# Enabling the IoT Machine Age with 5G: Machine-Type Multicast Services for Innovative Real-Time Applications

Massimo Condoluci, *Member, IEEE,* Giuseppe Araniti, *Senior Member, IEEE,* Toktam Mahmoodi, *Member, IEEE,* and Mischa Dohler, *Fellow, IEEE*

***Abstract*—The Internet of Things (IoT) will shortly be undergoing a major transformation from a sensor-driven paradigm to one that is heavily complemented by actuators, drones and robots. The real-time situational awareness of such active systems requires sensed data to be transmitted in the uplink to edge-cloud, processed and control instructions transmitted in the downlink. Since many of these applications will be mission critical, the most suitable connectivity family will be cellular due to the availability of licensed spectrum able to protect the offered communications service. However, while much focus in the past was on the uplink of machine-type communications (MTC), little attention has been paid to the end-to-end reliability, latency and energy consumption comprising both up and downlinks. To address this gap, in this paper we focus on the definition, design and analysis of machine-type multicast service (MtMS). We discuss the different procedures that need to be re-designed for MtMS and we derive the most appropriate design drivers by analyzing different performance indicators such as scalability, reliability, latency and energy consumption. We also discuss the open issues to be considered in future research aimed at enhancing the capabilities of MtMS to support a wide variety of 5G IoT use cases.**

***Index Terms*—IoT, 5G, MTC, E2E, Multicast, MtMS, LTE-M.**

## I. Introduction

The Internet of Things (IoT) [1], [2] is predicted to interconnect billions of devices over the next decades to come [3]. The resulting spatial and temporal data granularity is expected to yield significant business as well as consumer benefits that will be at par with today's Internet [4].

To date, most IoT applications pertain to some form of sensing. For instance, smart city IoT applications [5] would measure pollution, or the amount of cars in the streets, etc. That requires a specific sensor to be connected to a radio and transmit the information in the uplink, either regularly or when the event occurs. However, the IoT landscape is expected to change significantly with more "things" becoming active elements [1]. In the future, we will have sensors complemented by actuators, drones and other form of robots. In the context of smart cities, for instance, pollution and traffic sensors would gather real-time information about the city's traffic at high spatial resolution, which is then big data processed, thereupon actuators change traffic lights in order to minimize pollution and congestion.

Expanding the IoT into above mentioned capabilities, requires information to be transmitted with very high reliability and decisions to be taken almost in real-time. That, in turn, requires connectivity technologies able to offer service level agreements (SLAs), i.e., cellular 3GPP technologies [6] such as Long Term Evolution (LTE) [7] and beyond 5G [8] systems that are therefore the focus of this paper. Furthermore, it requires the uplink (UL) and downlink (DL) to be designed jointly with end-to-end delay minimization being a native part of the design. To date, however, mainly only the UL capabilities were studied whereas very little information is available on the downlink for the IoT.

In more details, milestone state of the art contributions for UL machine-type communications (MTC) are related to the improvement of the random access (RA) procedure [9]. Solutions such as [10], [11], [12] enhanced the RA by designing strategies able to guarantee low delays in the UL direction. From a downlink point of view, the contributions in literature mainly focused on the improvement of the paging procedure [13] to simultaneously send control messages toward a huge number of devices [14], [15]. Solutions to cut the delay in DL direction have not been properly investigated. In addition, strategies that efficiently support group-oriented (e.g., mulitcast) MTC traffic are still needed to be designed in order to cut delays and allow scalability when the number of receivers is huge. The core network introduces delays to either UL and DL in terms of lower delay in the data plane and higher delay in the control plane; this is due to the overhead in the control plane that is no longer negligible when considering MTC data traffic usually characterized by small packets.

In the light of above, our novel contribution is to focus on end-to-end design and analysis of MTC, including the UL access and DL machine-type multicast service (MtMS). We present the architecture, functionalities and different procedures that need to be re-designed to enable the efficient transmission of multicast data toward a large set of MTC devices. We analyse different performance indicators such as scalability, reliability, latency and energy consumption in order to evaluate the effectiveness of the design solutions defined for our proposed MtMS.

The paper is structured as follows. Section II analyses all the

M. Condoluci, T. Mahmoodi and M. Dohler are with the Department of Informatics, King's College London, UK. E-mail: massimo.condoluci@kcl.ac.uk, toktam.mahmoodi@kcl.ac.uk, mischa.dohler@kcl.ac.uk.

G. Araniti is with the Department of Information Engineering, Infrastructure and Sustainable Energy, University Mediterranea of Reggio Calabria, Italy. E-mail: araniti@unirc.it.

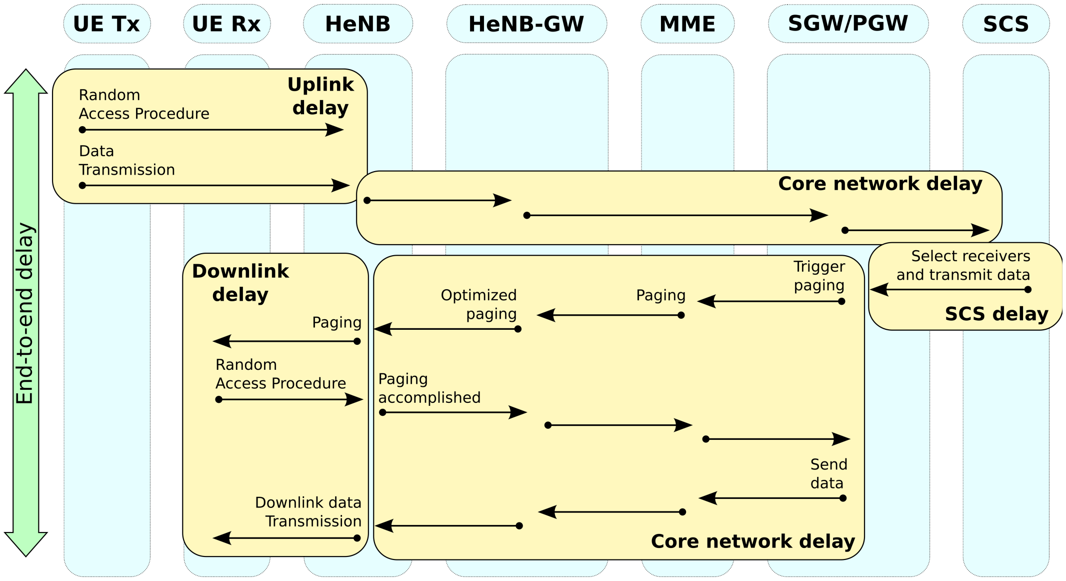

Fig. 1. The different components of the end-to-end delay in MTC environments.

different sources of delay in the end-to-end MTC traffic and summarizes the limitations of legacy group-oriented services when dealing with the unique features of MTC traffic. In Section III, we present the MtMS by describing the architectural components and the procedures to be adopted to efficiently support MTC multicast traffic. In Section IV, we provide the system model exploited to evaluate the performance of MtMS. The achieved results are shown in Section V. Section VI summarizes the results of our proposed MtMS and discusses possible research directions to further optimize MtMS and the end-to-end delay of MTC traffic. Section VII provides the conclusive remarks of the paper.

## II. The end-to-end delay in MTC environments

The end-to-end path in MTC environments is depicted in Fig. 1. We can denote the presence of three main segments: UL, core network, DL. The key features of each segment are discussed in the remainder of this Section, where we also survey the solutions in literature to improve each segment of the end-to-end path.

### *A. Delay in the uplink direction*

The UL is exploited by MTC devices to send data to remote servers. The UL data transmission is accomplished through the random access (RA) procedure, which is performed when the user equipment (UE) is not synchronized with the network (i.e., idle state) [16]. This is due to the exploitation of the discontinuous reception (DRX) [17] mode, where devices turn off the radio interface to save energy and wake up periodically to send data or to check for incoming traffic.

The RA procedure defined by 3GPP is an ALOHA-based radio access. Every RA slot, UL radio resources (Physical Random Access Channel, PRACH) are reserved to initiate the RA procedure by means of the transmission of an orthogonal preamble (Msg1). In case two devices send the same preamble in the same RA slot, then a collision occurs. At the reception of Msg1, the base station sends the Random Access Response (RAR, a.k.a. Msg2), which contains information about the detected preamble, uplink timing alignment, and the grant for the transmission of the Msg3 on the Physical Uplink Shared Channel (PUSCH). Finally, the Msg4 terminates the RA procedure and confirms the grant for the subsequent data transmission on the PUSCH.

The UL direction represents the most studied aspect of MTC traffic. Studies in literature investigate the issues of the 3GPP RA, mainly related to the limited number of preambles (i.e., 54) available to perform the transmission of Msg1 [9], [18]; this aspect strongly limits the scalability of the RA and thus introduces delay as devices that experience collisions on the preamble transmission need to re-schedule a novel RA attempt.

To overcome the limitations of the legacy 3GPP RA, the access class barring (ACB) [19] has introduced the idea of exploiting a backoff mechanism before the transmission of Msg1 to avoid a high number of collisions in case of network overload. The backoff value is obtained by considering some parameters (in particular, a probability factor and the barring timer relevant to the pre-defined ACB classes) broadcasted by the base station. Similarly, the extended access barring (EAB) [14] has introduced backoff mechanisms for delay-tolerant services in order to guarantee the availability of a higher number of preambles for delay-constrained devices. Although ACB and EAB approaches may guarantee short access delays to high-priority devices, this is paid with the introduction of higher delays for other devices. In order to cut delays, other solutions have proposed to send data directly in the Msg3 of the RA procedure [20] or without performing the RA procedure [21]. Nevertheless, these approaches lack in terms of flexibility as they require fixed size and modulation and coding schemes for the transmitted packets.

A novel approach for the RA is represented by the code expanded [11], which is based on the idea of sending an access code-word composed of several preambles instead of sending only one Msg1. To this end, several RA slots are grouped to form one RA frame. This approach allows to drastically increase the number of contention resources and, therefore, is able to potentially support huge MTC load; furthermore, it could allow to cut the delay by avoiding the need of RA re-attempts. The code-expanded can also be exploited to enable the transmission of low-latency messages (e.g., alarms). In [12], the reception of mission critical messages at the base station is associated with the reception of some predefined access code-words. This approach is able to avoid the transmission of Msg3 and Msg4 and thus cuts the delay on the UL direction.

### B. Delay in the core network

The core network introduces delays due to the procedures performed to manage control and data traffic. The LTE core network has a significantly lower latency compared to earlier 3GPP releases [7], thanks to the use of a flat architecture composed of one entity in the control plane (i.e., the mobility management entity, MME) and one entity in the data plane (i.e., the serving gateway, SGW). However, in either of downlink and uplink directions, the core network introduces additional delays to the end-to-end communication. These delays will be different for the data plane and the control plane. On the data plane, the round-trip data delay is measured by the time it takes for a small packet to travel from an IoT node to the service capability server (SCS, which gathers data from the sensors and then sends commands to the actuators) and back. On the control plane, the latency is measured as the time required for an IoT node to transit from idle state to active state in order to send or receive data traffic.

Taking multiple contributing factors such as the scheduler, frame size, retransmission delay and waiting time for the next transmission frame, round-trip delay in the LTE network can add up to 10-20 ms, while the additional element by the core network in this is minimal and in the order of 1 ms [22]. However, the round-trip delay is reported ten to hundred fold higher based on the measurements [23]. Solutions to activate the UEs for UL transmission without (or minimal) intervention of the core network in order to cut overhead and delay are currently under investigation [16].

In the direction of cutting overhead and delay in the mobile core, softwarization and virtualization paradigms are gaining importance in the design of future 5G system architecture [24], [25], [26], [27]. These paradigms allow flexibility of network functions deployment by decoupling network functionalities from the underlying hardware. This could allow to properly configure data ad control planes in order to achieve control overhead (for instance, by avoiding mobility management mechanisms in scenarios with fixed MTC device) and delay (e.g., by moving functionalities close to the edge to avoid to contact entities in the core network) reductions when handling MTC traffic. Nevertheless, further studies are still required to exploit softwarization and virtualization in the mobile core.

### C. Delay in the downlink direction

The DL data transmission over mobile cellular networks is performed by means of the paging procedure [13], triggered by the network to inform MTC devices about incoming data traffic. By considering that MTC UEs are usually in idle state, after the reception of the paging message the devices perform the RA procedure to acquire synchronization with the network in order to be scheduled for the reception of data traffic. As a consequence, the RA influences both DL and UL segments in the end-to-end communication.

The main issue relevant to the paging procedure defined by 3GPP is the capacity: only 16 devices can be paged in each paging occasion. Furthermore, only two paging occasions are available in a radio frame of 10ms. In addition to the scalability, the paging introduces a high overhead (in terms of amount of control messages) when the number of devices to be paged is huge. To overcome above issues, the group-paging [14], [15] has been proposed to simultaneously page a group of devices. Instead of transmitting one paging message for each UE, the group-paging identifies multiple devices with a group identity (GID) and thus performs the paging on a per-group basis. This drastically reduces the overhead. It is worth noticing that the group-paging strategy has been studied only when coupled with the legacy 3GPP RA procedure, which is not able to handle a high number of simultaneous accesses. As a consequence, the drawback of the group-paging is the high collision probability when the devices perform the RA procedure. The enhancements in literature focused on the introduction of back-off mechanisms in order to scatter the RA attempts of devices [28], [29] by extending the ACB/EAB approaches designed for the UL direction [19]. These approaches are thus able to reduce the collision probability during the RA phase, as they reduce the number of devices contending in the same RA frame, at the expense of a delay increase.

In the analysis of DL segment, it is worth considering that several IoT scenarios (such as smart cities, smart homes, industrial plants, intelligent transportation systems, etc.) could benefit from the exploitation of group-oriented (e.g., multicast) services [1]. This scenario poses additional issues, as it deals with data transmission toward a group of devices. The Multimedia Broadcast Multicast Services (MBMS) [30] represents the current standard to support group-oriented services (e.g., mobile TV, video streaming, multimedia content download) over mobile networks [31]. When considering the applicability of this standard to MTC traffic, two aspects need to be considered. The former aspect is that MBMS is a session-oriented standard: the network operator is in charge of advertising a specific MBMS session (e.g., the availability of mobile TV services) provided in specific areas of its own network. Differently, MTC traffic needs to be delivered only to a specific group of devices belonging to a specific customer/tenant of the network. The second aspect to consider is that MBMS is a human-based standard as the creation of multicast groups is performed with the transmission of advertising and joining requests to all the users in a specific area: this means that the human interaction is fundamental for MBMS in order to create the MBMS group. From an MTC point of view, the receivers

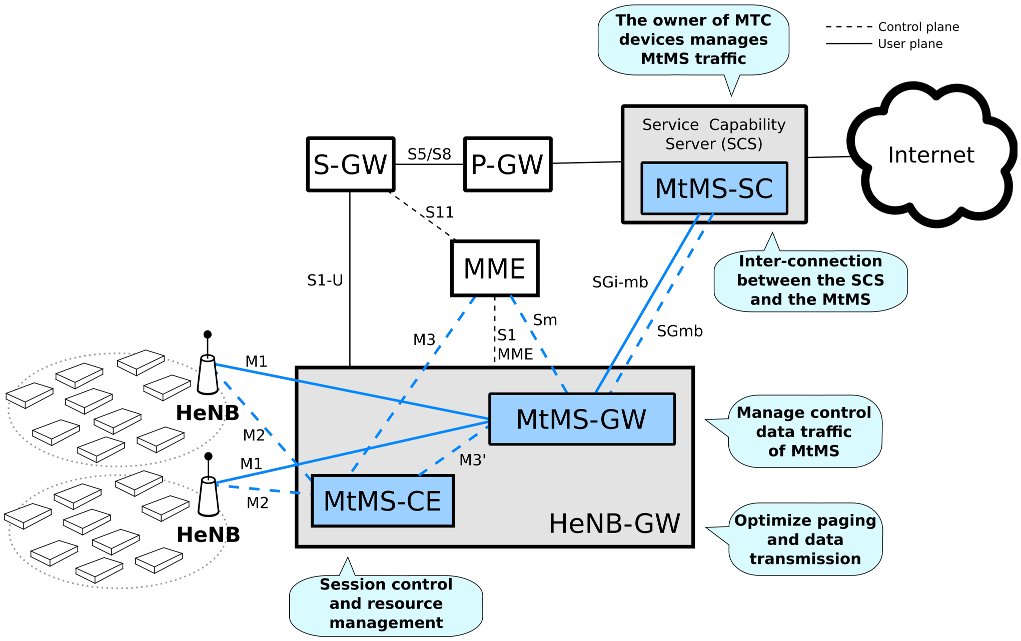
The owner of MTC devices manages MtMS traffic
Control plane
User plane
Service Capability Server (SCS)
MtMS-SC
Internet
S-GW
S5/S8
P-GW
S11
MME
S1-U
SGi-mb
Inter-connection between the SCS and the MtMS
SGmb
Sm
M3
S1 MME
M1
HeNB
M2
M1
M2
HeNB
MtMS-GW
Manage control data traffic of MtMS
M3'
MtMS-CE
HeNB-GW
Optimize paging and data transmission
Session control and resource management

Fig. 2. MTC architecture enhanced by machine-type multicast service (MtMS).

of the multicast content are already defined (e.g., a specific set of actuators that need to perform a specific task) and this pushes the need of re-thinking announcement and joining procedures for multicast MTC traffic. Finally, an additional aspect to be considered is related to the control overhead. As MBMS usually delivers multimedia content characterized by a huge amount of data and a long session duration, the control traffic needed to manage the MBMS session is considerably low compared to the amount of data traffic. On the contrary, MTC traffic is usually composed of few bytes that could need to be delivered under strict delay requirements. Consequently, the control traffic needs to be re-designed in order to reduce the overhead and thus to cut delays and energy consumption of MTC devices.

In the following Section, we present our solution to support MTC multicast traffic.

## III. Machine-type multicast service (MtMS)

In this paper we design the machine-type multicast service (MtMS), which aims to define the architecture and the related control and transmission procedures to efficiently handle MTC multicast traffic. MtMS takes advantage of the observation that *small-cells* [32], [33], [34] are expected to provide meaningful benefits (such as latency and energy consumption reductions and improved coverage and reliability) for MTC traffic compared to the use of traditional macro-cells [10], [12], [35]. As shown in Fig. 2, MtMS traffic is provided through home-evolved NodeBs (HeNBs), i.e., femto-cells, which are connected to the core network through the HeNB gateway (i.e., HeNB-GW). The role of this entity is to aggregate control and data traffic for the sake of overload reduction toward the core network [10].

The different phases of the MtMS session are depicted in Fig. 3. The MtMS session is initiated by the *MtMS serving center (MtMS-SC)*, which is the source of the MtMS content and is implemented at the SCS. The reason behind this choice is that the SCS is the anchor point of MTC devices with the mobile core, i.e., the SCS receives/sends data from/to MTC devices [36]. In case of an end-to-end communication between sensors and actuators, the SCS triggers a multicast session request to the MtMS-SC. The SCS provides the MtMS-SC with the parameters of the MtMS session, i.e., the set of devices to be served and the data content to be delivered. These parameters are properly selected according to the type, value, and the transmitter identity of the data received in the UL direction. The implementation of the MtMS-SC at the SCS thus allows to realize the shift from a network-based to a customer-based group initialization, where the owner of the MTC devices is in charge of selecting the proper group of machines to be served. This means that the service announcement procedure, which involves interactions with the users in the legacy MBMS sessions, can be now avoided. This design choice cuts delays, energy consumption, and overhead from an MTC point of view, with consequent benefits in terms of resource utilization from a network point of view.

The MtMS-SC initiates the MtMS session by forwarding the multicast content (through the SGi-mb user plane interface) and the list of devices to be served (through the SGmb control plane interface) at the *MtMS gateway (MtMS-GW)*. This entity initiates the MtMS session within the 3GPP network by triggering the joining procedure at the MME through the Sm control interface. Once the joining procedure is triggered, the MME provides the *MtMS coordination entity (MtMS-CE)* with the tracking area information relevant to the devices to be paged through the M3 control interface. The joining procedure is thus locally handled at the MtMS-CE. Once the joining procedure is accomplished, the MtMS-GW performs the data delivery by conveying data packets to involved cells through

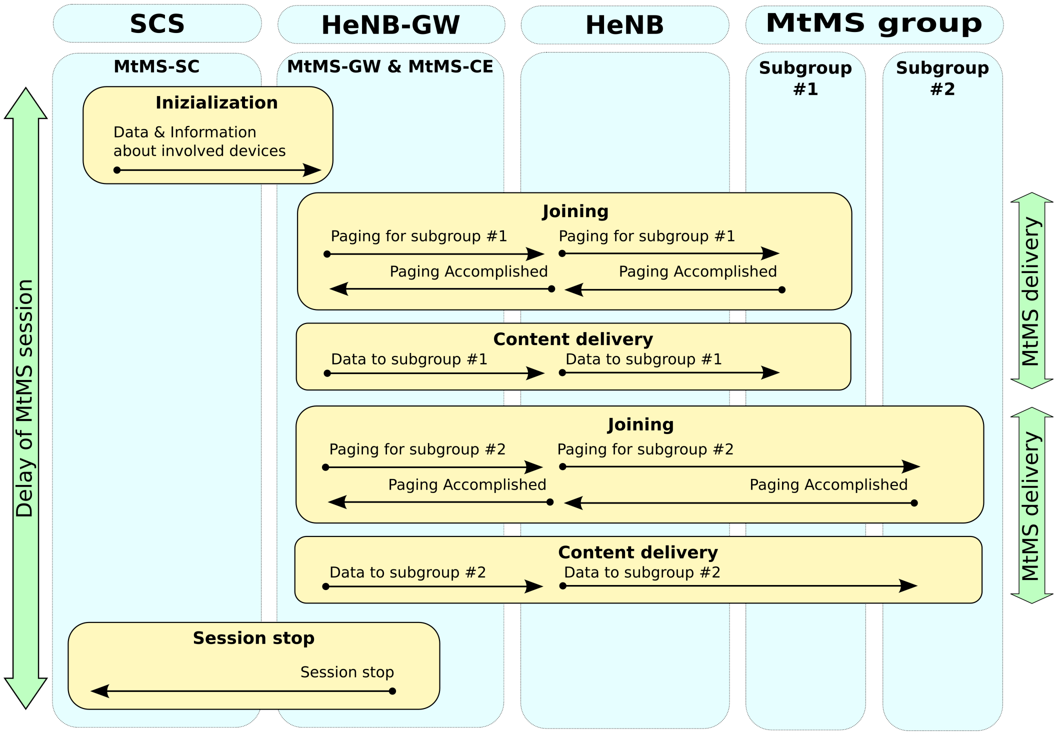

Fig. 3. Procedures of the MtMS session with the enhanced group-paging.

the M1 user plane interface. During the content delivery, the MtMS-CE controls the set of cells involved in the MtMS session by means of handling the allocation of time/frequency resources and the related transmission parameters (e.g., power, modulation and coding scheme). The information relevant to these parameters is sent by the MtMS-CE to the involved cells through the M2 control interface.

As depicted in Fig. 2, MtMS-GW and MtMS-CE are logically implemented at the HeNB-GW as this node acts as aggregator point for the femto-cells providing connectivity to the MTC devices. In case the HeNB-GW implements part of the MME functionalities (e.g., updating tracking area information of devices), the joining procedure can be handled without the intervention of the MME. To this aim, the M3' interface is designed as a logical control interface allowing a direct communication between the MtMS-GW and MtMS-CE.

The joining phase is accomplished through the paging procedure, exploited to inform MTC devices in idle state about the incoming MtMS data. As stated above, the paging consists of two phases, i.e., the transmission of the paging messages sent by the HeNB(s) and the subsequent RA procedure. The paging can be accomplished with different strategies, i.e., the legacy 3GPP procedure [13] and the group-paging scheme [14], [15], where the latter approach guarantees scalability and overhead reduction [15]. In terms of RA, to avoid the high collision probability of the legacy 3GPP RA [16], [9], the code-expanded scheme represents a viable solution to manage the large group of devices involved in the MtMS session [11].

According to the above considerations, the joint exploitation of group-paging and code-expanded RA seems a viable solution for the joining phase of MtMS. Nevertheless, it is important to take into consideration the current trends in the design of MTC devices. Indeed, the need of reducing the cost of IoT devices [37] pushes the idea of exploiting deployment with small channel bandwidth in order to reduce the hardware complexity of MTC equipment. Examples of this trend are LTE-M and Narrow-Band IoT (NB-IoT), which aim to reduce the bandwidth down to 1.4MHz and 180kHz, respectively [38], [39]. The exploitation of small channel bandwidth could involve a delay (and consequently energy consumption) increase due to the limited amount of time/frequency resources available on the radio interface. To overcome this limitation without losing the benefits in terms of low overhead offered by the group-paging strategy, we propose the *enhanced group-paging* procedure where the MtMS group is split into different subgroups and the paging is thus performed on a per-subgroup basis. The subgroup size and the time interval between paging messages are determined by the amount of resources available on the radio interface. For instance, if we consider a channel bandwidth of 1.4MHz (i.e., 6 resources blocks) with symmetric UL/DL configuration and with a RA slot periodicity equal to 5ms, we can assume the availability of 12 and 24 resources for UL and DL data channels, respectively, in a frame of 10ms (due to the fact that a subset of 6 UL resources is reserved for the transmission of the RA preambles in each RA slot). This means that the bottleneck to be considered for the selection of the subgroup size of the enhanced group-paging is the UL direction, which has a reduced capacity compared to the DL. Furthermore, to avoid a high overhead as in the 3GPP paging due to the transmission of paging messages every RA frame, the time interval between two paging messages needs to be properly tuned according to the average delay of data reception. By considering this aspect, the enhanced group-paging has been designed to page 36 UEs every 30ms, where 30ms is the expected average delay of the MtMS session and 36 UL resources are available in an interval of 30ms. The effectiveness of these choices will be testified by the results in Sec. V.

Once the joining procedure is terminated, the MtMS-GW triggers the data transmission to the involved UEs. It is worth to underline that, in legacy MBMS sessions, the content is delivered simultaneously to all the involved devices. When considering MtMS sessions, this would involve that all MTC devices need to wait until the whole set of devices has been successfully paged and this means high delays and energy consumptions. In addition, an MtMS session could require to accomplish data delivery within a specific time window; if a subset of devices is not able to terminate the joining procedure within the pre-defined time window, MtMS data will not be delivered. As a consequence, a more viable approach for MtMS is to schedule a data content transmission every time a subset of UEs accomplishes the joining phase in order to increase the number of devices successfully served within the MtMS time window.

Concerning the delivery of MtMS data, unicast and multicast transmission modes [30], [31], [40] represent two solutions to be considered for MtMS, in order to evaluate their pros and cons when applied to the transmission of small data content in deployments with small channel bandwidth.

In the remainder of this paper, we will evaluate the impact of above addressed paging, RA, and data transmission strategies on the performance of MtMS. To this end, in the next Section we will present an analytical model designed to evaluate different performance metrics for the MtMS session.

## IV. Analytical model

In the following, the analytical model to estimate the performance of an MtMS session is presented. Our model takes into consideration the joining and the content delivery phases. Table I lists the notations used in the paper.

TABLE I
List of Notations

| Notation | Definition | Value |
|---|---|---|
| $K$ | MtMS group size | 50-500 |
| $T^{TOT}$ | Total time interval to accomplish content delivery | $1s$ |
| $T^{RA}$ | Interval between two RA slots | $5ms$ [14] |
| $A$ | Number of RA slots in the RA frame | 1 (S-RA) [14];<br>2 (CE-RA) [11] |
| $\alpha_{i,n}$ | Number of devices that perform the $n$-th RA attempt in the $i$-th RA frame | (4) |
| $N$ | Maximum number of RA attempts | 10 [14] |
| $I$ | Maximum number of RA frames | (1) |
| $R$ | Number of preambles | 54 [14] |
| $C$ | Number of code-words | $R^A$ |
| $\rho(C, \alpha_{i,n})$ | Success probability in the $i$-th RA frame | (2) |
| $T^{RAR}$ | Processing time to detect a preamble | $2ms$ [15] |
| $W^{RAR}$ | RAR window | $5ms$ [15] |
| $M^{RAR}$ | Number of code-words addressed in a single RAR message | 6 [14] |
| $K^{RAR}$ | Maximum number of devices that can be acknowledged in a RA frame | $M^{RAR} \cdot W^{RAR} \cdot A$ |
| $W^B$ | Backoff window size | $20ms$ [14] |
| $\alpha^S_{i,n}$ | Number of devices that performed the $n$-th RA attempt in the $i$-th RA frame and successfully received the RAR message | (3) |
| $\beta_i$ | Number of devices to be scheduled for Msg3 transmission in the $i$-th RA frame | (9) |
| $\beta^S_i$ | Number of devices that successfully transmitted the Msg3 in the $i$-th RA frame | (8) |
| $U$ | Resources for Msg3 in the RA frame | 6 [14] |
| $\gamma_i$ | Number of devices to be scheduled in the $i$-th RA frame to receive the Msg4 | (11) |
| $\gamma^S_i$ | Number of devices that successfully received the Msg4 in the $i$-th RA frame | (10) |
| $D$ | Resources for Msg4 and data in the RA slot | 12 |
| $D^{data}$ | Resources needed to deliver the content | 1-5 |
| $\delta_i$ | Number of devices to be scheduled in the $i$-th frame to receive the data content | (14) |
| $\delta^S_i$ | Number of devices that successfully received the data content in the $i$-th frame | (12), (13) |
| $T^\alpha_n$ | Delay after $n$ RA attempts | (17) |
| $T^\beta$ | Delay to send the Msg3 and to receive the related acknowledgment | $5ms$ [15] |
| $T^\gamma$ | Delay to receive the Msg4 and to send the related acknowledgment | $5ms$ [15] |
| $T^\delta$ | Delay to receive the data content | $5ms$, $1ms$ [15] |
| $p^{idle}$ | Power consumption in idle state | $25mW$ [41] |
| $p^{Tx}$ | Power consumption in transmitting state | $100mW$ [38] |
| $p^{Rx}$ | Power consumption in receiving state | $100mW$ [38] |
| $p^\alpha_n$ | Power consumption after $n$ RA attempts | (19) |

### A. The joining procedure

We indicate with $K$ the total number of devices involved in the MtMS session. Such devices need to receive the MtMS content within a time interval equal to $T^{TOT}$. The HeNB can send paging messages with a periodicity equal to $T^{RA}$, which also represents the duration of the RA slot. At the reception of the paging message, a UE starts the RA procedure with the aim to acquire synchronization with the network as well as to be scheduled for data reception. In order to model 3GPP and code-expanded RA procedures, we assume that the RA is performed in a RA frame composed of $A$ RA slots; $A = 1$ refers to the 3GPP RA while $A > 1$ refers to the code-expanded RA. The duration of the RA frame is thus $T^{RA} \cdot A$. We indicate with $\alpha_{i,n}$ the number of devices performing the $n$-th attempt of RA procedure in the $i$-th RA frame, where $i = 1, \ldots, I$. It is worth noticing that the values of $\alpha_{i,n}$ when $n = 1$ are given by the considered paging procedure. The parameter $I$ is computed according to $T^{TOT}$. By considering that $T^{TOT}$ is a value expressed in milliseconds, we thus have:

$$I = \frac{T^{TOT}}{T^{RA} \cdot A} \tag{1}$$

We consider that $n = 1, \ldots, N$, where $N$ is the maximum number of allowed attempts before declaring a RA failure.

To perform the RA procedure, each device sends a preamble randomly selected among the $R$ defined for contention-based RA in every RA slot of the RA frame. The sequence of preambles chosen by a given device is defined as an access code-word. According to $R$ and $A$, the overall number of code-words in the RA frame is given by $C = R^A$. By considering the devices attempting to access in the $i$-th RA frame, we can define a success probability[1] for these devices as follows [11]:

$$\rho(C,\alpha_{i,n}) = \left(1-\frac{1}{C}\right)^{(\sum_{n=1}^{N}\alpha_{i,n})-1} \tag{2}$$

After the transmission of the randomly selected access code-word, a device waits a time interval $T^{RAR}$ before to start the RAR window, which lasts $W^{RAR}\cdot A$: in this window the device expects to receive a RAR addressing the chosen access code-word. As a RAR message can list up to $M^{RAR}$ different code-words, the maximum number of code-words that can be acknowledged per RA frame is given by $K^{RAR}=M^{RAR}\cdot W^{RAR}\cdot A$. By considering this, we can model the number of devices that successfully receive the RAR message before the RAR window expiration as follows:

$$\alpha_{i,n}^{S} = \begin{cases} \alpha_{i,n}\cdot\rho(C,\alpha_{i,n}), & \text{if } \sum_{n=1}^{N}\alpha_{i,n}\le K^{RAR} \\ \dfrac{\alpha_{i,n}\cdot\rho(C,\alpha_{i,n})\cdot K^{RAR}}{\sum_{n=1}^{N}\alpha_{i,n}\cdot\rho(C,\alpha_{i,n})}, & \text{otherwise} \end{cases} \tag{3}$$

In case the RAR window expires without the reception of a RAR message, then the UE declares a failure in the RA procedure and schedules another RA code-word transmission by considering a backoff interval equal to $W^{B}\cdot A+1$. By considering (3), we can derive the number of devices performing the $n$-th attempt (with $n>1$) of the RA procedure in the $i$-th RA frame as suggested by [15]:

$$\alpha_{i,n} = \sum_{j=j_i^{min}}^{j_i^{max}} \varphi_{j,i}\cdot(\alpha_{j,n-1}-\alpha_{j,n-1}^{S}) \tag{4}$$

The values $j_i^{min}$ and $j_i^{max}$ represent the possible indexes of the RA frames with failure that need to be considered for possible retransmission in the $i$-th frame, i.e., a novel RA attempt can be scheduled in the $i$-th frame only if the RA failure happened in a frame $j^{min}\le j\le j^{max}$. The value $\varphi_{j,i}$ represents the portion of devices performing a new RA attempt in the $i$-th RA frame after the failure of the $(n-1)$-th RA attempt in the $j$-th frame; this means that $\varphi_{j,i}$ is the portion of devices that failed at the $j$-th RA frame and the related backoff window expires at the $i$-th RA frame. According to [15], we can compute these parameters as follows:

$$j_i^{min} = \left\lceil (i-1)+\frac{T^{RAR}+(W^{RAR}+W^{BO})\cdot A}{T^{RA}\cdot A}\right\rceil \tag{5}$$

$$j_i^{max} = \left\lfloor i-\frac{T^{RAR}+W^{RAR}\cdot A+1}{T^{RA}\cdot A}\right\rfloor \tag{6}$$

while $\varphi_{j,i}$ can be derived as in (7).

The RAR message also carries the uplink grant to transmit the Msg3 in the following RA frame. We indicate with $\beta_i$ the number of devices to be scheduled for Msg3 transmission in the $i$-th frame. Among the $\beta_i$ devices, the HeNB schedules the ones (denotes by $\beta_i^S$) that will transmit the Msg3 in the $i$-th RA frame according to the amount of resources available in the UL direction. $U$ denotes the number of resources available in the UL direction every RA slot to transmit the Msg3; consequently, the overall number of resources for Msg3 in the $i$-th RA frame is given by $U\cdot A$. The parameter $\beta_i^S$ can be computed as follows[2]:

$$\beta_i^{S} = \begin{cases} \beta_i, & \text{if } \beta_i\le U\cdot A \\ U\cdot A, & \text{otherwise} \end{cases} \tag{8}$$

Consequently, by considering that at the generic $i$-th frame the HeNB could also schedule devices that did not transmit the Msg3 in the previous RA frame, $\beta_i$ can be computed as follows:

$$\beta_i = \sum_{n=1}^{N}\alpha_{i-1,n}^{S}+(\beta_{i-1}-\beta_{i-1}^{S}) \tag{9}$$

At the reception of Msg3, the HeNB transmits the Msg4 that contains the information relevant to the data delivery. We indicate with $\gamma_i$ the number of devices to be scheduled for Msg4 reception in the $i$-th frame. Among the $\gamma_i$ devices, the HeNB schedules the ones (denotes by $\gamma_i^S$) that will receive the Msg4 in the $i$-th RA frame according to the amount of resources available in the DL direction. We indicate with $D$ the number of DL resources available to transmit the Msg4; consequently, the overall number of resources for Msg4 in the $i$-th RA frame is given by $D\cdot A$. By assuming that one resource can carry the Msg4, we can derive $\gamma_i^S$ as follows:

$$\gamma_i^{S} = \begin{cases} \gamma_i, & \text{if } \gamma_i\le D\cdot A \\ D\cdot A, & \text{otherwise} \end{cases} \tag{10}$$

The parameter $\gamma_i$ can be computed by considering that at the generic $i$-th frame the HeNB could also schedule devices that did not receive the Msg4 in the previous RA frame:

$$\gamma_i = \beta_i^{S}+(\gamma_{i-1}-\gamma_{i-1}^{S}) \tag{11}$$

### *B. Content delivery*

Once a device receives the Msg4, it also receives the information on the scheduled opportunity to receive the data content. We consider that the data content needs an amount of resources equal to $D^{data}$ to be delivered[3]. We indicate with

[1] For the sake of completeness, preamble detection probability should be considered to compute the success probability. This can be modelled by considering the power ramping as suggested in [14], where $1-(1/e^{n})$ is the probability of successful preamble transmission at the $n$-th RA attempt. Nevertheless, as analysed in [10], a successful preamble reception at the first attempt can be considered realistic in scenarios where coverage for MTC devices is provided through the exploitation of femto-cells. Consequently, for the sake of simplicity, we assume that a preamble is always decoded with success by the base station as we will consider a femto-cell deployment in our performance evaluation scenario.

[2] For the sake of completeness, the error probability of Msg3 transmission should be taken into account to compute $\beta_i^S$. Nevertheless, by considering a deployment environment based on femto-cells, it is realistic to assume that channel conditions for MTC devices allow a successful data transmission/reception, as analyzed in [10]. This allows to simplify our model. The same assumption holds for Msg4 and data content delivery. In addition, in (8) we assume that one resource can carry the Msg3.

[3] For the sake of simplicity, we assume that the amount of resources is the same for unicast and multicast transmissions. For completeness, a larger $D^{data}$ could be necessary for multicast transmissions due to the exploitation of a more robust modulation and coding scheme to guarantee successful data reception to the UEs with poor channel conditions [31]. Nevertheless, as explained in Sec. V, this assumption does not limit the effectiveness of our proposed model.

$$\varphi_{j,i} = \begin{cases} \frac{(j-1)\cdot T^{RA}\cdot A+T^{RAR}+[W^{RAR}+W^{BO}-(i-2)\cdot T^{RA}]\cdot A}{W^{BO}\cdot A}, & \text{if } j_i^{min} \leq j \leq i - \frac{T^{RAR}+(W^{RAR}+W^{BO})\cdot A}{T^{RA}\cdot A} \\ \frac{T^{RA}}{W^{BO}}, & \text{if } i - \frac{T^{RAR}+(W^{RAR}+W^{BO})\cdot A}{T^{RA}\cdot A} < j < (i-1) - \frac{T^{RAR}+W^{RAR}\cdot A}{T^{RA}\cdot A} \\ \frac{(i-1)\cdot T^{RA}\cdot A-[(j-1)\cdot T^{RA}\cdot A+T^{RAR}+W^{RAR}\cdot A]}{W^{BO}\cdot A}, & \text{if } (i-1) - \frac{T^{RAR}+W^{RAR}\cdot A}{T^{RA}\cdot A} \leq j \leq j_i^{max} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$\delta_i$ the number of devices scheduled for data reception in the $i$-th frame.

In case of unicast transmission, the number of devices that successfully receive the data content can be computed as follows:

$$\delta_i^S = \begin{cases} \delta_i, & \text{if } \delta_i \cdot D^{data} \leq D \cdot A - \gamma_i^S \\ \left\lfloor \frac{D \cdot A - \gamma_i^S}{D^{data}} \right\rfloor, & \text{otherwise} \end{cases} \tag{12}$$

where (12) takes into consideration the fact that a portion of DL resources has been exploited to transmit the Msg4 to $\gamma_i^S$ devices.

In case of multicast transmission, $\delta_i^S$ can be computed as follows:

$$\delta_i^S = \begin{cases} \delta_i, & \text{if } D^{data} \leq D \cdot A \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

The parameter $\delta_i$ can be computed by considering that at the generic $i$-th frame the HeNB could also schedule devices that did not receive the data content in the previous frame:

$$\delta_i = \gamma_i^S + (\delta_{i-1} - \delta_{i-1}^S) \tag{14}$$

### C. Performance metrics

*1) Number of paging messages:* The first metric under consideration in our analysis is the number of paging messages, i.e., $\Lambda$, needed to perform the content deliver. It is computed as follows:

$$\Lambda = \sum_{i=1}^{I} 1_{\alpha_{i,n}}, \text{with } n = 1 \tag{15}$$

where $1_x$ is the indicator function equal to 1 if $x > 0$ and 0 otherwise.

*2) Delay of paging procedure:* This metric, denoted by $\Delta_{paging}^{TOT}$, indicates the overall delay needed to perform the paging procedure.

$$\Delta_{paging}^{TOT} = i^* \cdot T^{RA} \cdot A, \text{with } i^* = \max\left\{ i | \gamma_i^S > 0 \right\} \tag{16}$$

*3) Average delay of paging procedure:* The parameter $\Delta_{paging}^{avg}$ indicates the average delay that MTC devices experience from the instant they are paged to the instant the paging procedure is accomplished. This value takes into consideration the overall delay between the first transmission of the first preamble of the access code-word and the reception of Msg4. We model the delay that a device experiences to receive the RAR message after $n$ RA attempts as follows:

$$T_n^{\alpha} = n \cdot \left( \frac{T^{RA} \cdot A}{2} + 1 + T^{RAR} + \frac{W^{RAR} \cdot A}{2} \right) + (n-1) \cdot \left( \frac{W^{BO} + W^{RAR}}{2} \right) \tag{17}$$

By also considering the delay to transmit the Msg3 and to receive the related acknowledgment (i.e., $T^{\beta}$) as well as the delay to receive the Msg4 and to transmit the related acknowledgment (i.e., $T^{\gamma}$), we can define the average delay of the paging procedure as follows:

$$\Delta_{paging}^{avg} = \frac{\sum_{i=1}^{I} \left( \sum_{n=1}^{N} \alpha_{i,n}^S \cdot T_n^{\alpha} \right) + \beta_i \cdot T^{\beta} + \gamma_i \cdot T^{\gamma}}{\sum_{i=1}^{I} \gamma_i^S} \tag{18}$$

*4) Average energy consumption of paging procedure:* This metric is computed by considering the power consumption of device in idle, transmitting and receiving states, denoted by $p^{idle}$, $p^{Tx}$ and $p^{Rx}$, respectively. We can derive the energy consumption of the reception of the RAR message after $n$ RA attempt as follows:

$$E_n^{\alpha} = n \cdot \left( \frac{T^{RA} \cdot A}{2} \cdot p^{idle} + p^{Tx} + T^{RAR} \cdot p^{idle} + \frac{W^{RAR} \cdot A}{2} \cdot p^{Rx} \right) + (n-1) \cdot \left( \frac{W^{BO} + W^{RAR}}{2} \right) \cdot p^{idle} \tag{19}$$

By taking into account the energy consumption related to the Msg3 transmission (i.e., $E^{\beta}$) and that related to Msg4 reception (i.e., $E^{\gamma}$), then the average energy spent for the paging procedure can be obtained as follows:

$$E_{paging}^{avg} = \frac{\sum_{i=1}^{I} \left( \sum_{n=1}^{N} \alpha_{i,n}^S \cdot E_n^{\alpha} \right) + \beta_i \cdot E^{\beta} + \gamma_i \cdot E^{\gamma}}{\sum_{i=1}^{I} \gamma_i^S} \tag{20}$$

.

*5) Percentage of data delivery success:* This metric, denoted as $\Phi$, can be easily computed as the ratio of the UEs that received the data content within the MtMS deadline to the overall number of UEs to be served:

$$\Phi = \frac{\sum_{i=1}^{I} \delta_i^S}{K} \tag{21}$$

*6) Total delay of data delivery:* This metric, denoted by $\Delta_{data}^{TOT}$, indicates the overall delay needed to accomplish the data delivery by also considering the time needed for the paging procedure. It can be computed as follows:

$$\Delta_{data}^{TOT} = i^* \cdot T^{RA} \cdot A, \text{with } i^* = \max\left\{ i | \delta_i^S > 0 \right\} \qquad (22)$$

*7) Average delay for data delivery:* The parameter $\Delta_{data}^{avg}$ indicates the average delay for an MTC device from the moment the device is paged to the moment the device receives the data content. This value can be defined as follows:

$$\Delta_{data}^{avg} = \Delta_{paging}^{avg} + \frac{\sum_{i=1}^{I} \delta_i \cdot T^{\delta}}{\sum_{i=1}^{I} \delta_i^S} \qquad (23)$$

where $T^{\delta}$ is the time needed to receive the data content[4].

*8) Average energy consumption for data delivery:* The average energy spent by devices for data delivery can be obtained as follows:

$$E_{data}^{avg} = E_{paging}^{avg} + \frac{\sum_{i=1}^{I} \delta_i \cdot E^{\delta}}{\sum_{i=1}^{I} \delta_i^S} \qquad (24)$$

where $E^{\delta} = E^{\gamma}$ for the unicast mode while $E^{\delta} = T^{\delta} \cdot p^{Rx}$ for the multicast case.

## V. Performance evaluation

Computer simulations were conducted on top of a 3GPP-calibrated Matlab®simulator to verify the effectiveness of the proposed analytical model. The dashed lines of the plots in the remainder of this section indicate the results obtained with the above mentioned 3GPP-calibrated simulator. Our simulations consider a scenario where MTC devices are attached to one LTE-M femto-cell (1.4MHz bandwidth at 2GHz with symmetric UL/DL radio frame). The cell layout, radio channel and interfering model are set according to the [42], while system level parameters are set in accordance to [43], [19], [14]. A detailed list of simulation parameters can be found in Table I. The aim of our analysis is to evaluate the impact of different paging, RA, and data transmission strategies on the performance of MtMS when considering MTC deployments with small channel bandwidth. This aspect is crucial when evaluating the pros and cons of solutions designed for MTC traffic, as recent standardization activities state the importance of reducing the bandwidth for MTC traffic in order to reduce the complexity, cost, and energy consumption of MTC devices [38]. For the sake of simplicity, we don't consider background traffic in UL and DL directions.

The first analysis focuses on the number of paging messages, i.e., $\Lambda$, of the approaches in consideration in this paper: *(i)* 3GPP paging, hereinafter standard paging (SP), *(ii)* group-paging (GP); *(iii)* enhanced group-paging (eGP). The results are shown in Fig. 4. Obviously the approach that allows to minimize the number of paging messages is the GP, as it always needs only one paging message to page all MTC devices; from an overhead point of view, the GP thus allows meaning savings. The SP is obviously the approach with the highest $\Lambda$: up to 32 paging messages are needed to address 500 devices (due to the fact that a maximum number of 16 UEs every 5ms can be paged with the legacy 3GPP procedure). With respect to SP, the eGP has a reduction of about 56% in terms of $\Lambda$. It is worth noting that the eGP could introduce further reductions compared to SP by increasing the number of devices paged in every paging occasion. Nevertheless, dimensioning this number in eGP influences the performance in terms of paging latency, as analysed in the following of this Section.

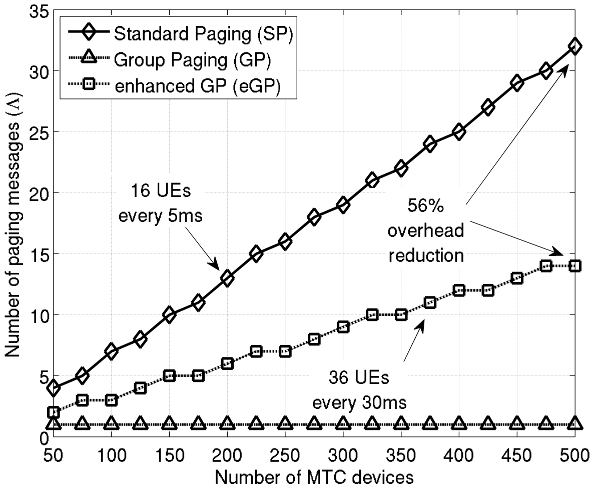

Fig. 4. Number of paging messages by varying the MtMS group size.

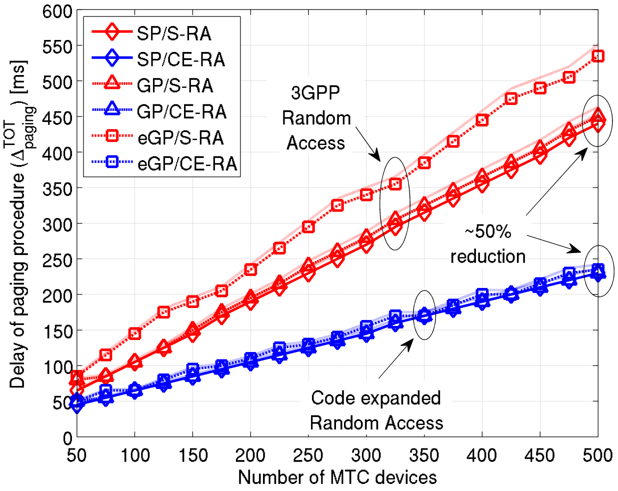

Fig. 5. Total delay of paging procedures with 3GPP and code-expanded random access.

The second parameter analyzed is the total delay of the paging procedure, i.e., $\Delta_{paging}^{TOT}$, which is shown in Fig. 5. The aim of this analysis is to evaluate the impact of the RA on the paging procedure. To this end, we consider the performance of SP, GP and eGP when coupled with the 3GPP RA procedure (hereinafter standard RA, S-RA) and the code-expanded RA (CE-RA). From Fig. 5, it clearly appears that the S-RA procedure involves meaningfully delays compared

[4] This value is equal to $T^{\gamma}$ in the case of unicast transmission (as it considers the reception of the data and the transmission of the related acknowledgement) while it is equal to 1ms in the case of multicast transmission (i.e., only data delivery without acknowledgement) [31].

to the CE-RA; this behavior is given by the fact that the S-RA procedure has a limited set of access code-words and this is translated to collisions and thus delays. It is worth noting that the solution GP/S-RA is the worst performing one: this is due to the fact that all devices are paged with a single paging message, and consequently a huge number of devices performs the S-RA procedure at the same time, with consequent low success probability due to collisions of access code-words. This phenomenon is less accentuated when considering SP/S-RA and eGP/S-RA, as in these cases the number of devices that perform the S-RA procedure in the same RA frame is reduced compared to the use of GP. According to the above results, we can thus conclude that the limited capacity of the S-RA procedure defined by 3GPP represents the limiting aspect when providing multicast service to huge number of MTC devices. When considering the performance of the CE-RA, the behaviors of SP, GP and eGP are similar. To understand this, we need to remind that the total delay of the paging procedure when varying the paging strategy is influenced by two aspects: *(i)* the percentage of collision in the RA procedure as this could potentially involve retransmissions and thus delays; *(ii)* the lack of available resources that could potentially introduce delay during the RA procedure. By considering the former aspect, it is worth noticing that the CE-RA is characterized by a large set of access code-words and this guarantees a high success probability in the transmission of Msg1. In a scenario with 500 devices paged simultaneously, only 16% of devices experience a collision as it can be verified with (2). Consequently, the collision does not meaningfully influence the paging procedure delay when coupled with CE-RA. The parameter that effectively influences $\Delta^{TOT}_{paging}$ is thus the number of available resources as the lack of resources involves delay in scheduling the transmission of Msg2, Msg3 and Msg4. The GP/CE-RA and eGP/CE-RA would experience reduced delays in case of a larger channel bandwidth.

Fig. 5 highlights the benefits of CE-RA in reducing the total delay of the paging procedure; in addition, it shows that the paging strategy does not meaningfully influence the paging delay in scenarios with limited channel bandwidth. Nevertheless, it is interesting to consider the average delay, i.e., $\Delta^{avg}_{paging}$, of CE-RA when varying the paging strategy. This metric is shown in Fig. 6. The GP/CE-RA has the highest average delay. This behaviour is due to the fact that a large set of devices perform a successfully Msg1 transmission, but due to the lack of resources the Msg2 is not received within the RAR window: these devices thus need to perform another RA attempt after the backoff period. When focusing on the performance of SP/CE-RA, we can note that it allows a drastic reduction of $\Delta^{avg}_{paging}$ compared to GP-CE-RA, as in every paging occasion the SP manages a reduced set of devices compared to the GP strategy. Nevertheless, The SP pages 16 UEs every 5ms, and this value still represents a huge load (i.e., 32 UEs every 10ms) when considering systems with low channel bandwidth such as LTE-M. We can note that the eGP/CE-RA provides a reduction of about 35% compared to SP/CE-RA in terms of average delay. This is due to the design choice at the basis of eGP. The eGP has been designed to page 36 UEs every 30ms, as 36 UL resources are available in an interval of 30ms (please, refer to Sec. III). Allowing more users (as for instance done by GP and SP) increases $\Delta^{avg}_{paging}$, as testified by the results in Fig. 6. It is interesting to note that the eGP shows an average delay that does not increase with the MtMS group size as the number of devices paged in each paging opportunity remains fixed. The benefits introduced by eGP/CE-RA in terms of a reduced $\Delta^{avg}_{paging}$ involve meaningful benefits in terms of energy consumption (plotted in Fig. 7), with reductions up to 40% compared to SP/CE-RA. According to the results in Figures 6, 7 and above, the eGP/CE-RA can be considered a viable solution to handle the joining phase of MtMS for a large set of devices as it offers the lowest average delay and energy consumption while reducing the overhead of the paging procedure.

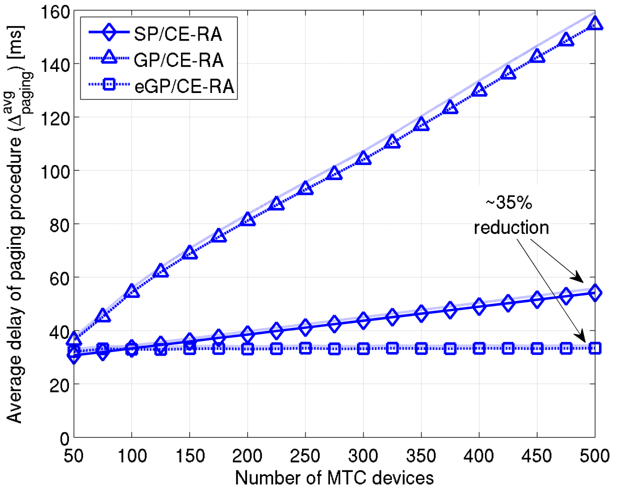

Fig. 6. Average delay of paging procedures with CE-RA.

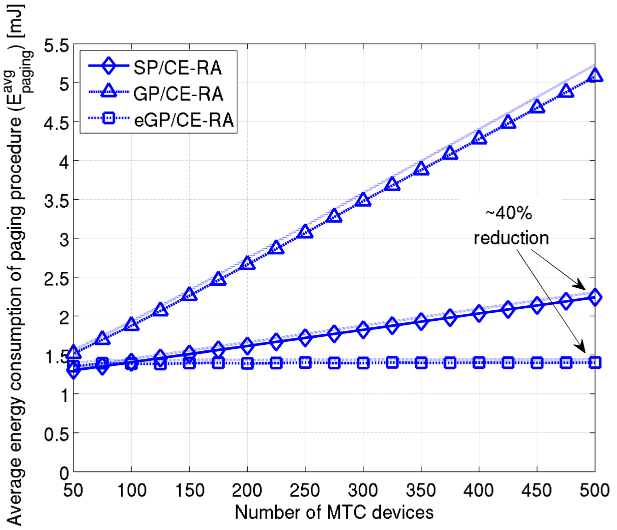

Fig. 7. Average energy consumption of paging procedures with CE-RA.

In the remainder of this Section, we will focus on the performance relevant to data transmission. We consider the eGP/CE-RA when coupled with unicast and multicast modes. In order to evaluate the impact of the message size, we vary

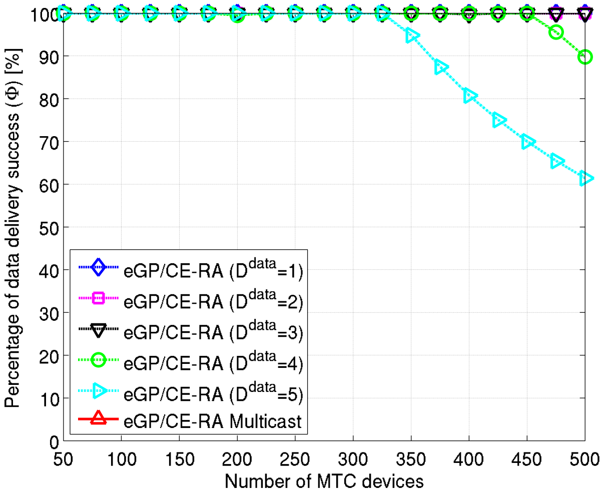

Fig. 8. Percentage of UEs which successfully received the content.

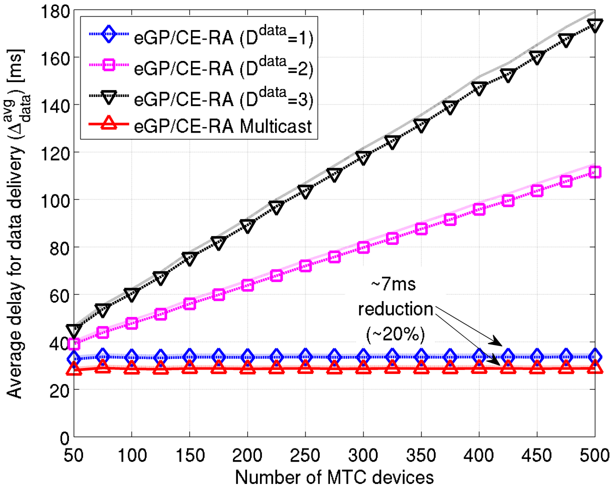

Fig. 10. Average delay for content delivery.

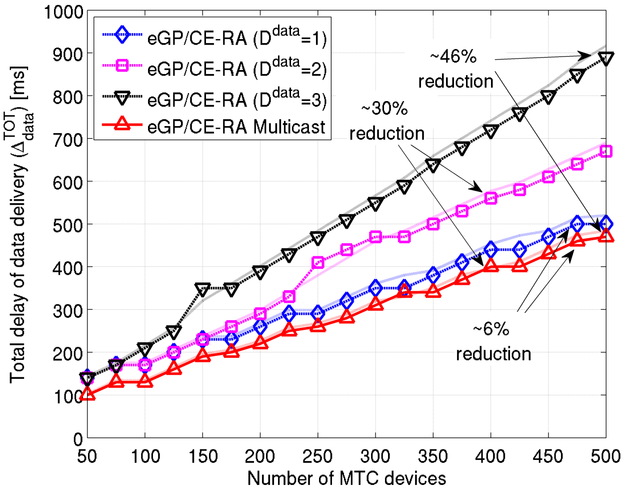

Fig. 9. Total delay for content delivery.

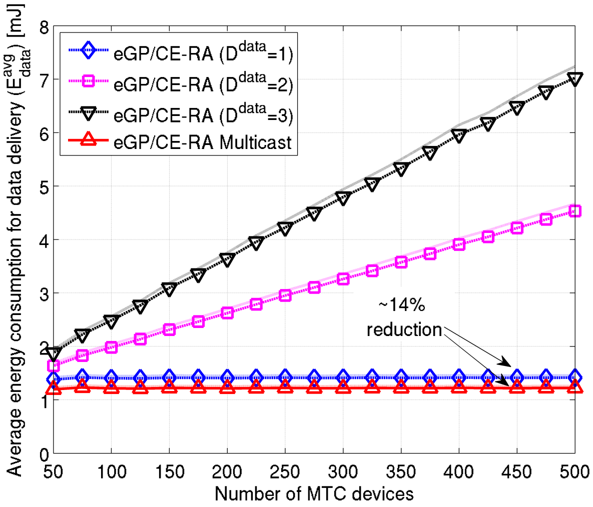

Fig. 11. Average energy consumption for content delivery.

$D^{data}$ from 1 to 5. The first analysis, shown in Fig. 8, focuses on the percentage $\Phi$ of devices receiving the data content before the expiration of the deadline $T^{TOT}$. From Fig. 8, we can note that $\Phi$ goes below 100% when the number $K$ of devices is higher than 450 for unicast transmissions with $D^{data} = 4$, while this behavior can be observed from $K = 325$ when considering $D^{data} = 5$. This is due to the high amount of resources needed in the DL direction to serve a large group of devices via unicast transmissions. As a consequence, in the remainder of this Section we only focus on $D^{data} = 1, 2, 3$ for the unicast mode. The exploitation of multicast transmissions guarantees full coverage until the amount of resources for data transmission is equal to $D^{data} = 25$ (these results are not shown for the sake of clarity of the plot).

The second parameter under investigation is the total delay in the data transmission procedure, i.e., $\Delta_{data}^{TOT}$, plotted in Fig. 9. The best performance for the unicast mode is achieved when $D^{data} = 1$, where $\Delta_{data}^{TOT}$ is approximately equal to 500ms in the huge load case of $K = 500$ devices. It is worth analyzing the trend of the unicast mode: $\Delta_{data}^{TOT}$ grows of about 200ms with the increase of $D^{data}$. This aspect underlines the strong limitation of unicast transmissions even if considering small data content requiring few resources to be transmitted. The only case when $D^{data}$ does not influence $\Delta_{data}^{TOT}$ for the unicast mode is when the size of the MtMS group is lower than 75 devices for $D^{data} = 1, 2, 3$, and lower than 150 for $D^{data} = 1, 2$. The exploitation of multicast transmissions allows a reduction of about 6% compared to the unicast mode with $D^{data} = 1$. It is worth noting the trend of the multicast mode when increasing $D^{data}$ (these results are not shown for the sake of clarity of the plot): when $D^{data}$ is equal to 20, $\Delta_{data}^{TOT}$ matches the value obtained by the unicast mode with $D^{data}$ equal to 1. This result is interesting because allows to understand that the increase in the size of data content does not involve meaningful increase in the overall delay when the data content is delivered with the multicast mode. As a consequence, this behavior testifies that the assumption to consider in our system model the same value $D^{data}$ for unicast and multicast transmission modes does not meaningfully influence the validity of our model.

When analyzing the average delay ($\Delta_{data}^{avg}$) and energy consumption ($E_{data}^{avg}$) in Fig. 10 and Fig. 11, respectively, we can observe that multicast transmissions involve a reduction of about 20% in terms of $\Delta_{data}^{avg}$ and 14% in terms of $E_{data}^{avg}$ compared to the unicast mode with $D^{data} = 1$. Finally, it is worth to underline the following aspect. From Fig. 10, we can note that the average delay to accomplish the MtMS session when using the multicast mode is approximately equal to 30ms. This testifies the settings chosen for the eGP, which is properly tuned to page 36 UEs every 30ms: within this time interval, all the paged devices are able to accomplish the data reception and, consequently, all the UL/DL resources are now free to allow a novel subgroup of devices to receive the MtMS content.

## VI. Lessons learned and future research trends

### A. Lessons learned

The performance evaluation provided in the previous Section shows that our proposed MtMS is able to perform multicast content delivery in LTE-M deployments with an average delay close to 30ms (please, refer to Fig. 10) . It is worth noting that this delay is not influenced by the number of UEs belonging to the MtMS group. Indeed, the average delay to receive data content does not vary when the MtMS group size increases as devices are split into different subgroups for the sake of optimizing the utilization of network resources and minimizing delay and energy consumption.

The results achieved with MtMS are useful to understand the overall delay of the end-to-end path. On the UL direction, recent advances in literature allow to cut the transmission delay of high-priority messages (e.g., alarms) down to 10ms [12]. Thus, the DL transmission is the segment that currently introduces the highest source of delay in the end-to-end path. Thanks to the analyses conducted in this paper, we can estimate that the end-to-end delay can be considered $\sim$50ms in case the core network introduces a delay $<$10ms, or $<$100ms otherwise. This testifies the applicability of our proposed MtMS in enabling end-to-end communications in smart environments, such as management of traffic lights to handle emergency services [44] or to balance car traffic to reduce congestion and pollution [1], [2] in smart cities.

Further enhancements to reach $<$10ms end-to-end delay are still required. Reaching this goal would enable the provisioning of end-to-end mission critical applications, with consequent novel business opportunities for telco operators [1], [37]. Possible strategies to be investigated to reach this target delay are discussed in the following of this Section.

### B. Future research trends

With the aim of supporting 5G IoT use cases with strict delay requirements, several aspects need to be further investigated in order to cut the delay of MtMS sessions as well as in the UL and core network segments.

A reduction in the overall end-to-end delay could be achieved through the exploitation of softwarization and virtualization paradigms [24], [25], [26], as mentioned in Sec. II-B. When considering the procedures of MtMS, a source of delay is related to the joining procedure as it requires to contact the mobile core (specifically, the MME) to acquire the information about the tracking area(s) of devices to be paged. This delay could be potentially reduced by migrating into the HeNB-GW the information relevant to the UEs involved in the MtMS session. Another aspect to be considered in this field is related to the benefits introduced by the exploitation of edge clouds [37], [45], [27], which bring computational capabilities close to the edge in order to avoid to contact the mobile core. One edge cloud could host the functionalities of HeNB-GW, MtMS-GW, MtMS-CE, SCS and MtMS-SC: this could drastically cut the end-to-end delay by allowing to trigger the MtMS session directly within the edge cloud. In addition, virtualization may take advantage of the analysis of big data generated by MTC devices in order to optimize the network utilization and to improve the performance [46]. For instance, the best placement of network functions in order to cut delays and/or to reduce the overhead can be obtained by considering the knowledge (e.g., traffic patterns, end-to-end paths, network load) extracted from the data received from the uplink direction.

Additional delay reductions of MtMS sessions could be achieved through a proper selection of the UEs to be paged. Although recent studies focused on this direction [47], [48], [49], additional research is needed to tailor these procedures for MtMS and for deployments characterized by small channel bandwidth. In addition, mechanisms to perform the joining procedure by considering DRX [17] cycles are needed. Under this point of view, devices with the same DRX cycle could be gathered together in the same subgroup in order to be paged simultaneously. Furthermore, the coexistence of UL/DL interfering traffic should be considered in order to design solutions handling traffic priority and so on.

Finally, another aspect to be considered is related to the resource allocation for MtMS data. Even though the resource allocation represents a topic widely investigated for multicast services [40], the research always focused on the maximization of human/network-oriented goals, such as throughput, in order to enhance the QoS experienced by the users as well as to improve the resource utilization. In case of MtMS, additional parameters (such as the level of residual battery charge of MTC devices as considered for instance in the RA procedure [50]) could be considered in the resource allocation step with the aim of allocating transmission parameters (e.g., power, modulation and coding schemes) in order to minimize the energy consumption of devices.

## VII. Conclusion

The transformation of IoT from a sensor-driven paradigm to one heavily complemented by actuators, drones and robots dictates the design of solutions to cut the latency of MTC traffic in both uplink and downlink directions. In this paper we analyzed the sources of delay in the end-to-end path of

MTC traffic. We surveyed the current literature focusing on the performance optimization for the uplink direction of MTC and for the mobile core of 5G systems. We further proposed the machine-type multicast service (MtMS), with the aim to enable the simultaneous transmission of data toward a large set of MTC devices. We presented the architectural components and the procedures to be adopted by MtMS to cut delay, energy consumption and control overhead. The effectiveness of our proposal has been testified through analyses conducted by considering LTE-M deployments. Finally, we discussed the open challenges to be further investigated for MtMS and we provided some guidelines to drive the future research on this topic.

## Acknowledgment

This work has been partially supported by the 5GPP VirtuWind (Virtual and programmable industrial network prototype deployed in operational wind park) Project.

## References

[1] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, "A Survey on Internet of Things From Industrial Market Perspective," *IEEE Access*, vol. 2, pp. 1660–1679, 2014.

[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, pp. 2347–2376, Fourthquarter 2015.

[3] Ericsson, "More than 50 billion connected devices." White Paper, 2011.

[4] C. Anton-Haro and M. Dohler, *Machine-to-machine (M2M) Communications: Architecture, Performance and Applications*. Elsevier, 2014.

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet of Things Journal*, vol. 1, pp. 22–32, Feb 2014.

[6] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 510–527, March 2016.

[7] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description." TS 36.300.

[8] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, pp. 74–80, February 2014.

[9] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 4–16, First 2014.

[10] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and K. Zheng, "Toward 5G densenets: architectural advances for effective machine-type communications over femtocells," *Communications Magazine, IEEE*, vol. 53, pp. 134–141, January 2015.

[11] H. Thomsen, N. K. Pratas, C. Stefanović, and P. Popovski, "Code-expanded radio access protocol for machine-to-machine communications," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 355–365, 2013.

[12] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced radio access and data transmission procedures facilitating industry-compliant machine-type communications over LTE-based 5G networks," *IEEE Wireless Communications*, vol. 23, pp. 56–63, February 2016.

[13] 3GPP, "Mobile radio interface layer 3 specification, core network protocols; Stage 2." TS 23.108.

[14] 3GPP, "RAN improvements for machine-type communications." TR 37.868.

[15] C. H. Wei, R. G. Cheng, and S. L. Tsao, "Performance Analysis of Group Paging for Machine-Type Communications in LTE Networks," *IEEE Transactions on Vehicular Technology*, vol. 62, pp. 3371–3382, Sept 2013.

[16] 3GPP, "Study on Enhancements to Machine-Type Communications (MTC) and Other Mobile Data Applications; Radio Access Network (RAN) Aspects." TR 37.869.

[17] C. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Communications Magazine*, vol. 47, pp. 48–55, June 2009.

[18] K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, and H. Zhu, "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Communications*, vol. 21, pp. 12–18, June 2014.

[19] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC)." TS 36.331.

[20] Y. Chen and W. Wang, "Machine-to-Machine Communication in LTE-A," in *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*, pp. 1–4, Sept 2010.

[21] S. Andreev, A. Larmo, M. Gerasimenko, V. Petrov, O. Galinina, T. Tirronen, J. Torsner, and Y. Koucheryavy, "Efficient small data access for machine-type communications in LTE," in *Communications (ICC), 2013 IEEE International Conference on*, pp. 3569–3574, June 2013.

[22] T. Blajić, D. Nogulić, and M. Družijanić, "Latency Improvements in 3g Long Term Evolution," in *Mipro CTI, svibanj*, 2006.

[23] N. Larson, D. Baltrunas, A. Kvalbein, A. Dhamdhere, k. claffy, and A. Elmokashfi, "Investigating Excessive Delays in Mobile Broadband Networks," in *SIGCOMM workshops, All Things Cellular*, Aug 2015.

[24] Y. Li and M. Chen, "Software-Defined Network Function Virtualization: A Survey," *IEEE Access*, vol. 3, pp. 2542–2553, 2015.

[25] T. Wood, K. K. Ramakrishnan, J. Hwang, G. Liu, and W. Zhang, "Toward a software-based network: integrating software defined networking and network function virtualization," *IEEE Network*, vol. 29, pp. 36–41, May 2015.

[26] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 236–262, Firstquarter 2016.

[27] A. Aissioui, A. Ksentini, A. M. Gueroui, and T. Taleb, "Toward Elastic Distributed SDN/NFV Controller for 5G Mobile Cloud Management Systems," *IEEE Access*, vol. 3, pp. 2055–2064, 2015.

[28] O. Arouk, A. Ksentini, and T. Taleb, "Group paging optimization for machine-type-communications," in *Communications (ICC), 2015 IEEE International Conference on*, pp. 6500–6505, June 2015.

[29] O. Arouk, A. Ksentini, and T. Taleb, "Group Paging-based Energy Saving for Massive MTC Accesses in LTE and Beyond Networks," *IEEE Journal on Selected Areas in Communications*, vol. PP, no. 99, pp. 1–1, 2016.

[30] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements," *IEEE Communications Magazine*, vol. 50, pp. 68–74, November 2012.

[31] M. Condoluci, G. Araniti, A. Molinaro, and A. Iera, "Multicast Resource Allocation Enhanced by Channel State Feedbacks for Multiple Scalable Video Coding Streams in LTE Networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2015.

[32] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Communications Magazine*, vol. 52, pp. 44–51, May 2014.

[33] D. Muirhead, M. A. Imran, and K. Arshad, "Insights and Approaches for Low-Complexity 5G Small-Cell Base-Station Design for Indoor Dense Networks," *IEEE Access*, vol. 3, pp. 1562–1572, 2015.

[34] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, Present, and Future," *IEEE Journal on Selected Areas in Communications*, vol. 30, pp. 497–508, April 2012.

[35] M. Chen, J. Wan, S. Gonzalez, X. Liao, and V. C. M. Leung, "A Survey of Recent Developments in Home M2M Networks," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 98–114, First 2014.

[36] F. Ghavimi and H. H. Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, pp. 525–549, Secondquarter 2015.

[37] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-Enabled Tactile Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 460–473, March 2016.

[38] Nokia Networks, "LTE-M - Optimizing LTE for the Internet of Things." White Paper, 2015.

[39] R. Ratasuk, N. Mangalvedhe, A. Ghosh, and B. Vejlgaard, "Narrowband LTE-M System for M2M Communication," in *Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th*, pp. 1–5, Sept 2014.

[40] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey," *IEEE Communications Surveys Tutorials*, vol. 15, pp. 240–254, First 2013.

[41] M. Gerasimenko, V. Petrov, O. Galinina, S. Andreev, and Y. Koucheryavy, "Energy and delay analysis of LTE-Advanced RACH performance under MTC overload," in *Globecom Workshops (GC Wkshps), 2012 IEEE*, pp. 1632–1637, Dec 2012.

[42] 3GPP, "Physical Layer Aspect for Evolved Universal Terrestrial Radio Access (UTRA)." TR 25.814.

[43] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC)." TS 36.321.

[44] M. Condoluci, F. Sardis, and T. Mahmoodi, "Softwarization and Virtualization in 5G Networks for Smart Cities," in *EAI International Conference on CYber physiCaL systems, iOt and sensors Networks*, Oct 2015.

[45] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating While Computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Processing Magazine*, vol. 31, pp. 45–55, Nov 2014.

[46] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Network*, vol. 30, pp. 44–51, January 2016.

[47] R. Harwahymiscu, R.-G. Cheng, and R. F. Sari, "Consecutive group paging for LTE networks supporting machine-type communications services," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, pp. 1619–1623, Sept 2013.

[48] O. Arouk, A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "On improving the group paging method for machine-type-communications," in *Communications (ICC), 2014 IEEE International Conference on*, pp. 484–489, June 2014.

[49] R.-G. Cheng, F. M. Al-Taee, J. Chen, and C.-H. Wei, "A Dynamic Resource Allocation Scheme for Group Paging in LTE-Advanced Networks," *IEEE Internet of Things Journal*, vol. 2, pp. 427–434, Oct 2015.

[50] M. Condoluci, G. Araniti, M. Dohler, A. Iera, and A. Molinaro, "Virtual code resource allocation for energy-aware MTC access over 5G systems," *Ad Hoc Networks*, vol. 43, pp. 3 – 15, 2016. Smart Wireless Access Networks and Systems for Smart Cities.