Group-based Delivery of Critical Traffic in Cellular IoT Networks

Olga Vikhrova[a,*], Sara Pizzi[a], Antonella Molinaro[a], Antonio Iera[b],
Konstantin Samuylov[c], Giuseppe Araniti[a]

[a]*DIIES Dept., University Mediterranea of Reggio Calabria, Italy*
[b]*DIMES Dept., University of Calabria, Italy*
[c]*Peoples' Friendship University of Russia (RUDN University), Russia*

Abstract

Fifth generation (5G) networks are expected to connect a huge number of
Internet of Things (IoT) devices in many usage scenarios. The challenges
of typical massive IoT applications with sporadic and short packet uplink
transmissions are well studied, while not enough attention is given to the
delivery of content of common interest, such as software/firmware updates
and remote control, towards IoT devices in emerging point-to-multipoint sce-
narios. Moreover, the delivery of delay-sensitive IoT traffic is not sufficiently
addressed in the literature. In this work we (i) identify the drawbacks of the
current Single-Cell Point-to-Multipoint (SC-PTM) solution for unplanned
critical traffic delivery in cellular IoT (cIoT) networks, and (ii) propose pag-
ing and multicast schemes for a fast distribution of critical updates after,
e.g., bug fixes or system failures. We benchmark the performance of the pro-
posed paging scheme against similar solutions available in the literature. Our
extended SC-PTM framework is energy efficient and guarantees low service
latency, as demonstrated both analytically and by simulations.

Keywords: IoT, MTC, 5G, point-to-multipoint, MBMS, SC-PTM, paging,
energy efficiency, multicast.

*Corresponding author
Email addresses: olga.vikhrova@unirc.it (Olga Vikhrova),
sara.pizzi@unirc.it (Sara Pizzi), antonella.molinaro@unirc.it (Antonella
Molinaro), antonio.iera@dimes.unical.it (Antonio Iera), samuylov-ke@rudn.ru
(Konstantin Samuylov), araniti@unirc.it (Giuseppe Araniti)

1. Introduction

Fifth-generation (5G) networks are expected to connect a huge number of heterogeneous devices. Differently from previous generations of cellular networks, 5G strongly focuses on massive Machine-Type Communications (MTC) and Internet of Things (IoT), addressing both massive MTC (mMTC) and Ultra-Reliable and Low-Latency Communication (URLLC) use cases [1]. Many of the emerging IoT use cases move the focus from sporadic data transmissions in the uplink (UL) direction – such as smart gas-metering devices that wake up once a day to send the consumption reports to the gas-metering network – to simultaneous data delivery from network to multiple receivers in the downlink (DL). The latter case includes software/firmware updates, system configuration changes, and remote device control [2].

Point-to-Multipoint (PTM) communication is the key technology in such scenarios, because of its capability to feed a theoretically unlimited number of devices in a single transmission [3, 4]. The 3rd Generation Partnership Project (3GPP) specified the subscription-based Multimedia Broadcast Multicast Service (MBMS) architecture to provide a way for the network to deliver the content of interest towards multiple receivers over a large number of cells [5]. Successively, the Single-Cell Point-to-Multipoint (SC-PTM) operation mode was introduced in Release 13 to support multicast data delivery in a single cell. In Release 14, it was enabled for Narrowband IoT (NB-IoT) and Long Term Evolution for Machines (LTE-M), which are recognized as 5G solutions that meet technical requirements of large-scale mMTC scenarios [1] and ensure coexistence with the 5G New Radio (NR) [6].

In conventional multicast scenarios, devices create a *multicast group* by subscribing to the content of interest and wait for the service announcement when the content is available for download. The service announcement stage usually runs for a long time to ensure that all devices in the group get ready for the content reception when multicast transmission starts.

In this paper, we focus on the challenging use case of a critical update dissemination towards a large number of IoT devices as a consequence of critical bug fixes or system reconfiguration because of a failure. Since devices are not aware of message arrival, network needs to send a paging message first to notify them of incoming data. The multicast group can not be created in advance and multicast transmissions can not be scheduled as in the example above because the critical content must be delivered to IoT devices as soon as possible.

1.1. Related work and contribution of this paper

The need for a customer-driven group formation for PTM services in cellular IoT (cIoT) has been early discussed in [3]. MTC devices usually operate in a limited regime to save battery, they sporadically wake up to perform routine tasks and upload only few bytes to an application server. The eMBMS is a Human-Type Communication (HTC) oriented technology that assumes all end-points being under human control, which is not the case of MTC.

The trade-off between device availability for network-originated data and device energy consumption is well covered in the literature. For instance, the work in [7] discusses the impact of device active and sleep periods on the expected battery life cycle. In [8], device energy consumption under different active and sleep intervals and when varying traffic rates is analyzed by assuming unicast DL transmissions. The results demonstrate that both very short and very long intervals between paging indication and DL traffic arrival lead to an increase in device energy consumption. Similar results have been reported in [9] for more types of traffic and use cases. Device grouping is exploited in [10] to improve the energy consumption of IoT devices with similar UL traffic pattern and Quality of Service (QoS). The grouping algorithm helps to avoid congestion in the UL when a huge number of devices try to access the network after receiving paging indication. However, mentioned works are mainly focused on the issue of *paging*, either to improve long-term device energy consumption with regular traffic or, alternatively, to reduce device collision rate in the UL. In our proposal, we address both paging and multicast traffic delivery aspects.

In a previous work [11], we proposed three different strategies to group IoT devices for the reception of multicast traffic. The first strategy is meant to group all relevant devices into a single group and schedules SC-PTM transmission when the last device of the group enters the Radio Resource Control (RRC) connected state joining the multicast group. According to the second strategy, devices are split into multicast groups of equal size; connected devices wait for the SC-PTM transmission until the group is formed. In the last strategy, we proposed to schedule identical multicast transmissions any moment when devices are ready for the data reception, i.e. any number of devices may fall into the multicast group. We considered only legacy paging strategy, defined by 3GPP, to notify devices of the multicast service; according to it, not more than 16 devices can be reached by one paging transmission [12].

In [13], we discussed the necessary improvements of the SC-PTM service announcement and proposed a new grouping solution for the multicast reception of critical content, considering the drawbacks of the strategies from [11]. In the new strategy, the network schedules SC-PTM transmissions in a fixed interval named *critical interval*. However, we did not discuss how this interval should be adjusted. We extended the analysis with two enhanced paging strategies from the reference literature, namely *Group paging* (GP) [14], which allows addressing any number of devices in one paging message, and *enhanced Group paging* (eGP) [15] where paging is sent out over fixed intervals to a group of devices.

Solution for paging in [16] improves device's battery life cycle at the expense of a very long service delay that is unacceptable for critical applications. Authors in [17] obtained the optimal size for a paging group based on the limited capacity of the *Random Access* (RA) followed by paging. However, none of the mentioned works, except for [15], takes into account the impact of paging on multicast efficiency. For this reason, we propose a new paging solution that leaves from the general idea of the paging approaches proposed in [15] and [17], but reinforces our SC-PTM transmission scheme for the delay critical IoT applications.

Before us, authors in [18] analysed the performance of the firmware updates over unicast and PTM links for NB-IoT. The work [19] deals with the resource allocation problem for the multicast transmission in the presence of unicast traffic. Both works lack an analytic approach and solutions for paging, which are contributions of our work. Our paging and device grouping solutions have been evaluated analytically and validated by extensive simulations.

The main contributions of this work are:

– A multicast framework for critical cIoT services that helps to avoid long legacy service announcement procedure, efficiently pages devices and schedules SC-PTM transmissions.

– A new paging strategy that properly adjusts the paging interval and size of the paging groups to improve the probability of content reception and reduce delay of SC-PTM services.

– An analytical framework that accurately models all the phases involved in SC-PTM service provision, such as paging, system configuration for the SC-PTM reception and multicast transmission itself.

– An extensive numerical analysis with device and network oriented metrics and different payloads of the multicast traffic that may represent very short commands, alerts and small bug fixes.

We also discuss minor but necessary changes in some messages of the RA stage, not addressed in [13].

The rest of the paper is organized as follows. In section 2, we give the background on paging and RA procedures and explain the necessary changes for SC-PTM to make delivery of critical traffic in cIoT feasible. The details of our proposal are given in section 3, while numerical results are discussed in section 4. Conclusive remarks are given in the last section.

2. Setting the scene

2.1. Paging and Random access procedures

The individual activity pattern of cellular IoT devices is determined by their duty cycle, alternating short *connected* and long *idle* periods. Therefore, *paging* is needed to notify the arrival of DL data when device is in idle mode. The duration of active and idle intervals is defined by the discontinuous reception (DRX) strategy. In 3GPP Release 13, an extended DRX (eDRX) strategy has been introduced, which, compared to the Power Saving Mode (PSM), allows IoT devices to remain idle for longer period, save energy, and improve their response time in applications with network-originated traffic.

After an inactivity period since the last transmission, defined by the *Inactivity Timer*, the device turns the receiver circuitry off and only periodically listens to the Paging Radio Network Temporary Identifier (P-RNTI) indication in the Physical Downlink Control Channel (PDCCH). Note that LTE-M and NB-IoT use ad-hoc designed MTC PDCCH (MPDCCH) and Narrowband PDCCH (NPDCCH) [20]. For the sake of brevity, we omit to specify the exact name of the different physical LTE-M and NB-IoT channels. It wakes up for the *onDuration* time to receive the paging message and to look up for its identifier (ID) in the *paging records* list. If the device finds the appropriate record then it follows the instruction from the paging message, otherwise it turns back to sleep [21].

Two parameters help to define when a device is available for a PTM service: the Paging Frame (PF) and the Paging Opportunity (PO) indicating the radio frame and subframe when the device must listen to the paging

indication in the PDCCH. For NB-IoT devices, the concept of Paging Narrowband (PNB) replaces PO to indicate not only the subframe but also the narrowband where paging indication can be received. For simplicity, we refer to PO only, including also PNB in this term. Network can address several devices at a time if they listen to the same PO at the same PF including their IDs into the paging record list. However, the number of paging records in one message is limited [12]. Alternatively, it may address devices by their Group ID (GID) [14] if assigned previously.
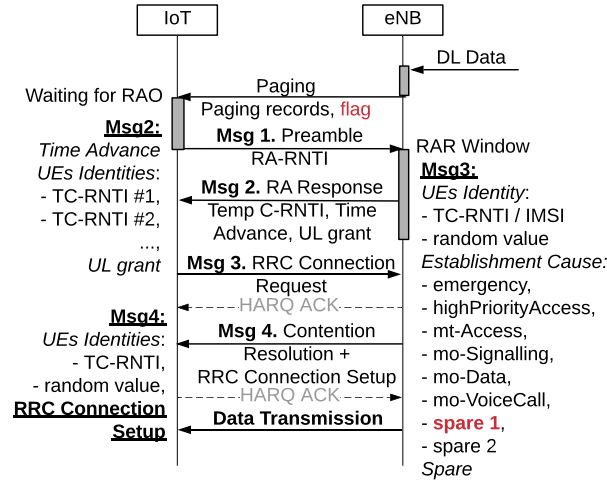

Figure 1: 3GPP complaint RA procedure.

Once devices are awake, they are immediately ready for the data reception, but they are unable to transmit. In order to send a feedback to the network, devices must synchronize with the BS and request resource allocation for the subsequent transmission in uplink. Without feedback from devices, the BS has to continuously broadcast data at the lowest rate over a long period of time to ensure that all devices get the content.

To synchronize with the BS, a device initiates the RA procedure, as illustrated in Fig. 1, by sending a randomly chosen preamble (*Msg1*) over the physical random access channel (PRACH) scheduled at specific random access opportunities (RAOs), defined by the PRACH configuration index. If the BS successfully decodes Msg1, then it replies with the RA response (RAR) message (*Msg2*), including the Temporary Cell-Radio Network Temporary Identifier (TC-RNTI), the timing advance information for synchro-

6

nization purpose, and a UL grant for the next message transmission in the physical uplink shared channel (PUSCH). Then the device sends the Radio Resource Control (RRC) connection request (*Msg3*) and specify the *Establishment Cause*. If the BS decodes Msg3 it replies with Contention Resolution message (*Msg4*) using identifiers from the Msg3. If both TC-RNTI and UE Identity equal to the TC-RNTI and UE Identity that the device included in Msg3, the RA stage is successfully completed.

Preamble retransmissions can be triggered due to the lack of resources for Msg2 transmission or due to collisions upon Msg3 transmission. These retransmissions are the events that contribute most to the access delay and may cause device access failure. After a failed RA attempt, the device waits for a backoff interval and then retries with a preamble transmission. When the maximum number of retransmissions is reached, the device is considered to be unable to connect to the network due to poor link conditions, and may go back to the idle mode.

2.2. Multicast Framework for critical IoT applications

SC-PTM reuses the MBMS architecture but utilizes supplementary radio bearer service. SC-PTM control and data are transferred in the dedicated Single-Cell Multicast Control Channel (SC-MCCH) and Single-Cell Multicast Traffic Channel (SC-MTCH) respectively. These two channels dynamically are mapped to the Physical Downlink Shared Channel (PDSCH) with prior indication in the PDCCH [5], [22]. Each multicast session has a unique Temporary Mobile Group Identity (TMGI) in core and radio access segments. Similar to paging, SC-PTM control and traffic transmissions are indicated by SC-PTM RNTI (SC-RNTI) and Group-RNTI (G-RNTI) in DCI respectively. Once a device gets TMGI, G-RNTI and scheduling information for the SC-PTM transmission (i.e., scheduling period, scheduling window and start offset), it can receive the content, as shown in the Fig. 2.

3GPP-based SC-PTM for cIoT is only supported in idle mode. To this end, a new System Information Block Type 20 (SIB-20) message was introduced to carry the scheduling information for one SC-MCCH per cell, that contains scheduling information for one SC-MTCH per each multicast service. When a new SC-PTM service becomes available in a cell, SC-MCCH is changed, therefore devices have to read SIB-20 to update the SC-MCCH. To inform devices about the changes in the SIB-20 network needs to broadcast SIB-1 messages (Option A in Fig. 2).
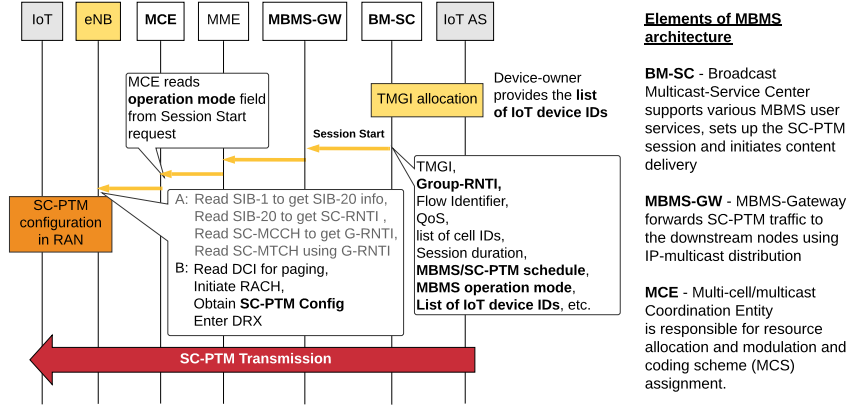
Figure 2: Standard (Option A) and proposed (Option B) scheme to deliver SC-PTM traffic towards cIoT devices.
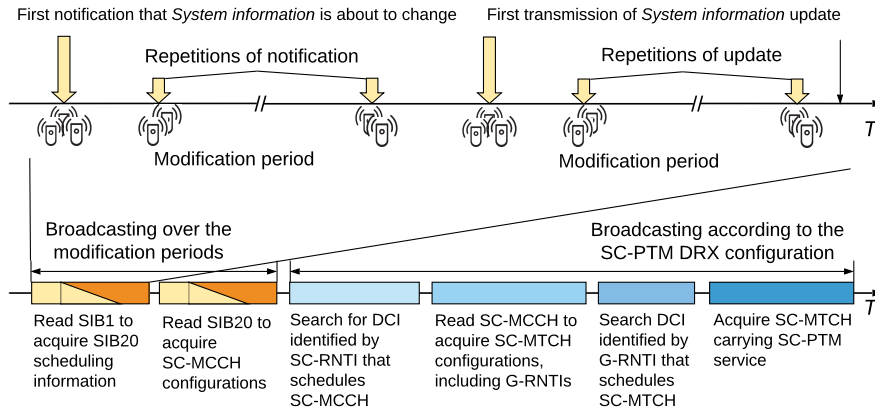

Figure 3: Delay of the standard SC-PTM transmission.

The transmission of one SIB message takes 64 frames or 640 ms [20]. Notifications of SIB changes apply the concept of *modification period*. It means that the system information content is not supposed to change within a modification period, and the same information can be repeated within a modification period. In the next modification period, the content is allowed to change. Hence, during the first modification period, the BS informs devices that the information is about to change, but the updated information itself is transmitted only in the next modification period, as shown in Fig. 3.

As we discussed in our previous work [13], the payload of critical IoT applications is relatively small and content must be delivered to devices with a minimal delay. The wait-for-all approach fails to fit such a requirement when the number of involved devices is high. We propose to send paging messages to small subgroups of devices and schedule multicast transmission in a short interval after paging as illustrated in Fig. 4.
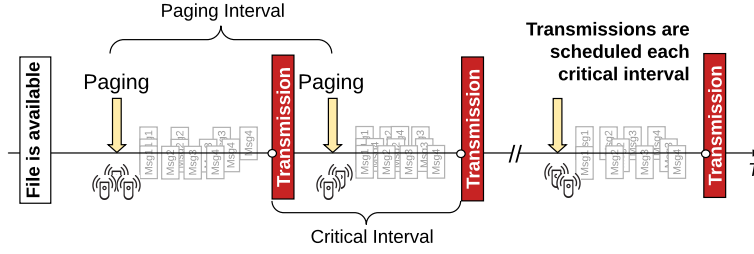

Figure 4: Paging and Multiple-subgroups Multicast Transmissions.

Upon receiving the list of relevant IoT devices, the network starts paging. All successfully paged devices have to initiate the RA procedure. The SC-PTM configuration information will be piggybacked on the Msg4 replacing the *RRC Connection Setup/Resume* message. Fig. 5 illustrates the necessary modifications to the paging message and to the 3GPP compliant RA procedure to enable the proposed solution. A *flag* in the paging message should be set to 1 to inform devices of the SC-PTM related paging. To emphasize that the SC-PTM configuration is requested, also Msg3 is extended to let device specify a new establishment cause in the corresponding spare field of Msg3 that we define as *mt-Multicast*. When carrying SC-PTM configuration in Msg4, IoT devices benefit from the hybrid automatic repeat request (HARQ) mechanism that improve the reliability of the multicast service. However, the RA stage could be a bottleneck. Paging a large number of IoT devices may

9

cause preamble retransmissions due to the limited opportunities for sending *Msg2*, and may delay the RA completion. The less devices complete RA before the next scheduled SC-PTM transmission, the less devices join the multicast group. When the multicast subgroups are small, radio spectrum is not efficiently utilized and the total SC-PTM service delay increases.
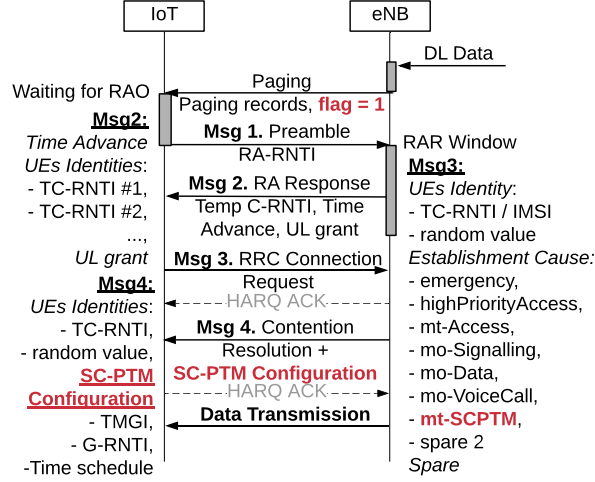

Figure 5: Enhanced RA procedure for the group-based critical communication.

We propose to page a relatively small number of IoT devices to ensure that all of them complete the RA stage before the SC-PTM transmission. Moreover, the next group of devices is paged only at the end of the RA stage of the previous group. The interval between two successive SC-PTM transmissions depends on the expected access delay and SC-PTM transmission delay. More details are given in the next section.

3. System Model

We consider a single-cell scenario with N uniformly distributed devices. Let us define a *virtual frame (VF)* composed of T_{VF} subframes as the time interval between two successive RAOs. The system time T is slotted into $I = \lceil T/T_{VF} \rceil$ VFs, where $\mathcal{I} = \{1, \dots, I\}$ denotes VF indexes. We assume that each VF has one PO and one RAO, as illustrated in Fig. 6.

Let Q denote the number of paging subgroups, $\mathcal{Q} = \{1, \dots, Q\}$. If paging subgroup $q \in \mathcal{Q}$ has n_q devices, then $n_1 + \dots + n_Q = N$ and $n_q \leq N_j$, where
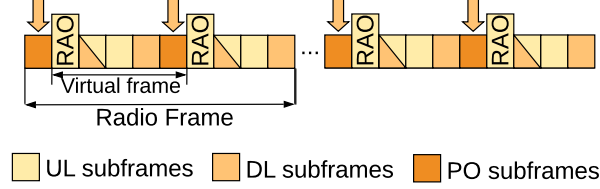
Figure 6: System time model.

$j \in \mathcal{J}$ denotes one of the paging schemes under consideration.

Let $\mathbf{P} = (\vec{P}_1, \ldots, \vec{P}_I)^T$ be the paging matrix composed of vectors $\vec{P}_i = (p_{i,q})_{i \in \mathcal{I}, q \in \mathcal{Q}}$, whose element $p_{i,q}$ denotes the number of devices in the paging subgroup q at the VF i. For a paging scheme $j \in \mathcal{J}$, we define $\mathcal{I}_j \subset \mathcal{I}$ as the subset of VF indexes in which paging messages should be sent. In particular, $\mathcal{J} = \{SP, GP, eGP, NeGP\}$. For the SP scheme, $\lceil N/N_{SP} \rceil$ VFs carry paging messages, where $N_{SP} = 16$ and paging interval is equal to one VF, therefore $\mathcal{I}_{SP} = (1, 2, \ldots, \lceil N/N_{SP} \rceil)$. According to the GP scheme, all $N_{GP} = N$ devices can be reached by one paging message [14], so $\mathcal{I}_{GP}$ consists of only one element. The eGP scheme claims that a new paging group ($N_{eGP} = 36$) can be formed every $T_{eGP} = 30$ ms, i.e., every $i_{eGP} = \lceil T_{eGP}/T_{VF} \rceil$ VFs, thus $\mathcal{I}_{eGP} = \{1, 1 + i_{eGP}, \ldots, 1 + ([N/N_{eGP}] - 1)i_{eGP}\}$.

In our proposed NeGP, we define $\mathcal{I}_{NeGP}$ by taking into account the RA and SC-PTM transmission delays. Specifically, F VFs are needed to complete the 4-message handshake for the RA when N_{NeGP} devices contend at the preamble transmission stage. Then, let W denote the number of VFs required for the SC-PTM transmission. Thus, a new group of devices can be paged every paging interval $T_{NeGP} = (F + W) \cdot T_{VF} = i_{NeGP} \cdot T_{VF}$ ms, and $\mathcal{I}_{NeGP} = \{1, 1 + i_{NeGP}, \ldots, 1 + ([N/N_{NeGP}] - 1)i_{NeGP}\}$. The optimal number of devices in a paging group is equal to the maximum number of devices that can be acknowledged in *Msg2* during the RAR window, i.e. $N_{NeGP} = N_{RAR}$. By considering the RA control overhead of $\sigma = 30\%$ and the RAR message format [22], the maximum number of devices that can be acknowledged during the RAR window is computed as $N_{RAR} = [(1-\sigma)D_0]\lceil T_{RAR}/T_{VF} \rceil$, where D_0 is the number of RBs available for the DL transmission in a VF, and T_{RAR} the RAR window duration. For a given system configuration $D_0 = 12$ and $T_{RAR} = T_{VF}$, which yields $N_{NeGP} = 8$.

An IoT device that receives a paging message in VF i initiates the RA

at the same VF. If the first RA attempt fails, the device may take up to R attempts, $\mathcal{R} = \{1, \ldots, R+1\}$. Let vector $\vec{\alpha}_{i,r}$ denote the number of devices having the RA attempt r in VF i, where $i \in \mathcal{I}$, $r \in \mathcal{R}$.

When devices make the first RA attempt, i.e. $r = 1$,

$$\vec{\alpha}_{i,1} = \vec{P}_i, i \in \mathcal{I}. \tag{1}$$

The total number α_i of devices having Msg1 transmission in VF i can be obtained as follows:

$$\alpha_i = \left(\sum_{r=1}^{R} (\vec{\alpha}_{i,r}) \right) \cdot \mathbf{1}, i \in \mathcal{I}, \tag{2}$$

where $\mathbf{1} = (1,1,\ldots,1)^T$, $|\mathbf{1}| = Q$.

The random access to C preambles by α_i devices is an instance of the occupancy problem. The probability to pick a preamble by a device from C available preambles is equal to $1/C$. If α_i devices contend at VF i the probability $q_i(c)$ of using exactly c out of C preambles at least by one device can be given as in [23]:

$$q_i(c) = \binom{C}{C-c} \sum_{j=0}^{c} (-1)^j \binom{c}{j} \left(1 - \frac{C-c+j}{C} \right)^{\alpha_i}. \tag{3}$$

The expected number of used preambles C_i in VF i, $i \in \mathcal{I}$, can be calculated as follows:

$$C_i = \left[\sum_{c=1}^{C_i^*} c q_i(c) \Big/ \sum_{c=1}^{C_i^*} q_i(c) \right] \tag{4}$$

where $C_i^* = \min(C, \alpha_i)$. We normalize $\sum_{c=1}^{C_i^*} c q_i(c)$ because the sum of probabilities $q_i(c)$ for $c = \{1, \ldots, C_i^*\}$ does not hold 1 when the number of contending devices α_i is less than C. The probability p_i of choosing a unique preamble in VF i depends on the number of contending devices α_i:

$$p_i = \left(1 - \frac{1}{C} \right)^{\alpha_i - 1}, i \in \mathcal{I}. \tag{5}$$

Collided devices which have received the same UL grant in Msg2 collide again in Msg3 transmission and can repeat the RA attempt after the Contention Resolution Time (CRT) window expiration. We denote $M = \lceil T_{CRT}/T_{VF} \rceil$ as the CRT window T_{CRT} in number of VFs.

The expected number of contending devices in VF i is the total number of devices that make the first RA attempt after paging, devices that failed to receive Msg2, and devices that collided at step 3 of the RA procedure. Let $\vec{\alpha}_{i,r}^{*}$ denote the number of devices that successfully received Msg2 in VF i after r RA attempts. Vectors $\vec{\beta}_{i,r}$ and $\vec{\beta}_{i,r}^{*}$ stand for the number of devices scheduled for the Msg3 transmission in VF i and for the number of devices that successfully sent Msg3 in VF i after r RA attempts, respectively. Finally, let $\vec{\gamma}_{i,r,m}$ denote the number of devices that receive Msg4 in VF i after m VFs of the contention resolution time and r RA attempts, while $\vec{\gamma}_{i,r,m}^{*}$ stands for number of devices that successfully received Msg4 in VF i.

Devices that failed the RA attempt retry after the back-off window (BW) T_{BW} or j VFs, $j = \overline{1,B}$, where $B = \lceil T_{BW}/T_{VF} \rceil$. Let $\varphi_j = 1/B$ be the probability of randomly choosing the back-off time. The expected number of devices contending in VF i yields:

$$\vec{\alpha}_{i,r} = H[i-1]\left(\vec{\gamma}_{i-1,r-1,M} - \vec{\gamma}_{i-1,r-1,M}^{*}\right) + H[i-k-M]p_{i-k-M} \cdot \vec{\beta}_{i-M,r-1} +$$

$$+ \sum_{j=1}^{B} H[i-j-1]\left(\vec{\alpha}_{i-j-1,r-1} - \vec{\alpha}_{i-j-1,r-1}^{*}\right)\varphi_j, i \in \mathcal{I}, r \in \mathcal{R}, j = \overline{1,B} \qquad (6)$$

where $H[x]$ is a Heaviside function; it equals to 1 if $x > 0$ and takes 0 if $x \leq 0$.

The BS needs T_{RA} ms to detect and decode transmitted preambles before sending Msg2. Thus, a device waits for $k = \lceil ((A-1)T_{VF} + T_{RA})/T_{VF}A \rceil$ VFs for the Msg2 reception. Let N_{RAR} denote the system capacity for Msg2 transmissions in numbers of preambles that can be acknowledged by the BS. If devices contending in VF $(i-k)$ used less than N_{RAR} preambles, then all devices receive Msg2. Otherwise, only a portion of them receives Msg2, that is given as follows:

$$\vec{\alpha}_{i,r}^{*} = \begin{cases} \vec{\alpha}_{i-k,r}, & C_{i-k} \leq N_{RAR} \\ \left[\vec{\alpha}_{i-k,r}N_{RAR}/C_{i-k}\right], & C_{i-k} > N_{RAR}. \end{cases} \qquad (7)$$

The expected number of devices to be scheduled for the Msg3 transmission in VF i can be given as follows:

$$\vec{\beta}_{i,r} = \vec{\alpha}_{i-1,r}^{*} + \left(\vec{\beta}_{i-1,r} - \vec{\beta}_{i-1,r}^{*}\right), \qquad (8)$$

where $\left(\vec{\beta}_{i-1,r} - \vec{\beta}_{i-1,r}^{*}\right)$ counts for the devices that failed to send Msg3 in VF $i-1$ due to the lack of UL resources.

Let U_0 be the total number of UL resources available in VF i. Since the PRACH occupies a fixed number U_P of RBs in the UL, the number of available UL resources in VF i for Msg3 transmission equals to $U_i = U_0 - U_P$. The expected number of devices scheduled for the Msg3 transmission in VF i can be given as follows:

$$\vec{\beta}_{i,r}^* = \begin{cases} \vec{\beta}_{i,r}, & \vec{\beta}_{i,r}\mathbf{u}^T \leq U_i \\ \left[\vec{\beta}_{i,r}U_i/\vec{\beta}_{i,r}\mathbf{u}^T \right], & otherwise, \end{cases} \tag{9}$$

where $\mathbf{u}^T$, $|\mathbf{u}| = Q$, denotes the average number of RBs required for the Msg3 transmission.

The expected number of devices to be scheduled for the Msg4 transmission in VF i is either the number of devices that successfully sent Msg3 in the previous VF or the number of devices that failed to receive Msg4 in the previous VF due to the lack of the DL resources:

$$\vec{\gamma}_{i,r,m} = \begin{cases} \vec{\beta}_{i-1,r}^*, & m = 1 \\ \vec{\gamma}_{i-1,r,m-1} - \vec{\gamma}_{i-1,r,m-1}^*, & otherwise. \end{cases} \tag{10}$$

where $i \in \mathcal{I}$, $r \in \mathcal{R} \setminus \{R+1\}$.

Let D_0 and D_{RAR} be the total number of DL resources available in VF i and the average number of resources required for the Msg2 transmission, respectively. The number of DL resources D_i after the Msg2 transmission can be calculated as:

$$D_i = \begin{cases} D_0 - D_{RAR}, & \left(\sum_{r=1}^{R} \vec{\beta}_{i,r} \right) \mathbf{1}^T > 0 \\ D_0, & otherwise. \end{cases} \tag{11}$$

Therefore, the expected number of devices that successfully sent Msg4 in VF i yields:

$$\vec{\gamma}_{i,r,m}^* = \begin{cases} \vec{\gamma}_{i,r,m}, & \vec{\gamma}_{i,r,m}\mathbf{d}^T \leq D_i \\ \left[\vec{\gamma}_{i,r,m}D_i/\vec{\gamma}_{i,r,m}\mathbf{d}^T \right], & otherwise, \end{cases} \tag{12}$$

where $\mathbf{d}^T$ denotes the average number of DL resources required for the Msg4 transmission, $|\mathbf{d}| = Q$.

After receiving Msg4 in VF i, devices can receive SC-PTM transmission scheduled in one of the next VFs. We assume that up to S multicast transmissions can be scheduled within I VFs, $\mathcal{S} = \{1, \ldots, S\}$. Let i_s be the first

VF of the SC-PTM transmission s. Then, the expected number $\vec{\delta_s}$ of devices ready for the SC-PTM transmission s yields:

$$\vec{\delta_s} = \sum_{k=i_{s-1}}^{i_s-1} \sum_{r=1}^{R} \sum_{m=1}^{M} \vec{\gamma}_{k,r,m}^{*}, s \in \mathcal{S}. \tag{13}$$

Let z define the critical interval between two successive SC-PTM transmissions. The first transmission should be scheduled with an offset to ensure that all devices of the first paging subgroup receive Msg4, while all next multicast transmissions are scheduled in z VFs.

Let Θ be the multicast payload in terms of resources needed for the SC-PTM transmission. The residual number of resources θ_{l_s} required to complete transmission s after the first $l_s - 1$ VFs is given as follows:

$$\theta_{l_s} = \begin{cases} \Theta, & l_s = 0 \\ \theta_{l_s-1} - D_{i_s^*+l_s}, & \theta_{l_s-1} > D_{i_s^*+l_s} \\ 0, & otherwise. \end{cases} \tag{14}$$

Let l_s^* stands for the last VF of the SC-PTM transmission s such that $\theta_{l_s^*} = 0$, i.e. denotes the duration of the SC-PTM transmission s. The expected number of devices $\vec{\delta_s^*}$ that successfully receive the multicast service after l_s^* VFs equals to $\vec{\delta_s}$. We now can calculate the metrics of interests.

Access success probability P_A is a ratio of the number of devices that completed the RA stage to the overall number of devices reached through paging

$$P_A = 1 - \left(\sum_{i=1}^{I} \vec{\alpha}_{i,R+1} \right) \mathbf{1}^T / \left(\sum_{i=1}^{I} \vec{\alpha}_{i,1} \right) \mathbf{1}^T. \tag{15}$$

Average access delay D_A corresponds to the time to complete the RA:

$$D_A = \frac{1}{Q} \sum_{q=1}^{Q} \left(i_q^* - i_q \right) T_{VF}, \tag{16}$$

where i_q stands for the VF at which group q receives paging and i_q^* is given as follows

$$i_q^* = \left[\left(\sum_{i=1}^{I} i \sum_{r=1}^{R} \sum_{m=1}^{M} \vec{\gamma}_{i,r,m}^{*} \right) \mathbf{e}_q^T / \left(\sum_{i=1}^{I} \vec{\alpha}_{i,1} \right) \mathbf{e}_q^T \right]. \tag{17}$$

15

Average idle delay D_{Idle} is the time that elapses from the end of the RA stage until the beginning of the multicast transmission, therefore

$$D_{Idle} = \frac{1}{Q} \sum_{q=1}^{Q} \left(i_q^{**} - i_q^* \right) T_{VF}. \tag{18}$$

where i_q^{**} is given as follows

$$i_q^{**} = \left[\sum_{s \in \mathcal{S}} i_s^* \left(\vec{\delta}_s \mathbf{e}_q^T \right) \bigg/ \left(\sum_{s \in \mathcal{S}} \vec{\delta}_s \right) \mathbf{e}_q^T \right] - 1 \tag{19}$$

because not all devices of the same paging subgroup will be members of the same multicast subgroup for the SC-PTM reception.

Average total delay D_{Total} includes the average access delay D_A, average idle delay D_{Idle}, and average SC-PTM transmission delay D_{TX}:

$$D_{Total} = D_A + D_{Idle} + D_{TX}, \tag{20}$$

where the average SC-PTM transmission delay can be computed as

$$D_{TX} = \frac{1}{S} \sum_{s=1}^{S} l_s^* \cdot T_{VF}. \tag{21}$$

Total service delay $D_{Service}$ is the total time to wake up all relevant devices and deliver the content of interest. Having i_{S^*} and l_{S^*} of the very last multicast transmission S^*, we compute the metric as follows

$$D_{Service} = (i_{S^*} + l_{S^*}) T_{VF}. \tag{22}$$

Average access energy consumption E_A can be given as an arithmetic mean of the average energy consumption per paging subgroup E_{A_q}:

$$E_A = \frac{1}{Q} \sum_{q=1}^{Q} E_{A_q}. \tag{23}$$

Let t_1, t_2, t_3 and t_4 be the average transmission delay of Msg1, Msg2, Msg3 and Msg4. The device energy consumption in transmission mode equals to

e_{TX} mW, in reception mode - e_{RX} mW, devices in idle mode consume e_{Idle} mW on average. In the access stage, devices of subgroup q consume:

$$E_{A_q} = (e_{TX}t_1 + e_{RX}t_2)r_q^2 + (e_{TX}t_1 + e_{RX}t_2 + e_{TX}t_3)(r_q^3 + 1) + \\ + e_{Idle}T_{BW}r_q^2 + e_{RX}t_4, \tag{24}$$

where r_q^2 and r_q^3 denote the average number of retransmission attempts due to failure after Msg2 and Msg3 transmission, respectively. The average number of RA attempts due to Msg2 or Msg3 failure is computed as the weighted mean:

$$r_q^2 = \frac{\left(\sum_{r=1}^{R} r \sum_{i=1}^{I} \left(\vec{\alpha}_{i,r} - \vec{\alpha}_{i,r}^*\right)\right) \mathbf{e}_q^T}{\left(\sum_{r=1}^{R} \sum_{i=1}^{I} \left(\vec{\alpha}_{i,r} - \vec{\alpha}_{i,r}^*\right)\right) \mathbf{e}_q^T}. \tag{25}$$

$$r_q^3 = \frac{\left(\sum_{r=1}^{R} r \sum_{i=1}^{I} \vec{\alpha}_{i,r}^s \left(1 - p_i\right)\right) \mathbf{e}_q^T}{\left(\sum_{r=1}^{R} \sum_{i=1}^{I} \vec{\alpha}_{i,r}^s \left(1 - p_i\right)\right) \mathbf{e}_q^T}. \tag{26}$$

Average device energy consumption is the total energy consumed during the access, idle and SC-PTM transmission stages by a device on average:

$$E_{Total} = (E_A + e_{Idle}D_{Idle} + e_{TX}D_{TX}). \tag{27}$$

Resource utilization R_{UL} and R_{DL} is the ratio between the number of occupied resources and the total number of available resources in I VFs in the UL and DL, respectively:

$$R_{UL} = 1 - \frac{\sum_{i=1}^{I} U_i}{IU_0}, \tag{28}$$

$$R_{DL} = 1 - \frac{\sum_{i=1}^{I} D_i}{ID_0}. \tag{29}$$

Table 1: Reference system model parameters

Notation	Definition	Value
C	Number of available preambles	54
R	Maximum number of preamble retransmissions	10
N_j	Paging group size, $j = \{SP, GP, eGP, NeGP\}$	$\{16, N, 36, 8\}$
T_j	Paging interval, $j = \{SP, GP, eGP, NeGP\}$	$\{5, 0, 30, 25\}$ ms
A	Number of RA subframes in a radio frame	2
d	Interval between two consecutive POs	5 ms
z	Critical interval	25 ms
T_{VF}	Virtual frame duration	5 ms
T_{RA}	Delay for the preamble detection and decoding	5 ms
T_{RAR}	RAR window	5 ms
T_{BW}	Back-off window	20 ms
T_{CRT}	Contention resolution time	48 ms
N_{RAR}	Number of devices that may receive RAR within T_{RAR}	8
U_0	Amount of resources available for the uplink transmission in each VF	12 RBs
U_P	Amount of resources occupied by PRACH in the UL	12 RBs
D_0	Amount of resources available for the downlink transmission in each VF	12 RBs
D_{RAR}	Amount of resources required for the RAR message transmission in DL VF i	6 RBs
$\mathbf{u}$	Vector of the average number of resources for Msg3 transmission	$(1,\ldots,1)$ RBs
$\mathbf{d}$	Vector of the average number of resources for Msg4 transmission	$(1,\ldots,1)$ RBs
Θ	Multicast traffic payload	$\{3,12,32\}$ RBs
e_{Tx}	Average device power consumption in the transmit mode	500 mW
e_{Rx}	Average device power consumption in the receive mode	80 mW
e_{Idle}	Average device power consumption in the idle mode	3 mW

4. Selected numerical results

We compare our paging solution, named *New enhanced Group Paging (NeGP)*, over three reference paging strategies, namely *Standard Paging (SP)* [12] (i.e. legacy 3GPP solution), *Group Paging (GP)* [14], and *enhanced Group Paging* (eGP)[15].

We consider a symmetric radio frame configuration (with the same number of UL and DL subframes) with $A = 2$ RAOs, as shown in Fig. 6. The mentioned paging strategies have different number of devices per paging subgroup and different paging intervals. For the reader's convenience, we give

Table 2: Simulation parameters

Parameter	Value
Cell radius	500 m
Carrier configuration	1.4 MHz carrier bandwidth at 800 MHz
PHY numerology	TDD frame type 1, TTI 1 ms
RA capacity	2 RAOs per radio frame
Resource allocation	PDSCH, PDCCH: $1 - 6$ PRBs
	PUSCH, PUCCH: $1 - 6$ PRB,
	PRACH: 6 PRBs
Device power class	23 dBm
BS transmit power	46 dBm
Power consumption	500 mW (TX), 80 mW (RX), 3mW (Idle)
Traffic payload	$\{392, 1608, 4584\}$ bits

definitions of the system model parameters and their corresponding values in Table 1. The analytic results have been validated by simulations in MAT-LAB. Simulation parameters are set according to [20] and [24], for radio interface, and to [25], for device energy consumption, as reported in Table 2. Data packets arriving in a burst of a given size are transmitted over a set of continuous subframes.

In the following figures, analytical results are shown as solid lines with markers, and simulation results only as markers; an almost perfect match is observed. Results are plotted for a cluster of up to 500 devices camping on a single LTE-M narrowband. As explained in [12], the device arrival rate of 40.3 access attempts per second with a target outage probability below 1% corresponds to the LTE-M traffic capacity per narrowband equaled to $0{,}36{\cdot}10^6$ devices/km^2, the higher capacity of 10^6 devices/km^2 can be achieved if three or more narrowbands are configured in a cell. Our NeGP paging solution allows 320 device arrivals per second with outage probability less than 1%, which ensures more than 10^6 device/km^2 of supported connection density.

Fig. 7 shows the average access delay (a) and average device energy consumption (b) for different paging strategies. The GP scheme introduces a significant delay and energy usage at the RA stage with respect to other schemes due to the high number of contending devices. For the SP and GP schemes both metrics grow almost linearly when the number of devices increases due to the preamble collisions and lack of radio resources. On the contrary, both metrics tend to saturate in the cases of the eGP and NeGP schemes. The eGP solution exploits the code-expanded preamble transmis-

19

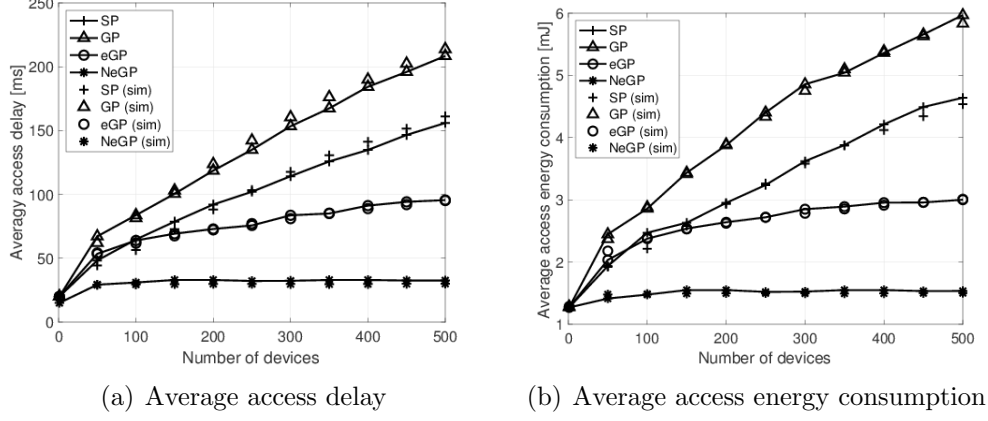(a) Average access delay   (b) Average access energy consumption

Figure 7: Average access delay and energy consumption.

sion technique that decreases collision rate and, consequently, the number of preamble retransmission attempts [15]. However, our NeGP solution shows more than 50% reduction of both the average access delay and the average device energy consumption compared to the eGP scheme. The reason behind such performance gain is that the size of the paging groups and paging intervals in NeGP are well customised in such a way that devices complete the RA without any additional delay caused by preamble collisions or shortage of the radio resources.

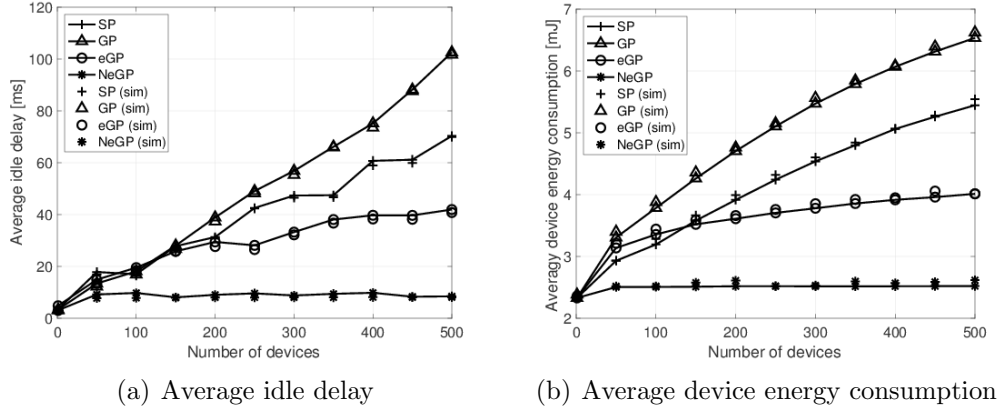(a) Average idle delay   (b) Average device energy consumption

Figure 8: Average idle delay and average device energy consumption.

Devices that complete the RA procedure remain in idle mode while wait-

ing for the SC-PTM transmission but keep listening to the DL since the last transmission until the end of the Inactivity timer defined by the DRX. If the timer expires before the SC-PTM transmission, devices switch off their receiving antenna and become unavailable until the next PO. Fig. 8(a) shows the average idle delay, i.e. the time to wait for the SC-PTM transmission after the reception of SC-PTM configuration parameters. The idle delay of the GP scheme grows fast under increasing number of devices. In the case of SP and eGP, the metric increases mainly due to the short paging interval or high number of devices per group. To ensure that all paged devices receive the multicast transmission, the Inactivity timer should be higher than the idle delay. Fig. 8(b) illustrates the average device energy consumption under the assumption that the Inactivity Timer is set according to the experienced idle delay. The metric constantly grows under GP, SP and eGP strategies but it is almost constant for the NeGP scheme. This is an important result for battery-powered IoT devices.

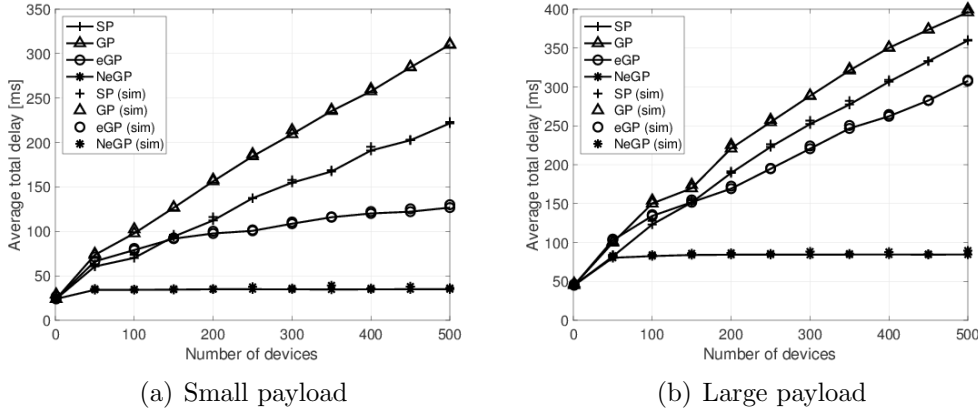(a) Small payload   (b) Large payload

Figure 9: Average total delay in case of: (a) small payload, and (b) large payload.

Fig. 9 shows the average total delay for the variable SC-PTM payload. In particular, the size is set to 392, and 4584 bits. For simplicity, we refer to these values as small (a) and large (b) payload, respectively. The total delay includes access delay, idle delay and the time to transmit SC-PTM payload. The system performance is sensitive to the payload size because long multicast transmissions may overlap with the RA stage. Our NeGP paging and SC-PTM transmission design has been designed in order to avoid such an overlapping. As shown in Fig. 9, the increase of SC-PTM payload

does not lead to the significant performance degradation in the case of NeGP and results only in an additional deterministic delay.

The access success probability is shown in Fig. 10(a). This metric also can be used as the *service probability* if necessary assumptions on the Inactivity Timer are made, as previously discussed. The failures are not only caused by preamble collisions but also by retransmissions after Msg2 and Msg3 failures. When the number of devices in the SP and GP schemes is increased not all devices can successfully complete the RA. For a cluster of 500 devices, from 5% to 10% of devices fail the RA in the case of SP and GP strategies. Very few devices lose the SC-PTM transmission if the eGP scheme is applied, while the NeGP guarantees the successful completion of the RA procedure by all devices.
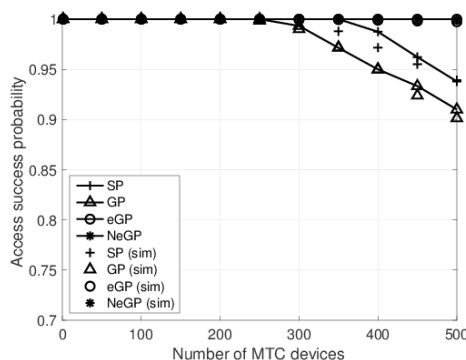

Figure 10: Access success probability.

We compare the performance of our proposal with reference schemes in terms of radio resource consumption in the UL and DL for different payloads as reported in Fig. 11. Regarding the UL utilization, the NeGP scheme requires less resources than SP, GP and eGP solutions, because it does not incur retransmissions of the RA messages. On the contrary, GP requires more UL resources than any other paging strategy due to the higher collision rate. Having more UL resources available is advantageous for the system that can support other background traffic (e.g., from other IoT devices). The DL resource utilization depends on the number of multicast transmissions required to service all relevant devices. As expected, the NeGP solution requires more DL resources because it induces more SC-PTM transmissions. The difference in required DL resources becomes more evident when the payload size is larger and more devices wait for the multicast service.
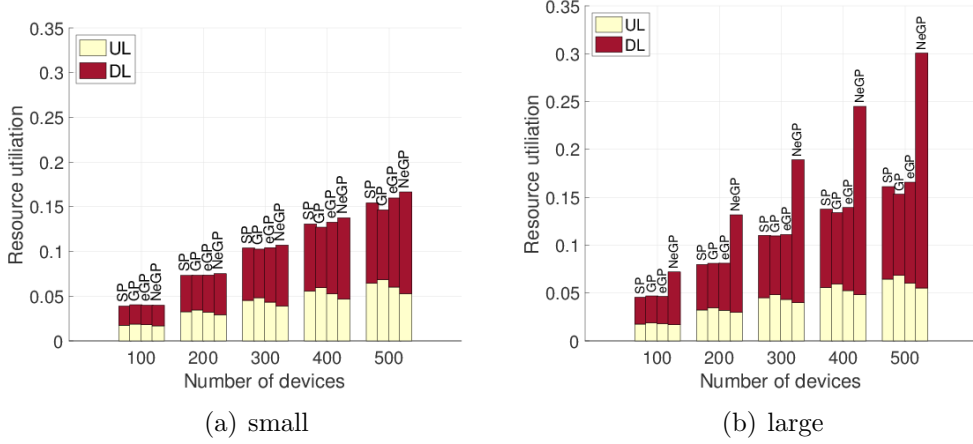
22

Figure 11: UL (yellow) and DL (red) resources utilization in the case of (a) small and (b) large payload.

5. Conclusions

In this paper, we investigated a wide set of performance metrics to evaluate the proposed multicast framework for the delivery of initially unplanned critical multicast traffic towards bandwidth- and power-limited cIoT devices. We proposed to schedule identical SC-PTM transmissions over an finely tuned interval to improve the service probability and reduce device energy consumption. We extensively compared our solution over similar reference schemes, both analytically and via simulations. We highlighted that paging significantly impacts the performance of critical SC-PTM communication when the arrival of multicast traffic can not be predicted. The optimal configuration of paging and SC-PTM scheduling guarantees 100% of the service delivery and stable device total delay irrespective of the number of receivers but at the expense of a long service delay. However, a short device total delay is more preferable than a short service delay in critical applications.

References

[1] ITU-R, "ITU-R M. Minimum Requirements Related to Technical Performance for IMT2020 Radio Interface(s)," 3rd Generation Partnership-Project (3GPP), Report M.2410-0, 2017.

[2] 3GPP, "Technical Specification Group Services and System Aspects; MBMS for IoT," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.850, 2018, version 16.0.0.

[3] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over Emerging 5G Networks: Challenges and Perspectives," IEEE Network, vol. 31, pp. 80–89, 2017.

[4] F. Rinaldi, S. Pizzi, A. Orsino, A. Iera, A. Molinaro, and G. Araniti, "A novel approach for MBSFN Area Formation aided by D2D Communications for eMBB Service Delivery in 5G NR Systems", IEEE Transactions on Vehicular Technology, vol. 69, No. 2, pp. 2058–2070, 2020.

[5] 3GPP, "Technical Specification Group Services and System Aspects; MBMS; Protocols and codecs," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.346, 2018, version 16.0.0

[6] GSMA, "Mobile IoT in the 5G Future - NB-IoT and LTE-M in the context of 5G," 2008. [Online]. Available: https://www.gsma.com/iot/wp-content/uploads/2018/05/GSMA-5G-Mobile-IoT.pdf

[7] H. Ferng and T. Wang, "Exploring Flexibility of DRX in LTE/LTE-A: Design of Dynamic and Adjustable DRX," IEEE Transactions on Mobile Computing, vol. 17, no. 1, pp. 99–112, 2018.

[8] S. Oh, K. Jung, M. Bae, and J. Shin, "Performance analysis for the battery consumption of the 3GPP NB-IoT device," in 2017 International Conference on Information and Communication Technology Convergence (ICTC), 2017, pp. 981–983.

[9] A. K. Sultania, P. Zand, C. Blondia, and J. Famaey, "Energy Modeling and Evaluation of NB-IoT with PSM and eDRX," in 2018 IEEEGlobecom Workshops (GC Wkshps), 2018, pp. 1–7.

[10] S. Xu, Y. Liu, and W. Zhang, "Grouping-Based Discontinuous Reception for Massive Narrowband Internet of Things Systems," IEEE Internet of Things Journal, vol. 5, no. 3, pp. 1561–1571, 2018.

[11] O. Vikhrova, S. Pizzi, A. Molinaro, K. Samouylov, and G. Araniti, "Group-Oriented Services for Critical Machine Type Communications in 5G Networks," in 2018 IEEE 29th Annual International Symposium

on Personal, Indoor and Mobile Radio Communications (PIMRC), 2018, pp. 824–828.

[12] O. Liberg, M. Sundberg, Y.-P. E. Wang, J. Bergman, and J. Sach, "Cellular Internet of things: technologies, standards, and performance", Academic Pres, 2018.

[13] O. Vikhrova, S. Pizzi, A. Iera, A. Molinaro, K. Samuylov, and G. Araniti, "Performance Analysis of Paging Strategies and Data Delivery Approaches for Supporting Group-Oriented IoT Traffic in 5G Networks," in 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2019, pp. 1–5.

[14] C. Wei, R. Cheng, and S. Tsao, "Performance Analysis of Group Paging for Machine-Type Communications in LTE Networks," IEEE Transactions on Vehicular Technology, vol. 62, no. 7, pp. 3371–3382, 2013.

[15] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IoT Machine Age With 5G: Machine-Type Multicast Services for Innovative Real-Time Applications," IEEE Access, vol. 4, pp. 5555–5569, 2016.

[16] E. Kurniawan, P. H. Tan, K. Adachi, and S. Sun, "Hybrid Group Paging for Massive Machine-Type Communications in LTE Networks," in GLOBECOM 2017 - 2017 IEEE Global Communications Conference, 2017, pp. 1–6.

[17] 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. 34, no. 5, pp. 1086–1102, 2016.

[18] L. Feltrin, G. Tsoukaneri, M. Condoluci, C. Buratti, T. Mahmoodi, M. Dohler, and R. Verdone, "Narrowband IoT: A Survey on Downlink and Uplink Perspectives," IEEE Wireless Communications, vol. 26,no. 1, pp. 78–86, 2019.

[19] G. Tsoukaneri, M. Condoluci, T. Mahmoodi, M. Dohler, and M. K. Marina, "Group Communications in Narrowband-IoT: Architecture, Procedures, and Evaluation," IEEE Internet of Things Journal, vol. 5, no. 3, pp. 1539–1549, 2018.

[20] 3GPP, "Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 45.820, 2014, version13.1.0.

[21] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.304, 2018, version15.0.0.

[22] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); RadioResource Control (RRC); Protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.331, 2018, version 15.0.0.

[23] V. Savaux, A. Kountouris, Y. Loüet, and C. Moy, "Modeling of Time and Frequency Random Access Network and Throughput Capacity Analysis," EAI Endorsed Transactions on Cognitive Communications, vol. 3,no. 11, p. e2, 2017.

[24] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physicallayer procedures," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.213, 2019, version 15.7.0.

[25] R. Ratasuk, N. Mangalvedhe, D. Bhatoolaul, and A. Ghosh, "LTE-MEvolution Towards 5G Massive MTC," in 2017 IEEE Globecom Workshops (GC Wkshps), 2017, pp. 1–6.