRC 2007*, Sep. 2007.

Paper 5

Massive M2M Access with Reliability Guarantees in LTE Systems

Germán Corrales Madueño, Nuno K. Pratas, Čedomir Stefanović, and Petar Popovski.

Accepted in IEEE International Conference on Communications (ICC), 2015.

Massive M2M Access with Reliability Guarantees in LTE Systems

Germán Corrales Madueño, Nuno K. Pratas, Čedomir Stefanović, Petar Popovski

Department of Electronic Systems, Aalborg University, Denmark
Email: {gco,nup,cs,petarp}@es.aau.dk

***Abstract*—Machine-to-Machine (M2M) communications are one of the major drivers of the cellular network evolution towards 5G systems. One of the key challenges is on how to provide reliability guarantees to each accessing device in a situation in which there is a massive number of almost-simultaneous arrivals from a large set of M2M devices. The existing solutions take a *reactive* approach in dealing with massive arrivals, such as non-selective barring when a massive arrival event occurs, which implies that the devices cannot get individual reliability guarantees. In this paper we propose a *proactive* approach, based on a standard operation of the cellular access. The access procedure is divided into two phases, an estimation phase and a serving phase. In the estimation phase the number of arrivals is estimated and this information is used to tune the amount of resources allocated in the serving phase. Our results show that the proactive approach is instrumental in delivering high access reliability to the M2M devices.**

I. Introduction

Among the major drivers for the evolution of current cellular networks towards the fifth generation (5G) is the efficient support of Machine-to-Machine (M2M) communications and services. Different from human-centric services (H2x), which are mainly characterized by the ever-increasing data rates, M2M services pose a different set of challenges, associated with the support of a massive number of users exchanging small amounts of data, often with requirements in terms of reliability and availability. A model for a particularly demanding M2M scenario is the one where the cellular network access should be offered with reliability guarantees in the case of massive almost-simultaneous arrivals. An example is correlated reporting of an alarm event by tens of thousands of devices in a cell [1]. The main concern in such scenarios is the overload of the cellular access infrastructure, i.e., the collapse of the random access channel (RACH), which happens due to the signaling overhead associated with each individual transmission [2]. We note that the RACH overload precludes any service operation, i.e., blocks the system, and it is therefore of paramount importance to prevent it.

Several methods have been recently proposed to prevent the RACH overload in LTE [3], in the context of M2M communications. Specifically, two main solutions are the extended class barring (EAB) [4] and dynamic allocation [5]. EAB is valid only for delay-tolerant M2M traffic and is an extension of the standard access class barring method. On the other hand, dynamic allocation is a straightforward approach: upon detection of RACH overload the number of random access opportunities (RAOs) per second is increased. However, both schemes have inherent limitations, as they are both reactive and triggered upon RACH overload detection. Once the overload is detected, there is an additional delay until the EAB or the dynamic allocation feedback messages are delivered from the BS to the M2M devices, which can take up to 5 s [6], as these messages are typically broadcasted periodically over the paging channel. Therefore, these two methods cannot ensure timely and reliable operation in M2M scenarios with massive synchronous arrivals, as it becomes apparent further in this text.

Motivated by the deficiencies of the reactive approaches, in this paper we propose a *proactive* approach for the reliable support of M2M service. The proposed approach consists of two phases, an *estimation phase* and a *serving phase*, which reoccur periodically. In the first phase, the BS estimates how many M2M devices are attempting to access. We show that by using an estimator that is tuned to the LTE access mechanisms this can be done in a simple and, more importantly, fast manner, requiring just a single RAO to estimate the number of accessing users in the order of tens of thousands. Following the estimation phase, the parameters of the access mechanism are tuned such that the RAOs of the serving phase are used in an efficient way, providing a reliable service. The proposed solution can be easily incorporated in the standard LTE access mechanism, leaving the radio interfaces intact and used both for the case of massive synchronous arrivals as well as the asynchronous traffic with Poisson arrivals. In this way the mobile operators can provide M2M service in a controlled manner, with guaranteed reliability and no overload, i.e., the operators can be provided with a technical data-sheet indicating the performance of the system for a given number of devices and the associated latency. This is a significant step towards reliable M2M services in LTE, which are currently based on the best effort approach.

The rest of the paper is organized as follows. In Section II we present a brief overview of the standard LTE RACH operation. Section III is the central part of the paper, where we describe and analyze the proposed solution, as well as outline its practical implementation. Section IV presents a case study involving two M2M traffic classes, presenting the performance results and a comparison with competing methods for M2M access. Finally, Section V concludes the paper.

II. LTE RACH Overview

The uplink time in LTE is divided in frames, where every frame is composed of ten subframes whose duration is 1 ms. The LTE frequency band is organized in subcarriers, where 12 subcarriers of 15 KHz over a subframe constitute a resource block (RB). The bandwidth of LTE ranges between 6 RBs (i.e., 1 MHz) and 100 RBs (20 MHz). In LTE, a random access opportunity (RAO) requires 6 RBs in a subframe. The number of RAOs per frame is a system parameter, ranging from one RAO every 20 subframes to one RAO every subframe. A typical configuration foresees one RAO every 5 ms [6]. Further, up to 64 orthogonal preamble sequences are available in each RAO, which can be detected simultaneously by the base station (BS). The actual number of available preambles depends on the system configuration, where a typical configuration foresees 54 preambles [6]. The System Information Blocks (SIB)s, where all announcements including where each RAO occurs, are broadcasted periodically via the paging procedure that occurs from every 80 ms up to every 5.12 s [4].

The LTE random access procedure, denoted as Access Reservation Procedure (ARP), consists of the following four stages. (1) First, a device (UE), selects one of the preambles and transmits it in one of the RAOs. (2) In the case a single UE has transmitted the preamble, the eNodeB decodes it and responds by sending a random access response (RAR) message. (3) This RAR message indicates the RBs where the device shall send its request consisting of a temporary ID together with the establishment cause, e.g., call, data, report, etc. If two or multiple devices have selected the same preamble within the same RAO a collision occurs, the eNodeB detects this and does not send back any response. (4) In the last stage, the eNodeB acknowledges the information received from the device and allocates the required data resources. If the UE does not receive a response to a preamble or a request, it restarts the procedure until it is successful or the maximum number of preambles retransmissions is reached.

When the number of devices attempting access is high, most of the RACH preambles are selected by multiple devices and end in collisions. Consequently, most devices are not granted access and therefore retry again. There reattempts coupled with the new arrivals lead to an even higher amount of attempted accesses, further overloading the RACH and with the end result of almost no device being granted access. The general load control mechanism in LTE is the access class barring (ACB), which works by assigning access probabilities to different access classes [7]. However, as the ACB does not distinguish between H2x and M2M traffic, the EAB was defined in [4] to deal with potential burst of M2M traffic arrivals. EAB is used to explicitly restrict access from devices configured as delay tolerant. The core network can also trigger the admission control at the radio access network [8], via dynamic blocking according with the load.

Another mechanism proposed to overcome the RACH overload is the dynamic allocation mechanism [5]. Here, whenever the eNodeB detects the occurrence of overload, it increases the number of RAOs per frame. Due to the system limitations, this increase is up to one RAO per subframe, announced to the devices via the paging procedure. This mechanism can be further enhanced through the expansion of the LTE contention space to the code domain [9].

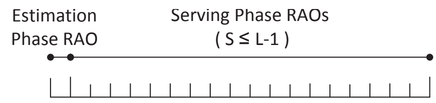

Fig. 1. Proposed access frame consisting of an estimation RAO followed by $S \leq L-1$ serving RAOs.

III. Proposed Solution

The core of the proposed solution consists of a reoccurring access frame, which is composed of RAOs that are dedicated to M2M devices.[1] It is assumed that the arrival process is gated, i.e., new arrivals are accepted at the frame beginning and all arrivals during the frame wait for the beginning of the next one. The frame time duration is assumed to be fixed and limited to half of the maximum allowed delay τ guaranteed by the network operator. The frame is then composed by up to L M2M dedicated RAOs within $\tau/2$[2]. Obviously, a larger L implies a longer delay, but it also accommodates more devices.

The frame consists of two parts, dedicated to the *estimation* and *serving* phase, as depicted in Fig. 1. We design the estimation part such that it consists just of a single RAO and describe in Section III-A the proposed estimation technique, showing that a huge range in the number of accessing M2M devices N can be reliable estimated.[3] The length of the serving phase S is determined by the estimated number of arrivals $\hat{N}$, with the constraint that $S \leq L-1$. The access algorithm in the serving S is based on the standard LTE RACH operation, but tuned to $\hat{N}$ such that its resources, i.e., RAOs, are used so that the required reliability R_{req} is met. Particularly, we distinguish two modes of operation in the serving phase. In the first mode, the length required by the target reliability S_{req} is lower or equal to $L-1$ and the actual length is set to $S = S_{\text{req}}$. In the second mode, $S_{\text{req}} > L-1$, which implies that there are not enough resources to provide required service. In this case, the length of the serving phase is set to $S = L-1$, and a barring factor is introduced to prevent RACH overload. Further details on the operation and dimensioning of the serving phase are presented in Section III-B.

A. Estimation Algorithm

We assume the estimation takes place in a single RAO with J preambles.The preambles are ordered from 1 to J

[1]The use of dedicated resources for M2M has been proposed previously in [5], [10], in an attempt to prevent M2M RACH accesses from affecting H2x services.

[2]Assuming the H2x dedicated RAOs occur every 5 ms [6], then within a $\tau/2 = 0.5$ seconds, there will be up to $L = 400$ available RAOs for M2M access, i.e., 8 RAOs per LTE frame.

[3]We note that the approach grants straightforward extension to cover the cases when the estimation phase consists of two or more RAOs.

(in a arbitrary way) and the active devices (i.e., devices with traffic arrivals) choose one of preambles with a predefined probability. The probability of selecting preamble j is given by:

$$p_j = \frac{p_0}{\alpha^j},\ j = 1, 2, \ldots, J, \tag{1}$$

where $p_0 \leq 1$ and $\alpha > 1$ are a priori determined parameters, whose choice depends on the expected range of the number of users N.

The eNodeB observes a ternary outcome[4] for each preamble - a preamble can be in the *idle* state (no devices transmitted it), *singleton* state (a single device transmitted it) or *collision* state (two or more devices transmitted it). Based on the observed outcomes, the eNodeB estimates how many users are present in the frame. The main idea behind varying the preamble activation probability is to obtain a favorable mix of collision, singleton and idle preambles, which will allow a reliable estimation. The same idea is standardly used in framed slotted ALOHA-based estimation algorithms [11]–[13]. Here we use a modification of a simple technique first proposed in [14], characterized by a large estimation range. The main difference with respect to [14] is that devices are limited to a single transmission due to the physical layer constraints.

Let a_j denote the probability that a device has not transmitted any of the previous $j-1$ preambles:

$$a_j = \prod_{i=1}^{j-1}(1-p_i),\ 1 < j \leq J, \tag{2}$$

with initial condition $a_1 = 1$. Denote the observed state of the preamble j preamble as s_j, where $s_j = 0$ if the state is idle, $s_j = 1$ if singleton and $s_j > 1$ if collision. The conditional probability mass function $f(s_j|N=n)$ is given by:

$$f(s_j|N=n) = \begin{cases} (1-p_j)^{a_j \cdot n} & s_j = 0, \\ a_j \cdot n \cdot (1-p_j)^{a_j \cdot n} & s_j = 1, \\ 1 - [1 + a_j \cdot n] \cdot (1-p_j)^{a_j \cdot n} & s_j > 1. \end{cases} \tag{3}$$

We note that the above expression is an approximation, as it assumes only the expected number of users capable of transmitting preamble j, i.e., $a_j \cdot N$. However, this approximation allows for an elegant solution that yields accurate results, as demonstrated further. The estimation of N is performed using the sequence of observations $\{s_j, j = 1, \ldots, J\}$, using the maximum likelihood approach:

$$\begin{aligned} \hat{N} &= \arg\max_n \prod_{j=1}^{J} f(s_j|N=n) \\ &= \arg\max_n \sum_{j=1}^{J} \ln f(s_j|N=n), \end{aligned} \tag{4}$$

which is obtained by solving for n the following equation:

$$\frac{\partial}{\partial n}\left(\sum_j \ln f(s_j|N=n)\right) = 0, \tag{5}$$

[4]In Section III-C is described how the collision detection is performed.

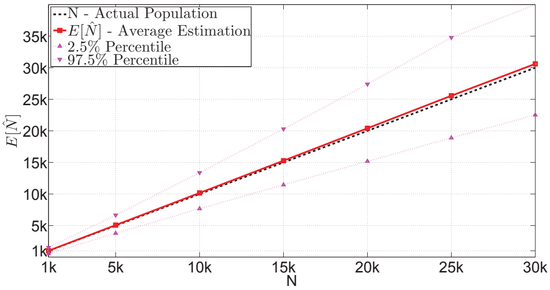

Fig. 2. Proposed estimator performance when the expected arrival rate is not a priori known. Through exhaustive numerical search it was found that for a dynamic range between $N \in [1, 30000]$ the optimal values of the estimator parameters are $p_0 = 0.001$ and $\alpha = 1.056$.

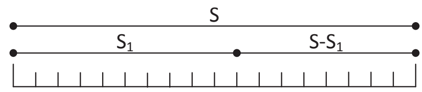

Fig. 3. Serving scheme structure, where the serving phase is composed by S RAOs grouped into two frames of lenght S_1 and $S - S_1$.

using a root-finding method. We conclude by presenting the estimator performance in Fig. 2, where it can be observed that the $E[\hat{N}]$ follows closely the actual value of N.

B. Serving Phase

The number of RAOs in the serving phase S should be, if possible, dimensioned according to N such that the required reliability R_{req} is met. On the other hand, S also depends on the access scheme employed in the serving phase, which is based on the LTE RACH operation, i.e., based on framed slotted ALOHA. In the further text, we assume that the serving phase consists of two frames, as depicted in Fig. 3. In the first frame the devices attempt access by transmitting a single randomly selected preamble in a randomly selected RAO, while in the second frame all devices that collided in the first frame reattempt access in the same way.[5] We show that in this way we can achieve close-to-one reliability for a huge range of accessing devices.[6]

We define reliability $R(N)$ as the probability of a device successfully obtaining a data resource when there are N contending device, at the completion of the Access Reservation Procedure discussed in Section II. For this to occur, the device has to be the only one to select a preamble from the RAOs available in either of the frames in the serving phase. We then

[5]We assume that the number of preambles is constant for all RAOs and equal to J.

[6]In principle, it could be argued that variants in which more than one retransmission per collided device is allowed could provide a higher reliability with the same number of RAOs. However, we demonstrate that the proposed approach shows rather favorable performance and allows for tractable modeling and analysis.

model the reliability $R(N)$ as:

$$R(N) = P_1(N) + [1 - P_1(N)]\, P_2(N), \tag{6}$$

where $P_1(N)$ and $P_2(N)$ denote the probabilities that a device does not collide in the first and second frames, respectively. In the first frame, the success probability is the probability that a device is the only one to select one of J preambles in one of S_1 RAOS, when there are N contending devices, which is:

$$P_1(N) = \left(1 - \frac{1}{S_1 J}\right)^{N-1}. \tag{7}$$

The success probability in the second frame depends on the number of collisions N_C in the first frame. Denote by $\Pr[N_C = k|N, S_1]$ the probability mass function (pmf) of the number of collisions in the first frame, conditioned on N and S_1, where:

$$\Pr[N_C = k|N, S_1] = \Pr[N_S = N - k|N, S_1], \tag{8}$$

where N_S denotes the number of successful devices in the first frame. The pmf $\Pr[N_S = N - k|N, S_1]$ can be modeled as a balls and bins problem, where the balls and bins represent respectively the devices and the contention resources (i.e., preambles and RAOs). In [15] this distribution is provided in a closed form expression as follows:

$$\Pr[N_S = s|N, S_1] = \frac{\binom{S_1 J}{s} \prod_{k=0}^{s-1}(N-k)G(S_1 J - s, N - s)}{(S_1 J)^N}, \tag{9}$$

where:

$$G(u, v) = u^v + \sum_{t=1}^{v} (-1)^t \prod_{j}^{t-1} [(v-j)(u-j)](u-t)^{v-t} \frac{1}{t!}. \tag{10}$$

The probability of a device being successful in the second frame $P_2(N)$, from the law of total probability, is given by:

$$P_2(N) = \sum_{k=2}^{N} \left(1 - \frac{1}{(S - S_1)J}\right)^{k-1} \cdot \Pr[C = k|N, S_1]. \tag{11}$$

Using (6), (8) and (11) it is possible to find the optimum S_1 that maximizes (6) and the minimum S_{req} that meets R_{req} through a numerical search.

When the number of required contention resources is higher than the maximum available $S_{\text{req}} > L-1$, a barring factor Q is introduced. This barring probability is then used independently by each device in a Bernoulli trial with probability $1 - Q$ to decide if the device should attempt to access the serving phase. To account with the barring probability, (6) is redefined as follows:

$$R_Q(N, Q) = (1 - Q) \sum_{k=0}^{N-1} B(N - 1, k, Q) R(k) \tag{12}$$

where $B(x, y, z) = \binom{x}{y}(1 - z)^y z^{x-y}$ is the binomial pmf. The optimal Q that maximizes (12) is found via:

$$\arg\max_{Q} R_Q(N, Q) := \{Q \mid \forall y : R_Q(N, y) \le R_Q(N, Q)\}. \tag{13}$$

We summarize the dimensioning of the serving phase in Algorithm 1.

Algorithm 1: Dimensioning of the size of the serving phase frame S and the associated barring probability Q.

```
Input N̂, R_req, L;
S_req is computed from (6);
if S_req ≤ (L − 1) then
    Q = 0; S = S_req;
else
    Q is computed from (13); S = L − 1;
Output S, Q;
```

C. Practical Implementation

All the information required by the devices to attempt access is broadcasted, similarly to the EAB, in a new system information message (SIB) [16] that takes place in each access frame, immediately after the estimation RAO. This SIB message includes the following information: First it indicates in which subframe the upcoming estimation RAO will take place together with the values of p_0 and α and the number of preambles J. Further, it informs the contending devices of the number of RAOs in the serving phase and S_1. Finally, a bitmap is included which indicates in which subframes these RAOs will occur. If the load exceeds the amount of capacity pre-reserved by the operator, the barring factor Q is also included in the SIB, to prevent the RACH overload.

The proposed scheme operation is then as follows. Assume that N contending devices become active prior to start of the access frame. When the estimation RAO occurs, each of these N devices attempt access, according with the procedure defined in Section III-A, enabling the eNodeB to obtain the estimation of the number of arrivals $\hat{N}$. The detection of collisions in the estimation phase, is performed during the execution of the Access Reservation Procedure. Namely, after the devices that have selected the same random access preamble, transmit their UE request, which will result in a collision as described in Section II. Based on $\hat{N}$, the eNodeB then defines how many RAOs are required in the serving phase to reach the contracted R_{req} and informs the devices where these RAOs will occur by broadcasting the corresponding SIB. Then, the contending devices select randomly between the serving RAOs, using the ARP described in Section II. In the meantime, other contending devices become active, which will wait until the start of the next access frame before proceeding in the same way.

We note that the proposed scheme requires minimal changes to the current LTE protocol, with no modifications to the physical layer at all.

IV. Case Study for Two M2M Traffic Classes

We now consider a case study with two traffic classes characterized by different requirements and serving probabilities. Let traffic class 1 (TC1) and traffic class 2 (TC2), have a respective reliability requirement $R^{(1)}_{req}$ and $R^{(2)}_{req}$. Further, let TC1 have priority access to the available serving RAOs over

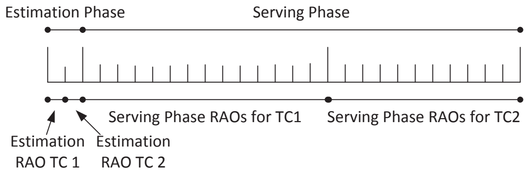

Fig. 4. Proposed access scheme for two traffic classes TC1 and TC2, where TC1 has priority over TC2.

TC2, e.g., alarm reports take priority over periodic reporting in the context of smart metering. Specifically, we try first to reach as close as possible to $R_{req}^{(1)}$ and only then as close as possible to $R_{req}^{(2)}$. Furthermore, we assume that each class has separate estimation and serving phases, as depicted in Fig. 4. At the beginning of the frame there is one estimation RAO for each traffic class, where the number of contending devices of each class is estimated to be $\hat{N}_1$ and $\hat{N}_2$ respectively. With the knowledge of $\hat{N}_1, \hat{N}_2$ we define a resource allocation strategy based on the scheme described in Section III.

The access frame duration – demarcated by the estimation phase RAOs occurrence – is constrained by the traffic class with the most stringent latency requirement, here given by TC1. Although, in this study we consider that both TCs have an estimation phase in each access frame, we note that in the case where TC2's latency requirement is much larger than TC1's, it might be worthwhile to consider the case where TC2 estimation RAO only occurs in some of the access frames, in order to optimize the amount of RAOs dedicated for estimation.

A. Serving Phase Size and Barring Factor for Two Traffic Classes

The extension of the analysis in Section III to two traffic classes is straightforward. Denoting as $S^{(1)}$ and $S^{(2)}$ the amount of serving RAOs respectively required to serve TC1 and TC2 to meet the reliability requirements of each class, $R_{req}^{(1)}$ and $R_{req}^{(2)}$. The main distinction from the case with a single traffic class, is that now there are three different operation regimes: (i) $S_{req}^{(1)}+S_{req}^{(2)} \leq L-2$; (ii) $S_{req}^{(1)}+S_{req}^{(2)} > L-2$ with $S_{req}^{(1)} < L-2$; and (iii) $S_{req}^{(1)} > L-2$. In (i) each traffic class receives the number of required serving RAOs. In (ii) a barring factor $Q^{(2)}$ is introduced to the lower priority class TC2, while no barring is necessary for the high reliability class TC1. Finally in (iii), TC2 is completely barred ($Q^{(2)}=1$) and a barring factor $Q^{(1)}$ is introduced for the high reliability class TC1. This procedure is described in detail in Algorithm 2.

B. Performance Results and Discussion

The performance results are obtained from a LTE event-driven simulator implemented in MATLAB, which models the complete access reservation procedure described in Section II. For the same network conditions, we compare the performance of the legacy LTE with dynamic allocation[7] with the performance of the proposed scheme. The system parameters of interest for the legacy system are listed in Table I; we assume an ideal, best-case dynamic allocation, where the network overload is detected instantaneously and there is no delay to change the parameters of the system such as the number of available RAOs. The incoming traffic is classified into two traffic classes: (TC1) alarm and (TC2) periodic reporting; where the alarm reporting takes priority over periodic reporting.

The alarm reporting case is modeled by a Beta distribution with parameters $\alpha = 3$ and $\beta = 4$ [17], which trigger N_1 smart meters within the cell to access the same access frame with latency requirement τ_1. The periodic reporting is modeled as a Poisson process with total arrival rate $\lambda = N_2/RI$, where N_2 denotes the number of M2M devices and $RI = \tau_2 = 60$ s, chosen so to match the arrival rate and latency requirement τ_2 of a typical M2M application such as smart metering [17].

The performance comparisons are done using different access frame L lengths, obtained from half of the maximum allowed delay for alarm reporting $\tau_1/2 = \{0.5, 2.5, 5\}$ seconds.[8] The performance evaluation is performed with the focus on the reliability achieved within the duration of the access frame. Specifically, we illustrate the performance during the peak of

Algorithm 2: Dimensioning of the size of the serving phase frame $S^{(1)}$ and $S^{(2)}$ and associated barring probabilities $Q^{(1)}$ and $Q^{(2)}$.

```
Input N̂1, N̂2, R_req^(1), R_req^(2), L;
S_req^(1) and S_req^(2) computed from (6);
if S_req^(1) + S_req^(2) ≤ L − 2 then
    Q^(1) = Q^(2) = 0; S^(1) = S_req^(1); S^(2) = S_req^(2);
else if S_req^(1) + S_req^(2) > L − 2 and S_req^(1) < L − 2 then
    Q^(1) = 0; S^(1) = S_req^(1);
    S^(2) = L − 2 − S^(1); Q^(2) computed from (13) ;
else
    Q^(1) computed from (13); S^(1) = L − 2;
    S^(2) = 0; Q^(2) = 1;
Output S^(1), S^(2), Q^(1), Q^(2);
```

TABLE I
LEGACY LTE SYSTEM PARAMETERS.

Parameter	Value	Parameter	Value
Preambles per RAO (J)	54	MSG 2 Window	5 ms
Max. RAOs per LTE frame	8	MSG 4 Timer	24 ms
Max. Retransmissions	9	Contention Timer	48 ms
System BW	20 MHz	Backoff	20 ms
eNodeB Processing Time	3 ms	UE Processing Time	3 ms

[7]We do not include a numerical comparison with EAB, as the algorithm that controls the blocking of M2M traffic is not standardized.

[8]Thus, taking into account the 2 RAOs per frame reserved for other purposes (e.g., H2x), the maximum amount of RAOs in each frame is then $L = \{400, 2000, 4000\}$.

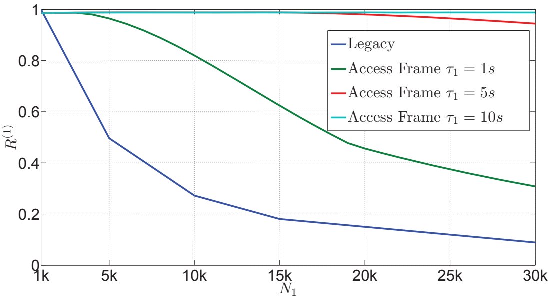

Fig. 5. Achievable transient $R^{(1)}$ within the access frame by the legacy and access frame solutions, with $N_2 = 10k$ and $R^{(1)}_{req} = R^{(2)}_{req} = 0.99$.

traffic due to the alarm reporting. The achievable reliability of TC1, $R^{(1)}$, for different number of active TC1 devices is shown in Fig. 5. We first observe that the LTE legacy with dynamic allocation, is not able to provide reliable access for $N_1 > 1k$ (in the legacy solution TC1 and TC2 are treated in the same way). On the other hand, the proposed mechanism is able to provide a reliable service for a considerably higher range of simultaneously accessing devices. Specifically, the proposed scheme provides service with a reliability guarantee of $R^{(1)}_{req} = 0.99$ for up to $N_1 = 30k$ smart meters if the tolerable delay is $\tau_1 = 10$ s. For TC2, the offered reliability will be constrained by the amount of TC1 arrivals in the same access frame. However, due to TC1 bursty nature and the less restrictive TC2 latency requirement (i.e. $\tau_1 < \tau_2$), we have observed that, after the "storm" caused by the alarms is over, our solution is able to met the set $R^{(2)}_{req}$.

We emphasize, that beyond this specific example, our proposed solution is tailored to offer the traffic reliability requirements, as long as the allowed latency constraints are in accordance with the number of devices to be served. Furthermore, it enables to achieve a trade-off between latency and reliability.

V. Conclusions

One of the key challenges associated with machine-to-machine (M2M) communications in cellular networks is to be able to offer service with reliability guarantees, particularly when a massive amount of simultaneous M2M arrivals occurs. While current solutions take a reactive stance when dealing with massive arrivals, by either imposing barring probabilities or increasing the contention space, they do so without knowledge of the volume of incoming traffic.

Here we propose a proactive approach, based on dedicated access resources for the M2M traffic, combined with a novel frame based serving scheme composed by an estimation and a serving phase. In the estimation phase the volume of arrivals is estimated and then used to dimension the amount of resources in the serving phase, such that reliable service guarantees are provided. The provided framework can be extended for more than two traffic classes, which is one of the future work directions. Other directions include combination of the proposed approach with the existing access control mechanisms, such as the EAB.

Acknowledgment

The research presented in this paper was partly supported by the Danish Council for Independent Research (Det Frie Forskningsråd), grants no. 11-105159 "Dependable Wireless Bits for Machine-to-Machine (M2M) Communications" and no. DFF-4005-00281 "Evolving wireless cellular systems for smart grid communications", and partly funded by the EU project SUNSEED, grant no. 619437.

References

[1] FP-7 METIS, "Requirements and General Design Principles for new air Interface," Deliverable D2.1, 2013.
[2] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A suitable for M2M Communications? A Survey of Alternatives," *IEEE, Communication Surveys Tutorials*, vol. 16, no. 1, pp. 4–16, First 2014.
[3] 3GPP, "Overview of 3gpp release 12," 3rd Generation Partnership Project (3GPP), Tech. Rep.
[4] *Radio Resource Control (RRC); Protocol specification*, 3GPP TS 36.331.
[5] 3GPP, "MTC simulation results with specific solutions," 3rd Generation Partnership Project (3GPP), TR R2-104662, Aug. 2010.
[6] ——, "MTC simulation assumptions for RACH performance evaluation," 3rd Generation Partnership Project (3GPP), TR R2-105212, Aug. 2010.
[7] 3GPP, "Service accessibility," 3rd Generation Partnership Project (3GPP), TS 22.011, Aug. 2010.
[8] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Cellular-based machine-to-machine: overload control," *Network, IEEE*, vol. 26, no. 6, pp. 54–60, 2012.
[9] H. Thomsen, N. K. Pratas, C. Stefanovic, and P. Popovski, "Code-expanded radio access protocol for machine-to-machine communications," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 355–365, 2013.
[10] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, "PRADA: Prioritized Random Access With Dynamic Access Barring for MTC in 3GPP LTE-A Networks," *Vehicular Technology, IEEE Transactions on*, vol. 63, no. 5, pp. 2467–2472, Jun 2014.
[11] M. Kodialam and T. Nandagopal, "Fast and reliable estimation schemes in rfid systems," in *Proceedings of the 12th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '06. New York, NY, USA: ACM, 2006, pp. 322–333.
[12] M. Rivero-Angeles, D. Lara-Rodriguez, and F. Cruz-Perez, "Random-Access Control Mechanisms Using Adaptive Traffic Load in ALOHA and CSMA Strategies for EDGE," *Vehicular Technology, IEEE Transactions on*, vol. 54, no. 3, pp. 1160–1186, May 2005.
[13] J.-B. Eom and T.-J. Lee, "Accurate Tag Estimation for Dynamic Framed-Slotted ALOHA in RFID Systems," *Communications Letters, IEEE*, vol. 14, no. 1, pp. 60–62, January 2010.
[14] C. Stefanovic, K. Trilingsgaard, N. Pratas, and P. Popovski, "Joint estimation and contention-resolution protocol for wireless random access," in *Communications (ICC), 2013 IEEE International Conference on*, June 2013, pp. 3382–3387.
[15] H. Vogt, "Efficient Object Identification with Passive RFID Tags," in *Pervasive Computing*. Springer, 2002, pp. 98–113.
[16] 3GPP, "Study on RAN Improvements for Machine-type Communications," 3rd Generation Partnership Project (3GPP), TR 37.868 V11.0, Aug. 2010.
[17] G. Corrales Madueño, C. Stefanovic, and P. Popovski, "Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering," in *Proceedings of IEEE Internation Conference on Smart Grid Communications (SmartGridComm 2014)*, 2017, pp. 1–5.

Paper 6

Efficient Coexistence of Regular M2M Traffic and Alarm Reporting in IEEE 802.1ah Networks

Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski.

Submitted to IEEE Internet of Things Journal (IoT-J), 2015.

Efficient Coexistence of Regular M2M Traffic and Alarm Reporting in IEEE 802.1ah Networks

Germán Corrales Madueño, Čedomir Stefanović, Petar Popovski

***Abstract*—IEEE 802.11ah protocol is a brand new WiFi based protocol targeting M2M communications in the sub 1 GHz band and extended range operation (up to 1 km). In contrast to the other 802.11 protocols, IEEE 802.11ah is designed according to the requirements of M2M services, i.e., offering service availability for large number of devices that sporadically exchange low amounts of data. We introduce a novel access mechanism that is capable to support traffic due to periodic reporting, but it also adapts to the regime of emergency, when alarm reporting dominates the traffic. We show that alarms are detected while efficiently allocating massive number of stations with restrictive deadlines. As a side result, we also provide a justification for modeling the inter-arrival time in alarm events as Beta-distributed.**

I. Introduction

At the time of writing, the IEEE 802.11ah task group is still working on the complete protocol definition, where the second draft was initially expected by mid-2014 [**?**]. Although the final specifications are still unknown, several authors have already studied its performance and proposed improvements. The number of devices supported for different uplink-downlink traffic ratios with different data rates was investigated in [1]. The performance of the grouping strategy proposed in 802.11ah, denoted as restricted access window (RAW) was studied in [2]. It was shown that limiting the number of contending stations through use of RAW could produce significant throughput improvements. In [3], the authors propose to estimate the number of contending devices in order to determine the adequate length of the RAW.

However, the main figure of interest in the aforementioned studies is the throughput, and, more importantly, the analysis is based on a full-buffer approach (i.e., stations always have data to transmit). In contrast, in this paper we assume two characteristic M2M operating regimes, which differ significantly from the full-buffer approach. The first operating regime is the *asynchronous periodic* reporting, characterized with small payloads and reporting intervals of several minutes [4]. The second operating regime is *synchronous alarm* reporting, where potentially thousands of stations are triggered almost simultaneously. Smart metering is a showcase example of an M2M application in which stations (i.e., meters) could operate in both regimes. Under normal conditions, smart meters asynchronously and periodically report the energy consumption to a remote server. However, in special occasions, such as in a power outage, thousands of devices almost simultaneously try to report the failure before the battery dies (i.e., last-gasp reporting) [5]. In cases like this, the throughput is not the main parameter of interest, but rather the reliability of the service.

The topic of this paper is design of an allocation technique tailored for 802.11ah standard, whose main topic is to provide a reliable and efficient service both for periodic and alarm reporting. The proposed allocation technique is inspired by the work presented in [6], whose main characteristic is a provision of a periodic pool of resources to the devices in the system. We extend the approach, by introducing a proactive (i) dimensioning of the pool of resources and (ii) operation of the access mechanism, such that matches the dominant operation regime in the network (i.e., periodic vs. alarm reporting). We show that it is possible to detect all alarms reports while efficiently allocating massive number of stations.

The rest of the paper is organized as follows. Section II provides an overview of the 802.11ah protocol. Section III present the reporting model and the proposed allocation mechanism. Section V analyzes the performance of the proposed solution, and the corresponding results are presented in Section VI. Finally, Section VII concludes the paper.

II. IEEE 802.11ah Access Mechanism

802.11ah is the wireless local area network (WLAN) approach for M2M operating below 1 GHz. The operating frequency allows a single access point (AP) to provide service to a area of up to 1 km. Additionally, 802.11ah is designed to support very large number of stations (i.e., up to 8K), in comparison to the other standards from the family.

Every station connected to the AP receives an unique identifier, denoted as an association identifier (AID). The AID is a 13 bit word that follows a four-level hierarchical structure as depicted in Fig. 1a). The first two bits are used to organize the stations in four pages. The next 5 bits split a page into 32 blocks, where every block is then sub-divided in 8 sub-blocks using the following 3 bits. Finally, the last 3 bits are used to determine the station's index within the sub-block, i.e., there are 8 stations per sub-block. Obviously, this hierarchical structure allows for a straightforward grouping of stations.

The AID can be used to determine which stations are allowed to access the medium. Specifically, 802.11ah introduces the concept of restricted access window (RAW), during which only certain stations are allowed to contend based on

The authors are with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark (e-mail: {gco, cs, petarp}@es.aau.dk).

The research presented in this paper was supported by the Danish Council for Independent Research (Det Frie Forskningsråd), grant no. 11-105159 "Dependable Wireless bits for Machine-to-Machine (M2M) Communications" and grant no. DFF-4005-00281 "Evolving wireless cellular systems for smart grid communications".

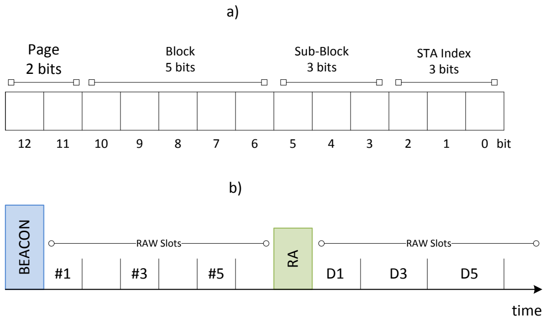

Fig. 1. a) AID Format. b) Beacon frame announcing a RAW with 6 slots followed by a Resource Allocation (RA) frame allocates data slots for stations 1, 3 and 5.

their AIDs (i.e., groups), with the main aim of enhancing the medium access efficiency and limit the number of contending stations. The structure of a RAW is depicted in Fig. 1b), where it can be seen that RAW is subdivided in slots. The information concerning the the number of slots in the RAW, the slot duration and the slot assignment is indicated in the beacon frame sent by the AP to the stations. The RAW assignment can be also done in periodical manner and it is denoted as PRAW. Further, every slot can be assigned to a single or multiple stations, according to the number of stations and number of available slots [7]. The amount of information that can be sent within a slot depends on the slot duration and the modulation used; the available modulation schemes range from BPSK to 256-QAM and the corresponding data rates from 150 kbps to 4000 kbps [8].

The two main RAW operation modes for data exchange between the AP and the stations within the group are generic RAW and triggered RAW, the main difference being in the access method [9]. In case of generic RAW, the stations from the group send their data directly in the RAW slot.[1] Note that stations might be idle, with the consequent idle slots; rendering this procedure inefficient. In the triggered RAW, an initial RAW takes places preceding the actual data transmission. In this initial RAW, every active station transmit a single poll frame. Therefore in triggered RAW, idle or collided slots are less costly due to the shorter slot length compared to the previous case. Based on the received polls, the AP can efficiently allocate a second RAW to receive/transmit data. The resource allocation for the second RAW is notified to the stations through a resource allocation frame (RA) sent immediately after the first RAW. For a better understanding we provide an example in Fig. 1b). The AP provides an initial RAW composed of 6 RAW slots for 6 stations, i.e., a slot per station. It is assumed that only stations #1, #3 and #5 have data to transmit and thus send their polls, whereas #2 and #6 are idle and do not transmit. Immediately after the RAW slots, the AP sends the RA frame providing three additional slots so stations #1, #3 and #5 can transmit their data, denoted as D1, D2 and D3 in Fig. 1b).

[1]In this work we consider the no cross-boundary slot option, implying that a station is only allowed to contend and transmit only within the boundaries of the allocated slot.

III. Reporting Model and Proposed Allocation Method

We assume a general reporting model that comprises periodic, on-demand and alarm reporting in 802.11ah. The periodic reporting is modeled by a Poisson arrival process (on a station basis) with rate $\lambda_p = 1/T_{RI}$, where T_{RI} denotes the duration of the reporting interval (RI). On-demand reporting (e.g., a consumption report trigged by the customer) is also characterized as Poisson arrival process with rate λ_d; it could be expected that $\lambda_p > \lambda_d$ [10]. Finally, alarm reporting corresponds to traffic generated by an event in which affected devices are activated simultaneously; the characteristics of this traffic type are given in Section IV.

The parameter of paramount importance is the maximum allowed delay from the report generation until its delivery. A report whose delay exceeds the maximum allowed value is considered as outdated and dropped by the station. Typically, it is assumed that periodic reports are delay tolerant, i.e., a report in the current RI could be delivered in the beginning of the following RI and thus their maximum allowed delay is $\tau_p = T_{RI}$ [4]. On the other hand, on-demand and alarm reporting have stricter deadlines, i.e., shorter maximum allowed delays, denoted as τ_a and τ_d, respectively [10]:

$$\tau_a \leq \tau_d \leq \tau_p. \tag{1}$$

Allocation Method

The proposed allocation technique consists of a periodically reoccurring pool of resources that is composed of two parts, denoted as preallocated and common pool, as depicted in Fig. 2a). The period by which the pool is reoccurring T_R is determined by the most restrictive maximum allowed delay of the reporting classes, denoted as τ. In the assumed reporting model, $\tau = \tau_a$. Further, in the worst case, a station is triggered by an alarm right after its RS in the current pool and manages to deliver the report at the end of the next pool. Therefore, $\tau > T_R + 2\max\{T_{\text{pool}}\}$, where T_{pool} denotes the pool duration.

The first part of the pool is fixed in size, and consists of preallocated resources in the form of reservation slot (RS). A RS is a RAW slot dimensioned to fit station's AID, including a single bit indicating if there is data waiting to be transmitted in the uplink.[2] In a straightforward approach, one could design preallocated pool such that there is a RS dedicated to every station, implying that all the users have a chance to report. However, this approach, although reliable, is very inefficient, as typically the stations are only periodically reporting with $T_{RI} >> T_R$, and most of the RS will therefore be wasted. The key idea is to preallocate the same RS to a group of stations, i.e., a RS is dedicated to $\Omega \geq 1$ stations, such that the probability of a RSs being idle is kept at the tolerable level. Therefore, the access in the preallocated pool is contention based, where in the worst case there could be Ω

[2]The station also accesses the RAW in case when there are is a pending downlink transmission for it, but this scenario is out of the paper scope.

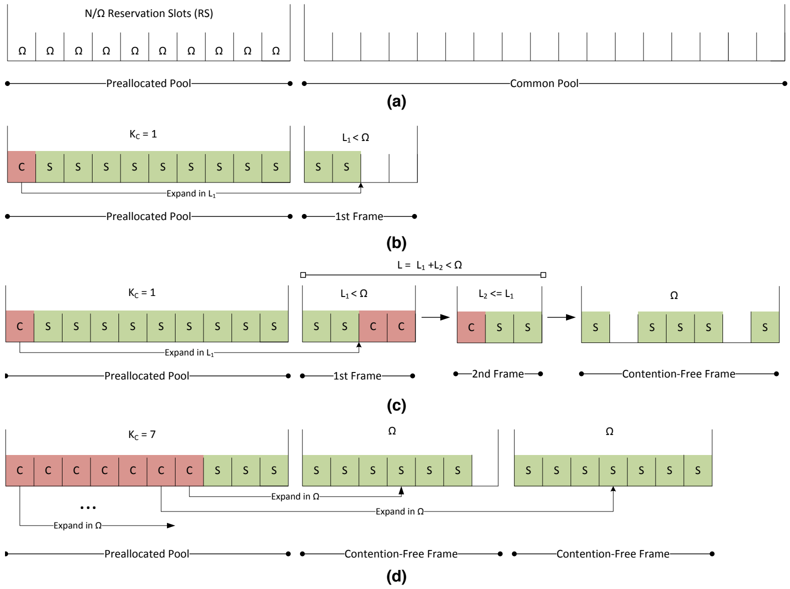

Fig. 2. a) Structure of the pool of resources: preallocated and common pools. b) Preallocated pool working under normal regime, where RSs in collision split in L_1 extra RS. c) Preallocated pool working under normal regime where not all collision where resolved. d) Preallocated pool under alarm regime where every collided RS expand into Ω extra RS.

stations contending per RS. However, as shown in Section V, dimensioning of Ω is also done such that, under periodic reporting, there is a low probability that two or more stations from the group will transmit in the same RS, i.e., that there will be a collision in the RS. If collisions do occur, the AP resolves them in the common pool, but the actual resolution mechanism depends on the number of observed collisions k_C. Specifically, k_C is compared to a predefined threshold Δ_C and then the collisions are resolved either using contention based or contention-free approach. If $k_C < \Delta_C$, the AP assumes that the dominant mode of operation is periodic reporting and every collided RS is resolved by providing a frame with L_1 additional slots, where $L_1 < \Omega$, see Fig. 6b). If all stations are not resolved in the frame with L_1 slots, a second frame with L_2 slots is provided, see Fig.6c).[3] If collisions are still present after these two frames, the AP ultimately provides a new frame with a dedicated slot for every station, i.e., Ω slots, as shown in Fig. 2c). On the other hand, if $k_C \geq \Delta_C$, the AP opts for a contention-free strategy, assuming that the collisions are due to an ongoing alarm event affecting many stations. In that case, Ω slots are provided for every collided RS, see Fig. 2d). It should be noted that RSs of periodic pool are used only to identify active stations and not for the actual data transmission. Finally, we note that by design all active stations are always identified, i.e., all collision are ultimately resolved. Once all the stations have been identified, the actual data transmission starts, where AP assigns data slots from the common pool for each station.

[3]Note that due to slotted Aloha, higher efficiency can be achieved dividing the frame into two different frames than providing one single frame with $L = L_1 + L_2$ slots [11].

IV. The Alarm Model

We consider a circular cell of radius R, where the AP is residing in its center. The deployment of the stations within the cell is uniformly random and the probability density function (pdf) of the distance d with respect to the cell center is:

$$h(d) = \frac{2d}{R}, \, d \in [0, R]. \quad (2)$$

It is expected that in case of an alarm event, e.g., a power outage, a large number of stations will be affected and their reporting becomes correlated in time. In [12], 3GPP proposed a model for highly correlated traffic arrivals, where the inter-arrival time follows a Beta distribution:

$$p(t) = \frac{t^{\alpha-1}(T-t)^{\beta-1}}{T^{\alpha+\beta-1}Beta(\alpha,\beta)} \text{ for } 0 \leq t \leq T, \quad (3)$$

where $\alpha > 0$ and $\beta > 0$ are shape parameters, $Beta(\alpha, \beta)$ is the Beta function [13] and T indicates the activation period, i.e., time period from the activation of the first station until the last station. The number of stations triggered during an interval T_I is given by:

$$N_a = N \int_0^{T_I} p(t)dt. \quad (4)$$

Further, in [12] the suggested values of parameters are $\alpha = 3$, $\beta = 4$ and $T = 10$ s, but no evidence is supplemented either for the use of Beta distribution or for the choice of

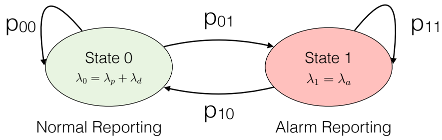

Fig. 3. State Diagram for a given station, where two states can be observed: normal and alarm.

the parameters' values. In the following text, we model the periodic/alarm reporting behaviour of the stations, provide a justification for the use of Beta distribution, provide insights into the choice of the activation period and shape parameters and comment on their relation to the values suggested by 3GPP.

A. Model of Stations' Reporting Behavior

As shown Fig. 3, every station can be in two states, where the first state, denoted as state 0, corresponds to periodic reporting, and the second state, denoted as state 1 corresponds to alarm reporting. The arrival rates for each state are $\lambda_0 = \lambda_p + \lambda_d$ and $\lambda_1 = 1$ arrivals per time step (Δt). In other words, a station in alarm reporting generates one report per Δt. The transition probability matrix of station n, denoted as $\mathbf{P}_n$, and vector of steady state probabilities, denoted as π_n, are:

$$\mathbf{P}_n = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}, \qquad \pi_n = \begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix},$$

where π_n can be obtained from the balance equation as $\pi_n = \pi_n \cdot \mathbf{P}_n$. However, this model does not take into account the correlation between stations in case of an alarm event and the variations over time. In order to model the correlated behavior of the stations, we base our work on [14]. The main idea is that a background process Θ influences the transition probability matrices $\mathbf{P}_n(t)$, $1 \leq n \leq N$, for all N stations:

$$\mathbf{P}_n(t) = \theta_n(t) \cdot \mathbf{P}_1 + (1 - \theta_n(t)) \cdot \mathbf{P}_0. \tag{5}$$

where the samples of the background process $\theta_n(t)$ are in the range $[0, 1]$, and where $\mathbf{P}_1$ and $\mathbf{P}_0$ are the transition probabilities for a station in normal regime and a station in alarm regime. We assume that in normal regime a station never reports alarm, while in alarm regime it reports a single alarm and goes back to normal regime:

$$\mathbf{P}_0 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{P}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Finally, the aggregated arrival rate λ_g of station n since the appereance of the last pool, i.e., τ, is given by:

$$\lambda_g = \sum_{t}^{t+\tau} \lambda_0 \cdot \pi_{n,0}(t) + \lambda_1 \cdot \pi_{n,1}(t). \tag{6}$$

We assume that the background process $\theta_n(t)$ is determined by a physical model of an alarm propagation, instead of a Beta distribution [14], as described in the following subsections.

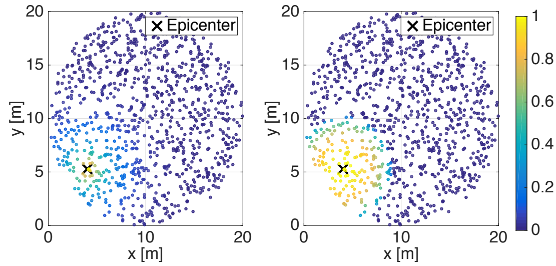

Fig. 4. a)$\Psi_n^{Exp}(d_n)$ for $a = 0.1$ and $N = 1000$ in a radius-cell of 10 m with uniform deployment. b) $\Psi_n^{Sq}(d_n)$ for $d_{max} = 4$ m and $N = 1000$ in a radius-cell of 10 m with uniform deployment.

1) Spatial Correlation Among Stations: An important aspect of the alarm model is the spatial correlation factor Ψ among stations in the cell, which expresses the probability that a station is affected by an alarm event that took place at distance d to the epicenter. Specifically, we analyze three different cases. The first case correspond to the situation where all stations in the cell are affected by the event:

$$\Psi_n = \Psi = 1, \; 1 \leq n \leq N. \tag{7}$$

In the second case, we consider a correlation exponentially decaying with the distance, which is given by:

$$\Psi_n^{Exp}(d_n) = \begin{cases} 0, & d_n < 0, \\ e^{-a \cdot d_n}, & d_n \geq 0, \end{cases} \tag{8}$$

where d_n represents the distance of station n to the epicenter and a the decay constant. In the third case we consider a square root function that provides higher correlation factor in the proximity of the epicenter:

$$\Psi_n^{Sq}(d_n) = \begin{cases} 0, & d_n < 0 \text{ or } d_n > d_{max}, \\ \sqrt{d_{max}^2 - d_n^2}, & 0 \leq d_n \leq d_{max}, \end{cases} \tag{9}$$

where d_{max} is the maximum distance with respect to the epicenter that the event is expected to reach. The examples of the exponential and the square root correlation factor are shown in Fig. 4. The parameters have been chosen such that in both cases, the correlation factor beyond d_{max} can be neglected. It can be seen that the squared root function provides higher correlation values in the are of interest.

2) Alarm Propagation Model: We assume that the event triggering the stations corresponds to a physical phenomenon that propagates at speed v. This means that a station n at distance d_n from the epicenter will be triggered after d_n/v time units, which is modelled by a Dirac delta function centered at d_n/v:

$$\delta_n(t) = \delta\left(t - \frac{d_n}{v}\right), \tag{10}$$

assuming that the event occurred at $t = 0$. By taking into account the correlation factor and the propagation model, the probability of station n being triggered by an alarm at instant t is:

$$\theta_n(t) = \Psi_n(d_n) \cdot \delta_n(t) = \Psi_n(d_n) \cdot \delta\left(t - \frac{d_n}{v}\right). \tag{11}$$

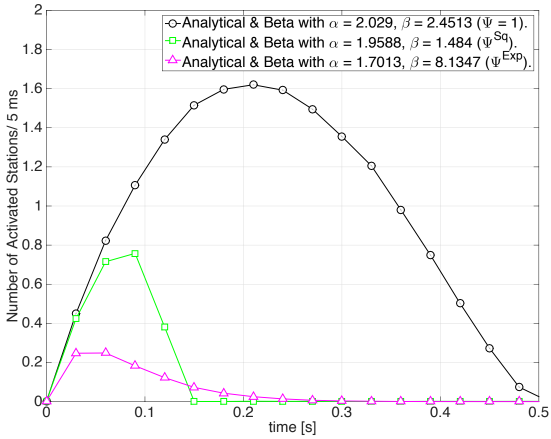

Fig. 5. Number of activations over time for three correlation factors: $\Psi = 1$, Ψ^{Sq} and Ψ^{Exp}. Scenario corresponds to a 1000 m radius cell, where the event propagates at 4000 m/s.

B. Comparison

We provide a short comparison of the alarm reporting patterns for the different correlation factors. The example alarm scenario corresponds to a power outage in the grid due, for example, to an earthquake, where the primary wave propagates at an average speed $v = 4000$ m/s [12], the cell radius is set to $R = 1000$ m and the number of stations is $N = 1000$.

The number of stations activated each 5 ms during an interval of 0.5 s for three correlation factors , i.e., $\Psi = 1$, Ψ^{Sq} and Ψ^{Exp}, is shown in Fig. 5. The results presented in the figure has been validated through simulations. Obviously, the number and spread of activated devices depends on the correlation factor; for $\Psi = 1$ all stations are triggered and the activation period is the largest. The total number of triggered stations for Ψ^{Sq} is larger than for Ψ^{exp}; however, the activation period is longer for Ψ^{Exp}. We further note that all activation patterns in Fig. 5 can be described by a Beta distribution with the appropriate choice of the shape parameters and activation periods, whose values significantly differ from the ones proposed by 3GPP [12]. Specifically, the activation period T is one order magnitude shorter than the suggested value, whereas the number of activated devices over time for the model proposed in [12] is negligible in comparison to the values depicted in Fig. 5.

V. Analysis

In this section we analyze dimensioning of the design parameters: RS degree Ω, number of extra slots in case of collision L and the alarm threshold Δ_c. The aim is find their optimal values such that the proposed access mechanism is able to detect the dominant mode of reporting in the cell, operates reliably and uses the resources efficiently.

Let H_0 denote the hypothesis that the majority of the stations are operating under periodic and on-demand reporting, i.e., in a normal regime. Further, let H_1 denote the alternative hypothesis, i.e., that an alarm event has taken place, affecting a considerable number of stations in the cell. The threshold Δ_c is used to determine what is the dominant operation mode, by comparing it to the number of observed collisions in the periodic pool:

$$H_0 : k_c < \Delta_c \text{ normal regime,} \tag{12}$$
$$H_1 : k_c \geq \Delta_c \text{ alarm regime.} \tag{13}$$

Let C_{ij} denote the cost of deciding H_i when H_j is true, $i, j = 0, 1$. The expected cost is:

$$\begin{aligned} E[C] &= \sum_i \sum_j C_{ij} \cdot P(H_i|H_j) \cdot P(H_j) \quad (14)\\ &= C_{00} \cdot P(H_0|H_0) \cdot P(H_0) + C_{01} \cdot P(H_0|H_1) \cdot P(H_1)\\ &+ C_{10} \cdot P(H_1|H_0) \cdot P(H_0) + C_{11} \cdot P(H_1|H_1) \cdot P(H_1), \end{aligned}$$

where $P(H_0)$ and $P(H_1)$ denote the a priori probabilities of the hypotheses. These probabilities can be calculated using, for example, the knowledge of the history of alarm events:

$$P(H_1) = \frac{\text{no. of alarms/day}}{\text{no. of pools/day}}, \tag{15}$$
$$P(H_0) = 1 - P(H_1). \tag{16}$$

The probability of detecting the normal regime is:

$$\begin{aligned} P(H_0|H_0) &= P(k_C < \Delta_C|H_0) \quad (17)\\ &= \sum_{k=0}^{\Delta_c - 1} \binom{\lceil N/\Omega \rceil}{k} \cdot P_{C_{H0}}^k \cdot (1 - P_{C_{H0}})^{\lceil N/\Omega \rceil - k}, \quad (18) \end{aligned}$$

where $\lceil N/\Omega \rceil$ is the number of RSs and $P_{C_{H0}}$ denotes the probability of collision in a RS in the normal regime. Note that we assume that the contention outcomes are independent over RSs, this assumption is verified by the results presented in Section VI. The probability of false alarm detection $P(H_1|H_0)$ is simply:

$$P(H_1|H_0) = 1 - P(H_0|H_0). \tag{19}$$

The probability of detecting the alarm regime is:

$$\begin{aligned} P(H_1|H_1) &= P(k_C \geq \Delta_C|H_1) \quad (20)\\ &= \sum_{k=\Delta_c}^{\lceil N/\Omega \rceil} \binom{\lceil N/\Omega \rceil}{k} \cdot P_{C_{H1}}^k \cdot (1 - P_{C_{H1}})^{\lceil N/\Omega \rceil - k}, \quad (21) \end{aligned}$$

where $P_{C_{H1}}$ denotes the probability of collision in a RS in the alarm regime. Again, we assume that the contention outcomes are independent over RSs, and verify this assumption in Section VI. Finally, the probability of missing the alarm $P(H_0|H_1)$ is:

$$P(H_0|H_1) = 1 - P(H_1|H_1). \tag{22}$$

A. Collision Probabilities

As already noted, we assume that the contention outcomes are statistically identical over the RSs, and that they depend only on the reporting regime. The probabilities of collision $P_{C_{H0}}$ and $P_{C_{H1}}$ are:

$$P_{C_{H0}} = 1 - (1 - P_{ap})^{\Omega} - \Omega \cdot P_{ap} \cdot (1 - P_{ap})^{\Omega-1}, \quad (23)$$

$$P_{C_{H1}} = 1 - (1 - P_{ag})^{\Omega} - \Omega \cdot P_{ag} \cdot (1 - P_{ag})^{\Omega-1}. \quad (24)$$

where P_{ap} indicate the probability of a station being active due to periodic reporting and P_{ag} due to a combination of normal reporting and alarm reporting as given in:

$$P_{ap} = 1 - e^{-\lambda_0 \cdot \tau}, \quad (25)$$

$$P_{ag} = 1 - e^{-\lambda_g \cdot \tau}. \quad (26)$$

In other words, the probability of collision is given by the probability of at least two stations active in the same RS.

B. Expected Number of Collided RSs

We proceed by deriving the expected number of collided RSs in the four possible cases $\overline{k}_{C_{00}}$, $\overline{k}_{C_{10}}$, $\overline{k}_{C_{01}}$ and $\overline{k}_{C_{11}}$, which are needed for the derivation of the costs in Section V-C. It can be shown that:

$$\overline{k}_{C_{00}} = \frac{\sum_{k=0}^{\Delta_C - 1} k \cdot \binom{\lceil \frac{N}{\Omega} \rceil}{k} \cdot P_{C_{H0}}^k \cdot (1 - P_{C_{H0}})^{\lceil \frac{N}{\Omega} \rceil - k}}{F_0(\Delta_C - 1)}, \quad (27)$$

$$\overline{k}_{C_{10}} = \frac{\sum_{k=\Delta_C - 1}^{\lceil \frac{N}{\Omega} \rceil} k \cdot \binom{\lceil \frac{N}{\Omega} \rceil}{k} \cdot P_{C_{H0}}^k \cdot (1 - P_{C_{H0}})^{\lceil \frac{N}{\Omega} \rceil - k}}{1 - F_0(\Delta_C - 1)}, \quad (28)$$

where $F_0(\cdot)$ denotes the Binomial cumulative density function (cdf) with parameters $B(\lceil \frac{N}{\Omega} \rceil, P_{C_{H0}})$. Similarly:

$$\overline{k}_{C_{01}} = \frac{\sum_{k=0}^{\Delta_C - 1} k \cdot \binom{\lceil \frac{N}{\Omega} \rceil}{k} \cdot P_{C_{H1}}^k \cdot (1 - P_{C_{H1}})^{\lceil \frac{N}{\Omega} \rceil - k}}{F_1(\Delta_C - 1)}, \quad (29)$$

$$\overline{k}_{C_{11}} = \frac{\sum_{k=\Delta_C - 1}^{\lceil \frac{N}{\Omega} \rceil} k \cdot \binom{\lceil \frac{N}{\Omega} \rceil}{k} \cdot P_{C_{H1}}^k \cdot (1 - P_{C_{H1}})^{\lceil \frac{N}{\Omega} \rceil - k}}{1 - F_1(\Delta_C - 1)}, \quad (30)$$

where $F_1(\cdot)$ denotes the Binomial cdf with parameters $B(\lceil \frac{N}{\Omega} \rceil, P_{C_{H1}})$.

C. Costs

Typically, it is to be expected that C_{00} and C_{11} are zero, i.e, that there is no associated cost when a right decision is made. However in our case, since the RS degree Ω plays an important role in the efficiency of the system, we include the number of RS used even when the right decision is made. First we consider C_{00}, i.e., the cost of deciding normal regime H_0 when that hypothesis is true, which is given by:

$$C_{00} = \lceil \frac{N}{\Omega} \rceil + \overline{k}_{C_{00}} [L_1 + C_{RA} + (L_2 + C_{RA}) \cdot (1 - R_{L1}) \quad (31)$$
$$+ (C_{RA} + \Omega) \cdot (1 - R_T)],$$

where C_{RA} denotes the cost in number of RS for the resource allocation frame (RA) sent by the AP to provide additional

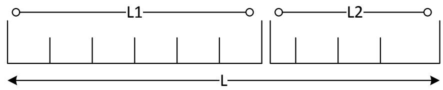

Fig. 6. Two frames with L_1 and L_2 RS ($L = L_1 + L_2$), where the second frame is activated if there was at least one collision in previous frame.

RS; $\overline{k}_{C_{00}}$ indicates the average number of RS in collision. In addition, R_{L1} and R_T denote the probability of resolving all contenting stations within the first frame and within the first or second frames respectively. The above expression can be easily understood as follows (see also Fig. 6. The first term on the right indicates the total number of RS required to provide a transmission opportunity to all the stations in the system. The second term on the right side term takes into account the cost of the average number of RS with collisions, where every collision expands at least in L_1 slots, followed by L_2 if there are not resolved in the first expansion. If after the second frame there are still collisions the system switches to no contention mode providing Ω slots.[4]

We proceed by computing R_{L1} and R_T. Let N_S denote the number of successfully resolved stations. The probability R_{L1} of resolving all contending stations k within the first frame with L_1 slots is given by:

$$R_{L1} = \sum_{k=0}^{\Omega} P(N_S = k | k, L_1) \cdot P(k), \quad (32)$$

where $P(k)$ denotes the probability of k stations contending:

$$P(k) = \begin{cases} \frac{\binom{\Omega}{k} \cdot P_{ap}^k \cdot (1 - P_{ap})^{\Omega - k}}{1 - F(1)}, & 2 \leq k \leq \Omega, \\ 0, & \text{otherwise}, \end{cases} \quad (33)$$

where $F(\cdot)$ denotes the Binomial cdf with parameters $B(\Omega, P_{ap})$, and where the denominator is due to the fact that there were at least two active stations, i.e., there was a collision. The probability $P(N_S = k | N, L_1)$ can be computed as [15]:

$$\Pr[N_S = k | N, L_1] = \frac{\binom{L_1}{k} \prod_{i=0}^{k-1} (N - i) G(L_1 - k, N - k)}{(L_1)^N}, \quad (34)$$

where:

$$G(u, v) = u^v + \sum_{t=1}^{v} (-1)^t \prod_{j}^{t-1} [(v - j)(u - j)](u - t)^{v - t} \frac{1}{t!}. \quad (35)$$

Therefore, taking into account that $N_S = k$ and that $N = k$, (32) can be rewritten as:

$$R_{L1} = \sum_{k=2}^{\Omega} \frac{\binom{L_1}{k} \cdot k!}{L_1^k} \cdot P(k), \quad (36)$$

In a similar manner, R_T is given by the sum of R_{L1} and R_{L2}, where the last term denotes the probability of resolving the remaining stations after the first frame with L_1 slots:

$$R_T = R_{L1} + \sum_{k=2}^{\Omega} R_{L2}(k) \cdot P(k), \quad (37)$$

[4]Recall that we aim to resolve all active stations.

where

$$R_{L2}(k) = \sum_{k_2}^{\Omega} P(N_S = k_2|k_2, L_2) \cdot P(k_2); \quad (38)$$

and where $P(k_2)$ denotes the probability of k_2 stations contending in the second frame, i.e., that there were k_2 collisions in the first frame:

$$P(k_2) = P(N_S = k - k_2|k, L_1). \quad (39)$$

C_{10}, i.e., the cost of deciding the system in alarm regime when it is not true, is:

$$C_{10} = \lceil \frac{N}{\Omega} \rceil + \bar{k}_{C_{10}}(\Omega + C_{RA}). \quad (40)$$

Further, C_{01}, i.e., the cost of deciding that the system is in normal regime when it is not true, is:

$$C_{01} = \lceil \frac{N}{\Omega} \rceil + \bar{k}_{C_{01}} \cdot ((k_1 + 2C_{RA}) + (\Omega + C_{RA})). \quad (41)$$

It should be noted that the likelihood of resolving the stations triggered by the alarm within L slots in negligible. In other words, the contention space is not adequate to the large number of stations active as the AP is not expecting alarm. Therefore, after the two frames there will be always collisions and the access point will provide Ω additional slots.

Similarly, C_{11} is given by average of collided RS, which each of them expands in Ω slots:

$$C_{11} = \lceil \frac{N}{\Omega} \rceil + \bar{k}_{C_{11}} \cdot (\Omega + C_{RA}). \quad (42)$$

VI. Results

In this section we present the results of the proposed solution in terms of the expected costs $E[C]$ and the probability of alarm detection $P_D = P(H_1|H_1)$. The investigated scenario corresponds to heavily loaded case with 8000 smart meters in a 1000-m radius cell, close to the maximum capacity of the protocol (8100 stations).[5] The traffic model of smart meters comprises periodic reporting, on-demand reporting and alarm reporting. For the periodic reporting we select reporting interval (RI) of 5 mins; i.e., the corresponding arrival rate is $\lambda_p = 300^{-1}$ reports/s, which is typical value [4]. For the on-demand reporting, RI is set to 15 mins, i.e., $\lambda_d = 1500^{-1}$ reports/s, which is a rather demanding configuration in comparison with typical values [10]. For the alarm reporting we assume a scenario where the alarm propagates at an average speed $v = 4$ km/s [12], which corresponds to a power outage or power fluctuation in the grid due to, for example, an earthquake, while the correlation between stations with respect to the distance is given by the function Φ^{Sq} eq. (9). According to an exhaustive description of smart grid traffic messages [10], we assume that a maximum tolerable delay for alarm and on-demand reporting of 5 s, and set the pool reoccurring period to $\tau = 2.5$ s, to account for the worst case; i.e., report is generated just after the allocated RS and has to wait to the following pool. Further, RS duration is set to 200 μs, which

[5] We have selected 8k so that Ω is always common divisor to the number of stations.

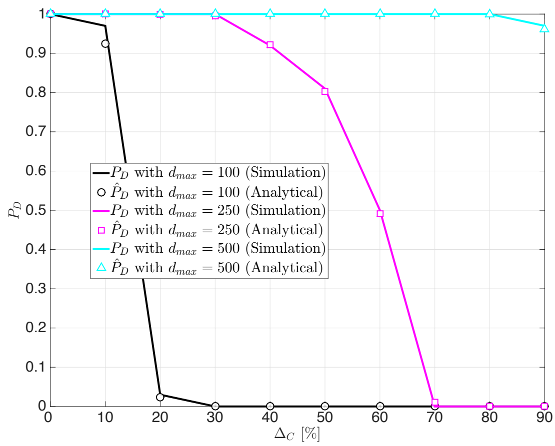

Fig. 7. Probability of alarm detection (P_D) in a 1000-m radius cell with 8000 stations uniformly distributed and $\Omega = 40$. The alarm threshold (Δ_C) is given as percentage of RSs in collision.

is enough to fit the transmission at lowest modulation of the 13 bits AID, including one bit to indicate uplink data, and leaves room for the transmission of the additional 23 bits that may carry other physical layer overheads (e.g., CRC). Finally, the length (i.e., the cost) of the resource allocation frame is set to $C_{RA} = 2$ RS.

Fig. 7 illustrates the probability of detection P_D as function of the alarm detection threshold Δ_C, for varying maximum propagation distance from the epicenter d_{max}. Obviously, Δ_C has to be selected according to d_{max}. For example, if we aim to detect an alarm event with $d_{max} = 250$ m, Δ_C should be set below 30% of RSs in collision, and when $d_{max} = 500$ m, the alarm is always detected if Δ_C is set below 80% of RSs in collision.

Next, we investigate the efficiency of the proposed mechanism as function of the alarm threshold Δ_C and RS degree Ω. We assume that the parameters of the additional frames for collision resolution are $L = 0.5\Omega$, $L_1 = 0.6L$ and $L_2 = 0.4L$, where the selected values can be found to be optimal for the average number of active stations due to periodic reporting. We consider $\Omega = \{1, \cdots, 200\}$, $\Delta_C = \{0\%, \cdots, 90\%\}$ of total RSs in collision and $P(H1) = 5 \cdot 10^{-3}$. Fig. 8 illustrates the expected cost per pool, i.e., the expected number of RS allocated per pool when there are 8000 stations in the system.

Obviously, if Ω is too low, too many RSs needs to be allocated just to detect if a station is active, where the worst case is for $\Omega = 1$ (i.e., every station have a reserved RS per pool). Increasing Ω sharply reduces the cost, until the breakpoint at roughly around $\Omega = 100$ after which too many collisions start to occurs, requiring an increased number of resource to resolve them. The alarm threshold Δ_C also plays an important role in the efficiency of the system. If Δ_C is set too low, probability of false alarm increases, incurring waste of resources in the common pool. On the other hand, with a high Δ_C the probability of miss-detection increases, which it is not a desired effect. Therefore, the optimal configuration is

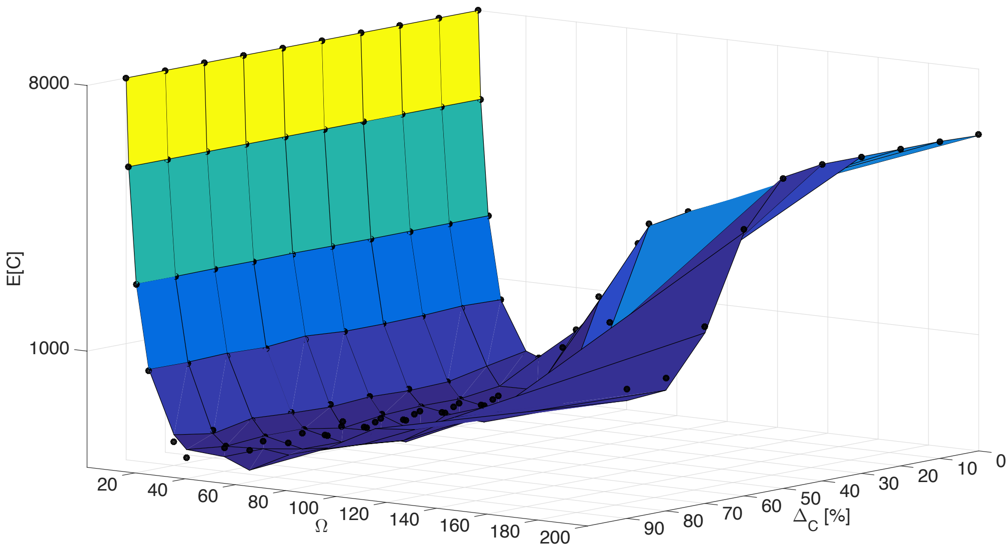

Fig. 8. Expected cost $E[C]$ with respect to the alarm threshold Δ_C as percentage of total RSs in collision and the slot degree Ω, where $L_1 = 0.6\Omega$, $L_2 = 0.4\Omega$ (black dots indicate analytical results).

set to $\Omega = 40$ and $\Delta_C = 50\%$ of total RSs in collision, where $E[C] \approx 400$ RSs.

A closer view of the optimal configuration is shown in Fig. 9. It can be seen that as the number of alarms increases, $E[C]$ increases, which is to be expected as we need to expend more RS in the alarm case (more stations are active). For the sake of comparison we have also included a naive approach where every collision in the initial RS are resolved by providing Ω additional RS (i.e., no contention). It can be seen that for the optimal configuration, the naive approach requires 1.8x times more RS per pool than the proposed solution. It can be shown that decreasing P_{H1} or increasing the reporting interval will make this difference even larger. We also note that as the duration of RS increases (for example to include short message identification), the more important this reduction of cost is. Further, note that for $E[C] \approx 400$ RSs per pool, an average of 6 RSs per RI and station are required ($\eta \cdot E[C]/N$). This means that with only 6 RS, every station is provided with $\eta = 120$ transmissions opportunities during a 5 mins interval. Finally, we note a tight match between the analytical models and simulations presented in this Section.

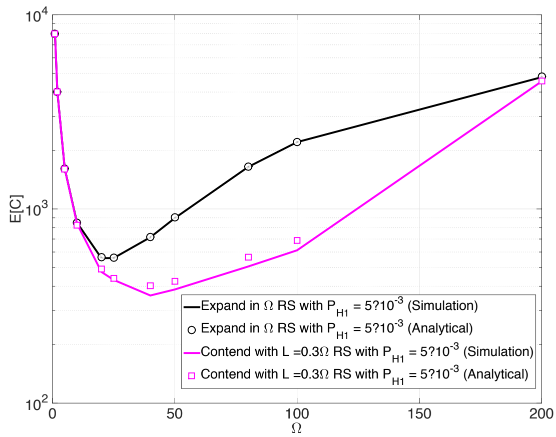

Fig. 9. Expected cost $E[C]$ vs. Ω, where the alarm threshold Δ_C has been set to 50% of RSs in collision.

VII. Conclusion

We have implemented a novel mechanism that (i) proactively dimensions the pool of resources and (ii) operates the access mechanism according to the dominant operation in the network (i.e., periodic vs. alarm reporting). We show that it is possible to detect all alarm reports, while efficiently allocate a massive number of stations with restrictive deadlines. We provide the analytical model for dimensioning the design parameters to optimally operate the system. We note that no modifications are required to the protocol as we exploit the already existing RAW mechanism. In addition, we provide a justification on the Beta distribution used by 3GPP to model alarm events.

References

[1] T. Adame, A. Bel, B. Bellalta, J. Barcelo, J. Gonzalez, and M. Oliver, “Capacity analysis of IEEE 802.11 ah WLANs for M2M communications,” in *Multiple Access Communications*. Springer, 2013, pp. 139–155.

[2] L. Zheng, L. Cai, J. Pan, and M. Ni, “Performance analysis of grouping strategy for dense ieee 802.11 networks,” in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 219–224.

[3] C. W. Park, D. Hwang, and T.-J. Lee, “Enhancement of ieee 802.11ah mac for m2m communications,” *Communications Letters, IEEE*, vol. 18, no. 7, pp. 1151–1154, July 2014.

[4] G. Corrales Madueño, C. Stefanovic, and P. Popovski, "Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering," in *Proceedings of IEEE Internation Conference on Smart Grid Communications (SmartGridComm 2014)*, 2017, pp. 1–5.

[5] IEEE 802.16p, "IEEE 802.16p Machine to Machine (M2M) Evaluation Methodology Document (EMD)," IEEE 802.16 Broadband Wireless Access Working Group (802.16p), EMD 11/0005, 2011.

[6] G. Madueño, C. Stefanovic, and P. Popovski, "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling," *Wireless Communications Letters, IEEE*, vol. 3, no. 4, pp. 429–432, Aug 2014.

[7] Task Group ah (TGah), "Uplink channel access general procedure 0831r00," 2012.

[8] Qualcomm, *White Paper 802.11ah*, 2015 (accessed February 15, 2015). [Online]. Available: https://www.qualcomm.com/invention/research/projects/wi-fi-evolution/80211ah

[9] Y. Zhou, H. Wang, S. Zheng, and Z. Lei, "Advances in IEEE 802.11ah standardization for machine-type communications in sub-1GHz WLAN," in *Communications Workshops (ICC), 2013 IEEE International Conference on*, June 2013, pp. 1269–1273.

[10] Open Smart Grid SG-Network task force, "Smart grid networks system requirements specification, release version 5," 2010.

[11] G. Corrales Madueño, N. K. Pratas, C. Stefanovic, and P. Popovski, "Massive M2M Access with Reliability Guarantees in LTE Systems," in *IEEE ICC 2015*.

[12] 3GPP, "Study on RAN Improvements for Machine-type Communications," 3rd Generation Partnership Project (3GPP), TR 37.868 V11.0, Aug. 2010.

[13] G. Andrews, R. Askey, and R. Roy, *Special Functions*. Cambridge University Press, 2000.

[14] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp, "Traffic models for machine type communications," in *Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on*, Aug 2013, pp. 1–5.

[15] H. Vogt, "Efficient Object Identification with Passive RFID Tags," in *Pervasive Computing*. Springer, 2002, pp. 98–113.

Paper 7

What Can Wireless Cellular Technologies Do about the Upcoming Smart Metering Traffic?

Germán Corrales Madueño, Nuno K. Pratas, Čedomir Stefanović,
René B. Sørensen and Petar Popovski.

Submitted to IEEE Communications Magazine (Special issue on Internet of Things), 2015.

What Can Wireless Cellular Technologies Do about the Upcoming Smart Metering Traffic?

Jimmy J. Nielsen, Germán C. Madueño, Nuno K. Pratas, René B. Sørensen, Čedomir Stefanović, Petar Popovski

Department of Electronic Systems, Aalborg University, Denmark

Email: {jjn,gco,nup,cs,petarp}@es.aau.dk

Abstract

The introduction of smart electricity meters with cellular radio interface puts an additional load to the wireless cellular networks. Currently, these meters are designed for low duty cycle billing and occasional system check, which generates a low-rate sporadic traffic. As the number of distributed energy resources increases, the household power will become more variable and thus unpredictable from the viewpoint of the Distribution System Operator (DSO). It is therefore expected, in the near future, to have an increased number of Wide Area Measurement System (WAMS) devices with Phasor Measurement Unit (PMU)-like capabilities, thus allowing the utilities to monitor the low voltage grid quality while providing information required for tighter grid control. From a communication standpoint, the traffic profile will change drastically towards higher data volumes and higher rates per device. In this paper, we characterize the current traffic generated by smart electricity meters and supplement it with the potential traffic requirements brought by introducing enhanced Smart Meters, which have PMU capability. Our study shows how GSM/GPRS and LTE cellular system performance behaves with the current and next generation smart meters traffic, where it is clearly seen that the PMU data will seriously challenge these wireless systems. We conclude by highlighting the possible solutions for upgrading the cellular standards, in order to cope with the upcoming smart metering traffic.

What Can Wireless Cellular Technologies Do about the Upcoming Smart Metering Traffic?

I. Introduction

Smart power grids represent an important group of devices and applications within the umbrella of internet of things (IoT). Especially, the large number of network-connected smart electricity meters that already are or will be located in all households and commercial/industrial locations are representative examples of IoT devices. At present, smart electricity meters are primarily used by electricity providers only for availability monitoring and billing. However, with the increasing number of distributed energy resources (DERs), such as wind turbines, solar panels, and electric vehicles, strong and sometimes unpredictable variations in the power quality are introduced, prompting for an increased need of monitoring and control. Specifically, distribution system operators (DSOs) need to be able to observe the circumstances in the low voltage (LV) power grid by introducing more frequently-sampled measurement points. Such wide area measurement systems (WAMS) exist already in the transmission grid, whereas in the distribution grid the DSOs rely mainly on open loop control beyond the substation level, i.e., without real-time feedback from consumers. As the number of DERs increases, this control loop must be closed by providing the feedback from measurements in the LV grid, enabling the state estimation and prediction of the grid behavior, and ultimately ensuring stable operation [1]. It is expected that in the future LV grid, in addition to the traditional smart meter (SM) that is so far primarily used for billing purposes with hourly or daily reports; another more advanced monitoring node will be needed, here referred to as Enhanced Smart Meter (eSM). The eSM is largely similar to a Wide Area Measurement System (WAMS) node, as integrates Phasor Measurement Unit (PMU) capabilities; in other words, the eSM measures more frequently and more detailed quality parameters (such as power phasors) compared to SMs [2]. While it is generally expected that not all smart meter locations need to be equipped with eSM devices, the fraction of eSMs needed in the distribution grid to achieve satisfactory state estimation is still an open research question [3].

Today, traditional smart meters are often connected via Power Line Communications (PLC) to a concentrator that collects measurements from a group of smart meters, and then uses a cellular GPRS connection to communicate with the DSO backend system. While this works well for collecting billing information on an hourly or daily basis, current PLC based systems, where blocks of smart meters can be connected in long chains and their messages therefore may experience large forwarding delays, may not be able to satisfy communication requirements of future applications and monitoring functions, e.g., real-time control or advanced flexibility scheduling. An option to enable connectivity to the DSO backend system, is to equip smart meters with cellular interfaces, so as to eliminate the potential delays, ease deployment and reduce maintenance costs associated with the network connectivity.

The traffic profile generated by smart meters falls into the category of Machine-to-Machine (M2M) traffic. A main characteristic of M2M traffic is that it consists of transmissions of small amounts of data from a very high number of devices, differing significantly from the bursty and high data rate traffic patterns in human-oriented services, and instead requiring network reliability and availability. Further, M2M traffic is more demanding in the uplink and less focused on downlink performance, as typical use cases encompass monitoring and control functions.

With LTE gaining an increasing market share, it is expected that within a number of years one of the existing 2G or 3G systems will be taken out of service in order to re-harvest the spectrum to use for newer technologies. Current reports on the active M2M cellular devices indicate that $64\ \%$ of them are GSM/GPRS-only, $25\ \%$ both 3G and GSM/GPRS compatible, 10% 3G-only, and only 1 % is LTE capable [4]. It is clear that GSM/GPRS dominates the M2M industry, and therefore in this paper investigate how well this technology can support the connectivity demands of SM and eSM devices. Specifically, in this paper we have the following four contributions: 1) extraction and classification of smart meter traffic models from relevant specifications, as well as predicted future traffic growth; 2) comprehensive simulation model of radio access systems that includes all phases in the access, in contrast to [5], [6] that use simplified models; 3) quantitive assessment of how many smart meter devices can be supported in cellular systems, comparing the simplified and comprehensive simulation model results; and 4) recommendations for standardization and future roadmap of the radio access technologies.

The rest of this paper is organized as follows. In Section II and III we characterize the traffic models of the SM and eSM devices. In Section IV we describe the access bottlenecks in the cellular access reservation protocol and provide numerical results that show how the proposed traffic models affect the performance of a GPRS and LTE network. Then, in Section V we provide directions which the future cellular network standards should take in account when designing the network system. Finally, we wrap-up the paper with the main take home points.

II. SMART METER TRAFFIC MODEL

In the literature there are different examples of traffic models for traditional smart meters, where the most often cited references are [7]–[9]. Of these, the OpenSG *Smart Grid Networks System Requirements Specification* [9] from the Utilities Communications Architecture (UCA) user group is the most coherent and detailed network requirement specification, and it has therefore been used in this paper as input for the SM traffic model. The UCA OpenSG is a relevant consortium of 190 companies and the considered smart grid use cases are in line with those studied by other standization organizations such as ETSI and USEF. Since there are differences in which use cases and applications are offered by the DSO or electricity retail company and which of those the individual costumers are using, a one size fits all traffic model does not exist. In the following we consider a comprehensive configuration where all use cases that involve communication from the smart meters to the core network will be in operation and note that actual deployments with different configurations may lead to different results. For calculating the message frequency in the uplink SM traffic model the event occurrence frequencies listed in Table I have been used. Besides the values listed in Table I we assume that a commercial/industrial SM sends a 2400 bytes meter reading packet every hour, whereas a residential SM sends a 1200 bytes report every 4 hours.

Resulting from these assumptions is the SM uplink traffic model presented in Figure 1. The gray boxes represent the different use cases and the boxes span the latency and payload size requirements of the corresponding messages. The white box represents the eSM traffic (defined in Section III). Nearly all use cases have reliability requirements of 98%, with the exceptions being two alarm messages in the IDCS use case requiring 99%, and the periodic meter reading, which has time-dependent reliability requirements ranging from 98% to 99.5%. In relation to the figure, Table II shows the average estimated uplink/downlink bandwidth for each use case.

The λ-values in the figure shows the number of generated messages per day per SM. The use cases grouped in the dash-dotted box transmit very infrequently with a combined rate of only ~ 0.5 messages per day. Further they are relatively similar in terms of latency and payload size. In addition to this group, two other OpenSG use cases from the figure stand out, namely

Event	Frequency [events per meter]
On-demand meter read requests	25/1000 per day
Meter capped energy mode request	5 per year
DR load management request to HAN devices	15/1000 per day
HAN device join/unjoin	5 per year
Real-time price (RTP) update	96 per day
Metrology firmware update	4 per year
Metrology program update	4 per year
NIC firmware update	4 per year
NIC program update	4 per year

TABLE I

ASSUMPTIONS FOR DERIVING TRAFFIC MODEL.

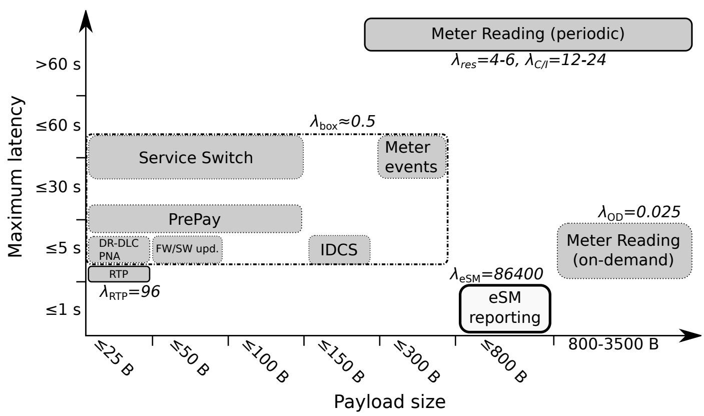

Fig. 1. Classification of OpenSmartGrid traffic originating from an SM. λ-values show the number of generated messages per day per device. Use case short names: Demand Response - Direct Load Control (DR-DLC), Premise Network Administration (PNA), Firmware and Software updates (FW/SW upd.), Real-Time Price (RTP), Islanded Distributed Customer Storage (IDCS).

the real-time pricing (RTP) that causes 96 messages per day and the periodic meter reading on the top right. For periodic meter readings, a commercial/industrial (C/I) SM sends reports more often than a residential SM. Notice for the eSM reporting that in addition to the stricter latency requirement of ≤ 1 sec, the number of generated messages per day is many orders of magnitude higher than any of the SM use cases.

Use case	Avg. bandwidth uplink	downlink
Meter Reading	30K	1.25
Service Switch	6	3
PrePay	8	3.5
Meter Events	50	0
Islanded Distr. Cust. Storage	5	2
DR-DLC	0.5	400
Premise Network Admin	1	1
Price	2.4K	10K
Firmware / Program Update	5	30K
Total	32.5K	40.4K

TABLE II

AVERAGE UPLINK/DOWNLINK BANDWIDTH AS [BYTES/METER/DAY] FOR THE CONSIDERED USE CASES.

Table II shows that the bandwidth requirements of SMs are overall quite balanced but relatively modest, with an average uplink bandwidth of appr. 32.5KB per day per SM and an average downlink bandwidth of appr. 40.4KB per day per SM. While the total downlink bandwidth is actually higher than the uplink, it is constituted primarily of software updates, which are large low-priority data transfers that occur infrequently during the night, where it does not interfere with the day-to-day operation of the smart grid. Given the modest traffic requirements, it is interesting to see if GSM/GPRS networks, that are widely deployed in most places, but becoming less and less suitable for human originating network traffic, can satisfy SM traffic requirements. Further, an option to increase observability in the power grid, is to reduce the meter reading reporting interval. While considering that the report packet sizes are respectively 300 bytes and 600 bytes for residential and commercial/industrial and reporting intervals from 5 min, 1 min, 30 sec, to 15 sec. We investigate how capable the current cellular systems are to support this in Section IV.

III. ENHANCED SMART METER TRAFFIC MODEL

Since eSM devices, defined as PMU-equipped devices connected through cellular networks, have not yet been standardized, the eSM traffic model in this study will be inspired from the PMU and WAMS related standards, IEEE 1588, IEEE C37.118 and IEC 68150. Data from PMUs can be used on different time scales, ranging from a few milliseconds (e.g., for protective relays) up to several seconds (e.g., real-time monitoring and state estimation) [10]. Since the eSMs are primarily intended to improve observability and enable state estimation, a report interval of 1 second is considered for the eSM. Specifically, we will assume that every 1 second an eSM sends a measurement report that consist of concatenated PMU measurements (50 Hz sample rate) from the preceding 1 second measurement interval. The samples are, as specified in PMU standards IEEE 1588 and C37.118, timestamped using GPS time precision. Assuming that the floating poing PMU frame format from IEEE 1588 is used and each sample covers 6 phasors, 1 analog value and 1 digital value, each PMU frame accounts to 76 bytes. Adding UDP header (8 bytes) and IPv6 header (40 bytes) to each report of 50 PMU samples, an eSM packet is 3848 bytes, and a bit rate of 30 kb/s. Since it may be an exaggeration to send all 50 PMU samples per measurement interval, we also consider in our performance analysis the case of eSM reduced report sizes.

IV. Cellular Systems Performance

From the communications perspective, it is important to investigate which cellular technologies can support the current billing-only smart meter use cases, but also the use cases/services that go beyond the current ones. In [6] and [5] performance analyses were carried out to determine the number of smart grid devices supported by different wireless technologies, however, they only evaluated the data capacity of the systems and neglected to account for the bottlenecks in the access reservation protocol used in cellular systems. As it is shown in [11] the access reservation bottlenecks are particularly prone to exposure with M2M traffic such as smart grid traffic, meaning that a pure data capacity based analysis may lead to overly optimistic results. Therefore, our analysis will include all aspects of the access reservation procedure and compare those results to a data capacity only analysis. For the analysis we will consider the traffic patterns for SM and eSM devices described in Sections II and III. From those traffic models it is clear that the communication requirements of these two device types are orders of magnitudes apart in terms of message frequency and bandwidth, meaning that for eSM deployment a more capable technology than GSM/GPRS is needed. With its integrated PMU unit, the eSM is already a more complex and expensive device than the SM, and since fewer eSMs than SM will be needed, a higher unit price can be better tolerated, and thus we will assume that it is likely that the eSM uses LTE.

A. Access Reservation Protocol Operation and Limitations

In cellular networks, if a device needs to perform a data transmission and it does not already have an active connection to the network, then it first has to establish one. This is accomplished via an access reservation protocol, which in general consists of three main stages: random access, granting access and data access. In the first stage, the cellular devices perform a random access request in one of the random access opportunities (RAOs). In the second stage, the base station grants access to the network if: (i) the random access request is received without error by the base station; (ii) no other device has transmitted in the same RAO (i.e., collision free); and (iii) there are data resources available to the device. Otherwise, the access reservation procedure must be restarted and the device will transmit a new random access request until it is granted by the base station or until the maximum number of retransmissions is reached and the request fails. In GSM/GPRS there are 217 RAOs/s per carrier while in LTE there are 10.8k RAOs/s[1]. On the other hand, only 32 grants/s and 3k grants/s are offered in GPRS and LTE, respectively [11], [12]. Therefore, when the random access stage is heavily loaded, the grant stage becomes a decisive limitation in cellular networks. Furthermore, in GPRS and LTE the data stage is not only limited by the actual data resources, but also by the uplink identifiers used to coordinate transmissions from active devices, posing an important limitation to the amount of simultaneously active M2M communication links.

B. Outage Performance Evaluation

The individual device outage is used as the metric to evaluate the cellular system performance, since it measures the probability of a device not being served before reaching the maximum number of connection attempts. In other words, it measures the reliability of the communication service offered by the cellular network. The evaluation scenario is set in a single cell with 1000-m radius, which includes 4500 smart meters [11], from which 90% correspond to residential customers and the remaining 10% to commercial/industrial customers. The considered LTE bandwidth is 1.4-MHz (6 PRBs), in line with the reduced capabilities for LTE devices [13].

[1]Assuming the contention resources occur every 5 ms, each with 54 contention preambles available.

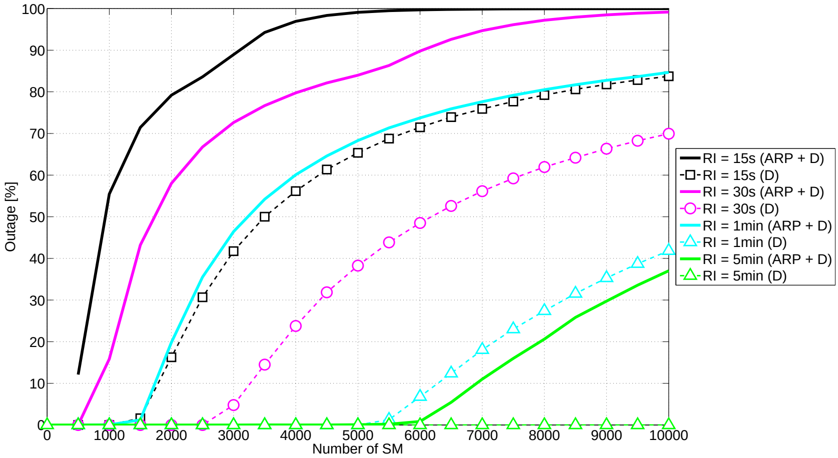

Fig. 2. GPRS outage evaluation for increasing number of SM with different report interval values and RS = 300 Bytes for residential and RS = 600 Bytes for commercial/industrial, where ARP+D denotes the access reservation protocol plus data phase, while D denotes only data phase.

In addition, the control channel and data channel probability of error, are respectively 10^{-2} and 10^{-1} [12]. In the case of GPRS, we consider a single carrier corresponding to a 200 KHz system. In both systems, we assume the devices always transmit with the highest modulation scheme available, in order to focus the evaluation on the performance of the access reservation protocol. In these conditions, we observed that the SM traffic, provided in Section II, is supported by both GPRS and LTE with near 0% outage, as the total number of messages per hour from each SM only amounts to approximately 125. We start by considering for GPRS the scenario of reducing SM Report Intervals (RI), as described in Section II. Fig. 2 depicts the outage probability for increasing number of SMs and different RIs. Taking as reference a cell population of 4500 SMs, we can see that for RI $>$ 5 min, GPRS can provide a significant increase on the distribution network observability from hourly intervals to every 5 minutes. For smaller report intervals to be supported in GPRS, then the options are either to reduce the cell size and/or increase the number of carriers.

We proceed by considering the cellular network outage as a function of the eSM penetration,

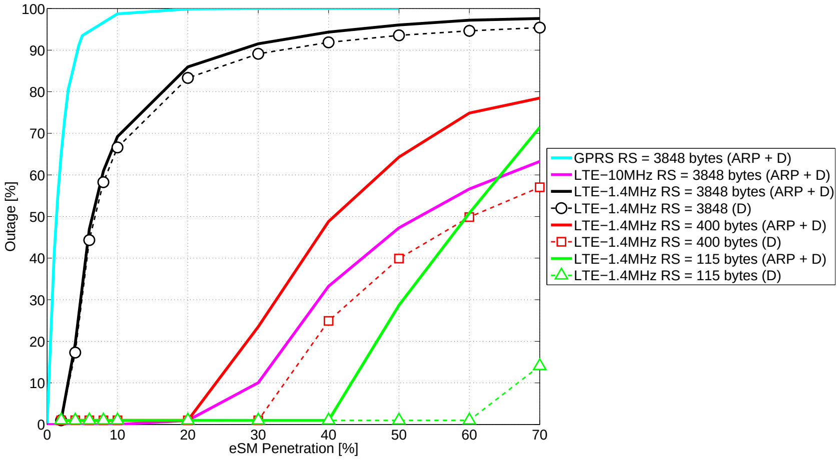

Fig. 3. LTE and GPRS outage evaluation for increasing penetration of eSMs, where ARP+D denotes the access reservation protocol plus data phase, while D denotes only data phase.

i.e., of how many eSMs are deployed per every 100 smart meter locations. As described in Section III, each eSM report contains 50 samples of the power phasors measured since the last report with an expected payload of 3848 bytes. Since this large payload has severe implications on the cellular network performance, we also consider the impact of smaller payloads on system performance, which can be motivated by the introduction of pre-processing to extract statistics, data compression and/or reduced number of samples. Specifically, we consider reduced report sizes (RS) of RS $=\{3848$, 400, and $115\}$ bytes, for which the last two values correspond respectively to a payload reduction of approximately 10% and 3% of the original payload size. The outage results for LTE and GPRS are shown in Fig. 3. We note that GPRS is not able to support eSM traffic irrespective of the chosen RS, while LTE for $RS = 3848$ only supports up to 2% eSM penetration. When a 10 MHz bandwidth is completely dedicated in LTE to serve the eSM traffic then it is possible to reach 30% of penetration with less than 10% of outage, which means a large amount of resources dedicated to a potentially low profit application. On the other hand, if we assume lower RS, already at 400 bytes LTE supports up to 20% of eSMs. Further, when comparing the results that correspond to the case when only data phase is taken into account with the results obtained by considering the access reservation phase as well, it can be observed that the access reservation protocol impacts the number of supported eSMs. Particularly, the limitations of the access reservation protocol become substantial as the report size decreases and it could shown that this is mainly due to the lack of access grant messages required to complete the access reservation procedure. Note that this effect has been overlooked in the previous works [5], [6].

The presented results allow us to conclude that the RS of the eSM nodes must be small to support a high percentage of nodes. In addition, we emphasize that small data traffic cannot be analyzed only in terms of the system data capacity, but that the bottlenecks of the access reservation protocol itself must be considered, as observed in the gap between the two types of analysis depicted in Fig. 3.

V. Standardization Outlook

Although the traffic resulting from smart meters can be easily accomodated into current cellular systems, the same is not observed for the traffic generated by the eSM. In the following, we discuss the challenges and possible solutions that need to be tackled by standardization bodies so to ensure that the observability of the distribution network can be improved.

A. Smart Meter

The inclusion of additional phasor measurement units into the distribution grid, so to increase its observability, is being discussed specifically at the last mile to the customer permises [1]. Currently, it is not yet clear if that will imply the same level of detail (in number of samples and report frequency) as in the WAMS nodes PMU's, where the reporting is done by SCADA over dedicated wired links.

As discussed in Section IV, if the eSMs generate the same amount of traffic as WAMS, then the cellular networks will require an extensive overhaul so that they can support both eSM and human centric traffic, leading to substantial investment in the cellular infrastructure. On the other hand, it is plausible, that eSM's PMUs will be lighter versions, which will both sample and report less frequently. Therefore, if local processing and compression of the monitoring data is allowed and/or the required level of detail lowered, then the amount of generated traffic will be much lower. Another viable option, as discussed in Section IV, is to increase the report frequency of current smart meters without introducing local PMU functionality. The generated small packets could then be handled by the network, as long as the bottlenecks at the access protocol level are addressed.

It seems likely that the standardization for the eSM's PMU functionality falls within the scope of the IEEE C37.118 and IEC 68150 standards, since these specify the measurement and communications requirements for traditional PMU units. Therefore, it is of paramount importance that standardization bodies reach a consensus on the eSM communication requirements allowing the affected stakeholders to take informed actions.

B. Cellular Network

In 3GPP, the standardization body responsible for the cellular air interface and core network functionality, there are two activities that will affect how the traffic from SM and eSM will be handled [13].

We start by noting, that although GRPS is seen as an outdated communication technology [13], there is an ongoing effort to continue to reenginer GPRS to serve M2M applications, in which the SM traffic can be classified into. One of the goals of this initiative is to achieve[2] 160 bps. Concurrently, there is a push from the industry (both utilities and vendors) to keep GPRS networks and their associated resources active, while facing the pressure to re-harvest the GPRS spectrum to be used in the next cellular network generation. Where a viable solution to keep the GPRS connectivity, is to *virtualize* its air interface into the next generation cellular systems.

The second effort is to define a low complexity LTE user equipment category with respect to the cellular interface, which supports reduced bandwidth and transmit power while extending coverage operation [13]. Specifically, the goal of reduced bandwidth is to specify 1.4 MHz operation within any LTE system bandwidth, allowing operators to multiplex reduced bandwidth MTC devices and regular devices within their existing LTE deployments. In terms of extended coverage the goal is to improve the coverage of delay-tolerant MTC devices by 15 dB, thereby allowing operators to reach MTC devices in poor coverage conditions, such as smart meters located in basements [13].

[2]Considering the minimum SDU size, i.e. 80 bytes, with 4 seconds latency.

To further improve the support of the traffic generated by SM and eSM with very low duty cycle and latency requirements in the order of seconds, the inclusion of *periodic reporting and discontinuous transmission* functionality into cellular standards should be considered. In here, the network provides periodic communication resources so that devices can perform their short data transmission. This allows devices to go to sleep and save energy, since they have prior knowledge of when the next transmission time slot can occur. A solution based on this concept [12], has been proposed through the reengineering of the LTE access protocol.

To cope with the eSM traffic demands and increase the network capacity, then *localized aggregation of traffic* should be considered. In this solution the traffic generated by multiple SMs and eSMs in a geographical area could be aggregated, at eSMs or cellular relays, and then trunked to the cellular network [14]. The use of aggregation and relaying would then allow to decrease the contention pressure at the base station, as well as to improve the single link connection, providing connectivity and coverage enhancements to SMs and eSMs with poor propagation conditions.

Finally, to support *massive asynchronous access of small packet transmissions*, access reservation protocols in cellular systems are just the first step of the asynchronous access to the network. After it as been completed, then the device starts exchanging signalling information via the higher layers with the entities in the core network, which leads to a high signalling overhead and possible air interface and core network congestion. Although there are already efforts to reduce the signalling exchanges with the core network [15], when the payloads are small enough, the facility to perform the data transmission already in the third step of the access reservation protocol should be in place.

VI. Conclusion

In this paper we have evaluated two approaches to increase the observability of the network: (1) decreasing the report interval of the meter reading and (2) introduction of enhanced smart meters with phasor measurement units (PMUs). We provided details on the characteristics of the traffic generated by smart meters and enhanced smart meters and have highlighted the associated challenges in supporting it from a cellular network point of view. The obtained results from GPRS have shown, that it can support traditional smart meter traffic, as well as more frequent measurements down to 5 min report intervals. It was further shown that LTE can support distribution grid PMUs, if the report payloads are appropriately dimensioned. These results can be used as input for both smart meter and cellular system standardization bodies to enable the introduction of current and future smart grid devices into the cellular networks. The current main open issue is the uncertainty associated with the eSM communication requirements, which will lead to different cellular systems optimizations.

ACKNOWLEDGMENT

This work is partially funded by EU, under Grant agreement no. 619437. The SUNSEED project is a joint undertaking of 9 partner institutions and their contributions are fully acknowledged.

REFERENCES

[1] Y.-F. Huang, S. Werner, J. Huang, N. Kashyap, and V. Gupta, "State estimation in electric power grids: Meeting new challenges presented by the requirements of the future grid," *Signal Processing Magazine, IEEE*, vol. 29, no. 5, pp. 33–43, 2012.

[2] "TR 102 935 "M2M Applicability of M2M architecture to Smart Grid Networks Impact of Smart Grids on M2M platform" V2.1.1," ETSI, Tech. Rep., 2012.

[3] L. Schenato, G. Barchi, D. Macii, R. Arghandeh, K. Poolla, and A. Von Meier, "Bayesian linear state estimation using smart meters and PMUs measurements in distribution grids," in *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*. IEEE, 2014, pp. 572–577.

[4] A. Ericsson, "Interim Ericsson Mobility Report, February 2014," 2014.

[5] C. Hagerling, C. Ide, and C. Wietfeld, "Coverage and capacity analysis of wireless M2M technologies for smart distribution grid services," in *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*. IEEE, 2014, pp. 368–373.

[6] NIST, "NIST PAP2 guidelines for assessing wireless standards for smart grid application," 2012.

[7] J. G. Deshpande, E. Kim, and M. Thottan, "Differentiated services QoS in smart grid communication networks," *Bell Labs Technical Journal*, vol. 16, no. 3, pp. 61–81, 2011.

[8] U. Department of Energy, "Communications requirements of smart grid technologies," Technical report, US Department Of Energy, 10 2010. Available at http://energy.gov/gc/downloads/communications-requirements-smart-grid-technologies, Tech. Rep.

[9] Open Smart Grid SG-Network task force, "Smart Grid Networks System Requirements Specification, Release Version 5," 2010.

[10] M. Adamiak, B. Kasztenny, and W. Premerlani, "Synchrophasors: definition, measurement, and application," *Proceedings of the 59th Annual Georgia Tech Protective Relaying, Atlanta, GA*, 2005.

[11] G. C. Madueno, C. Stefanovic, and P. Popovski, "Reengineering GSM/GPRS towards a dedicated network for massive smart metering," in *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*, Nov 2014, pp. 338–343.

[12] G. Madueno, C. Stefanovic, and P. Popovski, “Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling,” *Wireless Communications Letters, IEEE*, vol. 3, no. 4, pp. 429–432, Aug 2014.

[13] “Overview of 3GPP Release 13,” 3GPP, Tech. Rep., 2015.

[14] G. Rigazzi, N. Pratas, P. Popovski, and R. Fantacci, “Aggregation and Trunking of M2M Traffic via D2D Connections,” *Accepted in ICC 2015*, 2015.

[15] “TR 23.887 - Study on Machine-Type Communications (MTC) and other mobile data applications communications enhancements,” 3GPP, Tech. Rep., 2013.

Complete List of Publications

Conference Proceedings

[21] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "How Many Smart Meters can be Deployed in a GSM cell?" In: *International Conference on Communications (ICC) 2013 - Workshop Telecom R2S*. Budapest, Hungary, June 2013, pp. 1283–1288.

[22] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering". In: *International Conference on Smart Grid Communications (SmartGridComm) 2014*. 2014, pp. 338–343.

[24] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Efficient LTE Access with Collision Resolution for Massive M2M Communications". In: *IEEE Global Communications Conference 2014 (GLOBECOM) - Workshop Ultra2*. 2014.

[25] Germán Corrales Madueño, Nuno Pratas, Čedomir Stefanović, and Petar Popovski. "Massive M2M Access with Reliability Guarantees in LTE Systems". In: *International Conference on Communications (ICC) 2015*. 2015.

[28] P. Popovski, G.C. Madueño, L.C. Gimenez, L. Luque Sanchez, and N.-C. Gjerrild. "Protocol coding for reliable wireless bits under jamming: Concept and experimental validation". In: *Military Communications Conference, 2011 - MILCOM 2011*. 2011, pp. 113–118.

Journal Papers

[23] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling". In: *IEEE Wireless Communications Letters (WCL) 2014* 3.4 (2014), pp. 429–432. ISSN: 2162-2337.

[26] Germán Corrales Madueño, Čedomir Stefanović, and Petar Popovski. "Efficient Coexistence of Regular M2M Traffic and Alarm Reporting in IEEE 802.1ah Networks". In: *Internet of Things Journal, IEEE* (2015).

[27] Jimmy Nielsen, Germán Corrales Madueño, Nuno Pratas, Čedomir Stefanović, René B. Sørensen, and Petar Popovski. "What Can Wireless Cellular Technologies Do about the Upcoming Smart Metering Traffic?" In: *Communications Magazine - Special issue on Internet of Things, IEEE* (2015).

Posters

[29] Germán Corrales Madueño, Nuno K Pratas, Henning Thomsen, Čedomir Stefanović, and Petar Popovski. "Massive Machine-to-Machine Communications". In: *European Telecommunications Standards Institute (ETSI) M2M Workshop*. Sophia Antipolis, France, Oct. 2012.

Patents

[30] Germán Corrales Madueño, Nuno K Pratas, Čedomir Stefanović, and Petar Popovski. "2G Cellular Networks Reengineering for massive number of devices with Low-Rate". In: *J.No. 2012-521/10-0291*. 2012.