Digital Communications and Networks (2015) 1, 1–19

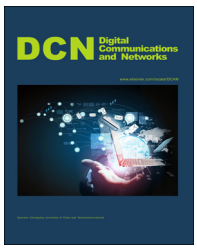
HOSTED BY

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/dcan



The challenges of M2M massive access in wireless cellular networks

Andrea Biral, Marco Centenaro, Andrea Zanella*, Lorenzo Vangelista, Michele Zorzi

Department of Information Engineering of the University of Padova, Via Gradenigo 6/B, 35131, Padova, Italy

Received 21 January 2015; received in revised form 16 February 2015; accepted 28 February 2015
Available online 27 March 2015

KEYWORDS
M2M;
MTD;
MTC;
Massive access;
5G

Abstract
The next generation of communication systems, which is commonly referred to as 5G, is expected to support, besides the traditional voice and data services, new communication paradigms, such as Internet of Things (IoT) and Machine-to-Machine (M2M) services, which involve communication between Machine-Type Devices (MTDs) in a fully automated fashion, thus, without or with minimal human intervention. Although the general requirements of 5G systems are progressively taking shape, the technological issues raised by such a vision are still partially unclear. Nonetheless, general consensus has been reached upon some specific challenges, such as the need for 5G wireless access networks to support massive access by MTDs, as a consequence of the proliferation of M2M services. In this paper, we describe the main challenges raised by the M2M vision, focusing in particular on the problems related to the support of massive MTD access in current cellular communication systems. Then we analyze the most common approaches proposed in the literature to enable the coexistence of conventional and M2M services in the current and next generation of cellular wireless systems. We finally conclude by pointing out the research challenges that require further investigation in order to provide full support to the M2M paradigm.

1. Introduction

As telecommunication technologies continue to evolve rapidly, fueling the growth of service coverage and capacity, new use cases and applications are being identified. Many of these new business areas (e.g., smart metering, in-car satellite navigation, e-health monitoring, smart cities) involve fully-automated communication between devices, without human intervention. This new form of communication is generally referred to as Machine-to-Machine (M2M) Communication, or Machine-Type Communication (MTC), while the involved devices are called Machine-Type Devices (MTD). Examples of common MTDs are environmental and biomedical sensors, actuators, meters, radio frequency tags

*Corresponding author.
E-mail address: zanella@dei.unipd.it (A. Zanella).
Peer review under responsibility of Chongqing University of Posts and Telecommunications.

http://dx.doi.org/10.1016/j.dcan.2015.02.001

(RFtags), but also smartphones and tablets, vehicles, cameras, and so on.

The number of MTDs is continuously growing, together with the set of M2M applications and services that they enable. As a matter of fact, MTDs are key elements in the emerging "Internet of Things" and "Smart City" paradigms [1,2], which are expected to provide solutions to current and future social-economical demands for sensing and monitoring services, as well as for new applications and business models in areas such as building and industrial automation, remote and mobile healthcare, elderly assistance, intelligent energy management and smart grids, automotive, smart agriculture, traffic management, and many others [3].

In the last years, the potential of the M2M paradigm has been recognized by both academia and industry, which have generated a number of studies, protocols and products oriented to the support of M2M services. Worth mentioning for their popularity and market penetration are the IEEE 802.15.4 standard for low-bitrate short-range transmissions [4], the 6LowPAN protocol suite for low power devices [5], the ZigBee solution for MTD interconnection in small wireless sensor networks [6], and other communication systems like EIB/KNX, LON and BACnet for home automation [7]. Another interesting paradigm that can be adapted to M2M services is the 3GPP Proximity-based Service (ProSe) protocol [8]. The basic idea is to offload the network by exploiting the physical proximity of the terminals, e.g., enabling direct communication between user devices, or limiting the signaling to the local area, without involving core network elements.

Despite the appeal of such solutions, the potential of the M2M vision can be fully unleashed only when MTDs connectivity will be possible everywhere, with no (or minimal) configuration, and without the need for deploying additional devices, such as gateways or concentrators. As a matter of fact, the ideal scenario is such that MTDs need just to be *placed* in the desired locations to get connected to the rest of the world. Taking inspiration from the well-known plug-&-play notion, we refer to this new connectivity paradigm as *place-&-play*.

Unfortunately, the technologies that can support some form of MTC are currently incapable of fulfilling the demand for ubiquitous access of MTDs to the communication systems. Local network solutions, such as ZigBee/6LoWPAN or IEEE 802.11ah extension for MTC, are suitable to interconnect MTDs in the same local area, but are able neither to offer coverage everywhere, nor to guarantee highly reliable coordinated control of the network. On the other hand, the ubiquitous coverage offered by satellite connections has prohibitively high cost, both economically and in terms of energy consumption, and pose significant challenges when used in indoor environments.

The place-&-play concept, hence, calls for terrestrial radio technologies that are capable of providing widespread (ideally ubiquitous) coverage, with extremely low energy consumption, low complexity at the end device, possibly low latency, and minimal cost per bit. A few proprietary solutions that satisfy some of these requirements have been recently commercialized, though the deployment of a new infrastructure network at a global scale is economically challenging. Therefore, the most natural and appealing solution is to add MTC to the services provided by the existing cellular networks. Indeed, cellular networks have a world-wide established footprint and are able to deal with the challenge of ubiquitous and transparent coverage. Furthermore, the wide-area mobile network access paradigm offers a number of other advantages over local-area distributed approaches, such as higher efficiency, robustness and security, thanks to locally coordinated control, coordinated infrastructure deployment, ease of planning, performance predictability and the possibility of deploying advanced MTC-tailored PHY/MAC schemes that shift complexity from MTDs to base stations (BSs).

The MTC paradigm is hence expected to play a significant role in current and future cellular networks, both as the enabler of potentially disruptive markets (e.g., Smart Cities [9]) and for the new challenges that M2M services will pose to the communication systems. Unfortunately, current cellular network technologies will likely be unable to cope with the expected growth of M2M services. Indeed, today's standards are designed to provide access to a relatively small number of devices that need to transfer a significant amount of data, so that the signaling traffic generated by the management and control plane is basically negligible. M2M services, instead, are generally expected to involve a huge number of devices that generate sporadic transmissions of short packets. The risk is then a collapse of current wireless cellular systems under the weight of the signaling traffic generated by MTDs [10]. In addition, although transmissions from MTDs are, in many cases, delay tolerant (smart metering, telemetry), there is also an important class of M2M applications that require ultra-low latency (e-health, vehicular communications). Furthermore, most MTDs are expected to be severely constrained in terms of computational and storage capabilities, and energy capacity.

For all these reasons, the M2M scenario is considered as a major challenge for next generation wireless cellular systems, commonly referred to as 5G [11]. In addition to increased bit rate and energy efficiency of the terminals and of the whole system, 5G will hence be required to provide minimal latency to critical services and seamless integration of Internet of Things (IoT) nodes, and to support massive M2M communication services, all without degrading the quality of services like voice, audio/video streaming, and web browsing, which are referred to as "conventional" services in the following.

In this paper we survey the main challenges offered to the current and next generations of wireless cellular standards by the expected massive diffusion of M2M services. To begin with, Section 2 discusses how M2M services are supported by current cellular standards and highlights the limits of such solutions. Section 3 addresses in greater detail the pivotal challenge of massive MTD access in LTE systems, describing the methods that have been proposed by 3GPP to counteract the service degradation in case of overload of the Random Access Channel (RACH) due to massive MTD access requests. Section 4, instead, discusses the schemes that have been studied to improve the energy efficiency and the quality of service (QoS) of MTC. After that, in Section 5 the issue of cell coverage extension is addressed. We then consider in Section 6 the studies that address the massive access problem with a more

fundamental approach. Section 7 wraps up the paper by discussing the current state-of-the-art in massive M2M cellular access and the gaps that need to be filled in the near future in order to fully support the M2M paradigm, with specific reference to the advanced features that are envisioned in 5G systems. Finally, Section 8 concludes the paper with some final considerations.

2. M2M support in current cellular networks

Considering the relatively low Average Revenue Per User that is expected from MTD-based services, the costs to provide ubiquitous coverage at global level will become sustainable only if the volume of the M2M market will enable economies of scale, which requires the utilization of the same resources for multiple services and, hence, the interoperability of the different enabling technologies. This necessity has been recognized by the "oneM2M" Partnership Project (PP)[1] that was created in 2012 with the goal of developing global, access-technology agnostic Service Layer specifications for M2M. While the oneM2M initiative can effectively contribute to enable the aforementioned place-&-play paradigm for what concerns the "play" aspect, i.e., the standard and zero-configuration access to M2M data and the interoperability of M2M services, the "place" aspect, which requires ubiquitous and seamless connectivity of MTDs, still lacks a satisfactory solution. In this respect, the current cellular systems are natural and appealing candidates for the support of widespread MTD wireless access, because of their capillary geographical coverage, their technological maturity, and the cost-effectiveness provided by conventional higher-revenue services, such as voice, video and wideband data connectivity. Unfortunately, current systems may not be able to support the expected growth of MTC, so that proper countermeasures need to be taken, as better explained in the remainder of this section.

2.1. Revamping GSM for M2M support

The second generation of standards for cellular systems, i.e., GSM, GPRS, and EDGE, is progressively being replaced by the third (UMTS/HDSPA) and the fourth (LTE) generations for what concerns Human-to-Human (H2H) services, which require higher data-rates, lower power consumption, and better support to user mobility. In this framework, GSM becomes an attractive candidate to provide ubiquitous connectivity to MTDs, which may exploit the pervasive presence of GSM coverage and the empty space left by the migration of the conventional services towards 3G and 4G networks. However, considering the scarcity of available radio frequencies and the always growing demand for new wireless services, the refarming of GSM frequency bands is being debated by governments, so that the remaining operational time for GSM is uncertain.

In addition, several studies show that the GSM radio access network faces serious capacity issues in the presence of the synchronized access of a massive number of MTDs [12–14]. As a consequence, there has been an effort in 3GPP to further enhance the GSM related standards to facilitate the support of MTC. The attention has been mostly focused on improvements in the core network and at the higher layers of the radio access network, with reference to the following aspects [15,14]:

- identification of MTDs within the network;
- Short-Messaging-Service in the Packet Switched domain;
- load control through the introduction of extended class barring;
- reduction of the signaling overhead by minimizing the occurrence of signaling procedures not essential to the MTD profile.

While these efforts can indeed make GSM a valuable solution for MTC support in the mid-term, there are a number of practical considerations and technical limits that will likely prevent GSM from becoming the ultimate access technology for MTDs in the long run. To begin with, the number of MTDs that can be connected to a single GSM base station (BS) is quite limited. This limit can be alleviated by tightening the granularity of the transmission resources, while keeping the backward compatibility with the original system, as proposed in [14]. In this way, the MTDs that can be served by a single BS can range between 10^4 and 10^5, depending on the requirements in terms of delay tolerance. However, the energy consumption and the delay of GSM access may still be prohibitively high for most M2M services. In addition, operators are reluctant to guarantee that GSM networks will be operational for the quite long lifetime of M2M applications. Furthermore, the always growing demand for mobile wideband services (e.g., video streaming) may push the operators to re-farm the spectrum currently allocated to GSM to further increase the capacity of LTE and/or 5G RANs.

2.2. M2M services & LTE

Acknowledging the shortcomings of GSM as a long-term access network for M2M services, some operators are slowly pushing M2M applications towards UMTS. Compared to GSM, however, UMTS has a number of disadvantages: because it is typically used in a higher frequency band, it is more difficult to get good (esp. indoor) coverage; furthermore, UMTS modules are more expensive than GSM modules.

A rather natural option is to resort to the latest cellular standard, i.e., LTE. However, LTE design was mainly intended to increase the data rate offered to mobile users, without considering the requirements and traffic characteristics of typical M2M services. Supporting MTC in the LTE architecture, hence, raises a number of challenges, including control overhead, energy efficiency, coverage extension, robustness to malfunctioning devices, security, and scalability. However, the most compelling problem regards the coexistence of M2M and conventional services and, in particular, the support of massive MTD access without hampering the quality of conventional services [16].

The problem originates from the Random Access Channel (RACH), i.e., the transport-layer channel defined by LTE

[1] http://www.onem2m.org.

standard to manage the channel access requests by end-devices, which is better explained below [17].

2.3. RACH: the random access procedure of LTE

Hereafter, we will use User Equipment (UE) and eNodeB to refer to end-device and BS, as per LTE terminology. According to LTE specifications, the random access procedure is triggered by a UE any time it needs to establish (or re-establish) a data connection with the eNodeB, e.g., for the association to the network or the synchronization with the eNodeB after a long idle period; after radio-link failure; or when changing the serving eNodeB because of a handover.

Depending on the purpose, the random access procedure can either be contention-free or contention-based. The first procedure is not much impacted by M2M traffic, being used for managing delayed-constrained access requests with high success requirements, such as those related to handover. The process is indeed under full control of the eNodeB that coordinates the access requests of the end-devices in order to avoid conflicts and to minimize the access delay. Conversely, the contention-based random access procedure is much more sensitive to M2M traffic. To better understand the origin of the problem, it is convenient to describe in more detail the random access procedure.

The RACH is formed by a sequence of time–frequency resources, called Random Access (RA) slots. UEs are allowed to transmit their access requests only in RA slots by using specific orthogonal preambles, which are called "signatures." The time–frequency resource on which the RA preamble is transmitted is known as the Physical Random Access Channel (PRACH), which is time and frequency multiplexed with the Physical Uplink Shared Channel (PUSCH), as illustrated in Fig. 2.

In each LTE cell, there are 64 preambles, created by the so-called Zadoff–Chu sequence. Some preambles are reserved for contention-free (or coordinated) RA, while the remaining ones are used for contention-based (or uncoordinated) RA. This second set of preambles is further divided into two groups: Group A preambles are intended for sending small packets and Group B preambles are intended for sending large packets.

In the frequency domain, the PRACH resource has a bandwidth corresponding to 6 resource blocks (1.08 MHz). Instead, the periodicity of RA slots, the total number of random access preambles available for contention-based random access, the total number of random access preamble sequences available within Group A, the maximum message size allowed for such preambles, and other parameters related to the RACH are broadcast to the UE in the system information block 2 (SIB2).

Fig. 1 illustrates the logical sequence associated to a contention-based RA procedure, which develops along the following steps .

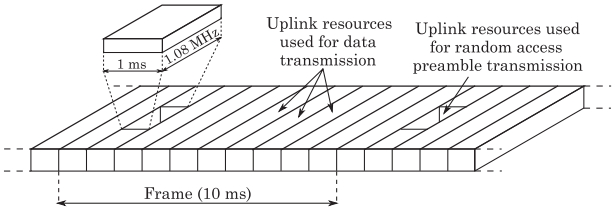


Fig. 2 Illustration of PRACH transmission resources in LTE.

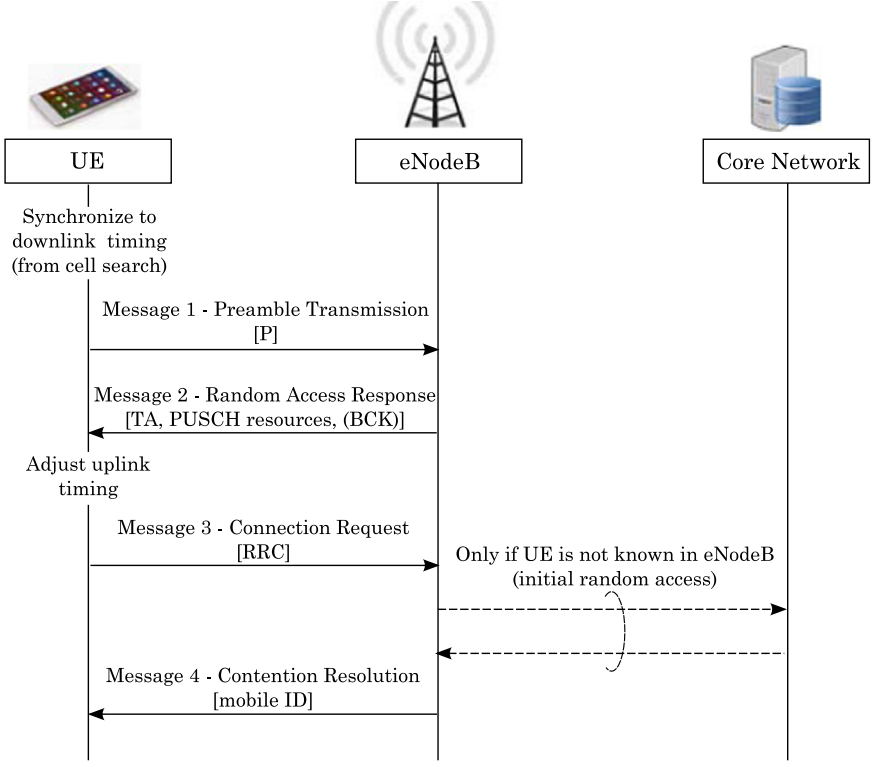


Fig. 1 Contention-based Random Access procedure in LTE.

- *Message* 1 [*Preamble Transmission*]: The UE randomly chooses one of the preambles (P) reserved for contention-based RA and transmits its request in the first available RA slot. Because of the orthogonality of the different preambles, multiple UEs can transmit their access requests in the same RA slot, using different preambles. In this case, the eNodeB is able to decode the requests and estimate the transmission timing of the terminals. Conversely, if two or more devices transmit the same preamble, a collision occurs and the corresponding requests will not be detected by the eNodeB (see Fig. 3). However, it is also possible that multiple UEs choose the same preamble and the eNodeB correctly detects it (e.g., because one is much stronger than the others or the different signals appear as a single transmissions going through multiple fading paths). In this case, the acknowledgement sent by the eNodeB will trigger a transmission by multiple UEs and a collision will occur at the third step of the handshake.
- *Message* 2 [*Random Access Response*]: For each successfully decoded request, the eNodeB transmits a RA response on the Physical Downlink Shared Channel (PDSCH), which includes a Timing Alignment (TA) command to adjust the terminal transmit timing, and the resources on the uplink channel (PUSCH resources) that have been assigned to UEs for the third step of the RA procedure. This message, furthermore, carries an optional backoff indicator (BCK) that is used to reduce the probability of further collisions in successive attempts by UEs that collided in the previous RA slot. Indeed, UEs whose Message 1 is not acknowledged by Message 2 within a specific time window wait for a random backoff interval before starting again the RACH procedure in the next available RA slot. Moreover, if the counter of consecutive unsuccessful preamble transmission attempts exceeds a certain threshold, a Random Access problem message is indicated to upper layers.
- *Message* 3 [*Connection Request*]: Upon receiving the random access response, UEs will transmit a Radio Resource Control (RRC) message on the reserved resources in the PUSCH. Note that UEs involved in an undetected preamble collision will transmit over the same PUSCH resource blocks, thus generating a new collision (see Fig. 4).
- *Message* 4 [*Contention Resolution*]: If the connection request of the UE is successfully detected by eNodeB, it replies with a contention-resolution message on the PDSCH, echoing the mobile terminal identity (mobile ID) in order to acknowledge the correct reception of its request. Conversely, UEs that do not receive a contention resolution message from the eNodeB assume that their access requests failed and, after waiting a random backoff time, perform a new preamble transmission attempt in the next RA slot. Again, when the number of unsuccessful attempts reaches a certain value, the network is declared unavailable by the UE and an access problem exception is raised to the upper layers.

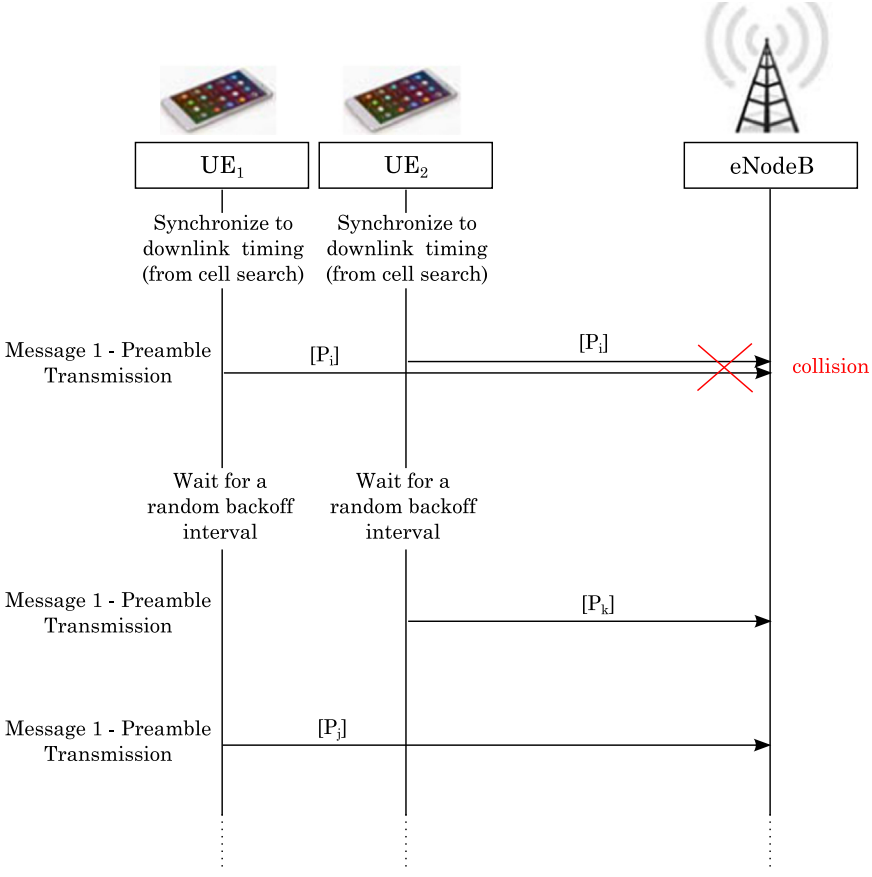


Fig. 3 Collision event in *Message* 1.

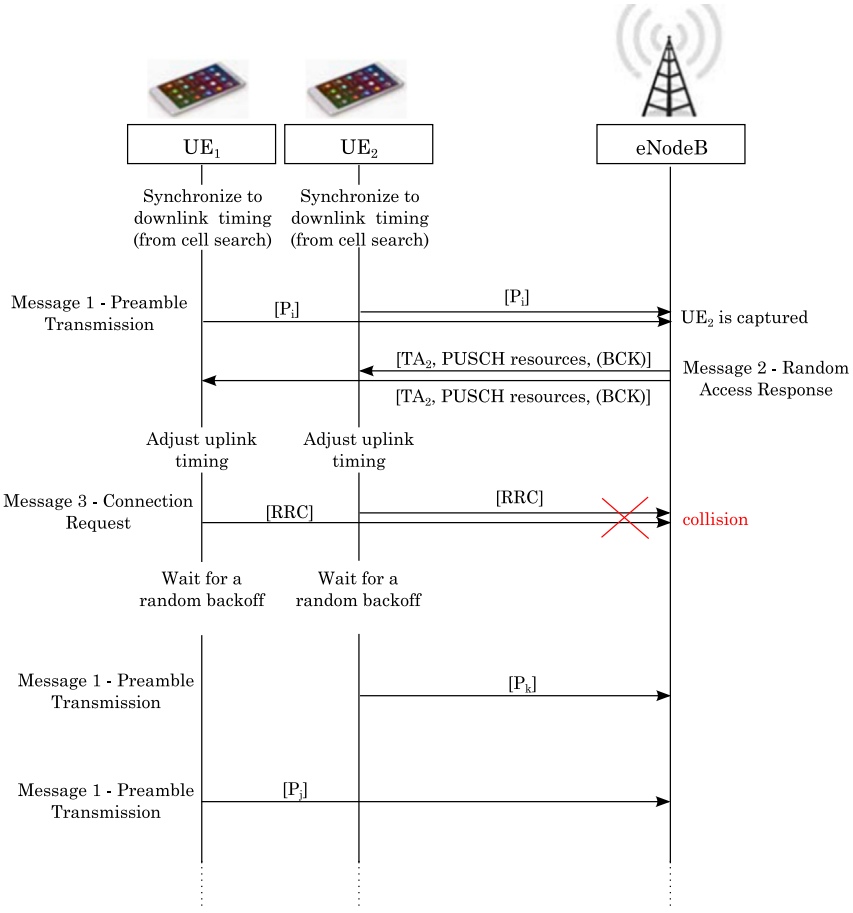


Fig. 4 Collision event in *Message* 3.

2.4. The threat of PRACH overload

The RACH procedure has been identified by 3GPP as a key challenging task [16] for M2M communications because of the signaling and traffic load spikes caused by a sudden surge of the number of M2M devices trying to access the same base station simultaneously (e.g., a huge number of smart meters becoming active almost at the same time after a period of power outage).

The risk, hence, is that massive access requests by MTDs can overload the PRACH, yielding an increase of the contention probability and, in turn, of the access delay and failure rate. A possibility to reduce the load of the PRACH is to increase the number of access opportunities scheduled per frame, but this determines a reduction of the amount of resources available for data transmission and, hence, a contraction of the data transport capacity of the uplink channel. Furthermore, the total amount of RA slots that can be allocated in an LTE frame is limited. Finally, the processing of the Zadoff–Chu sequence, which is employed in the LTE RACH preambles, is computationally demanding and can be a further issue for resource-constrained devices, such as MTDs.

Summing up, the standard LTE procedure for managing channel access requests by end-devices will not properly scale in the presence of massive access attempts by a large number of MTDs, which may result in a sharp degradation of the quality offered to conventional services because of long access delay and high access failure rate. Of course, M2M services are also affected by these impairments, though the impact may be less significant with respect to conventional services. Nonetheless, the overload of the PRACH impacts MTDs on other aspects, such as the energy consumption and computational effort, which are generally critical for MTD applications.

3. Standard schemes to alleviate the PRACH overload problem

Coexistence and management of human-triggered and machine-triggered traffic is a major challenge for next generation cellular networks.[2] In particular, the contention to access PRACH resources may lead to dramatic degradation of H2H services. For this reason, the PRACH overload problem has been attracting the attention of the scientific community, and many possible methods to improve the RACH procedure have been proposed [10]. Most of these

[2]Since the access requirements are mainly determined by the class of the service initiator, with a slight abuse of terminology in the following we use H2H and M2M to refer to human-triggered and machine-triggered services, respectively, whatever the actual nature of the destination.

methods provide some form of separation between access requests originated by H2H and M2M services, with the aim of shielding the former from the PRACH overload issues that can be generated by the latter. The various approaches differ in the way this separation is enforced. We hence distinguish between "strict" schemes, in which the pool of access resources is deterministically split between H2H and M2M, thus achieving perfect isolation between the two types of access requests; and "soft" schemes, where H2H and M2M share the same resources, but with different access probabilities. These two approaches can also be combined, giving rise to "hybrid" schemes.

The remainder of this section briefly discusses the different approaches that have been proposed to counteract PRACH overload, most of which appeared in 3GPP technical documents such as [18]. The interested reader is also referred to [10] for a different classification of (most of) these schemes.

3.1. Strict-separation schemes

As mentioned, strict-separation schemes achieve perfect isolation between H2H and M2M access requests by allocating different physical resources to UEs and MTDs. In this category, we can list the following schemes.

(1) *Resource separation*: The simplest and most immediate way to shield H2H from the risk of access request collisions due to massive MTC requests is to assign orthogonal PRACH resources to H2H and M2M devices. The separation of resources can be done by either splitting the preambles into H2H and MTC groups, or by allocating different RA slots in time and/or frequency to the two categories of terminals [18]. This solution, however, can yield suboptimal performance when the number of resources assigned to each class of devices does not reflect the actual demand.

To be effective, this scheme needs to be coupled with mechanisms to dynamically shift resources among the two classes, according to the respective access request rates. In some scenarios, the network can predict sharp increments of the access load due to MTDs, e.g., using the Self-Optimizing Overload Control (SOOC) scheme proposed in [16] and described later under the "hybrid" category.

(2) *Slotted access*: This scheme was proposed by 3GPP in [18]. It consists in defining access cycles (similar to paging cycles), which contain RA slots dedicated to MTD access requests. Each MTD can only access its dedicated RA slots, in a collision-free manner. The RA slots reserved to each MTD in a cycle are determined from the unique identifier of the devices (namely, the International Mobile Subscriber Identity – IMSI) and the RA cycle parameter broadcast by the eNodeB. While this scheme protects H2H devices from MTC, the allocation of dedicated RA slots to each MTD may yield very long RA cycles and, hence, long access latency, which may not be compatible with the service requirements of delay-constrained M2M applications (e.g., alarms).

(3) *Pull-based scheme*: This is a centralized mechanism that allows MTDs to access the PRACH only upon being paged by the eNodeB [18]. Paging is triggered by the MTC server that is assumed to know in advance when MTDs need to establish a radio link connection, to either send or receive data. The eNodeB can control the paging taking into account the network load condition, thus preventing PRACH overload. This is already supported by the current specification. The paging message may also include a back-off time for the MTDs, which indicates the time of access from the reception of the paging message. This approach is suitable to manage channel access of MTDs with regular traffic patterns. However, its centralized nature limits the number of MTDs and M2M services that can be managed by a single M2M server. Furthermore, the scheme cannot deal with an unexpected surge of MTD access requests.

3.2. Soft-separation schemes

In soft-separation schemes there is no neat separation of access resources between M2M and H2H, rather all devices can use the same resources but with different probabilities. Therefore, the separation between MTDs and classic UEs is achieved in a statistical sense. The main schemes based on this approach are described below.

(1) *Backoff tuning*: A way to smoothly decrease the rate of channel access requests by MTDs in case of congestion is to assign longer backoff intervals to MTDs that fail the transmission of *Message* 1 in the RACH procedure [18]. Although this method can alleviate the contention between H2H and M2M devices in case of peaks of MTD requests, it is not very effective when dealing with stationary MTDs massive access, due to the instability issue that characterizes ALOHA-like access mechanisms.

(2) *Access Class Barring*: The backoff tuning scheme is generalized by the Access Class Barring (ACB) method, which is actually part of LTE and LTE-A specifications. ACB makes it possible to define multiple access classes with different access probabilities [18]. Each class is assigned an access probability factor and a barring timer. The devices belonging to a certain access class are allowed to transmit *Message* 1 in a RA slot only if they draw a random number that is lower than the access probability factor. Otherwise, the access is barred and the devices have to wait for a random backoff time, which is determined according to the barring timer of that class, before attempting a new access. The ACB parameters are broadcast by the eNodeB as part of the system information. Furthermore, 3GPP proposed the Extended Access Barring (EAB) scheme, which is a method for the network to selectively control access attempts from UEs that can tolerate longer access delays or higher failure probability [18]. These devices will hence be barred in case of overload of the access and/or the core network, without the need to introduce any new Access Classes. These mechanisms can be used to alleviate the MTD massive access issue by defining a special class for MTDs, with lower access probability factor and/or longer barring timer, or labelling MTDs as EAB devices. However, MTDs with delay-constrained access requirements can be associated to classes with higher access probability and lower barring timer.

ACB mechanisms are quite effective in preventing PRACH overload, but at the cost of longer access delay for MTDs. Furthermore, ACB does not solve the access contention problem when many delay-constrained MTDs need to access the channel in a short time interval, as the result of certain

events (e.g., alarms triggered by unexpected events, such as failures of the power grid, earthquakes, and flooding). Nonetheless, ACB mechanisms can be combined with other techniques to counteract the PRACH overload due to massive MTD access.

3.3. Hybrid schemes and other solutions

Here we discuss the solutions that cannot be classified as either strict- or soft-separation schemes, since they include aspects from both families or are based on totally different approaches.

(1) *Self-Optimizing Overload Control*: SOOC is a composite scheme presented in [16] to counteract PRACH overload by combining many of the schemes described above, including PRACH resource separation, ACB, and slotted-access schemes. The fundamental feature of the SOOC scheme is the execution of a control loop to collect information for overload monitoring at each RA cycle. Then, based on such data, the eNodeB adapts the number of RA slots in the random access cycles.

More specifically, when a device is not able to get an access grant at the first attempt, it enters the overloaded control mode. In this status, the classical p-persistent mechanism is applied in order to regulate RA retries for collided terminals. Besides, in order to distinguish between time-tolerant MTDs and time-sensitive MTDs, two access classes are added to the LTE-A ACB scheme for M2M devices (namely, low access priority and high access priority) and different p parameters are set according to the access class of the terminal.

In order to monitor the congestion level of the system, when a terminal receives *Message* 2 in the contention-based RA procedure (see Section 2.3), it includes a PRACH overload indicator, which contains the number of RA retries attempted by the device, within *Message* 3. Based on this information, the eNodeB reacts by dynamically increasing or decreasing the number of PRACH RA slots in the successive cycle in order to maintain a target maximum collision probability for the system. Moreover, in the borderline case when the number of RA slots can not be further increased due to insufficient uplink radio resources, the eNodeB can deny access to low priority MTDs until the overload condition improves.

Unfortunately, although the goal of handling high traffic loads is clear and the proposed scheme surely goes in this direction, [16] only describes SOOC theoretically and no performance results have been presented.

(2) *Random access scheme for fixed-location M2M communication*: When MTDs are static, the fixed uplink Timing Alignment (TA) between the MTDs and the BS can be exploited in the resource allocation procedure, as proposed in [19], where the focus is on a large class of motionless MTDs (e.g., smart meters). In this context, authors propose an energy-efficient RA scheme to reduce collision probability and average access delay. The procedure consists of 5 steps:

1. The device randomly chooses a preamble out of M orthogonal ones and transmits it in a certain slot (*Message* 1).
2. The BS detects which preamble is transmitted, determines the proper TA value, and broadcasts a resource allocation response that contains several parameters, including TA information (*Message* 2).
3. The device compares the TA value carried by the response with its own TA value: in case of matching the handshake goes on with step 4, otherwise, the MTD performs a retransmission after a random backoff time.
4. The MTD synchronizes the uplink transmission time to the received TA information and sends a Radio Resource Control message on the allocated uplink resource (*Message* 3).
5. If the message is successfully decoded, the BS sends an acknowledgement message (*Message* 4). Otherwise, the MTD needs to repeat the access request procedure.

Note that the procedure resembles the conventional LTE resource allocation process described in Section 2.3 except for step 3, which takes advantages of the TA information to reduce the collision probability in the transmission of *Message* 3. Indeed, the TA value of a fixed-location MTD is expected to remain constant over time. If the TA received from the eNodeB in *Message* 2 does not match with the expected TA of the MTD, then there is a high probability that *Message* 2 is actually intended for a different MTD that transmitted on the same PRACH. In this case, according to step 3, the MTD will avoid transmission of *Message* 3, thereby reducing the probability of collision at step 4 and, in turn, the access delay.

(3) *Bulk MTC signaling scheme*: Another possible solution to congestion/overload events may be to enable bulk MTC signaling handling, as stated in [20], where the authors remark the lack of mechanisms to simultaneously handle the overhead generated from a group of MTDs. Therefore, under the assumption that signaling messages from MTDs are moderately delay tolerant, it may be convenient to minimize the overhead at the BS by exploiting bulk processing, i.e., aggregating signaling data coming from MTDs before forwarding them to the core network. For example, consider the case in which a group of MTDs are triggered to send a Tracking Area Update (TAU) message: the BS could wait a default timeout interval or until it gathers a sufficient number of signaling messages to forward a single aggregate message towards the Mobility Management Entity (MME). Indeed, since the MTDs are associated to the same MME, the TAU messages will differ only on the MME Temporary Mobile Subscriber Identity (M-TMSI). Considering an average of 30 TAU messages per second, and a message aggregation period of 10 s, 300 TAU messages can be aggregated in a single 1211 bytes message. Individual messages would instead require 4500 bytes. This approach can alleviate the traffic produced by massive channel access towards the MME, but it does not address the issue of batch MTD transmissions on the access network side.

(4) *Q-learning solution*: The standard RACH access is basically derived from the classical slotted ALOHA protocol, of which it inherits simplicity and limitations. In particular, the system may drift to an unstable region in the presence of massive M2M traffic. In this context, [21] suggests a solution based on Q-learning to enhance the throughput of RACH and shield H2H traffic from the performance loss that

can be caused by massive M2M access requests. According to the authors, users should be divided into two groups: a learning group containing all MTDs, and a non-learning group composed of H2H devices. M2M communication uses a virtual frame of RA slots called *M2M-frame*, whose length (in time slots) should be equal to the number of MTDs in the network. Every node keeps a Q-value for each slot in the M2M-frame, which records the transmission history on that slot in consecutive frames. Such value is updated after every transmission attempt as follows:

$$Q \leftarrow (1-\alpha)Q + \alpha r$$

where α is the learning rate, and r is the reward, which equals 1 if the transmission is successful or $-k$ otherwise, where k is known as penalty factor and is introduced to mitigate the effect of collisions with H2H devices [21].

Each MTD will transmit in the slot with the highest Q-value. The performance evaluation shows that, in case of high load from H2H traffic, Q-learning access stabilizes the total RACH throughput at 35% (approximately the maximum efficiency of Slotted ALOHA), as the M2M traffic increases. When the H2H traffic load is low, instead, the proposed solution provides a significant improvement, raising the total RACH normalized throughput to 55%. Moreover, delay is reduced and the learning convergence is quickly achieved.

(5) *Game theory scheme*: In [22], the problem of H2H and M2M coexistence is formulated by adopting a game theoretic approach. As mentioned, the standard LTE RACH procedure is characterized by a unique pool of preambles, from which each device randomly picks a preamble to be used for its channel access request to the BS. In the proposed solution, instead, different preamble pools for M2M and H2H usage are reserved: in particular there are R_H preambles for H2H, R_M for M2M, and R_B available for both. Then, MTDs are allowed to extract the preamble either in the M2M-dedicated pool (action $a_i = M$), in the shared one ($a_i = B$), or to stay silent ($a_i = S$) with a probability distribution that is determined according to the outcome of a game. The game formulation consists of a constant number H_B of H2H users that transmit preambles taken from the shared pool, and N MTDs, which are the players of the game. They play a mixed strategy $\sigma_i(a_i)$, choosing actions M, B or S with probability $p_{i,M}$, $p_{i,B}$ or $1-p_{i,M}-p_{i,B}$, respectively. All MTDs have a cost $\lambda \in [0,1]$ for preamble transmission (e.g., in terms of energy consumption), yielding the following gains:

$$g_i = \begin{cases} 1-\lambda & \text{if transmission is successful;} \\ -\lambda & \text{if transmission fails;} \\ 0 & \text{if transmission is not attempted.} \end{cases}$$

Denoting by $P_S(a_i)$ the RACH success probability if action a_i is taken by player i, the expected payoff is given by

$$\mathbb{E}[g_i] = \sum_{a_i \in \{M,B\}} \sigma_i(a_i) \cdot (P_S(a_i) - \lambda). \quad (1)$$

Simulations show that, following a mixed strategy Nash Equilibrium, every MTD has non-negative utility, and the throughput of both M2M and H2H users is improved with respect to the baseline scheme in the case of overloaded RACH. Moreover, the authors provide a procedure to estimate the actual number of H2H and M2M devices in real systems, which in practice may have imperfect knowledge of the exact values of H_B and N. The proposed approach is proved to converge quickly and to provide small estimation errors for N.

4. Tackling energy efficiency and QoS challenges

As already pointed out in Section 1, PRACH overload is just one of the issues that are raised by M2M massive access. Other relevant challenges regard energy efficiency, heterogeneous quality of service (QoS) support, coverage extension, robustness to malfunctioning devices, security, and scalability [23-26].

In this section, we will discuss some proposed approaches that explicitly address the aspects related to energy efficiency and QoS support for M2M services. We classify the different techniques according to the methodological nature of the proposed solutions, rather than their specific objectives, which can involve one or more performance indices, as explicitly indicated in Tab. 1, and discussed in Section 7.

Based on the adopted methodology, the different solutions are hence divided into Clustering techniques, Game Theoretic Approaches, and Machine Learning algorithm. The remainder of this section describes in greater detail such solutions.

4.1. Clustering techniques

One possible way to handle massive access to the BS is to appoint a few nodes, called *coordinators* or *cluster-heads*, as relays for the remaining terminals. In this way, the number of access requests to the BS is limited to the number of coordinators. Furthermore, a proper selection of the coordinators can also contribute to decrease the energy consumption of the system by exploiting multi-hop transmissions over high-gain links in place of direct transmissions over poor quality links. The problem now becomes the design of suitable policies for electing the relays and assigning the terminals to the different clusters. In the following we describe some solutions that have been recently proposed.

(1) *Energy efficient clustering of MTDs*: The massive access management and energy efficiency aspects are jointly addressed in [27], where authors proposed a clustering approach to limit the number of simultaneous accesses to the BS and the energy consumption of MTDs.

Specifically, the authors consider a scenario with N MTDs, which are randomly deployed in a single cell centered at the BS and have each one packet to transmit. The BS is assumed to know the channel conditions to each terminal. The idea is to group the MTDs in G clusters and, for each group, select a coordinator that is the only device allowed to communicate with the BS, relaying the communications of the other terminals in its cluster (see Fig. 5).

The total energy consumption of the system can hence be expressed as

$$\mathrm{EC} = \sum_{i=1}^{G} \sum_{j \in G_i \setminus \{c_i\}} \left(\frac{P_t L_s}{R(d_j, c_i)} + \frac{P_t L_s}{R(c_i, BS)} \right) \quad (2)$$

Tab. 1 Comparison of the proposed approaches.

Solution	Main challenge	3GPP	H2H & M2M	Performance indices		
				Delay	*Energy efficiency*	*Access probability*
Resource separation [18]	PRACH overload	✓	✓			
Slotted access [18]	PRACH overload	✓				✓
Pull-based scheme [18]	PRACH overload	✓		✓		✓
Backoff tuning[18]	PRACH overload	✓	✓			✓
Access Class Barring [18]	PRACH overload	✓	✓[a]	✓		
SOOC [16]	PRACH overload			✓		✓
RA for fixed-location [19]	PRACH overload			✓		✓
Bulk signalling [20]	PRACH overload				✓	✓
Q-learning [21]	PRACH overload		✓	✓		✓
Game theory scheme [22]	PRACH overload		✓		✓	✓
Energy-efficient clustering [27,28]	Energy consumption				✓	
QoS-based clustering [29-31]	QoS support for M2M			✓		
M2M-aware scheduling [32]	QoS support for H2H & M2M		✓	✓		✓
Matching theory scheme [33]	QoS support for H2H & M2M		✓			✓
Reinforcement learning [34]	BS selection			✓		✓
Physical layer design [35-38]	Coverage extension	✓			✓	✓
Cooperative coverage extension [42]	Coverage extension				✓	
Clean slate approaches [47,48,51]	Massive access					✓

Note: *Delay* is intended as the time from the first transmission attempt until the successful conclusion of the access procedure.
[a]The scheme can support H2H and M2M separation, though it is not specifically designed for this purpose.

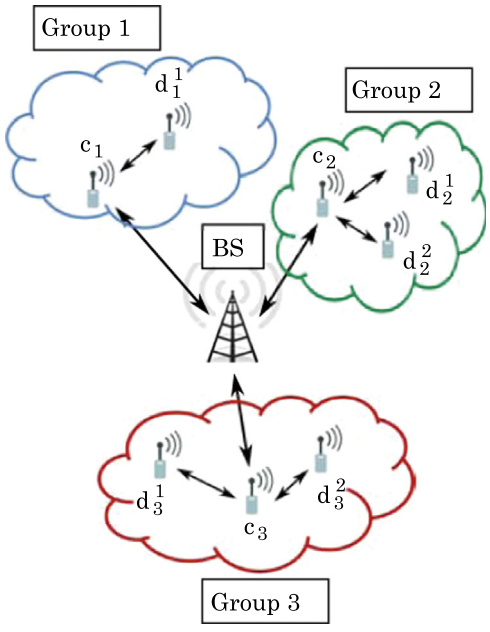


Fig. 5 Proposed grouping model for M2M system.

where G_i is the set of nodes in the ith cluster,[3] c_i and d_j denote the concentrator of cluster i and the jth MTD, respectively, $R(x, y)$ is the transmit bit rate from node x to node y with transmit power P_t, while L_s is the length of the packet to be transmitted.

Now, the objective is to minimize (2) while keeping the number of groups G below a certain threshold M, hence limiting the maximum number of access requests to the BS and reducing the redundant signaling of M2M services. To this end, the authors of [27] study a number of algorithms that combine different grouping and coordinator selection techniques, which are briefly described below.

Grouping: The clusterization of MTDs is obtained by the K-means algorithm: initially, G nodes are randomly selected as coordinators, and the other MTDs join the cluster based on their channel conditions to the respective coordinator. Then, a new coordinator is selected for each group, according to certain coordinator selection policies which will be presented in the next paragraph. Then, the remaining MTDs are clustered again with respect to the new

[3]The notation $G_i \backslash \{c_i\}$ indicates the set G_i without the element c_i.

coordinators. The whole procedure is iterated until the coordinators do not change between consecutive cycles.

Coordinator selection: The core of the algorithm is the policy adopted to elect the coordinator of each group. The authors of [27] consider various schemes for coordinator selection, based on different criteria to minimize the energy consumption of MTDs. Some policies are independent of the quality of the link between coordinator and BS, while others keep this parameter into account when selecting the coordinator of each group. The policies that emerge as more interesting from the simulation study are the *K Maximal Channel Gain* (K-Max-CG), which simply selects as cluster-heads the MTDs with highest channel gain towards the BS, and the *Optimum Energy Consumption* (Opt-EC), which implements an exhaustive search for the cluster-head that minimizes the energy consumption within each of the groups returned by the K-means algorithm.

Actually, the simulation results presented in the paper show that clustering techniques are effective in reducing the massive access issue and improving the energy efficiency of the MTDs. Indeed, almost all proposed schemes perform better than direct transmission between MTDs and BS in terms of both energy consumption and channel contention. More specifically, in a scenario with randomly distributed MTDs around the BS, the energy consumption tends to decrease with the number G of groups, until it becomes almost constant for $G \geq 10$. In this case, the best performance is attained by K-Max-CG. On the other hand, when the MTDs distribution over the cell is not uniform (e.g., nodes are concentrated in two or three smaller areas around the BS), the energy consumption is minimized by a lower number of groups, and the optimum coordinator selection algorithm turns out to be Opt-EC, while the performance of K-Max-CG dramatically deteriorates.

Another clustering technique to maximize the energy efficiency of MTC has been proposed in [28] with reference to an OFDMA-based cellular network. Once again, the authors propose to appoint some nodes as coordinators of a certain group of MTDs and use two-hop communication to connect the peripheral nodes to the BS. Differently from [27], here authors consider some details of the transmission channel. More specifically, they assume that communications between MTDs and coordinators are managed by means of a Time Division Multiple Access (TDMA) scheme, while the coordinators communicate with the BS by using an Orthogonal Frequency Division Multiple Access (OFDMA) channel (which is the basis for the Single Carrier Frequency Division Multiple Access adopted in LTE uplink channels). Due to the short delay spread and the narrow bandwidth, the first link is subject to flat fading, whereas the second link can be affected by either flat or frequency selective fading, i.e., the channel gain is either the same for all subchannels or varies across different subcarriers.

Clustering and coordinator selection are based on an energy-consumption model that accounts for both the energy spent by each terminal to transmit data and some additional energy expenditure due to the circuitry, i.e.,

$$\mathrm{EC} = \sum_{i=1}^{G} \left(\frac{(P_{c_i} + P_{\mathrm{cir}}) D_{c_i}}{R(c_i, BS)} + \sum_{j \in G_i \setminus \{c_i\}} P_{d_j} \cdot \frac{L_s}{R(d_j, c_i)} \right) \tag{3}$$

where the notation is as in (2), except for P_x that denotes the transmit power of node x, P_{cir} is the fixed circuitry power consumption, and D_{c_i} is the aggregate data received by the coordinator c_i from all the MTDs in its cluster. Therefore, the optimization problem can be formulated as follows:

$$\min_{G, G_i, c_i, P_{c_i}} \quad \{\mathrm{EC}\}$$

$$\text{subject to} \quad G \leq N$$
$$P_{d_j} = P_t, \quad i \in \{1, \ldots, G\},\ j \in G_i \setminus \{c_i\},$$

where N is the total number of MTDs and P_t is a fixed power value. Although this formulation holds under the assumption that all links are subject to flat fading, a similar problem can be defined in the case of frequency selective fading. The problem, however, is very complex; therefore, the authors of [28] propose a suboptimal solution that consists in an iterative algorithm that first clusters the MTDs into groups, and then selects the coordinator for each group. The algorithm actually starts from a random selection of the coordinators. Then, each MTD joins the group whose coordinator has minimum energy cost for delivering the packet from the MTD to the eNodeB. Successively, a new coordinator is selected among all nodes in the same cluster in order to minimize the group average energy cost. The procedure is iterated until the composition of the groups and the set of coordinators remain unchanged. After that, the transmit power of cluster-heads is also optimized by using another iterative algorithm still with the aim of minimizing the overall energy consumption.

Notably, neither [27] nor [28] account for the energy spent in reception. Furthermore, as for all cluster-based scheme, coordinators are subject to higher power consumption and may fail because of energy depletion before the other nodes. Hence, mechanisms for the rotation of the cluster-head role shall be considered. On the other hand, these countermeasures would require higher costs in terms of signaling and control traffic, which shall also be accounted for.

(2) *QoS-based clustering*: In [29-31], clustering is used as an effective solution to manage radio resource assignment to a large population of MTDs with small data transmissions and very disparate QoS requirements. Specifically, MTDs are grouped into G clusters based on their packet arrival rate (ρ) and maximum tolerable jitter (δ), in such a way that devices in the same group have very similar traffic characteristics and QoS requirements. By this grouping operation, the BS can manage radio resources at the cluster level rather than per single MTD, in the following way.

With reference to an LTE-Advanced network, eNodeB allocates an Access Grant Time Interval (AGTI) to the ith cluster every $1/\rho_i$ ms. Clusters are served in order of priority, dictated by their traffic rate ρ, so that the cluster with highest traffic rate is served first and if AGTIs for different clusters are arranged in the same subframe, the AGTI for the cluster with lower priority (i.e., lower ρ) is postponed to the subsequent subframe. Each AGTI has a fixed duration τ for all clusters. Based on such radio resource management, the jitter of packets in the ith cluster is upper bounded by

$$\delta_i^* = \tau + \sum_{k=1}^{i-1} \left\lceil \frac{\rho_k}{\rho_i} \right\rceil, \quad i = 2, \ldots, M \tag{4}$$

and $\delta_i^* = \tau$ for $i=1$ (see [30] for details). Then, if $\delta_i^* \leq \delta_i$ for all clusters, packets in all clusters can meet the jitter

constraints. Hence, this bound can serve as a call admission control scheme to guarantee that QoS constraints of all admitted MTDs transmissions can be satisfied. As a proof of concepts, a simulation-based evaluation is developed in [29] considering the system parameters of LTE-Advanced and various QoS characteristics of the MTDs that cover a rather general set of M2M applications. The results show that the grouping scheme proposed, and based on the worst case analysis given by (4), can achieve the desired QoS guarantees.

Moreover, in [30], the authors distinguish between deterministic (hard) and statistical (soft) QoS guarantees. The flexibility of the QoS requirements for cluster i is captured by means of the "QoS outage" probability ϵ_i, i.e., the probability that the jitter exceeds the threshold δ. Clearly, hard QoS constraints correspond to $\epsilon_i = 0$. In case the number n_i of MTDs in cluster i is less than a certain threshold, groups of resource blocks, called Allocation Units (AU), originally dedicated to MTC can be opportunistically reallocated by the BS, with probability q, to serve non-MTD users. The results show that the (soft) QoS requirement of cluster i is satisfied if, besides the condition $\delta_i^* \leq \delta_i$, q is such that

$$q \leq \min\left\{ \left(\frac{\epsilon_i}{1 - \frac{C_{x-1}^{n_i-1}}{C_x^{n_i}}} \right)^{1/x}, 1 \right\} \tag{5}$$

where $0 \leq x \leq n_i$ is the number of AUs still allocated to MTC and $C_b^a = \binom{a}{b} = a!/[(a-b)!b!]$. These two sufficient conditions ensure that the QoS constraints of MTDs are satisfied, providing at the same time a flexible solution that enables the BS to opportunistically schedule UE devices to achieve efficient resource allocation.

The management of the QoS requirements when M2M and H2H applications coexist is addressed in [32]. Here, the authors apply clustering to divide the devices into two classes: the high-priority class collects all classical H2H users and some delay-sensitive M2M service users, while all the other MTDs are grouped in the low-priority class. Then, the authors define the *M2M Aware Scheduling Algorithm* (M2MA-SA), which aims at preserving the performance experienced by high-priority services in case of massive presence of low-priority devices. The scheduler implements a compound 2-phase procedure. In the first phase, resources in the LTE uplink channel are allocated to the devices in the high-priority class, according to a rather sophisticated priority metric that accounts for multiple factors, including the rate achievable by the user in each available resource block, the remaining time before the delay threshold, the gap between the current rate and the rate required to satisfy the delay constraint, and the time spent waiting in the queue. When all the H2H devices are served, the system starts assigning the residual resources to the MTDs in the low-priority queue. In this case, the algorithm works with a timeline divided in periods of T seconds.[4] Within every interval T, resources are allocated according to two scheduling policies, namely max-utility and round robin. The max-utility policy assigns resources to MTDs with better channel quality. This approach, however, can lead to starvation of users with long-lasting poor channel conditions. To avoid this drawback, users that have not been served for a certain period of time are put in a "timeout" queue and scheduled in a round robin fashion before performing the max-utility scheduling.

The simulation study presented in [32] confirms that M2MA-SA can yield smaller delay, lower outage probability, and higher throughput for H2H users than a simple scheduling mechanism that does not discriminate between H2H and M2M queues. Of course, this result is obtained by penalizing the performance of M2M users.

[4]The length T is chosen according to the delay requirements of M2M services.

4.2. Game theoretic approaches

An alternative to QoS-based clustering approaches is given by the *distributed matching algorithm* introduced in [33]. Once again, the authors refer to a scenario in which both traditional UEs and MTDs are randomly deployed in the coverage area of a BS and share the same resources. However, in this work, the radio resources are originally assigned to the Cellular Users (CUs) only, while MTDs communication is subordinated to the availability of idle portions of such resources. More specifically, each MTD is coupled with a CU and exploits parts of its transmission resources. The coupling between MTDs and CUs results from a distributed algorithm, based on the matching theory, which negotiates the associations between CUs and MTDs, according to a specific metric. In particular, MTDs trade subchannel access in exchange for a reward that is proportional to the fraction of the CU resources utilized by the MTDs. At this stage, utility functions can be defined in order to optimally match the MTDs to the CUs in the network. In particular, the utility of the CU can be defined as the (weighted) sum of its own data-rate and the reward earned from the set of the MTDs it is coupled with. On the other hand, the utility for the MTD is given by the data-rate obtained by using the resources offered by the CU that is suitably scaled by the so-called "urgency factor," which is a decreasing function of the maximum tolerable delay, and then subtracted from the cost paid to the CU in return for such resources.

Now, the goal is to solve an optimization problem that finds the optimal matching between CUs and MTDs, such that the total sum of the utilities of all MTDs and CUs is maximized. The proposed solution, which makes use of a matching algorithm, can be described as follows: after initialization, each CU that has not yet been matched offers a price for the allocation of its own resources to the unmatched MTDs. Each MTD then bids for the CU that provides the highest positive utility (demand set). The CUs that get a bid from one MTD only will be matched with the corresponding device, while those that collect multiple bids increase their allocation price by a certain amount for the next iteration. Results show that the proposed algorithm achieves a stable matching state and an average aggregate utility comparable with a classical centralized algorithm, but with lower overhead for the BS.

4.3. Machine learning algorithm

Machine learning techniques are widely applied to protocol design thanks to their ability to deal with very complex systems in a relatively simple and efficient manner. The management of M2M massive access is no exception, and

several studies in this domain have adopted machine learning techniques to address different problems.

One such problem is the selection of the best serving BS by a MTD. Indeed, one of the expected characteristics of 5G systems is the proliferation of pico and femto cells, which will lead to the densification of the radio coverage. As a consequence, each MTD will likely be in the coverage range of multiple BSs that, however, may offer quite disparate QoS, depending on their distance, signal propagation conditions, traffic load, and so on. Therefore, the selection of the most suitable serving BS becomes an interesting problem.

In [34], focusing on an LTE-Advanced network, the authors propose a Q-learning algorithm to enable MTDs to choose the best serving BS. The algorithm consists of the following 5 steps:

(i) Every device initializes the Q value $Q(s,a)$, where s is the state (i.e., the current reference BS), and a is the action (i.e., the next BS to select), for all states and actions.
(ii) With probability $p \ll 1$ the device performs an *exploration* phase, i.e., randomly chooses a, while with probability $1-p$ it performs an *exploitation* phase, i.e., chooses $a = \arg\max_{a'}\{Q(s,a')\}$.
(iii) The MTD transmits its packet to the BS determined by the selected action a and it measures the current QoS performance $D_{s,a}$.
(iv) Q-value is updated as follows:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(D_{s,a} + \psi \cdot \max_{a'} Q(s',a'))$$

where α is the learning rate (typically $\alpha = 0.5$), ψ is the discount rate (typically $\psi = 0.1$), and s' is the next state if action a is taken when in state s.
(v) Repeat from step ii.

A scenario with two BSs is considered for the performance evaluation. The simulation results show that the proposed reinforcement learning method yields a balanced distribution of the MTDs between the two BSs. Indeed, when the number of nodes that choose one BS increases, the packet delivery delay also increases due to higher congestion at that BS, so that the algorithm will progressively move some devices to the other BS. Moreover, the higher the number of allocated resource blocks by a certain BS, the more the MTDs that select that BS. Finally, it is shown that random BS selection is outperformed by the proposed reinforcement learning algorithm in terms of average packet delivery delay. On the downside, the algorithm is not capable of promptly reacting to sudden changes of the M2M offered traffic, thus possibly yielding a suboptimal behavior for relatively long periods.

5. Cell coverage extension

The cell coverage of current cellular networks has been designed considering devices that, in general, are much more powerful (and expensive) than MTDs. The last generations of smartphones and tablets, for instance, are capable of considerable uplink power levels and are equipped with complex RF components and extremely powerful processing units. Moreover, users are generally satisfied when the battery charge of these devices lasts enough to cover a working day, so that daily recharging is acceptable and, actually, common practice. When dealing with MTDs, instead, the situation is dramatically different. MTDs are generally expected to be very cheap devices, with limited computational and storage capability and extremely long battery life (tens of years). Reducing the power in uplink, the transmit bandwidth, and the complexity of the baseband components usually decreases the cost and power consumption of the RF chain, but at the price of a lower coverage range. Therefore, the maximization of the cell coverage for M2M services is an important objective that calls for innovative solutions both at the physical and access layers. Some techniques that can be considered to reach this objective are discussed in the following.

5.1. Physical layer improvements

From the physical layer perspective, since the room for increase in power and bandwidth is limited due to the strict hardware and battery life constraints of MTDs, one remaining option consists in extending the time duration of the transmission. This can be achieved by time spreading, which can also be done using quasi-orthogonal codes, so that multiuser capabilities are exploited as well. Additionally, retransmissions at the MAC layer can further enhance the coverage for power- and bandwidth-limited transmission.

These approaches are investigated in [35-37]. One approach is based on repetition codes and consists in transmitting multiple copies of the same packet in different time instants, in order to increase the decoding probability of the BS and, in turn, the coverage range. Another technique consists in increasing the power spectral density of the resource blocks that are assigned to MTDs, reducing the power of those allocated to other devices. In this way, the coverage of the MTDs is clearly extended, to the detriment of the service offered to the other terminals.

Note that, in [35], these techniques are actually proposed to boost the coverage in downlink, i.e., from the BS to the MTDs. Nonetheless, repetition coding can also be applied in uplink, though at the cost of higher energy consumption of the MTD node. The latter technique, instead, cannot be used in uplink, because of the limited transmit power of the MTDs.

A physical layer approach to increase both uplink and downlink coverage is proposed in [38]. The proposed mechanism consists in enabling MTDs with low uplink signal strength to transmit over multiple transmission time intervals, which are bundled together, in order to increase the energy received by the BS. The transmission mode of each MTD is stored by the BS, and then used to provide adequate uplink resources to the MTDs during their wake-up time windows.

In summary, physical layer techniques are valuable candidates to increase both downlink and uplink cell coverage for MTDs. However, the solutions proposed in the literature will generally consist in increasing the resources (power,

time, or frequency) allocated to peripheral MTDs, thus exacerbating the massive access issues we discussed in the previous sections. A possible way around these problems is offered by some new technologies that offer long-range, low-bitrate connectivity for MTDs, such as LoRa™[39], Sigfox [40], and *Weightless*™[41], just to mention a few that have been gaining momentum in the past few years. These technologies, however, are generally based on proprietary communication standards, which are not natively compatible with the IP world, so that suitable gateways need to be provided to enable global connectivity.

5.2. Cooperative approach

Another technique to increase the cellular coverage for MTDs in downlink is presented in [42]. This solution leverages on dedicated devices, named Cooperative Gap Fillers (CGFs) [43], which act as relay nodes to MTDs and are equipped with additional functionalities, such as multiple communication interfaces (both long and short range) and larger energy resources. The basic idea is that CGFs forward BS messages through the short-range interface by using network coding techniques [44]. Each message is broken into several packets, which are linearly combined by the intermediate CGF nodes in a random fashion. A global encoding vector is inserted in packet headers to enable message decoding when a sufficient number of linearly independent combinations of packets received.

Although this technique can also be applied in uplink, its effectiveness is limited because messages generated by MTDs are typically short, so that the network coding overhead will become dominant.

6. Clean slate approaches

While the previous studies generally consider as a starting point the current cellular standards, though with different degrees of abstraction of the system components, a few works have investigated the massive access issues in a standard-agnostic manner, with the aim of finding more fundamental results and shading light on the intrinsic performance bounds of these types of systems. Some important results obtained from such a "clean slate" approach are described below.

6.1. Schemes based on Slotted ALOHA

The performance of coordinated and uncoordinated transmission strategies for multiple access is analyzed in [45], where it is shown that, for payloads shorter than 1000 bits (which are typical values in the M2M context), uncoordinated access schemes support more devices than coordinated access mechanisms, because of the lower signaling overhead. A well-known protocol for uncoordinated access is Slotted ALOHA. An enhanced version of this protocol, called Fast Adaptive Slotted Aloha (FASA), is proposed in [46]: taking into account the burstiness of M2M traffic, the knowledge of the idle, successful, or collided state of the previous slots is exploited in order to improve the performance of the access control protocol. In particular, the number of consecutive idle or collision slots is used to estimate the number of active MTDs in the network (the so-called "network status"), enabling a fast update of the transmission probability of the MTDs and, hence, reducing access delays. By means of drift analysis techniques, the authors prove the stability of the FASA protocol when the normalized arrival rate is lower than e^{-1}.

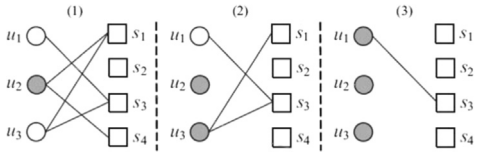


Fig. 6 Sample graph representation of SIC procedure.

Another improved version of Slotted ALOHA exploiting Successive Interference Cancellation (SIC), called Frameless ALOHA, is presented in [47]. A simple example illustrating the principle of such SIC-enabled Slotted ALOHA is shown in Fig. 6. It depicts the situation in which $N=3$ users contend to transmit within the same frame, composed by $M=4$ slots. The nodes on the left represent users, the nodes on the right stand for time slots, and the edges connect the users with the slots in which their respective transmissions take place. All transmissions made by a user in the frame are replicas of the same packet; moreover, every transmission includes pointers to all its replicas. With this in mind, SIC can be effectively exploited as follows.

First, slots containing a single transmission (*singleton* slots) are identified and the corresponding transmission resolved. Referring to Fig. 6, s_4 is recognized as a singleton slot and the associated packet of user u_2 is hence correctly decoded (Fig. 6.1). In the next step, using the pointers carried by the decoded packets, all their replicas are removed from the associated slots, i.e., the interference caused by such transmissions is canceled from the aggregate received signal, thus potentially leading to new singleton slots.[5] In the example of Fig. 6, the replica sent by u_2 in s_1 is deleted and, as a result, s_1 becomes singleton slot (Fig. 6.2). Such procedure iterates identically until either there are no new singleton slots, or all transmissions have been recovered (Fig. 6.3).

Generalizing this procedure and applying it to the M2M communication context, we can consider a scenario in which N MTDs contend to access the same BS. The protocol assumes that users are slot-synchronized and aware of the start of the contention period, which will be broadcast by the BS. For each slot in the contention period, each active device randomly decides whether or not to transmit a replica of the pending packet, according to a predefined probability:

$$p_a = \frac{\beta}{N} \tag{6}$$

where β is a suitably chosen parameter, subject to optimization. After each slot, the BS collects the received compound signal and tries to decode the transmitted packets using the above described SIC procedure. The key feature of the frameless ALOHA proposed in [47] is that the

[5] For the sake of simplicity, signal cancellation is assumed to be perfect, i.e., to completely remove the power of the cancelled signal without leaving any residual interference.

end of the contention period is dynamically determined in order to maximize the throughput. Users that have not successfully delivered their packet by the end of a contention period will keep performing the algorithm in the subsequent contention round. Note that, if the contention period is terminated at the Mth slot and the number of resolved MTCs is N_R, then the instantaneous throughput can be computed as $T_I = N_R/M$. The results presented in [47] show that the proposed algorithm can achieve extremely high throughput and very low loss rate, thus proving the effectiveness and efficiency of the described model in a M2M scenario.

The performance of the SIC-based Frameless ALOHA scheme can be further improved by considering the capture phenomenon, as done in [48]. Indeed, the BS can actually be able to decode colliding signals with a signal-to-interference-plus-noise-ratio (SINR) γ larger than a certain threshold, which is called capture ratio and denoted by b. Mathematically, signal j is captured if

$$\gamma_j = \frac{P_j}{\sum_{h \neq j} P_h + \eta_0} > b \tag{7}$$

where P_i denotes the power of the ith signal at the receiver, and η_0 represents the background noise power. Concerning the capture effect, the performance of the iterative decoding algorithm is further enhanced, since packet decoding can also occur in non-singleton slots, where capture phenomena take place. In [48] it is shown that, at least when the impact of noise is low, the capture effect may result in substantially higher throughput compared to the scenario without capture.

All in all, frameless ALOHA can ideally guarantee high performance in a M2M scenario in terms of throughput. Nonetheless, energy efficiency and complexity aspects have not been considered yet. In particular, the SIC mechanism sets quite high requirements to the BS in terms of storage and processing capabilities. The BS indeed has to store the raw samples of the compound received signal in all the slots of a contention period, and carry out SIC-based signal decoding on many slots in real time. In addition, the frameless ALOHA protocol has a strong impact on MTDs' energy consumption because, for each frame, the devices must transmit a possibly large set of replicas of the same packet to the BS. This aspect is a major issue in M2M communication since, as we already pointed out, many MTDs are constrained by the need to operate for years without any battery replacement/recharge.

6.2. Asymptotic analysis of massive access capacity

In the recent literature [49,50], it was observed that using SIC in combination with multi-packet reception capabilities makes it possible to dramatically increase the system throughput even when transmitters are not centrally coordinated. Let us then consider a cellular system where the base station (BS) is capable of decoding all signals that satisfy (7), and can also perform *perfect SIC*, completely removing the interference caused by the decoded signals to the remaining signals. Suppose that N nodes transmit simultaneously, and that the received signal powers at the BS, $\{P_j, j = 1, \ldots, N\}$, are iid random variables with Cumulative Distribution Function $F(\cdot)$. A simplified expression of the maximum achievable throughput of such a system is derived in [51], elaborating on the results provided in [50]. For the sake of completeness, we report here the derivation.

Denoting by $N_0 = N$ the initial number of overlapping transmissions, the average number of still undecoded signals after h interference cancellation cycles can be recursively estimated as

$$N_h = NF(I_{h-1}), \quad I_h = (N_h - 1)b\mathrm{E}[P|P < I_{h-1}], \quad h = 1, 2, \ldots \tag{8}$$

where I_h denotes the (approximate) residual mean aggregate interference at the hth cycle, with $I_0 = \infty$, while $\mathrm{E}[P|P < I_{h-1}] = \int_0^{I_{h-1}} (1 - F(x)/F(I_{h-1}))\, dx$ is the mean power of each still undecoded signal. Since N_h cannot increase in h, the throughput achieves it maximum when h grows to infinity, for which we get

$$S_\infty(N) = N - \lim_{h \to \infty} N_h = N(1 - F(I_\infty(N))) \tag{9}$$

when $I_\infty(N)$ is equal to the fixed-point of (8), i.e.,

$$I_\infty(N) = (NF(I_\infty(N)) - 1)b\mathrm{E}[P|P < I_\infty(N)] \tag{10}$$

provided it exists, and $I_\infty(N) = 0$ otherwise. Finally, given b, we can determine

$$S^*(b) = \max_N S_\infty(N)$$

Now, supposing that MTDs work in the low SNR regime, the per-user transmit rate will linearly scale with b. We hence define the *aggregate system capacity* as

$$C^* = bS^*$$

which is an approximate measure of the maximum spectral efficiency that can be achieved in the cell, assuming perfect SIC capabilities and capture threshold b. This value is shown in Fig. 7, for three of the scenarios defined in [50], namely pure path loss (PL2), Rayleigh fading channels (RF), and Shadowing channels with standard deviation of 3 dB (SH3) and 6 dB (SH6). From these results we observe that a BS capable of performing perfect SIC and MPR can theoretically decode an arbitrary large number of simultaneous transmissions by proportionally reducing the per-user data rate. Doing this, the aggregate system capacity remains almost constant. Furthermore, from this analysis it appears that the capacity of the cell depends on the statistical distribution of the signal powers, and the higher the variance, the more effective the SIC. Therefore, combining SIC, multi-packet reception, and coded random access techniques, it is possible to support massive access of sporadic transmitters to a common BS, provided that the receiver is capable of performing multi-slot SIC decoding. Once again, however, the analysis has not yet considered aspects related to the energy consumption of the MTDs.

7. Discussion

This paper provides a survey of the main challenges raised by massive M2M access in wireless cellular systems, and of the methodologies and approaches that have been proposed to improve the coexistence of human-initiated and machine-initiated communications and counteract the problems generated by signaling overload and channel access

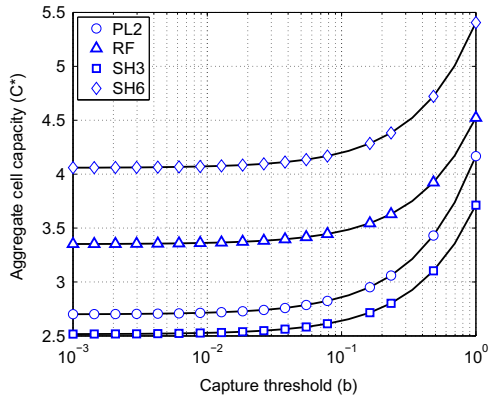


Fig. 7 Asymptotic optimal capacity of a cell with SIC and MPR capabilities.

contention. In this section, we sum up our study by providing a comparative analysis of the solutions described in the previous sections. Then, we discuss the potential of the most promising approaches that are being proposed for the next generation of wireless cellular systems to solve the challenges of M2M massive access.

7.1. Comparative analysis of current solutions for massive access

Tab. 1 offers a compound view of the solutions described in the paper, with an indication of the characterizing features and the main targeted performance indices. More specifically, we consider the following aspects:

- *Main challenge*: primary issue addressed by the scheme;
- *3GPP*: the scheme has been proposed in 3GPP technical report and is designed with explicit reference to 3GPP standards (LTE in particular);
- *H2H & M2M*: the scenario assumes the coexistence of H2H and M2M services;
- *Performance indices*: the scheme is designed to improve the following figures of merit
  - minimization of *access delay*;
  - minimization of *energy consumption* of MTDs;
  - maximization of *access probability/throughput* of UEs and/or MTDs.

A quick look at the table shows that much attention has been devoted to the PRACH overload problem, which does not only impact the performance of M2M services but, more critically, can severely degrade the quality of conventional H2H services, which have high *Average Revenue Per User* (ARPU). This risk is much feared by operators because the quality degradation of conventional services would impact the customer satisfaction and loyalty, thus jeopardizing the operators business. For this reason, great effort has been devoted to the study and test of solutions for enabling M2M services in the current and next cellular system architectures or, at least, for mitigating the possible impact of M2M traffic on conventional services. Besides this main trend of research, there have been several studies regarding other aspects of MTC and, in particular, energy efficiency, QoS, and coverage extension. Although the literature contains a fair number of works that apply different approaches and methodologies to explore the challenges posed by M2M massive access scenarios, the full realization of such scenarios requires much more work. Most of the proposed schemes, indeed, refer to specific use cases and are designed to optimize only few aspects of the system, such as the energy efficiency of MTC, the delay, or the coexistence with H2H traffic. However, M2M applications have extremely different characteristics in terms of generated traffic and service requirements, so that the full support of the M2M paradigm calls for flexible solutions, capable of discriminating among different types of MTDs and of providing differentiated services according to the specific requirements of the applications and the current conditions of the system.

Finally, we observe that, while the fundamental limits for broadband systems are well understood in the literature and considered by the standardization initiatives, a similar level of understanding for MTC-oriented systems is still lacking. Finding such limits will serve as a basis to understand what is possible and to design efficient and flexible MTC systems. Therefore, there is a need for more studies that, adopting a clean-slate and standard-agnostic approach, can provide insights on the fundamental aspects of MTC, thus contributing to the definition of radio protocols and architectural principles able to serve a large number of MTC connections.

7.2. Massive access in 5G

The support of M2M services has been identified as one of the most challenging objectives of 5G [11]. While the general requirements of 5G systems are progressively taking shape [52,53], the technological issues raised by the M2M paradigm are still partially unclear. General consensus has been reached on the importance of few, key approaches and technologies, including massive MIMO, small cells, millimeter wave (mmWave) communication, and virtualization of system elements and network functions. Nonetheless, whether these technologies will be able to provide efficient support to M2M services in 5G systems and to coexist with traditional broadband services is still an open question, as argued in the following.

(a) *Massive MIMO*: This technique consists in equipping the BS with much more antennas than the number of devices, so that the channels to the different devices will be quasi-orthogonal, thus making it possible to increase the spectral efficiency by using simple spatial multiplexing/demultiplexing procedures [54]. Therefore, massive MIMO can dramatically enlarge the number of simultaneous transmissions that can be successfully received by a (powerful) BS, without burdening the peripheral nodes. These characteristics, in principle, make massive MIMO extremely attractive for supporting massive M2M access. The limit of such an approach is that enabling *massive* MIMO for a *massive* number of MTDs may require an exceedingly large number of antennas at the BS, which can be infeasible due to logistic and technical problems. Furthermore, despite the huge interest in massive MIMO, there is still much to be learned, in particular regarding the propagation and cost

effectiveness, so that the actual performance and feasibility of this technique are still open to investigation.

(b) *Small cells*: One possible solution for dealing with the increase of the number of devices in hot spot areas are the densification of the network by employing small cells [55], a paradigm that is also known as Heterogeneous Networks (HetNets). Small cells are indeed deployed to reduce the distance between devices and access points, thus enabling higher bit rates (or lower transmit power and interference), while also improving the spatial reuse. In an M2M setting, however, the focus is not on high transmit rate, but rather on reliable and ubiquitous connectivity. MTDs are in fact expected to be spread across wide areas, also where human-generated broadband traffic may be light, e.g., along highways/road/railroads or in agricultural areas. Hence, providing access to MTDs will require uniform and ubiquitous coverage that is not economically sustainable by using microcells. Even in case of a high concentration of MTDs in relatively small areas, the Average Revenue Per User of MTD-based services is likely lower compared to conventional services, thus not justifying the deployment of small cells for the sake of MTD-coverage only. Finally, the densification of the network does not impact the signaling overhead at the PHY layer, which is inefficient due to the MTC transmission characteristics.

(c) *mmWave communication*: After years of striving to squeeze more spectral efficiency from the crowded bandwidth used by current microwave cellular systems, the huge bandwidth available at mmWave frequencies, from 3 to 300 GHz, represents an irresistible attraction for 5G systems. Although the signal propagation at these frequencies is not yet thoroughly understood, the measurements reported in [56] indicate that transmission can occur even in the absence of line of sight, though with a much higher path loss exponent. In combination with large antenna arrays, mmWave communication can make it possible to reach huge bitrates over short distances. However, the sensitivity to blockage, the rapid power decay with distance, and the higher power requirements of mmWave communications make this technology less attractive for MTDs that, instead, need long-range, low-power, and low bitrate connections.

(d) *Virtualization*: Software Defined Networking (SDN) and Network Function Virtualization (NFV) are two emerging paradigms that basically consist in abstracting low-level network functionalities to enable a much more flexible management of the network resources and a better and adaptive support of different types of services [57]. The accomplishment of these concepts would make it possible to differentiate the services offered to the different traffic flows and to dynamically instantiate network elements where and when needed. Ideally, these mechanisms shall deliver the illusion of "infinite capacity," giving to each application exactly the resources it needs to achieve the desired Quality of Experience (QoE). This vision is extremely appealing for what regards the support of massive M2M traffic, in that SDN can naturally provide separation between M2M and H2H traffic, while guaranteeing the desired QoS levels to each type of flow (both at the access network and across the core network). Moreover, the fine-grained and per-flow resource allocation paradigm enabled by SDN will result in a better utilization of the network resources, thus contributing to alleviate the massive access problem. NFV, on the other hand, can be used to dynamically shape the network architecture according to the traffic requirements. For example, NFV can instruct network elements in a certain area to act as concentrators to collect MTDs data, or as relays to extend the coverage range, or even as additional BSs to satisfy temporary peaks of access requests. While this virtualization principle can bring a disruptive change in the architectural design of next generation communication systems, major research efforts are still required to turn this vision into reality.

8. Conclusions

Today's cellular systems have not been designed to support M2M services, yet they are able to supply the present-day demand for M2M services, realizing the "place-&-play" concept that was described in the introduction. However, if the M2M market will fulfil the big expectations of the stakeholders, the limits of these technologies will soon become apparent. These considerations have driven academia, research institutes, industries, and standardization bodies to devise improvements of current standards and to design novel solutions to face the challenges posed by these new types of services. Meanwhile, the growing demand from the M2M market has fueled the proliferation of proprietary solutions, which are not natively compatible with the IP world and necessitate suitable gateways to interact with the rest of the world.

For the moment, then, the picture of M2M support appears quite composite and variegate, with no clearly emerging solution. The evolution of the M2M services in the coming years will likely rely upon a mix of proprietary technologies, explicitly designed for MTD connectivity, and legacy cellular standards, suitably enhanced to better scale with massive access requests from MTDs. Considering that many M2M systems are expected to remain operational for a long time with minimal intervention and maintenance, such hybrid architectures will probably endure for many years to come, and will be eventually absorbed by 5G that will offer native support to M2M services.

References

[1] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of Things for Smart Cities, IEEE Internet Things J. 1 (February (1)) (2014) 22–32.
[2] L. Atzori, A. Iera, G. Morabito, The internet of things: A survey, Comput. Netw. 54 (October (15)) (2010) 2787–2805.
[3] P. Bellavista, G. Cardone, A. Corradi, L. Foschini, Convergence of MANET and WSN in IoT urban scenarios, IEEE Sens. J. 13 (October (10)) (2013) 3558–3567.
[4] J. Zheng, M.J. Lee, A comprehensive performance study of IEEE 802.15.4, in: Sensor Network Operations, IEEE Press Book, Wiley Interscience, 2006, pp. 218–237 (Chapter 4).
[5] G. Montenegro, N. Kushalnagar, J. Hui, D. Culler, Transmission of IPv6 Packets over IEEE 802.15.4 Networks, Technical Report IETF Request for Comments 4944, September 2007. [Online]. Available ⟨http://tools.ietf.org/pdf/rfc4944.pdf⟩.

[6] P. Kinney, et al., Zigbee technology: Wireless control that simply works, in: Communications Design Conference, October 2003.
[7] H. Merz, J. Backer, V. Moser, T. Hansemann, L. Greefe, C. Hübner, Building Automation: Communication Systems with EIB/KNX, LON and BACnet, Springer-Verlag, Berlin, Heidelberg, June 2009.
[8] 3GPP, Feasibility Study for Proximity Services (ProSe), Technical Report TR 22.803 V12.2.0, June 2013.
[9] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, A. Oliveira, Smart cities and the future Internet: towards cooperation frameworks for open innovation, in: The Future Internet, Springer, Berlin, Heidelberg, vol. 6656, 2011, pp. 431–446.
[10] A. Laya, L. Alonso, J. Alonso-Zarate, Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives, IEEE Commun. Surv. Tutor. 16 (First Quarter(1)) (2014) 4–16.
[11] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, P. Popovski, Five disruptive technology directions for 5G, IEEE Commun. Mag. 52 (February (2)) (2014) 74–80.
[12] 3GPP, Bottleneck Capacity Comparison for MTC, Technical Report TSG GERAN 46 GP-100895, May 2010.
[13] R. Paiva, R.D. Vieira, M. Saily, Random access capacity evaluation with synchronized MTC users over wireless networks, in: IEEE Vehicular Technology Conference, May 2011, pp. 1–5.
[14] G. Madueno, C. Stefanovic, P. Popovski, How many smart meters can be deployed in a GSM cell? in: IEEE International Conference on Communications (ICC) Workshops, June 2013, pp. 1263–1268.
[15] P. Jain, P. Hedman, H. Zisimopoulos, Machine type communications in 3GPP systems, IEEE Commun. Mag. 50 (November (11)) (2012) 28–35.
[16] A. Lo, Y. Law, M. Jacobsson, M. Kucharzak, Enhanced LTE-advanced random-access mechanism for massive Machine-to-Machine (M2M) communications, in: 27th World Wireless Research Forum Meeting, October 2011.
[17] E. Dahlman, S. Parkvall, J. Sköld, 4G LTE/LTE-Advanced for Mobile Broadband, Elsevier, October 2013.
[18] 3GPP, Study on RAN Improvements for Machine-type Communications, Technical Report TR 37.868 V11.0.0, September 2011.
[19] K.S. Ko, M.J. Kim, K.Y. Bae, D.K. Sung, J.H. Kim, J.Y. Ahn, A Novel Random Access for Fixed-Location Machine-to-Machine Communications in OFDMA Based Systems, IEEE Commun. Lett. 16 (September (9)) (2012) 1428–1431.
[20] T. Taleb, A. Kunz, Machine type communications in 3GPP networks: potential, challenges, and solutions, IEEE Commun. Mag. 50 (March (3)) (2012) 178–184.
[21] L.M. Bello, P. Mitchell, D. Grace, Application of Q-learning for RACH access to support M2M traffic over a cellular network,in: European Wireless Conference, May 2014, pp. 1–6.
[22] Y.-C. Pang, S.-L. Chao, G.-Y. Lin, H.-Y. Wei, Network access for M2M/H2H hybrid systems: a game theoretic approach, IEEE Commun. Lett. 18 (June (5)) (2014) 845–848.
[23] W.H. Chin, Z. Fan, R. Haines, Emerging technologies and research challenges for 5G wireless networks, IEEE Wirel. Commun. 21 (April (2)) (2014) 106–112.
[24] H. Dhillon, H. Huang, H. Viswanathan, R. Valenzuela, On resource allocation for machine-to-machine (M2M) communications in cellular networks, in: IEEE Globecom Workshops, December 2012, pp. 1638–1643.
[25] S. Jayashree, B. Manoj, C. Murthy, A battery aware medium access control (BAMAC) protocol for ad hoc wireless networks, in: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), vol. 2, September 2004, pp. 995–999.
[26] R. Kim, Snoop based group communication scheme in cellular Machine-to-Machine communications, in: International Conference on Information and Communication Technology Convergence (ICTC), November 2010, pp. 380–381.
[27] C.-Y. Tu, C.-Y. Ho, C.-Y. Huang, Energy-efficient algorithms and evaluations for massive access management in cellular based machine to machine communications, in: IEEE Vehicular Technology Conference (VTC Fall), 2011, September 2011, pp. 1–5.
[28] C.Y. Ho, C.-Y. Huang, Energy-Saving Massive Access Control and Resource Allocation Schemes for M2M Communications in OFDMA Cellular Networks, IEEE Wirel. Commun. Lett. 1 (June (3)) (2012) 209–212.
[29] S.-Y. Lien, K.-C. Chen, Y. Lin, Toward ubiquitous massive accesses in 3GPP machine-to-machine communications, IEEE Commun. Mag. 49 (April (4)) (2011) 66–74.
[30] S.-Y. Lien, K.-C. Chen, Massive Access Management for QoS Guarantees in 3GPP Machine-to-Machine Communications, IEEE Commun. Lett. 15 (March (3)) (2011) 311–313.
[31] P. Si, J. Yang, S. Chen, H. Xi, Adaptive Massive Access Management for QoS Guarantees in M2M Communications, IEEE Trans. Veh. Technol., Available: ⟨http://dx.doi.org/10.1109/TVT.2014.2349732⟩, in press.
[32] S. Zhenqi, Y. Haifeng, C. Xuefen, L. Hongxia, Research on uplink scheduling algorithm of massive M2M and H2H services in LTE, in: IET International Conference on Information and Communications Technologies (IETICT), April 2013, pp. 365–369.
[33] S. Bayat, Y. Li, Z. Han, M. Dohler, and B. Vucetic, Distributed massive wireless access for cellular machine-to-machine communication, in: IEEE International Conference on Communications (ICC), 2014, June 2014, pp. 2767–2772.
[34] M. Hasan, E. Hossain, D. Niyato, Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches, IEEE Commun. Mag. 51 (June (6)) (2013) 86–93.
[35] Y. Qi, A. Ijaz, A. Quddus, M. Imran, P. Navaratnam, Y. Ma, R. Tafazolli, M. Webb, Y. Morioka, On the physical layer design for low cost machine type communication in 3GPP LTE, in: IEEE Vehicular Technology Conference (VTC Fall), September 2014, pp. 1–5.
[36] 3GPP, LTE Coverage Enhancements, Technical Report TR 36.824 V11.0.0, June 2012.
[37] General Dynamics Broadband UK, Coverage Extension for MTC UEs, Technical Report TSG RAN171 R1-125204, November 2012.
[38] Y. Chan, J. Yang, P. Zong, Machine-to-machine Communications over Fixed Wireless Networks, April 2013, US Patent 8,416,741. [Online]. Available ⟨http://www.google.com/patents/US8416741⟩.
[39] ⟨http://www.semtech.com/wireless-rf/lora.html⟩.
[40] ⟨http://www.sigfox.com⟩.
[41] ⟨http://www.neul.com⟩.
[42] G. Cocco, C. Ibars, N. Alagha, Cooperative coverage extension in heterogeneous machine-to-machine networks, in: IEEE Globecom Workshops, December 2012, pp. 1693–1699.
[43] EXALTED Project, First Report on LTE-M Algorithms and Procedures, Technical Report, August 2011. [Online]. Available: ⟨http://www.ict-exalted.eu/fileadmin/documents/EXALTED_WP3_D3.1_v2.0.pdf⟩.
[44] R. Ahlswede, N. Cai, S.-Y. Li, R. Yeung, Network information flow, IEEE Trans. Inf. Theory 46 (July (4)) (2000) 1204–1216.
[45] H. Dhillon, H. Huang, H. Viswanathan, R. Valenzuela, Fundamentals of Throughput Maximization With Random Arrivals for M2MCommunications, IEEE Trans. Commun. 62 (November (11)) (2014) 4094–4109.

[46] H. Wu, C. Zhu, R. La, X. Liu, Y. Zhang, FASA: Accelerated S-ALOHA Using Access History for Event-Driven M2MCommunications, IEEE/ACM Trans. Netw. 21 (December (6)) (2013) 1904–1917.
[47] C. Stefanovic, P. Popovski, ALOHA Random Access that Operates as a Rateless Code, IEEE Trans. Commun. 61 (November (11)) (2013) 4653–4662.
[48] C. Stefanovic, M. Momoda, P. Popovski, Exploiting capture effect in frameless ALOHA for massive wireless random access, in: IEEE Wireless Communications and Networking Conference (WCNC), April 2014, pp. 1762–1767.
[49] X. Wang, J. Garcia-Luna-Aceves, Embracing interference in ad hoc networks using joint routing and scheduling with multiple packetreception, Ad Hoc Netw. 7 (March (2)) (2009) 460–471.
[50] A. Zanella, M. Zorzi, M2M massive wireless access: Challenges, research issues, and ways forward, IEEE Trans. Commun. 60 (April (4)) (2012) 1058–1071.
[51] A. Zanella, M. Zorzi, A. dos Santos, P. Popovski, N. Pratas, C. Stefanovic, A. Dekorsy, C. Bockelmann, B. Busropan, T. Norp, M2M massive wireless access: challenges, research issues, and ways forward, in: IEEE Globecom Workshops, December 2013, pp. 151–156.
[52] NetWorld2020 ETP, 5G: Challenges, Research Priorities, and Recommendations, Technical Report, August 2014. [Online]. Available ⟨http://networld2020.eu/wp-content/uploads/2014/02/NetWorld2020_Joint-Whitepaper-V8_public-consultation.pdf⟩.
[53] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, et al., Scenarios for 5G mobile and wireless communications: the vision of the METIS project, IEEE Commun. Mag. 52 (May (5)) (2014) 26–35.
[54] Y.-H. Nam, B.L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, J. Lee, Full-dimension MIMO (FD-MIMO) for next generation cellular technology, IEEE Commun. Mag. 51 (June (6)) (2013) 172–179.
[55] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, S. Geirhofer, Network densification:the dominant theme for wireless evolution into 5G, IEEE Commun. Mag. 52 (February (2)) (2014) 82–89.
[56] T.S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G.N. Wong, J.K. Schulz, M. Samimi, F. Gutierrez, Millimeter-wave mobile communications for 5G cellular: It will work!, IEEE Access 1 (May) (2013) 335–349.
[57] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, J. Yao, 5G on the Horizon: Key Challenges for the Radio-Access Network, IEEE Veh. Technol. Mag. 8 (September (3)) (2013) 47–53.