The Random Access Procedure in Long Term Evolution Networks for the Internet of Things

Tiago P. C. de Andrade, Carlos A. Astudillo, Luiz R. Sekijima, and Nelson L. S. da Fonseca

The authors review the LTE random access procedure and its support for IoT applications. They also assess the performance of the RAN overload control schemes proposed by 3GPP, taking into consideration the interaction between the random access procedure and packet downlink control channel resource allocation.

ABSTRACT

Network connectivity is a key issue in the realization of IoT, and LTE cellular technology is the most promising option for the provisioning of such connectivity. However, in LTE networks, a large number of IoT devices trying to access the medium can overload the RAN. In this article, we review the LTE random access procedure and its support for IoT applications. We also assess the performance of the RAN overload control schemes proposed by 3GPP, taking into consideration the interaction between the random access procedure and packet downlink control channel resource allocation.

INTRODUCTION

In the Internet of Things (IoT), tens of billions of devices with sensing, computing, and communication capabilities will improve our daily life and create new business opportunities [1]. Various sectors will benefit from the information exchange in IoT, such as transportation, health care, and manufacturing, as well as the development of smart cities, smart grid, and smart home. Devices will be interconnected by a diverse communication infrastructure, with connectivity being a key issue for the realization of IoT.

Machine-to-machine (M2M) communication, or machine-type communication (MTC) technology, will enable the interaction of IoT devices. Technologies using unlicensed frequency bands and featuring low power consumption for a short transmission range, such as RF identification, Zigbee, Bluetooth Low Energy, and low-power WiFi, have been designed to support M2M applications. However, in order to provide coverage for wide areas, which is a key requirement for various IoT applications, these technologies rely on multihop packet forwarding, as well as the addition of backhaul links. Moreover, these technologies are prone to interference because of the use of the unlicensed spectrum, which reduces the reliability and availability of these systems, and increases communication delays [2].

To overcome some of these limitations, long-range, low-power communication technologies, known as low-power wide area, such as Sigfox, LoRa, Weightless, and Long Term Evolution (LTE),[1] are gaining momentum in the IoT connectivity landscape [3], with the LTE cellular technology being the most suitable solution for the interconnection of IoT devices due to its wide coverage, security, licensed spectrum, and simplicity of management. By using LTE technology for MTC, mobile network operators (MNOs) can leverage their investment in 4G LTE networks to provide IoT connectivity.

Traditionally, cellular networks were designed to support human-to-human (H2H) communications. However, the requirements of IoT M2M communication and the energy limitation of devices impose additional requirements for the cellular networks. For example, severe congestion can occur when a massive number of transmitting devices attempt to access the network simultaneously. Moreover, the connection-oriented communication in traditional cellular networks can generate excessive signaling overhead for transmitting small data packets generated by IoT applications. Consequently, quality of service (QoS) provisioning for human-type communication (HTC) and MTC can be jeopardized.

In order to make the LTE technology more suitable for M2M communications, 3GPP LTE-standard Releases 11, 12, and 13 included different features to support MTC applications, known as LTE for MTC (LTE-M), as well as a new technology, known as Narrowband LTE (NB-LTE).

This article focuses on the radio access network (RAN) overload problem, especially the problem of congestion arising from a massive number of MTC devices trying simultaneously to access the LTE network. The evolution of the LTE standard for the support of IoT applications is reviewed, especially a variety of proposals impacting the random access procedure. The performance of the main approaches for the amelioration of the RAN overload problem [4] is shown. This article considers the interaction of the random access procedure and packet downlink control channel (PDCCH) resource allocation, which has not been undertaken previously. Results derived via extensive simulations show that the RAN overload problem has been underestimated. We show that physical and PDCCH constraints strongly impact the network performance during the random access procedure. Based on these findings, we present key research directions for the improvement of the performance of the random access procedure of LTE to enhance the access by the massive number of devices expected in IoT scenarios.

[1] In this article, we use LTE to refer to all technologies based on Third Generation Partnership Project (3GPP) LTE standards (Release 8 and beyond).

Digital Object Identifier:
10.1109/MCOM.2017.1600555CM

The authors are with the State University of Campinas.

0163-6804/17/$25.00 © 2017 IEEE

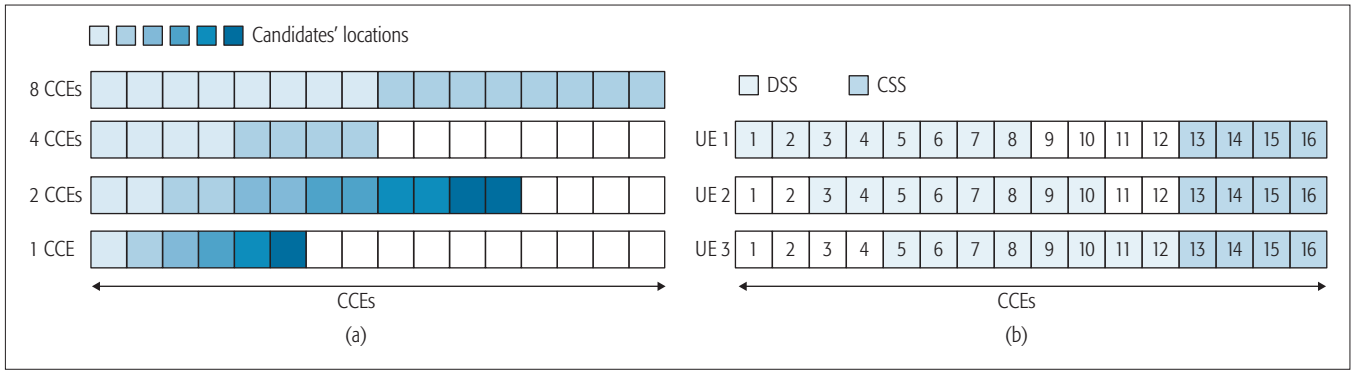

Figure 1. Constraints to the PDCCH resource allocation in LTE networks: a) PDCCH search space candidates; b) overlap on PDCCH for different UEs.

LTE Packet Downlink Control Channel

In the PDCCH, downlink control Information (DCI) messages are transmitted carrying downlink assignment, uplink grants, and random access related messages. Assignments are used to convey the information needed to receive data from the evolved NodeB (eNB) on the physical downlink shared channel (PDSCH), whereas grants allow user equipments (UEs) to transmit data to the eNB on the packet uplink shared channel (PUSCH). The PDCCH uses up to the first three orthogonal frequency-division multiplexing (OFDM) symbols of each subcarrier.

Each DCI message can use 1, 2, 4, or 8 control channel elements (CCEs) (aggregation levels), depending on the message format and channel quality. The DCI messages are sent on the PDCCH, but the UE does not know a priori information about the exact location of its messages. Each UE applies blind decoding on a specific set of CCEs in two regions of the PDCCH, the common search space (CSS) and the dedicated search space (DSS) to determine which, if either, contains DCI message(s) to the device. Each UE will monitor 6 candidate locations at aggregation levels 1 and 2, as well as 2 candidate locations at levels 4 and 8, as illustrated in Fig. 1a.

One problem with the design of the PDCCH is that the eNB cannot freely use all available CCEs to schedule DCI messages, which can only be scheduled on the specific PDCCH resources of the intended UE. Thus, there can be an overlap in the resources allocated for the UEs on the PDCCH if there are few CCEs or a large number of UEs in the cell, as illustrated in Fig. 1b.

LTE Random Access Procedure

The random access procedure is performed by a UE in the following cases:

- Upon initial access to the network
- Upon arrival of uplink data at the UE buffer if no radio resources have been assigned to request uplink resources
- During handover
- Upon radio failure to re-establish a connection
- When the UE is not synchronized with the eNB

There are two types of random access (RA) procedures: contention-free and contention-based. The former is used to perform handover, whereas the latter is used otherwise. The four-way handshake contention-based RA procedure is described below.

The UE first transmits a preamble (msg1) message on the random access channel during the first random access opportunity (RAO) after the triggering of the random access procedure. The eNB periodically informs the UEs about a set of up to 64 orthogonal preamble sequences from which the UE can make a choice. Collisions occur when two or more UEs transmit the same preamble sequence during the same RAO. However, the eNB does not detect such collisions during this step.

The second step is the transmission of a random access response (RAR) (msg2) message by the eNB, addressed to the random access temporary identifier (RA-RNTI) over the downlink shared channel. A DCI message is thus scheduled on the CSS of the PDCCH to indicate the PDSCH resources in which the RAR message was transmitted. This message contains a timing advance command and an uplink grant for the transmission of a message in the following step. If the UE device that sent a preamble sequence does not receive an msg2 message from the eNB within a certain period of time, it enters a backoff period again, trying to access the network once this period has expired.

Then the UE transmits a layer 2/3 (L2/L3) (msg3) message on the uplink shared channel. The message is addressed to the RA-RNTI and carries either the identity of the UE,if it is in the connected radio resource control (RRC) state, or a temporary UE identity (if the UE is in the idle RRC state). If two or more UEs have chosen the same preamble sequence in a RAO, they will receive the same grant in the RAR message, and thus, all their L2/L3 message transmissions will collide. A contention resolution (msg4) message is then sent to each UE on which msg3 message was successfully received by the eNB.

LTE RAN Enhancement for MTC and Its Implications for Random Access Procedure Performance

This section provides a brief review of the efforts by 3GPP to support MTC, highlighting UE categories and capabilities, as well as the implications of these efforts for the random access procedure.

Enhancement for MTC

3GPP Release 11 focuses on RAN overload control schemes [4], which can improve network reliability in the face of a massive number of simul-

The 3GPP Release 12 introduces a new LTE UE device category (Cat. 0) for MTC devices to potentiate LTE penetration into restrictive MTC markets. This new category decreases the cost and complexity of the LTE chipset. It features a single receiver, 1 Mb/s maximum bit rate, and half-duplex operation.

taneous attempts to access the network. Most of these solutions are based on the barring of access of devices or the splitting of random access radio resources between different UE device classes. Moreover, an enhanced PDCCH (ePDCCH) structure has been proposed. By using the ePDCCH, a portion of the resources dedicated to the PDSCH is used for conveying control resources. It can alleviate the shortage of control resources resulting from a massive number of devices [5].

3GPP Release 12 introduces a new LTE UE device category (Cat. 0) for MTC devices to potentiate LTE penetration into restrictive MTC markets. This new category decreases the cost and complexity of the LTE chipset. It features a single receiver, 1 Mb/s maximum bit rate, and half-duplex operation, which can achieve a 50 percent reduction in complexity and a 30 percent reduction in cost when compared to a Cat. 1 chipset.

Two important features for MTC devices are introduced in 3GPP Release 13: enhanced coverage and Cat. M1, a new low-complexity UE device category. The Cat. M1 reception bandwidth has been reduced to 1.4 MHz. This release also introduces coverage levels and physical channel repetitions to improve coverage and allows the relaxation of hardware requirements. These cost reductions and coverage enhancements allow MNOs to cover a larger number of IoT applications with LTE technology. Since the legacy PDCCH is spread across the entire bandwidth, a new PDCCH design, the MTC PDCCH, has been developed to support this new category. However, the uplink channel, including the physical random access channel (PRACH), remains the same as for UE Cat. 0 and above.

The final 3GPP enhancement for MTC is called NB-LTE [6], which will operate with a 200 kHz channel bandwidth. Consequently, other UE device categories will be introduced in future releases, further reducing the cost of the MTC device. Such a new category should reduce the complexity of hardware by at least 80 percent in comparison with the hardware of a Cat. 1 UE. This technology aims to support ultra-low-complexity and low-throughput IoT applications via LTE cellular systems. Although NB-LTE still makes extensive use of the higher-layer user plane, it represents a "clean slate" approach,[2] in which many aspects of the physical layer of the LTE technology will be changed. In the uplink, the duration of OFDM symbol, slot, and subframe will be six times longer than its LTE counterpart. Thus, an NB-LTE subframe (M-subframe) is now 6 ms rather than the 1 ms in the traditional LTE system. Even though NB-LTE has the possibility of 64 preamble sequences available as well as the random access procedure with four messages, it remains basically the same as for the standard LTE technology; the MTC PRACH occupies two M-subframes (12 ms) and uses different preamble sequence signals. Depending on the level of coverage, which determines the number of repetitions required, up to six M-subframes may be needed to transmit the preamble sequence. Six devices is the maximum number that can be scheduled in an M-subframe. The Narrowband IoT technology, which is similar to NB-LTE, was recently included in 3GPP Release 13.

[2] There is no agreement in 3GPP whether NB-LTE is a clean slate approach or not.

Implications on the Random Access Procedure

Although the ePDCCH was proposed to alleviate the shortage of control resources, this channel still has two important limitations during the random access procedure. One limitation is that the ePDCCH is configured in the UE only after the establishment of the RRC connection [7]. As a consequence, devices in idle RRC state (usually when the device performs its initial access) cannot use this channel during the random access procedure; moreover, the eNB must rely exclusively on the PDCCH to allocate the control messages to random-access-related messages. The second limitation is that the ePDCCH supports only UE-specific DCI allocations, which means it cannot be used to allocate control resources to RAR messages. Another issue is the reduction in the capacity to transmit data on the downlink as a consequence of resource sharing with the PDSCH. This can have an impact on the performance of downlink-intensive HTC users in a scenario with coexisting M2M/H2H communications.

The main difference between the ePDCCH and the MTC PDCCH of UE Cat. M1 is that the latter also supports CSS allocation, thus allowing the base station to allocate resources to RAR messages in the MTC PDCCH. The enhanced coverage feature in 3GPP Release 13 increases access delay due to repetition of the transmissions.

Although NB-LTE devices perform the random access procedure as described earlier, 3GPP considered that advanced RAN overload control schemes will not be required, due to the adoption of a simple mechanism based on an access class barring (ACB) bitmap. Moreover, the delay during the random access procedure may increase significantly since only a single device can be scheduled per millisecond on average. In traditional LTE networks, however, devices can transmit not only 3 msg3/ms on average but also conventional downlink/uplink data. However, this is not expected to affect performance, since Narrowband IoT applications are generally delay-tolerant, and the NB-LTE system will not share resources with legacy LTE users. Thus, those IoT applications with a QoS requirement or throughput constraints will use Cat. M1 chipsets or above.

Based on this analysis, we now focus on the interaction between the random access procedure and the PDCCH. Therefore, the insights arising from this article can be generalized to all LTE-M technologies, including Releases 11, 12, and 13.

State of the Art in RAN Overload Control for LTE Networks

Different approaches have been proposed to counteract the RAN overload problem, most of which were proposed by 3GPP in [4]. This section briefly discusses some novel solutions that use more robust approaches, solutions that do not use a combination of 3GPP proposed schemes, but rather more innovative ones. These approaches have emerged as a consequence of the limitations of the existing solutions in the LTE standards.

An approach for handling massive MTC traffic by using a dense network was proposed in [8]. Femtocells are used to decrease the access delay

Scheme	Benefits	Limitations	Challenges	Overhead
Access class barring	Different access probability values can be configured to deal with different PRACH loads.	Low flexibility to provide device differentiation	Determination of the barring probability based on the PRACH load	Low computational processing and low number of message exchanges
EAB	Access differentiation can be provided with fine granularity.	Access classes are either completely barred or unbarred and only delay-tolerant devices are supported.	Determination of the PRACH load and selection of the barred and unbarred classes	Moderate computational processing and moderate number of message exchanges
SB	Collisions are solved faster and it is backward compatible with traditional backoff scheme.	Inefficient under high PRACH load conditions	Definition of the scheme settings based on the device requirements	Very low computational processing and very low number of message exchanges
RRS	HTC is not affected by MTC.	Inefficient when there is unbalanced load between MTC and HTC	Dynamic allocation of RA resources for each device type	Very low computational processing and very low message exchanges
Slotted access	Dedicated RA slots for individual devices or group of devices	The number of unique RA slots is proportional to the RA cycle length The PRACH is overloaded when the number of devices is greater than the total number of unique RA slots.	Effective allocation of RA slots to devices/groups	Low computational processing and messages exchange
Pull-based	Overload on PRACH can be effectively mitigated.	Unexpected surge of access requests cannot be handled.	Mechanism to decrease the load in the paging channel	High number of message exchanges in the core network and paging channel
Distributed queuing	Infinite number of simultaneous devices can ideally be handled.	Only delay-tolerant devices are supported.	Access delay increases as the number of devices increases.	High computational processing and high number of message exchanges
Femto-cell-based	Low energy consumption; support for high number of simultaneous devices	Lack of access differentiation; low outdoor coverage	Differentiation of users' attempts	Huge cost of the approach for MNOs

Table 1. Summary of the RAN overload control schemes proposed by 3GPP.

as well as energy consumption. However, this is not a cost-effective solution and is not practical in real IoT scenarios.

Another solution, proposed in [3], is distributed queuing (DQ)-based. It supports an infinite number of contending devices over PRACH. It has clear advantages over the conventional RA procedure; delays and energy consumption are reduced more than they are in the 3GPP RAN overload control schemes. Nevertheless, the access delay is greater than that required by delay-sensitive IoT applications.

In [9], two methods for the management of critical and emergency alarm messages in LTE networks are proposed. They require dedicated preambles for alarm devices as well as specific modifications of both the eNB and UE. Each message is mapped on either a predefined index or a sequence of preambles, depending on the method used. Although such methods can significantly reduce the time required for notification of an alarm to the eNB, they require excessive PRACH resources (i.e., RAO) every millisecond as well as the reservation of a set of preambles for use only by these applications. Table 1 shows a summary of the RAN overload control schemes [4].

Performance Evaluation

To evaluate the performance of different RAN overload control schemes, a special module was developed for the LTE-Sim simulator, which includes detailed implementation of the random access procedure, described next. The collision of preambles can only be detected when a UE does not receive the msg4 message within the waiting time window. Thus, the UEs can send msg3 messages in the same PUSCH resources, even though this leads to collisions. In addition, whenever an msg3 message is retransmitted, the contention resolution timer is restarted. The PDCCH CSS and DSS mechanisms were also implemented. This inclusion reduces the region in which the eNB can allocate control information to each UE as well as increasing blocking on the PDCCH, when two or more UEs use the same region. The processing latency for each step of the RA procedure was introduced, following the specifications in [10].

We validated this new module by comparing its output with the metric values given by the 3GPP TR 37.868 MTC simulation model [4]. To provide a fair comparison, however, it was necessary to assume that the eNB does not decode simultaneous transmission of the same preamble, and, therefore does not send the uplink grant for those preambles as assumed in [4]. This comparison is displayed in Table 2, on the columns "LTE-Sim Module" and "3GPP TR 37.868," respectively. The last column of Table 2 shows the impact of the inclusion of the above-mentioned realistic assumptions in the enhanced simulation model. Simulations considered scenarios with 5000, 10,000, and 30,000 UEs, with the number of connection requests over a period of 10 s following a *Beta*(3, 4) distribution as proposed in [4].

Tables 3 and 4 show the configuration param-

Metric	3GPP TR 37.868			LTE-Sim module			Enhanced LTE-Sim module		
Number of devices per cell	5000	10,000	30,000	5000	10,000	30,000	5000	10,000	30,000
Access success probability	100%	100%	29.5%	100%	100%	29.6%	100%	87.95%	14.93%
Average access delay (ms)	29.06	34.65	76.81	29.67	35.95	80.43	46.05	108.59	156.12
10th percentile access delay (ms)	15	15.25	15.89	15.02	15.39	16.52	17.19	20.97	19.83
90th percentile access delay (ms)	51.61	65.71	174.39	52.80	66.56	176.64	98.13	247.04	336.72
Number of preamble transmission	1.56	1.77	3.49	1.62	1.83	3.56	1.66	2.92	3.86
Preamble collision probability	0.45%	1.98%	47.76%	0.46%	1.96%	47.70%	0.44%	7.30%	53.21%
msg2 blocking probability	–	–	–	0%	0.73%	4.52%	0.03%	31.60%	63.66%

Table 2. Validation of our simulation model and the impact of realistic considerations on the performance of a traditional random access scheme.

eters used in the simulations. The following metrics were considered in the analysis: access probability, defined as the ratio between a fully complete RA and the total number of RAs triggered; average delay, defined as the time elapsed from the transmission of the first msg1 message to the reception of an msg4 message, considering only successful accesses; preamble collision ratio, which is the ratio between the number of events when two or more devices send the same msg1 message (collision) and the overall number of msg1 messages available during the period; CCE utilization, which is the ratio between the number of CCEs used on the PDCCH and the overall number of available CCEs in the PDCCH; and msg2 blocking probability, which is the ratio between the number of dropped msg2 messages to send msg3 and the number of this type of msg2 messages that joined the eNB queue.

Table 2 shows that the results for the enhanced LTE-Sim module differ from those of the other models due to consideration of realistic assumptions in both the detection of preamble sequence collisions and the allocation of control resources, with the access success probability decreasing while the preamble collision probability and the access delay increase. One of the reasons for the increase in the average access delay is the consideration that preamble collisions will only be detected when the msg4 message is not received. This increases the number of detected preambles, thus increasing the msg2 blocking probability. The dropping of the msg2 messages due to timeout of the timer is the main factor for the decrease in the access probability. This blocking occurs mainly when various UEs are trying to access the network at the same time (or in a short period) and the eNB does not have enough resources during the time window to send the msg2 messages. Another reason for the increased delay is the delay in the downlink grant for msg4 message on the PDCCH. This happens when the PDCCH resources for the allocation of msg4 message destined to a UE are already allocated to other UE msg4 messages, resulting in the postponement of the transmission of the msg4 message despite the existence of available resources on the PDCCH.

Under light loads, all schemes achieved 100 percent access, except some losses when using the Fixed-EAB (F-EAB) scheme (Fig. 2a). Moreover, the F-EAB scheme imposed the greatest average access delay, some 2.8 times greater than those of the second slowest scheme (the fixed-ACB [F-ACB]), and 46 times greater than the smallest delay imposed by the LTE scheme (Fig. 2b). No scheme produced a preamble collision probability greater than 1 percent (Fig. 2c), which suggest that few attempts using the same preamble were sent. However, the CCE utilization exceeded 20 percent in 6 schemes, the F-ACB scheme being the one with lowest utilization, only 11 percent (Fig. 2d). Although the operation of both ACB schemes is quite similar, the fixed approach imposes a greater average access delay than does the adaptive one. Since the barring probability varies as a function of the network load, and few preamble collisions occur on the network, it is possible to conclude that the variation in blocking probability is of little relevance, remaining most of the time in 0, thus allowing the preamble transmissions. The F-EAB scheme produces high msg2 blocking probability values as a consequence of numerous simultaneous access attempts (Fig. 2e). The other schemes spread access attempts along the timeline, decreasing the intensity of access attempts.

Under medium loads, the F-ACB, adaptive-ACB (A-ACB), adaptive-EAB (A-EAB), and specific backoff (SB) schemes also achieved an access ratio of almost 100 percent, despite a decrease in this ratio for certain other schemes (Fig. 2a). Such a high access probability is due to the fact that these schemes spread in time the attempts to transmit the preambles, thus avoiding collisions. However, such a high access success ratio leads to a considerable increase in delay, which reaches as high as 35 times (Fig. 2b). The access ratio of F-EAB decreased to 68 percent due to the period required for altering the unbarred class. The longer the period, the greater is the number of devices waiting to attempt access after the change of unbarred class. This procedure degrades the performance when compared to the performance of the LTE scheme. Such an increase is due to the large number of devices trying to use the same preamble in an attempt to access the network. Moreover, the CCE utilization of all schemes increased, reaching 50 percent for the conventional scheme, which shows that more msg4 messages were transmitted (Fig. 2d). The msg2

LTE	
Backoff period	20 ms
F-ACB	
Barring factor	0.9
Barring time	4 s
A-ACB	
Barring factor	Adaptive
Barring time	4 s
Monitoring period	500 ms
Update period	500 ms
F-EAB	
Round period	500 ms
ON/OFF	Always ON
A-EAB	
Round period	500 ms
ON/OFF	Adaptive
Update period	500 ms
RRS	
# preambles to HTC	22
# preambles to MTC	30
SB	
Backoff period HTC	20 ms
Backoff period MTC	960 ms

Table 3. RAN overload control schemes configuration.

blocking probability of all schemes increased, but the ACB schemes produced the lowest probability value, only 0.5 percent (Fig. 2e). However, this low value of the F-ACB scheme is achieved at the expense of high access delay (Fig. 2b). Conversely, the A-ACB scheme obtained msg2 blocking probability as low as 2.95 and 1.8 percent, respectively, while providing low access delays (Fig. 2b).

Under heavy loads, the access ratio decreases dramatically for some schemes. For the adaptive ACB scheme, the access ratio decreased 25 percent, while for the RACH resource separation (RRS) scheme, it decreased more than 90 percent (Fig. 2a). The separation of preambles according to the type of device (MTC or HTC) implies a reduction in the number of preambles for MTC devices and a consequent increase in the preamble collision probability (Fig. 2c). The preamble collision probability of the conventional scheme is 45 percent, while the preamble collision probability of the RRS scheme is 60 percent. As a consequence of the variation of barring probability, the preamble collision probability of the adaptive ACB scheme is only 15 percent. Despite the F-ACB preamble collision probability of only 13 percent, the access probability achieves only 44 percent, and the CCE utilization is equal to 29 percent (Figs. 2b and 2d). Moreover, there is a slight increase in the CCE utilization over what is found for the scenario with 10,000 UEs for the F-ACB scheme, which suggests saturation of the networks. All the schemes produced an msg2 blocking probability lower than 40 percent, showing limitation of the capacity for sending RAR. For the ACB and EAB schemes, the fixed approach dropped more msg2 messages than did the adaptive one (Fig. 2e).

Parameter	Value
System bandwidth	5 MHz
Frame structure	FDD
PRACH configuration Index	6
Max. preamble retransmissions	10
Number of msg2 per subframe	3
Total preamble sequences	52
RAR window size	5 ms
Contention resolution timer	48 ms
Max. msg3 retransmissions	5
Number of CCEs	16

Table 4. Simulation parameters.

In summary, the performance of the RA procedure depends on the number of competing UEs and their generated traffic. For networks with a small number of UEs, the LTE scheme is more appropriate, since it provides a good trade-off between access ratio and delay. For instance, 100 percent of access is possible with a delay of only 47 ms. For networks with a large number of UEs generating delay-tolerant traffic, the recommendation is the employment of the adaptive ACB scheme, since it produces the greatest access ratio, although this access is delayed.

Challenges and Research Directions

Although progress has been made in reducing the impact of the RAN overload problem, several challenges remain to be overcome, especially those related to the support of delay-sensitive IoT applications. This section discusses three key challenges originating from the need to provide some kind of guarantee to limit delays during the random access procedure as these can reach several seconds under heavy load conditions and control resources' influence on the performance.

QoS-Aware RAN Overload Control

Existing RAN overload control mechanisms do not provide QoS guarantees for delay-sensitive IoT applications. It is necessary to investigate new ways to reduce the random access delay as well as increase the chances of access of IoT devices [11]. For example, the adaptation of the distributed queuing approach for QoS provisioning has great potential for handling a very large number of devices but reducing the access delay. The

Existing RAN overload control mechanisms do not provide QoS guarantees for delay-sensitive IoT applications. It is necessary to investigate new ways to reduce the random-access delay as well as increasing the chances of access of IoT devices.

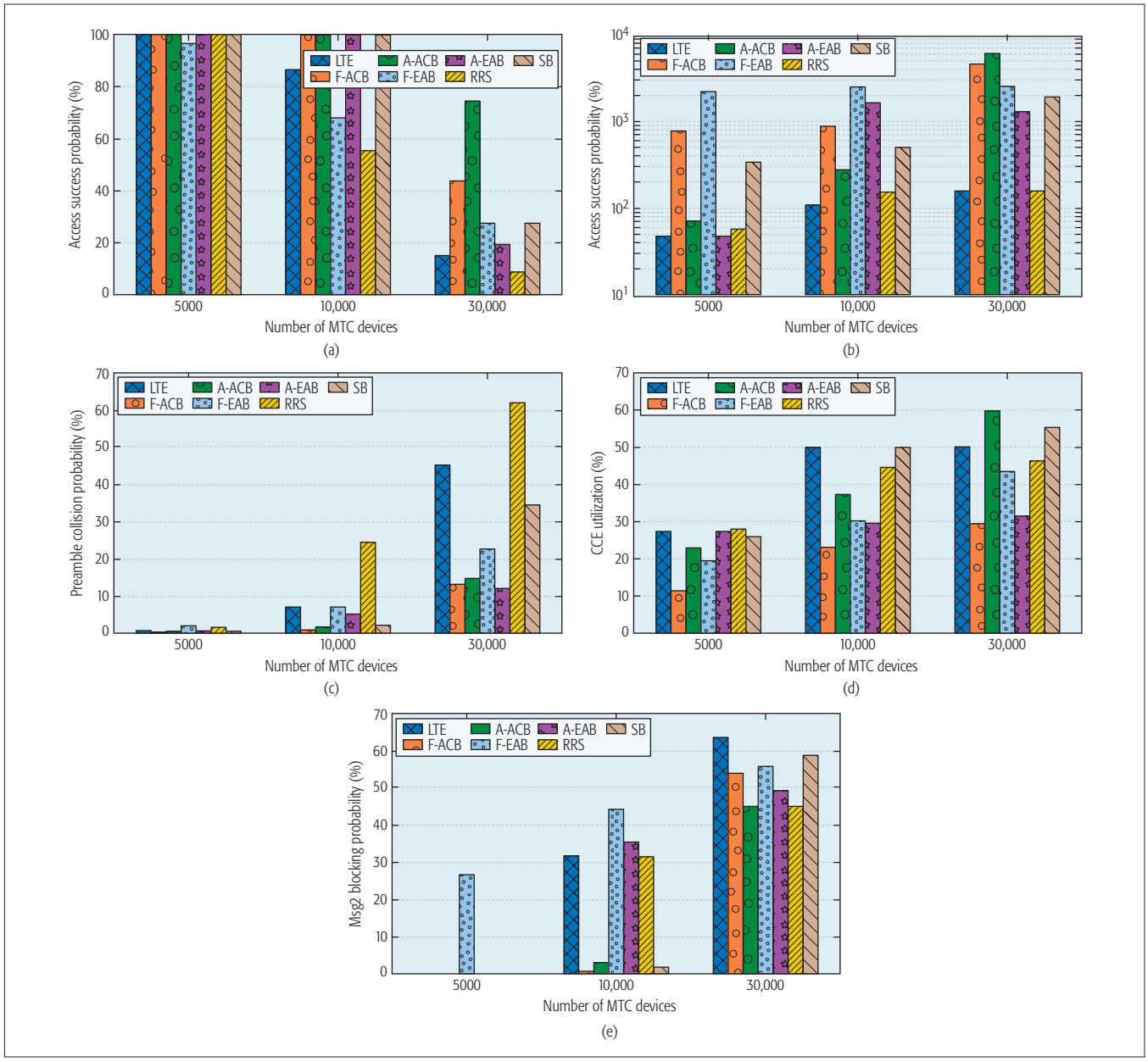

Figure 2. Performance of the 3GPP RAN overload control schemes vs. the number of MTC devices: a) access success probability; b) average access delay; c) preamble collision probability; d) CCE utilization; e) msg2 blocking probability.

challenge is finding a way to differentiate access on the basis of class. The state-of-the-art methods differentiate classes by means of preamble sequence reservation (e.g., the RRS scheme and the two methods proposed in [9]) or configuration of random access parameters (e.g., the SB scheme and the ACB mechanism in [12]), but these methods are not scalable and affect performance. An interesting option is to prioritize preamble transmissions by means of their transmit power level [13]. Another option that could be combined with various existing schemes would be the use of the QoS class identifier (QCI) available in LTE technology rather than the device type to provide greater flexibility to the random access procedure [14].

QoS-Aware PDCCH Resource Allocation

A QoS-aware PDCCH scheduler can also improve performance during random access. QoS awareness in the allocation of control resources can further improve the performance of a network during periods of access attempt by a massive number of users [15]. In fact, the PDCCH scheduling algorithm can have great impact on the network performance in MTC/HTC coexisting scenarios [15]. Moreover, 3GPP does not standardize any PDCCH scheduling algorithm, but rather leaves this option to the vendor to implement its own solutions. Thus, PDCCH schedulers can make a real difference in the IoT market. PDCCH schedulers typically give high priority to msg2 and msg4 messages regardless of the QoS required by a device with control resources shared by downlink assignments, uplink grants, and msg2 and msg4 messages. Thus, PDCCH policies taking QoS requirements of all control messages into consideration can improve network performance and maximize resource utilization.

Random-Access-Aware Packet Scheduling

With large delays in network access, the performance of delay-sensitive IoT applications is degraded. The packets generated by delay-sensitive applications do not receive adequate service differentiation. Even though various M2M schedulers reserve certain physical resource blocks (PRBs) for MTC devices and implement some sort of prioritization for them, existing schedulers do not take into consideration the time consumed for access of the channel. Typically, LTE schedulers estimate the delay of device packets on the basis of arrival time of the request at the base station, thus ignoring delays in random access in the production of schedules. However, this can lead to less urgent packets receiving grants unless random access awareness is considered.

Concluding Remarks

The RAN became the bottleneck of an LTE system when a very large number of MTC devices transmit within a short time interval. In this article, the performance of the LTE random access procedure for IoT connectivity has been analyzed. The impact of LTE enhancements on the random access procedure for MTC is highlighted. RAN overload control schemes standardized by 3GPP and novel approaches for supporting massive access to IoT devices over LTE networks have been reviewed, and the performance of those schemes standardized by 3GPP under realistic assumption in both the PRACH and control resource allocation are assessed. Extensive simulation results indicate that the RAN overload problem has been underestimated due to use of unrealistic assumptions in previous work. Based on these observations, directions for future research to ameliorate the RAN overload have been presented.

Acknowledgments

This work was supported in part by the CNPq Brazilian research agency and Motorola Mobility.

References

[1] A. Al-Fuqaha *et al.*, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, 4th qtr. 2015, pp. 2347–76.
[2] S. Andreev *et al.*, "Understanding the IoT Connectivity Landscape: A Contemporary M2M Radio Technology Roadmap," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 32–40.
[3] A. Laya *et al.*, "Goodbye, ALOHA!," *IEEE Access*, vol. 4, 2016, pp. 2029–44.
[4] 3GPP, "Technical Specification Group Radio Access Network; Study on RAN Improvements for Machine-type Communications," TR 37.868, Sept. 2011.
[5] S. Ye, S. H. Wong, and C. Worrall, "Enhanced Physical Downlink Control Channel in LTE Advanced Release 11," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 82–89.
[6] 3GPP, "Technical Specification Group GSM/Edge Radio Access Network; Cellular System Support for Ultra-Low Complexity and Low Throughput Internet of Things (CIoT)," TR 45.820, Nov. 2015.
[7] T. Tirronen *et al.*, "Telecommunications Apparatus and Method Relating to a Random Access Procedure," Dec. 11, 2014, wO Patent App. PCT/SE2013/050,647; http://www.google.com/patents/WO2014196908A1?cl=en.
[8] M. Condoluci *et al.*, "Toward 5G Densenets: Architectural Advances for Effective Machine-Type Communications Over Femtocells," *IEEE Commun. Mag.*, vol. 53, no. 1, Jan. 2015, pp. 134–41.
[9] M. Condoluci *et al.*, "Enhanced Radio Access and Data Transmission Procedures Facilitating Industry-Compliant Machine-Type Communications over LTE-Based 5G Networks," *IEEE Wireless Commun.*, vol. 23, no. 1, Feb. 2016, pp. 56–63.
[10] 3GPP, "LTE; Feasibility Study for Further Advancements for E-UTRA," TS 36.912, Jan. 2016.
[11] O. Arouk, A. Ksentini, and T. Taleb, "Group Paging-Based Energy Saving for Massive MTC Accesses in LTE and Beyond Networks," *IEEE JSAC*, vol. 34, no. 5, May 2016, pp. 1086–1102.
[12] J. S. Vardakas *et al.*, "Performance Analysis of M2M Communication Networks for QoS-Differentiated Smart Grid Applications," *2015 IEEE GLOBECOM Wksps.*, Dec. 2015, pp. 1–6.
[13] T. Kim, K. S. Ko, and D. K. Sung, "Prioritized Random Access for Machine-to-Machine Communications in OFDMA Based Systems," *IEEE ICC 2015*, June 2015, pp. 2967–72.
[14] T. P. C. de Andrade, C. A. Astudillo, and N. L. S. d. Fonseca, "Random Access Mechanism for RAN Overload Control in LTE/LTE-A Networks," *IEEE ICC 2015*, June 2015, pp. 5979–84.
[15] T. P. C. de Andrade, C. A. Astudillo, and N. L. S. da Fonseca, "Allocation of Control Resources for Machine-to-Machine and Human-to-Human Communications over LTE/LTE-A Networks," *IEEE Internet of Things J.*, vol. 3, no. 3, June 2016, pp. 366–77.

Biographies

Tiago P. C. de Andrade received his MSc. and BSc. degrees in computer science from the State University of Campinas (UNICAMP), Brazil, in 2013 and 2009, respectively. Currently, he is a Ph.D. student in computer science at the Institute of Computing, UNICAMP. His current research interests are in machine-to-machine communications, device-to-device communication, quality of service, and energy efficiency mechanisms for 4G/5G cellular networks.

Carlos A. Astudillo received his B.Sc. degree in electronics and telecommunications engineering from the University of Cauca (UNICAUCA), Popayán, Colombia, in 2009 and his M.Sc. degree in computer science from UNICAMP in 2015, and is currently working toward his Ph.D. degree at the Institute of Computing, UNICAMP. In 2010, he was a Young Researcher with the New Technologies in Telecommunications R&D Group (GNTT), UNICAUCA, supported by the Colombian Administrative Department of Science, Technology and Innovation. His current research interests include quality of service and energy-efficient mechanisms in machine-to-machine communications and mobile backhauling for 4G/5G cellular networks.

Luiz R. Sekijima is an undergraduate student in computer engineering at the Institute of Computing, UNICAMP. His current research interest is in energy-efficient mechanisms for 4G cellular networks and the Internet of Things.

Nelson L. S. da Fonseca received his Ph.D. degree from the University of Southern California in 1994. He is a full professor at Institute of Computing, UNICAMP. He is the IEEE ComSoc Vice-President of Publications. He has served as Vice-President of Member Relations, IEEE ComSoc Director of Conference Development, Director of Latin America Region, and Director of Online Services. He is past Editor-in-Chief of *IEEE Communications Surveys & Tutorials*. He is a Senior Editor of *IEEE Communications Magazine*.

An interesting option is to prioritize preamble transmissions by means of their transmit power level. Another option that could be combined with various existing schemes would be the use of the QoS class identifier available in LTE technology rather than the device type to provide greater flexibility to the random access procedure.