Optimal Group Paging Frequency for Machine-to-Machine Communications in LTE Networks With Contention Resolution

Wen Zhan, *Student Member, IEEE*, Xinghua Sun, *Member, IEEE*, Yitong Li, Feng Tian, *Member, IEEE*, and Hong Wang

***Abstract*—Group paging is a baseline solution proposed by the long-term evolution (LTE) standardization body for supporting machine-to-machine (M2M) communications in the current-generation and the next-generation cellular networks. Yet, in conventional group paging scheme, upon the reception of paging message, all machine-type devices (MTDs) in a group will simultaneously access the base station, leading to severe network congestion and intolerably low access efficiency. To handle this issue, in this article, we propose a dynamic group paging mechanism, where only the MTDs with packets to send will join the contention process, and the collisions in the random access channel are addressed by the contention resolution scheme. Explicit expressions of key performance measures including the mean access delay of each MTD are derived as functions of the length of waiting period (interval between two consecutive paging periods) T_W, where a smaller T_W indicates a higher frequency of group paging. It is shown that T_W is a key system parameter that determines the crucial tradeoff between the signaling overheads of the system during the paging period and the access delay performance of each MTD. To study how to properly tune the waiting period length, a utility-based analytical framework is established by taking the aforementioned tradeoff into consideration. The optimal waiting period length for maximizing the network utility is derived and verified by simulation results. The analysis in this article reveals that the network should increase the group paging frequency as the traffic becomes heavier or the number of preambles decreases. Providing more preambles can indeed improve the delay performance, while the gain becomes marginal if the number of preambles is large.**

***Index Terms*—Contention resolution, group paging, machine-to-machine (M2M) communications, random access.**

Manuscript received April 2, 2019; revised July 26, 2019; accepted August 26, 2019. Date of publication September 5, 2019; date of current version December 11, 2019. The work of X. Sun was supported in part by the Open Project of the Key Laboratory of Wireless Sensor Network and Communication, Chinese Academy of Sciences under Grant 20190901, and in part by the Fundamental Research Funds for the Central Universities. The work of Y. Li was supported by the National Natural Science Foundation of China under Grant 61801433. The work of F. Tian and H. Wang was supported in part by the National Natural Science Foundation of China under Grant 61772287 and Grant 61801246, in part by the Key University of Science Research Project of Jiangsu Province under Grant 18KJA510004, and in part by the Open Research Foundation of National Mobile Communications Research Laboratory of Southeast University under Grant 2018D09. *(Corresponding author: Xinghua Sun.)*

W. Zhan is with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong (e-mail: wzhan5-c@my.cityu.edu.hk).

X. Sun is with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510006, China, and also with the Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China (e-mail: xinghua.sun@ieee.org).

Y. Li is with the School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: ieytli@zzu.edu.cn).

F. Tian is with the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: tianf@njupt.edu.cn).

H. Wang is with the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: wanghong@njupt.edu.cn).

Digital Object Identifier 10.1109/JIOT.2019.2939667

I. Introduction

WIRELESS communications technologies are now evolving toward a better support for the upcoming era of Internet-of-Things (IoT), where everything that worth to be connected will be connected. One key enabling technology for IoT is the machine-to-machine (M2M) communications, which usually involves a large number of machine-type devices (MTDs) that operate without human intervention. M2M communications gives rise to many emerging applications, such as smart retail, smart agriculture, and intelligent transportation systems [1]. With the booming M2M market, it is expected that the number of MTDs will reach 14 billion by 2022 [2].

The third generation partnership project (3GPP) has defined M2M communications as an important service type to be supported by the long-term evolution (LTE) networks. By virtue of the LTE network coverage at a global scale, MTDs can be placed almost everywhere and maintain the connection to the rest of the world, which fully unleashes the potentials of M2M visions. However, the current cellular system will be unlikely to cope with the expected growth of M2M services. The key challenge originates from the radio access network, a part of the cellular system, which was designed for providing access to a relatively small number of devices that deliver a significant amount of data. In M2M context, instead, a huge number of MTDs request access but transmit small packets. The risk is then the severe congestion occurs at the random access channel in the radio access network when massive MTDs transmit access requests simultaneously [3]. This congestion may cause intolerable delay, packet loss, or even service unavailability for most of MTDs.

A. Overload Control Schemes

How to improve the access efficiency and control the overload at the random access channel has been one of the main

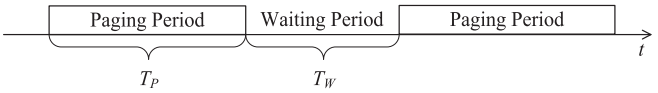


Fig. 1. Graphic illustration of the group paging.

issues that capture the attention from 3GPP. Various standard solutions have been proposed to handle the overload, which can be broadly divided into two categories [4]: 1) push-based schemes and 2) pull-based schemes. In push-based schemes, each MTD itself determines when to access the networks. In pull-based schemes, the network notifies MTDs to perform random access.

Backoff schemes, such as access class barring (ACB) or uniform backoff (UB), are examples of push-based schemes. In the ACB scheme, an ACB factor q is broadcasted by the evolved node B (eNB), and can be regarded as the access request transmission probability. In the UB scheme, once an MTD fails in the request transmission, it randomly selects a value from $\{0, 1, \ldots, W_s\}$, and counts down, where W_s is the UB window size in unit of milliseconds. The MTD retries the request transmission when the backoff counter reaches zero. It can be seen that by assigning a smaller ACB factor q or a larger UB window size W_s, bursty access requests could be dispersed over the time domain to reduce the number of concurrent requests, which may improve the access efficiency. It has long been observed that the effectiveness is crucially determined by the backoff parameters, including q and W_s [5]. To properly tune backoff parameters, lots of efforts have been devoted to developing algorithms to estimate traffic load on the channel [6]–[10]. Yet, tracking and estimating the time-varying number of access requests can be highly challenging and demanding for practical system.

Group paging is the typical representative of pull-based schemes. In group paging, MTDs are labeled with group identifiers according to various metrics such as service types, quality of service (QoS) requirements. MTDs with the same group identifier constitute a group. When the eNB sends a paging message enclosed with the corresponding group identifier, MTDs in this group will immediately initiate the random access procedure and send access requests during a period of time, which is dubbed paging period. Comparing to the ACB or UB scheme, group paging enables a more flexible control of M2M traffic [4], where the eNB defines a period of time, during which MTDs are permitted to access. Therefore, the group paging scheme is suitable for the applications, where the remote server requires the information collection periodically. When one paging process terminates, the system may initiate the next paging process after a period of time, if needed. Accordingly, the group paging process follows an on–off manner, as shown in Fig. 1. The waiting period denotes the interval between two neighboring paging periods, in which the MTDs cannot perform access request transmission. In this article, we denote the length of the waiting period as T_W. It can be seen that a smaller T_W indicates that the system performs the group paging more frequently.

Extensive works have been done to evaluate and improve the group paging performance of massive M2M access in the LTE networks. Most of them put the major focus on a singleton paging period by assuming that the number of MTDs involving the contention in this period is known. In [11], by recursively calculating the number of the current contending devices based on the observations in the previous time slots, an analytical framework was proposed to derive various performance measures, such as the access delay and the utilization ratio of random access resources. It reveals that the access efficiency may dramatically decrease when the number of MTDs being paged increases.

To improve the random access performance, in [12]–[14], clustering techniques were adopted to formulate hierarchical network architecture, in which a few MTDs, called cluster head, were appointed as relays for the remaining MTDs. The network congestion can be alleviated as the number of access requests to the eNB is limited to the number of cluster heads [3]. However, extra signaling overheads are induced in intracluster/intercluster coordination and demanding requirements on device functionalities, e.g., energy and buffer size, have to be imposed on cluster heads. Therefore, most of the related works have their roots in the network scenario in which all MTDs communicate directly with the eNB. For instance, in [15], a consecutive paging method was devised, where multiple consecutive paging periods are scheduled for one paging occasion. In [16]–[20], backoff schemes were adopted to scatter the access requests over one period to reduce the number of concurrent requests [16]–[18], [20], or to provide service differentiation [19]. To improve resource utilization ratio, in [21], the network assigns dedicated random access resources to MTDs. In [22], a dynamic resource allocation scheme was proposed to adaptively adjust the amount of resources for group paging based on the estimated number of MTDs in each time unit.

The development in the above studies has been substantial. Yet, the access performance degradation in group paging is still observed when the number of MTDs being paged is excessively large. The fundamental reason lies in that upon receiving the paging message, all MTDs in the group are required to initial the random access procedure and transmit access requests. As the group size could be large, severe congestion and intolerably low-access efficiency become inevitable. In practice, however, for many M2M use-cases such as information collection, the MTDs may have no data to report when they receive the paging message, yet, are still required to access. In this case, redundant access requests are generated, which leads to the congestion. Regarding this issue, in this article, we will consider a dynamic scenario where the MTD sends the access request only if it has packets to deliver upon the reception of the paging message. For many M2M applications, the packet arrival at each MTDs is a stochastic process, implying that the MTD may have an empty data queue when the eNB starts the paging process. Due to the randomness of the packet arrival, the eNB therefore does not know in advance the number of MTDs that has packets at the beginning of each paging round.

B. Contention Resolution Scheme

Note that although the conventional group paging mechanism can improve the access efficiency, it still relies on backoff schemes for spreading the access requests over time domain to avoid collision. This approach falls short of providing good performance under heavy-load case and is prone to suffer from congestion [5], [23]. To further improve the access performance, a promising strategy that has gained wide attention is the contention resolution scheme, which is also referred to as the tree-splitting algorithm.

Specifically, in contention resolution scheme, once a collision occurs, the eNB will attempt to resolve the collision actively by assigning new contention resources for those MTDs that are involved in the collision. By using the ternary feedback messages, their retransmissions will be coordinated in the newly assigned resource space. This process repeats until there is no collision and all those MTDs' requests are successfully transmitted. Comparing to the backoff schemes, where traffic estimation algorithms are required for adaptively tuning backoff parameters, the system with contention resolution scheme no longer needs to do so, and the performance, e.g., access delay, can still outperform a large body of backoff schemes [24].

The contention resolution scheme has a long line of research. It has been shown that in the case of Poisson arrivals, the maximum throughput of binary contention resolution scheme is 0.347 packets/slot [25]. To further improve the performance, [26] suggests that multiple resource units should be assigned at the beginning of the resolution process. By doing so, the maximum throughput is boosted to 0.43 packets/slot. Recently, the contention resolution scheme has been applied to addressing the congestion issue at the LTE random access channel for massive M2M access. Simulation results in [27] have shown that the contention resolution scheme can greatly improve the delay performance, comparing to the standard approaches such as the UB mechanism. In [24] and [28], prebackoff and dynamic resource allocation strategies were combined to further enhance the performance. In [23], [29], and [30], the distributed queueing mechanism was introduced, where collisions are addressed by the contention resolution scheme first, and then, data packets are transmitted in a contention free manner. Extensive results in above studies have validated that with contention resolution scheme, the network can achieve superior throughput and delay performance even in the massive access scenarios.

C. Contributions

In this article, we will enhance the access performance of MTDs in the LTE group paging by using the contention resolution scheme. That is, upon the reception of the paging message, a large number of MTDs may initiate the random access procedure simultaneously. To handle the massive concurrent access requests, the contention resolution scheme is then applied to accelerate the service process until all requests are successfully accommodated.

It is worth mentioning that to reduce the number of concurrent access requests, in this article, we will consider a dynamic case, where the MTD will join the contention process upon receiving the paging message only if it has packets to deliver.[1] Accordingly, the number of MTDs that join each paging period becomes a time-varying variable, which does not known in advance by the eNB. It is in sharp contrast to the conventional group paging mechanism, where such number is equal to the group size and is a known value.

Note that no matter in the dynamic case or the conventional case, how frequent should the group paging be performed is always a crucial issue that determines the network performance. Intuitively, if the eNB performs the group paging more often, i.e., a smaller waiting period length T_W, then, the mean access delay of each MTD can be effectively reduced as there are more frequent access opportunities. On the other hand, it will consume more system overheads as for each paging period, the eNB has to deliver a certain amount of signaling messages in coordinating the access and reserve the access resource. Therefore, there exists a tradeoff between the delay performance of each MTD and the signaling overhead of the system. Yet, it has long been ignored in existing literature, as most of them only consider one singleton paging period [11], [15]–[20], [22].

In this article, we will take such tradeoff into consideration and study how to tune the group paging frequency so that a proper balance between the delay performance of each MTD and the signaling overhead of the system can be achieved. Note that the study becomes especially challenging in the dynamic case, where the number of MTDs that join each paging period is no longer a deterministic value but a time-varying variable. Nevertheless, by focusing on the long-run network behavior, we are capable of deriving various performance measures, with which a utility-based analytical framework is further established for determining the optimal paging frequency.

The contribution of this article is summarized as follows.

1) A dynamic group paging scheme is proposed, in which upon the reception of the paging message, only the MTDs with nonempty data buffer would access the eNB.
2) Different from many existing solutions that try to avoid collisions by using backoff scheme, we use the contention resolution scheme to resolve synchronous access attempts during the LTE group paging process. The implementation details of the contention resolution scheme on the top of the existing LTE random access procedure is also presented.
3) We focus on the long-run access performance and derive key system measures, including the average number of MTDs that join the contention in each paging period and the mean access delay of each MTD, as explicit functions of the waiting period length T_W and the traffic arrival rate.
4) A utility-based analytical framework is established, with the signaling overhead of the eNB during the paging period and the mean access delay of each MTD taking

[1] According to the LTE standard [31], for an MTD, when the paging message is received at its radio resource control (RRC) layer, this message will be reported to upper layers, e.g., application layer, in which we assume that the MTD would further check its data buffer and perform the random access procedure only if the buffer is nonempty.

into consideration. The optimal waiting period length T_W^* for maximizing the network utility is obtained and validated by simulation results.

The rest of this article is organized as follows. Section II presents the system model and the implementation details of the contention resolution schemes in the LTE group paging process. Key performance measures and the optimal waiting period length T_W^* with one single preamble are characterized in Sections III and IV, respectively, and extended to the multipreamble scenario in Section V. Finally, concluding remarks are summarized in Section VI.

II. System Model

Consider a single cell LTE network and one paging group. We focus on a data collection scenario, in which the eNB initiates the paging process for collecting information from MTDs [33]. Upon the reception of paging message, each MTD checks its data buffer and only the MTDs that have data packets will perform the random access procedure to establish the connection with the eNB for data reporting. It is in sharp contrast to the conventional group paging scheme considered in existing literature, where all MTDs have to access the eNB upon the reception of paging message. Assume that the group size, i.e., the total number of MTDs in this group, is very large and the arrival process of MTDs that have data packets follows Poisson distribution with parameter λ.

In the random access procedure, each MTD randomly selects one out of M orthogonal preambles and transmits it to the eNB. The access request transmission is successful if and only if there is no concurrent transmission of a given preamble at the same time. Otherwise, a collision happens and all of them fail. Note that as MTDs contend with each other only when they choose the same preamble, in the following, we start from the single-preamble scenario, where all MTDs share one preamble, i.e., $M = 1$. The analysis will be extended to multipreamble scenario in Section V.

A. Contention Resolution-Based Random Access

To efficiently accommodate the massive access during the paging period, the contention resolution scheme[2] is used. Specifically, the eNB organizes the access request transmissions into frames, and each frame contains two slots, where the MTD can transmit its access request. Each MTD then randomly selects one slot from the frame and transmits its access request. If more than one MTDs transmit in the same time slot, then collision occurs. The involved MTDs could be aware of the failure based on the feedback from the eNB upon the completion of the frame. For each collision slot, the eNB assigns a new frame. Accordingly, if all two slots in the frame encounter collisions, then two new frames will be allocated. This leads to the formation of a tree with node degree being two in each level. The expansion of the tree stops at either empty slot, i.e., no MTD transmits access request, or a successful slot, i.e., only one MTD transmits its access request. The tree is completed when all MTDs successfully transmit their access requests.

[2]In this article, we consider the classic two-ary tree-splitting scheme for simplicity. With minor modification, the analysis can be extended to other variants of two-ary tree-splitting scheme [34].

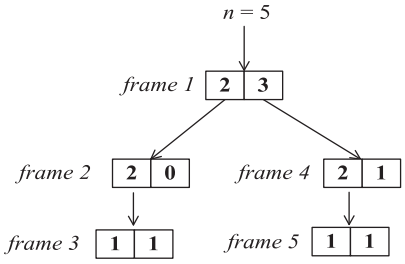


Fig. 2. Example of the contention resolution scheme with $n = 5$ MTDs. The number in each slot denotes the number of MTDs involving the contention. If this number is larger than 1 (a collision), then a new frame is assigned to those MTDs. If this number is 0 (an empty slot) or 1 (a successful transmission), the expansion stops.

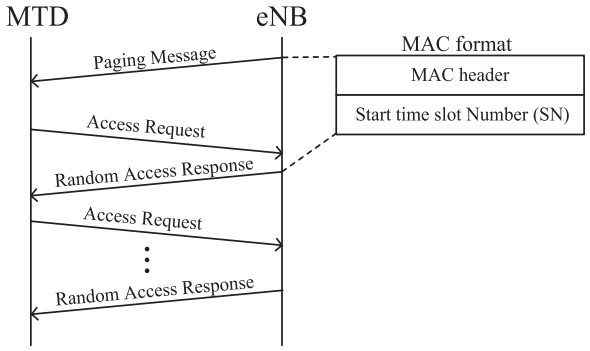


Fig. 3. System signaling procedures of the contention resolution-based LTE group paging.

A graphical example of a contention tree for $n = 5$ MTDs is presented in Fig. 2. It can be seen that the contention resolution process naturally forms a tree structure and the tree solves all n MTDs' access request transmissions at frame 5.

B. Implementation to the LTE Group Paging

In the following, we elaborate on how the contention resolution scheme can be implemented on the top of existing LTE paging and random access scheme.

Specifically, suppose that the eNB has broadcasted the paging message, which should contain not only the identifier of the desired group but also the resource allocation information for the first frame. In Fig. 3, a possible MAC format of the paging message is presented. Specifically, the MAC header part represents the typical paging message defined in the standard, enclosing the related group identifier, through which an MTD can know if this message is sent for itself or for other groups. Besides, one more information element, i.e., start time slot number (SN), is attached to identify the first random access slot that the eNB allocates for the first frame.

Upon receiving the paging message, all MTDs with nonempty data buffer in the paging group initiate the random access procedure, send their access requests to the eNB via the randomly selected slots in the first frame, and wait for the random access response (RAR). The eNB replies the RAR only at the end of each frame to acknowledge successfully transmitted access requests and assign new frames for collision slots.

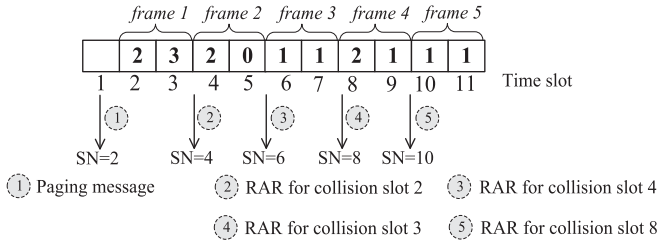


Fig. 4. Illustration of contention resolution based on the example in Fig. 2.

The MAC format of RAR is same to that for paging message, as shown in Fig. 3, except that the MAC header part is the typical RAR format defined in the standard and the SN is attached to identify new frame resources for collision slot. The access request and RAR exchanges between the eNB and an MTD come to an end when the access request is successful. If all MTDs have successfully transmitted their access requests, then, the group paging procedure is completed.

For a better demonstration, following the example in Fig. 2, we show how the contention resolution scheme operates in the LTE group paging process in Fig. 4. Specifically, as Fig. 4 illustrates, at the beginning, the paging message, which is sent in the first slot, assigns slot 2 to 3 to the first frame, i.e., SN = 2. In frame 1, 5 MTDs join the contention, inducing two collision events in slot 2 and 3, respectively. At the end of the first frame, the eNB assigns one new frame, i.e., frame 2 with SN = 4, for the collision in slot 2. On the other hand, the involved MTDs in the slot 3 at the first frame do not receive either the successful acknowledgment or newly assigned frame, and can therefore be aware of the collision and wait for the RAR that will be broadcasted by the eNB in the future. We can see from Figs. 2 and 4 that with two more frames, i.e., frames 2 and 3, the eNB addresses the collision at slot 2 in frame 1. Afterward, at the end of frame 3, the eNB replies the RAR, enclosing a new frame, i.e., frame 4, to the involved MTDs at collision slot 3 in frame 1, and continuously serve them until service completion at frame 5.[3]

Finally, we assume that for each paging period, the eNB just serves the access requests that join the contention at the beginning of this period. That is, no new requests will join the contention during the paging period. For those access requests that arrive during the paging period, they will be served in the next paging period. Accordingly, let n_{k+1} denote the number of MTDs join the $k+1$th paging period, where $k = 1, 2, \ldots$, and then, those n_{k+1} MTDs should arrival at the system either in the kth paging period or in the kth waiting period. Due to the randomness of the arrival process, the number of MTDs join each paging period is a random variable, which is crucially determined by the length of the waiting period T_W. Intuitively, with a larger T_W, more MTDs can be served in one paging period, which, however, sacrifices the delay performance. Below, we will first study the delay performance of MTDs and then investigate how to properly select the length of the waiting period T_W.

[3]The contention resolution process in this article is developed based on the depth-first-search tree-splitting scheme. The network may also use width-first-search tree-splitting scheme, with which access requests could be served in different chronological order.

III. Performance Analysis

In this section, we derive key performance measures, including the average number of MTDs that join the contention in each paging period $E[n]$, the average length of the paging period $E[T_P]$, and the mean access delay of each MTD $E[D]$, where the access delay is defined as the time spent from the arrival of the access request until its successful transmission.

A. Performance Measures

Let us first focus on the average number of MTDs that will join the contention in each paging period $E[n]$ and the average length of the paging period $E[T_P]$. Note that given the arrival rate λ, the waiting period length T_W and the length of the kth paging period $T_{P,k}$, the average number of MTDs that join the contention in the $k+1$th paging period can be written as

$$E\big[n_{k+1}\big] = \lambda\big(T_W + T_{P,k}\big). \tag{1}$$

On the other hand, given the number of MTDs that will join the $k+1$th group paging n_{k+1}, it has been shown in [25] that the average number of frames to complete their services is given by $[(n_{k+1})/\ln 2] - 1$. As each frame contains two slots, the average length of the $k+1$th paging period can then be written as

$$E\big[T_{P,k+1}\big] = 2\big(\tfrac{n_{k+1}}{\ln 2} - 1\big) + 1. \tag{2}$$

Suppose that the system will be in equilibrium in long run, that is, the number of MTDs with nonempty data buffer in the network does not grow unboundedly with time. Accordingly, the average number of MTDs that will join the contention in each paging period $E[n]$ is the long-run (steady-state) expected value of n_k, and the average length of the paging period $E[T_P]$ is the long-run expected value of $T_{P,k}$. By combining (1) and (2), we can have

$$E[n] = \tfrac{\lambda \ln 2(T_W - 1)}{\ln 2 - 2\lambda} \tag{3}$$

and

$$E[T_P] = \tfrac{2\lambda T_W - \ln 2}{\ln 2 - 2\lambda}. \tag{4}$$

Moreover, the access delay of each MTD is the length of time period from the arrival to the successful access request transmission. Recall that the MTDs arrival at the system either in the kth paging period or in the kth waiting period are served in the $k+1$th paging period. Accordingly, the access delay of each MTD should contain two parts. The first part is the length of time period from the arrival to the beginning of the next paging period. As MTDs arrivals uniformly in each paging period and waiting period. The average length of the first part can be written as $[(E[T_P] + T_W)/2]$. The second part is the length of time period from the beginning of the paging period until the access request is served. The length of this period is proportional to the length of paging period T_P, and the ratio is $\theta \approx 0.5$ [24]. Thus, by combining (3) and (4), the mean access delay of each MTD $E[D]$ can be obtained as

$$E[D] = \tfrac{E[T_P] + T_W}{2} + \theta E[T_P] = \tfrac{T_W}{2} + \Big(\tfrac{1}{2} + \theta\Big)\Big(\tfrac{2\lambda T_W - \ln 2}{\ln 2 - 2\lambda}\Big). \tag{5}$$

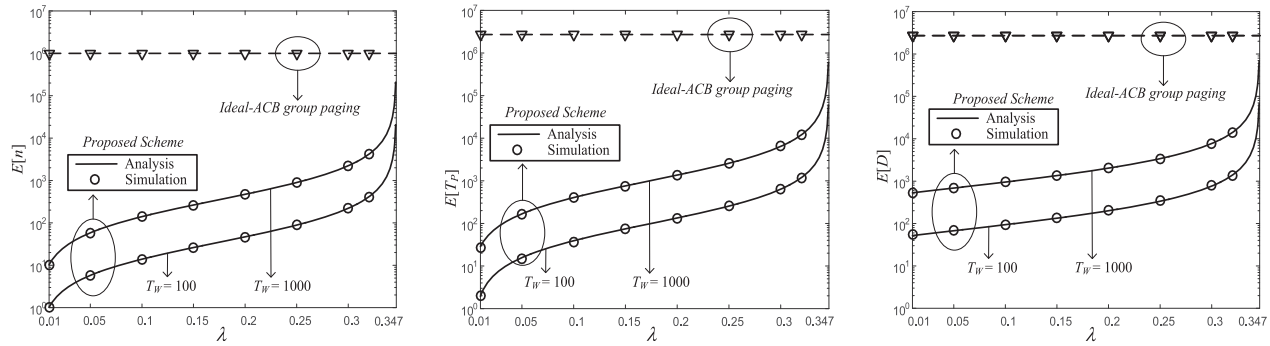


Fig. 5. Average number of MTDs that join the contention in each paging period $E[n]$, the average length of the paging period $E[T_P]$ (in unit of slots), and the mean access delay of each MTD $E[D]$ (in unit of slots) versus input rate λ with the proposed group paging scheme and the Ideal-ACB group paging scheme. $T_W \in \{100, 1000\}$. (a) $E[n]$ versus λ. (b) $E[T_P]$ versus λ. (c) $E[D]$ versus λ.

B. Simulation Results and Discussion

In this section, simulation results are presented to validate the analysis above and demonstrate the performance of the proposed group paging scheme by comparing it with that of the conventional group paging schemes.

The simulation setting is the same as the system model described in Section II, and we omit the details here. Each simulation is carried out for 10^3 waiting and paging rounds. In each paging period, we count the number of MTDs in each paging period, the length of each paging period, and the access delay of each MTD. Accordingly, the average number of MTDs that join the contention in each paging period is obtained as the ratio of the sum of number of MTDs in each paging period to 10^3, the average length of the paging period is obtained as the ratio of the sum of lengths of all paging periods to 10^3, the mean access delay is obtained as the ratio of the sum of access delay of all MTDs to the total number of MTDs in those 10^3 waiting and paging rounds.

For the performance comparison with existing literature, we focus on a large body of related works [16]–[20], in which backoff schemes are adopted. The backoff schemes in the LTE networks include the UB scheme and the ACB scheme. Without loss of generality, we consider the ACB-based group paging scheme as the benchmark, in which after receiving the paging message from the eNB, each MTD would attempt to access according to an ACB factor (i.e., the transmission probability of access request of each MTD). In simulations of this scheme, we assume that the total number of MTDs in the group is 10^6, and the network can always optimally tune the ACB factor and refer to such case as *Ideal-ACB group paging*.[4] In simulations of the Ideal-ACB group paging scheme, we also count the number of MTDs in each paging period, the length of each paging period, and the access delay of each MTD, and then obtain the corresponding performance measures.

[4]Note that it has been shown that if the number of MTDs attempting to access the eNB at time slot $t = 1, 2, \ldots$, denoted by n_t, is known, then, the optimal ACB factor for minimizing the access delay is given by $(1/n_t)$ [32]. However, due to the uncoordinated nature of MTDs, the real-time information of n_t is unavailable in practice. Therefore, the tuning of the ACB factor is inevitably suboptimal.

Simulation results of the Ideal-ACB group paging scheme and the proposed group paging scheme are presented in Fig. 5, which demonstrates how the average number of MTDs that join the contention in each paging period $E[n]$, the average length of the paging period $E[T_P]$, and the mean access delay of each MTD $E[D]$ vary with the input rate λ with the length of waiting period $T_W = 100$ or 1000. We can see from Fig. 5(a) that with Ideal-ACB group paging scheme, the number of MTDs that join the contention in each paging period is fixed at 10^6, regardless of the length of the waiting period T_W. Note that in conventional group paging scheme, upon the reception of the paging message, all MTDs in the group will access the eNB. Therefore, the number of MTDs that join the contention in each paging period is always same as the group size. On the other hand, in the proposed group paging scheme, only the MTDs with nonempty data queue would join the contention. As less MTDs join the contention, we can see from Fig. 5(b) and (c) that with the proposed group paging scheme, the average length of the paging period $E[T_P]$, and the mean access delay of each access request $E[D]$ are always smaller than those with Ideal-ACB group paging scheme. The performance gap between the proposed group paging scheme and the Ideal-ACB group paging scheme is significant especially when the data traffic input rate is small.

Note that the analytical results of $E[n]$, $E[T_P]$, and $E[D]$ of the proposed scheme are given in (3), (4) and (5), respectively, and are also shown in Fig. 5. A perfect match between simulation results and the analysis can be observed. We can see that for the proposed group paging scheme, as λ grows, the average number of MTDs in each paging period $E[n]$, the average length of each paging period $E[T_P]$, and the mean access delay of each MTD $E[D]$ increase accordingly. In particular, with heavy traffic, e.g., $\lambda \to (\ln 2/2) \approx 0.347$, we can observe from (3)–(5) and Fig. 5 that $E[n]$, $E[T_P]$, and $E[D]$ will tend to infinity, which implies that when $\lambda \geq 0.347$, the number of MTDs with nonempty data buffer in the network grows unboundedly and the system becomes unstable.

On the other hand, with light traffic and a short waiting period, e.g., $\lambda = 0.01$ and $T_W = 100$, as we can see from Fig. 5(a) and (b), both the average number of MTDs in each paging period $E[n]$ and the average length of each

paging period $E[T_P]$ are close to 1, indicating that in this case, performing group paging in a frequent manner could be ineffective as only a few MTDs are served in each paging period while the network has to consume a certain amount of signaling overheads. To improve the efficiency, intuitively, the network can increase the length of waiting period T_W, which, however, results in a larger mean access delay of each MTD, as shown in Fig. 5(c). It can therefore be seen that the length of waiting period T_W is a key system parameter that needs to be carefully selected.

IV. Utility Formulation and Optimization

In this section, we will investigate how to properly tune the length of waiting period T_W. To be more specific, based on the performance measures derived above, we formulate the network utility by considering the signaling overhead of the eNB during the paging period and the mean access delay of each MTD, and then maximize the network utility by properly tuning T_W.

Let us first focus on the signaling overhead of the eNB. Specifically, in a paging period, signaling messages that the eNB uses for coordinating the transmissions of MTDs during the paging process include the paging message and feedbacks. That is, at the beginning of the paging process, one paging message is sent, and at the end of each contention resolution frame, the eNB replies a feedback notifying the successful access, possibly a new assignment of a frame, etc. As the example in Fig. 4 indicates, the total number of signaling messages is proportional to the length of the paging period. Therefore, the total cost for sending signaling messages depends on the length of the paging period, which can be written as $c_{\text{eNB}}E[T_P]$, where c_{eNB} is the cost of sending one signaling message, i.e., the paging message or feedback. On the other hand, before the group paging, the eNB has to reserve certain amount of preambles for random access. Let K denotes the cost for reserving one preamble and we assume $K \geq c_{\text{eNB}}$. Accordingly, the average utility of the eNB is given by

$$U_{\text{eNB}} = -\frac{MK + c_{\text{eNB}}E[T_P]}{E[T_P] + T_W} \tag{6}$$

where M represents the number of preambles. In this section, we let $M = 1$, and the multipreamble case $M > 1$ will be considered in Section V.

It is worth mentioning that in this article, we consider the average utility in unit of time slots, i.e., the total utility in each waiting and paging round over the average length of one waiting and paging round, since the length of each paging period is a random variable that affects the total utility. Thus, by dividing the average length of one waiting and paging round, i.e., $E[T_P] + T_W$, the average utility qualifies as a fair criterion for comparing the utilities with different values of T_W.

Moreover, let c_{MTD} denotes the cost for an MTD to be delayed for one time slot, and r denotes the income for successfully serving one MTD. The average utility of all MTDs in one paging cycle is then given by

$$U_{\text{MTD}} = \frac{(r - c_{\text{MTD}}E[D])E[n]}{E[T_P] + T_W}. \tag{7}$$

By combining (3)–(8), the network utility can be written as

$$\begin{aligned} U &= U_{\text{eNB}} + U_{\text{MTD}} \\ &= \frac{2\lambda K - 2\lambda c_{\text{eNB}} T_W + (c_{\text{eNB}} - K)\ln 2}{(T_W - 1)\ln 2} \\ &\quad + \frac{\lambda c_{\text{MTD}}((4\theta\lambda + \ln 2)T_W - (2\theta + 1)\ln 2)}{4\lambda - 2\ln 2} + r\lambda. \end{aligned} \tag{8}$$

Typically, the fixed cost K, the delay cost c_{MTD}, the signaling cost c_{eNB}, the income unit r, and the input rate λ are system input parameters. Therefore, in this article, we aim to maximize the network utility U by optimally choosing the waiting period length T_W, i.e.,

$$\begin{aligned} \underset{T_W \geq 1}{\text{maximize}} \quad & U \\ \text{subject to} \quad & E[n] \geq 1 \\ & E[T_P] \geq 1 \end{aligned} \tag{9}$$

where the constraints are the average number of MTDs in each paging period and the average length of each paging period should be no smaller than one.

A. Optimal Length of the Waiting Period T_W^*

The following theorem presents the solution, i.e., the optimal length of the waiting period T_W^*, to the above optimization problem.

Theorem 1: The optimal length of the waiting period is given by

$$T_W^* = \begin{cases} \frac{\ln 2 - 2\lambda + \lambda \ln 2}{\lambda \ln 2} & \text{if } \lambda < \lambda_0 \\ 1 + (\ln 2 - 2\lambda)\beta & \text{otherwise} \end{cases} \tag{10}$$

where $\beta = \sqrt{\frac{2(K - c_{\text{eNB}})}{c_{\text{MTD}}\lambda(4\theta\lambda + \ln 2)\ln 2}}$ and $\lambda_0 = \frac{c_{\text{MTD}} \ln 2}{2(K \ln 2 - c_{\text{eNB}} \ln 2 - 2c_{\text{MTD}}\theta)}$.

Proof: See the Appendix. ∎

It is interesting to see from (10) that the optimal waiting period length T_W^* is independent of the income for successfully serving one MTD r. Intuitively, when the network is stable, i.e., $\lambda < (\ln 2/2)$, the input rate equals departure rate, regardless of the waiting period length T_W. Thus, according to (8), we can observe that the average revenue that the system receives in one paging round is given by $r\lambda$, which is solely determined by λ and does not relate to T_W.

However, the waiting period length T_W indeed affects the cost of eNB and MTDs in each paging round and therefore should be carefully selected. Fig. 6(a) illustrates the optimal waiting period length T_W^* under various values of the fixed cost K, the delay cost c_{MTD}, the signaling cost c_{eNB}, and the input rate λ. Specifically, we can see from Fig. 6(a) that T_W^* is a monotonic increasing function of K and decreasing function of c_{MTD} and c_{eNB}. It is clear that as the fixed cost for initialing one paging period K grows, the system should reduce the paging frequency to decrease the average cost on reserving preambles for the group paging; on the other hand, if the cost for an MTD to be delayed for one time slot c_{MTD} or the signaling cost c_{eNB} increases, the system should perform group paging more frequently to reduce the mean access delay of each MTD $E[D]$ and the average length of the paging period $E[T_P]$, so that the total cost for MTDs to be delayed, i.e., $E[D]E[n]c_{\text{MTD}}$, and the total cost for sending signaling messages in each paging period, i.e., $c_{\text{eNB}}E[T_P]$, would be

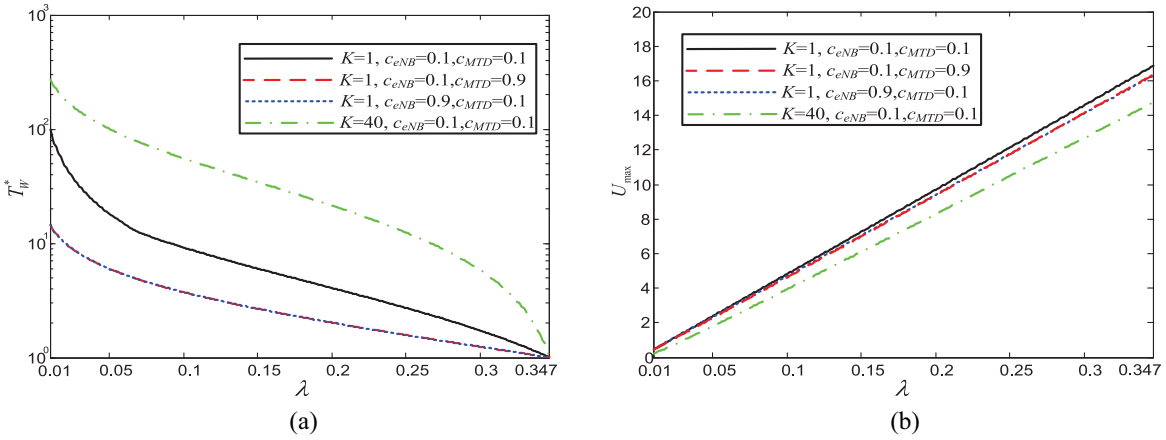


Fig. 6. Optimal waiting period length T_W^* (in unit of slots) and maximum network utility $U_{\max}$ versus the input rate λ. $r = 50$. $c_{\text{eNB}} \in \{0.1, 0.9\}$. $c_{\text{MTD}} \in \{0.1, 0.9\}$. $K \in \{1, 40\}$. $M = 1$. (a) T_W^* versus λ. (b) $U_{\max}$ versus λ.

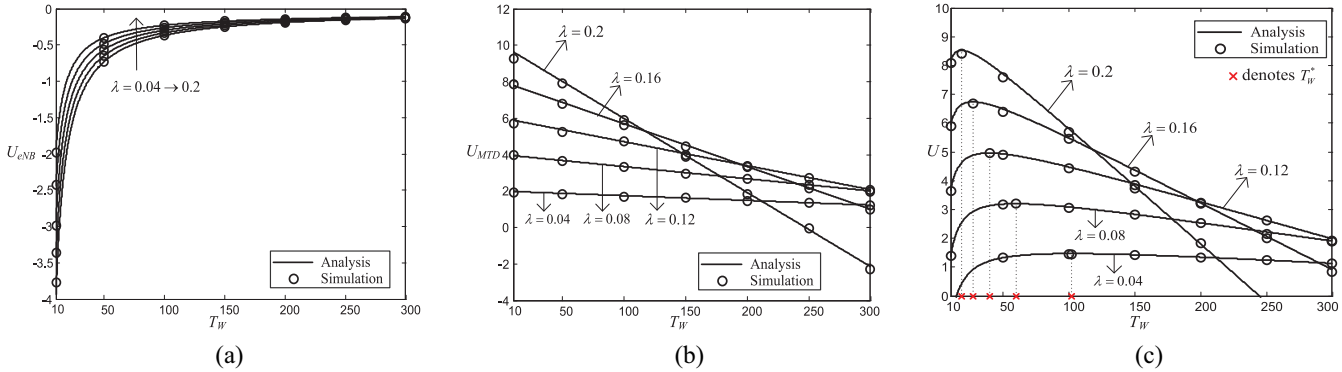


Fig. 7. Average utility of eNB U_{eNB}, the average utility of MTDs U_{MTD}, and the network utility U versus the waiting period length T_W. $r = 50$. $c_{\text{eNB}} = 0.1$. $c_{\text{MTD}} = 0.1$. $K = 40$. $\lambda \in \{0.04, 0.08, 0.12, 0.16, 0.2\}$. $M = 1$. (a) U_{eNB} versus T_W. (b) U_{MTD} versus T_W. (c) U versus T_W.

reduced. Moreover, when the traffic becomes heavier, the paging frequency should also be increased accordingly. Thus, we can observe from Fig. 6(a) that as the input rate λ increases, the optimal waiting period length T_W^* decreases.

Note that by substituting the optimal waiting period length T_W^* in (10) into (8), we can obtain the maximum network utility $U_{\max}$ in (11), shown on the bottom of this page. Fig. 6(b) illustrates how the maximum network utility $U_{\max}$ varies with the input rate λ under various values of K, c_{MTD} and c_{eNB}. We can see that $U_{\max}$ decreases as the fixed cost K, the delay cost c_{MTD} or the signaling cost c_{eNB} increases. On the other hand, as the input rate λ grows, $U_{\max}$ linearly increases, as the average revenue that the system receives in one paging round, $r\lambda$, rises.

B. Simulation Results

In this section, simulation results are presented to verify the proceeding analysis. Each simulation is carried out for 10^3 waiting and paging rounds. In each waiting and paging round, we count the utility of the eNB, the utility of MTDs, and the network utility for given the fixed cost K, the delay cost c_{MTD}, the signaling cost c_{eNB}, and the income unit r. Accordingly, by calculating the ratios of the sum of utilities in each waiting and paging period to 10^3, we obtain the average utility of the eNB, the average utility of MTDs, and the average network utility.

Fig. 7(a) and (b) demonstrate how the average utility of eNB U_{eNB} and the average utility of MTDs U_{MTD} vary with the waiting period length T_W with the input rate $\lambda \in \{0.04, 0.08, 0.12, 0.16, 0.2\}$. We can see from Fig. 7(a) that the average utility of eNB U_{eNB} increases as the waiting period length T_W or the input rate λ grows, while the gain becomes marginal, when T_W is large, in which case U_{eNB} converges to $-[(2c_{\text{eNB}}\lambda)/\ln 2]$, according to (3)–(5), and (6). On the other hand, it can be observed from Fig. 7(b) that the average utility of MTDs U_{MTD} decreases as T_W increases because the delay cost rises up. The decreasing rate of U_{MTD} would further increase with the input rate λ.

Moreover, we can conclude from Fig. 7(a) and (b) that the waiting period length T_W determines a crucial tradeoff

$$U_{\max} = \begin{cases} \frac{c_{\text{eNB}}\lambda\left(\lambda\left(2-\frac{4}{\ln 2}\right)-2\ln 2+2\right)+K\lambda(\ln 2-2\lambda)}{\lambda(2+\ln 2)-\ln 2} + \frac{c_{\text{MTD}}\lambda\left(\theta\left(\lambda\left(4-\frac{8}{\ln 2}\right)+4-4\ln 2\right)+\frac{\ln 2}{\lambda}-2-\ln 2\right)}{4\lambda-2\ln 2} + r\lambda & \text{if } \lambda < \lambda_0 \\ \frac{c_{\text{MTD}}\lambda((4\theta\lambda+\ln 2)((\ln 2-2\lambda)\beta+1)-2(\theta 2\ln 2+\ln 2))}{4\lambda-2\ln 2} + \frac{-2c_{\text{eNB}}\lambda((\ln 2-2\lambda)\beta+1)+c_{\text{eNB}}2\ln 2+2\lambda K-K\ln 2}{\ln 2((\ln 2-2\lambda)\beta-1)} + \lambda r & \text{otherwise} \end{cases} \quad (11)$$

between U_{eNB} and U_{MTD}. As T_W grows, the average utility of eNB U_{eNB} grows, indicating that the average cost on signaling message transmissions and preamble reservation in each paging period decreases. On the other hand, the average utility of MTDs U_{MTD} drops because the mean access delay of MTDs increases, which leads to a higher delay cost. Therefore, to maximize the average network utility U, where $U = U_{\text{eNB}} + U_{\text{MTD}}$, it is of great importance to properly tune the waiting period length T_W.

Fig. 7(c) illustrates how the network utility U varies with the waiting period length T_W. Note that the expression of the network utility U has been given in (8), which shows that it is crucially determined by T_W. As we can see from Fig. 7(c), the network utility U is sensitive to the setting of the waiting period length T_W especially in the heavy traffic case. To achieve the maximum network utility, T_W should be properly set. The optimal waiting period length T_W^* is given in Theorem 1, and verified by the simulation results presented in Fig. 7(c).

V. Extension to Multipreamble

So far, we have demonstrated how to properly tune the length of waiting period T_W to maximize the network utility. Note that the analysis reveals that the aggregate traffic input rate λ should be no larger than $(\ln 2/2)$. Otherwise, the system becomes unstable. In this case, a rather poor network utility U is expected, regardless of T_W. To accommodate massive access requests, the network should offer more random access resources by, intuitively, increasing the number of preambles M. In this section, the above analysis will be extended to multipreamble scenario $M > 1$. A closer look will be taken at the performance gain brought by increasing M, and how the optimal length of waiting period and the maximum network utility vary with M.

Note that in the previous sections, we have considered the special case of the number of preambles $M = 1$. For $M = 2, \ldots$, on the other hand, we can see that each contention frame will contain $2M$ orthogonal resource units. That is, each frame contains two time slots and in each time slot, M preambles can be chosen. If two or more MTDs choose the same preamble in the same time slot, then a collision occurs and another frame, i.e., one more package of $2M$ orthogonal resource units, will be assigned. Therefore, we can conceptually regard the contention resolution process spans as a tree, where the node degree is given by $2M$.

For a large M, e.g., $M \gg 1$, it has been derived in [25] that the average number of frames to complete n MTDs' service should be given by[5]

$$F(M, n) = \frac{n}{\ln 2M} - \frac{1}{2M-1} + \Psi(n, 2M) \tag{12}$$

where

$$\Psi(n, 2M) = n\lambda_1 \cos\left(2\pi \log_{2M} n + \theta_1\right) + \lambda_2 \sin\left(2\pi \log_{2M} n + \theta_2\right) \tag{13}$$

[5]If M is small, e.g., $M = 1$, then $\Psi(n, 2M) \approx 0$ and $F(M, n) \approx (n/\ln 2M) - (1/2M - 1)$, based on which (2) is obtained.

$$\lambda_1 = 2\sqrt{\frac{2\pi^2/\ln 2M}{\left(4\pi^2 + (\ln 2M)^2\right)\sinh\left(2\pi^2/\ln 2M\right)}} \tag{14}$$

$$\theta_1 = \arg\left(\frac{\Gamma(1 - 2\pi i/\ln 2M)}{1 + 2\pi i/\ln 2M}\right) \tag{15}$$

$$\lambda_2 = \frac{2}{(\ln 2M)^2}\sqrt{\frac{2\pi^2/\ln 2M}{\sinh\left(2\pi^2/\ln 2M\right)}} \tag{16}$$

and

$$\theta_2 = \arg(\Gamma(1 - 2\pi i/\ln 2M)). \tag{17}$$

Note that in above expressions, $\Gamma(\cdot)$ denotes the Gamma function, and $\arg(\cdot)$ denotes the function that returns the phase angles.

Following the similar analysis in Section III, we assume the system will be in equilibrium in long run. The average length of the paging period, in which n MTD join the contention, can then be written as

$$E\left[T_P^M\right] = 2F(M, n) + 1. \tag{18}$$

Accordingly, the average number of MTDs that join the contention in each paging period $E[n]$, the average length of the paging period $E[T_P^M]$, and the mean access delay of each MTD $E[D]$ can all be numerically obtained by jointly solving (1), (5), and (12)–(18).

Fig. 8(a) demonstrates how the mean access delay of each MTD $E[D]$ varies with the input rate λ with the number of preambles $M = 2, 4$, or 6. A perfect match between the analysis and the simulation results can be observed. Moreover, we can see from Fig. 8(a) that as the input rate λ rises, the mean access delay of each MTD $E[D]$ increases. For a given M, $E[D]$ may even grow unboundedly if λ is too large, e.g., $\lambda = 1$ and $M = 4$. To evaluate the network stability with regards to M, let us define $\bar{\lambda} = \max\{\lambda | E[D] \in (0, +\infty)\}$, that is, if the input rate $\lambda \geq \bar{\lambda}$, then, the network will be unstable and $E[D] = +\infty$. $\bar{\lambda}$ can be numerically obtained based on (1), (5), and (12)–(18). We can see from Fig. 8(b), as the number of preambles M increases, $\bar{\lambda}$ grows, indicating that the network can accommodate heavier traffic.

Moreover, Fig. 8(a) also shows that for a given input rate λ, the mean access delay can be effectively reduced with more preambles. Yet, the performance gain becomes marginal when the number of preambles M is large, as shown in Fig. 8(c). This is because the average number of frames to complete MTDs' service, or equivalently, the average length of the paging period, has been small. Therefore, in this case, the main contributor to the mean access delay of each MTD $E[D]$ could be the length of waiting period T_W, instead of the length of paging period T_P. For instance, with $\lambda = 0.6$ and $M = 4$, it can be seen from Fig. 8(c) that $E[D]$ is down from 2000 time slots to around 250 time slots with T_W deceasing from 1000 to 100. Such observation confirms that even when there are abundant resources in preamble domain, the proper tuning of T_W is still indispensable for improving the network performance.

Let $T_W^{*,M}$ denotes the optimal length of waiting period to maximize the network utility in multipreamble scenario $M > 1$. Yet, due to the implicit nature of (12), explicit expressions of the network utility U cannot be derived. Therefore, to derive $T_W^{*,M}$, we propose an exhaustive search algorithm in

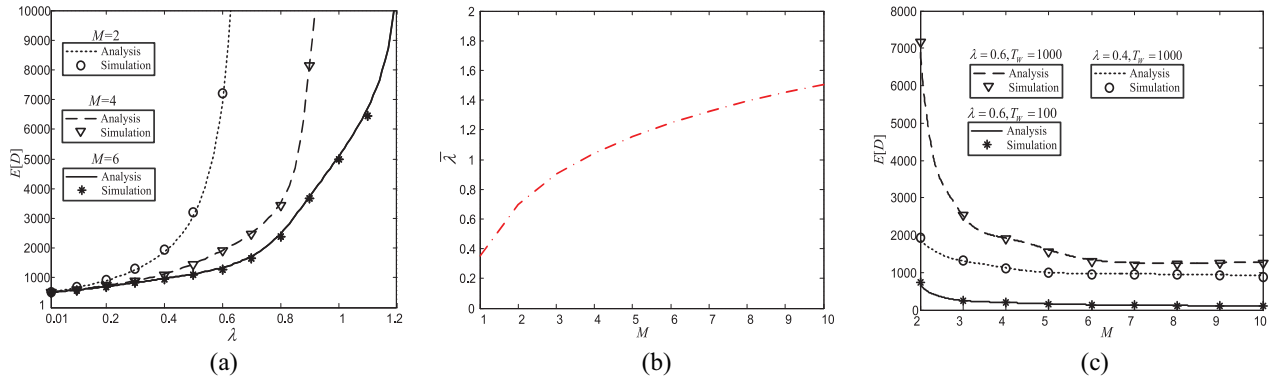


Fig. 8. (a) Mean access delay of each MTD $E[D]$ (in unit of slots) versus the input rate λ. $M \in \{2, 4, 6\}$, $T_W = 1000$. (b) $\bar{\lambda}$ versus the number of preambles M. $T_W = 1000$. (c) $E[D]$ versus M. $\lambda \in \{0.4, 0.6\}$. $T_W \in \{100, 1000\}$.

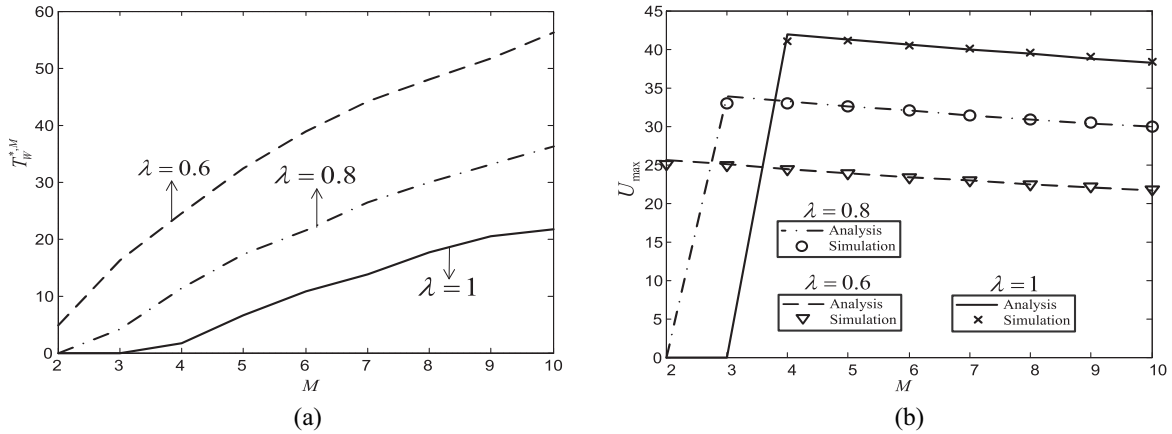


Fig. 9. Optimal length of waiting period $T_W^{*,M}$ (in unit of slots) and the maximum network utility $U_{\max}$ versus the number of preambles M. $\lambda \in \{0.6, 0.8, 1\}$. $r = 50$. $c_{\mathrm{eNB}} = 0.1$. $c_{\mathrm{MTD}} = 0.1$. $K = 40$. (a) $T_W^{*,M}$ versus M. (b) $E[D]$ versus M.

Algorithm 1 Calculation of $T_W^{*,M}$ and $U_{\max}$

Input $M, \lambda, K, \mathrm{c_{MTD}}, \mathrm{c_{eNB}}, \mathrm{r}$ and initialize $\mathrm{T_W} = 0$, $U_{\max} = 0$.
repeat
  Obtain $E[\mathrm{T_P^M}]$, $E[n]$ and $E[D]$ according to (1), (5) and (12)–(18).
  Calculate U_{eNB}, U_{MTD} and then U according to (6), (7) and (8), respectively.
  If $U > U_{\max}, E[n] \geq 1, E[\mathrm{T_P^M}] \geq 1$ **then**
    $U_{\max} = U$ and $\mathrm{T_W^{*,M}} = \mathrm{T_W}$.
  endif
  $\mathrm{T_W} = \mathrm{T_W} + 1$.
until $\mathrm{T_W} > \mathrm{T_{W,max}}$
Output $\mathrm{T_W^{*,M}}$ and $U_{\max}$.

Algorithm 1. The basic idea of Algorithm 1 is to obtain the network utility U for each value of T_W in the search space $\{0, \ldots, T_{W,\max}\}$, where $T_{W,\max}$ should be sufficiently large, and finally, choose the value of T_W that maximizes U while satisfies the constraints $E[n] \geq 1$ and $E[T_P^M] \geq 1$.

Fig. 9(a) demonstrates how the optimal length of waiting period $T_W^{*,M}$ varies with the number of preambles M for the input rate $\lambda = 0.6, 0.8$, or 1 and $T_{W,\max} = 10^4$. We can see that the optimal length of waiting period $T_W^{*,M}$ could be zero, when M is small and the traffic is heavy. For instance, with $M = 2$, as shown in Fig. 8(b), the maximum input rate that a stable network can support is $\bar{\lambda} \approx 0.69$. If the input rate $\lambda = 1 > \bar{\lambda}$, then the network will be unstable and the utility U decreases with time, regardless of T_W^M. Therefore, the optimal length of waiting period $T_W^{*,M} = 0$.

Moreover, we can also observe that as M increases, the optimal length of waiting period $T_W^{*,M}$ increases, indicating that the paging frequency is reduced. Intuitively, for given input rate λ, the length of the paging period shrinks as more preambles are provided. In this case, the length of waiting period should be properly enlarged to reduce the cost on reserving preambles in long run.

Fig. 9(b) demonstrates how the maximum network utility $U_{\max}$ varies with the number of preambles M for the cost of reserving one preamble $K = 40$. We can see that when the traffic is heavy, it is necessary for the network to provide sufficient number of preambles so that a positive utility can be achieved. However, as shown in Fig. 8(c), the performance gain in access delay becomes marginal when the number of preambles M is large. Since reserving more preambles increases the cost according to (6), an excess of preambles may further bring

down $U_{\max}$, as shown in Fig. 9(b). It corroborates that proper settings of the length of waiting period T_W and the number of preambles M are indispensable for maximizing the network utility.

VI. Conclusion

This article proposed a novel group paging scheme for massive M2M communications in the LTE networks, where the contention resolution technique is applied to address preamble collisions during the paging process. To reduce the number of concurrent access requests, we consider a dynamic scenario in which only the MTD that has packets to deliver will join the paging process.

Starting from the single-preamble case, we obtain explicit expressions of key system measures, including the average number of MTDs that join the contention in each paging period, the average length of the paging period, and the mean access delay of each MTD. The analysis is further extended to the multipreamble case and verified by simulation results. It is demonstrated that a smaller length of waiting period T_W indicates a higher frequency of group paging that can reduce the mean access delay of each MTD but may also incur more signaling overheads. To study how to properly tune T_W, a utility-based framework is established, where the optimal waiting period length for maximizing the network utility is derived. The analysis sheds important light on the practical system design for the LTE group paging in massive M2M access scenario. Specifically, it reveals that as the aggregate traffic becomes heavier, the network should perform the group paging more frequently, i.e., a smaller waiting period length T_W, or increase the number of preambles M. With more preambles, the network is capable of handling intensive traffic along with a lower mean access delay of each MTD. However, the gain in delay performance becomes marginal when the number of preambles is large.

Finally, we would like to note that in this article, it is assumed that there is no constraint on the access delay. In practice, however, some M2M applications are delay-sensitive and may set an upper-bound on the access delay of each MTD, with which if an access request fails to be delivered within the delay bound, then it could be dropped. It is of practical significance to further study how to optimally tune the waiting period length to maximize the network utility while satisfy the delay requirement. The extension of the analysis to incorporate the access delay constraint would be an important issue that deserves much attention in the future work.

Appendix
Proof of Theorem 1

To derive the optimal waiting period length T_W^* for maximizing the network utility U with the constraints that $E[n] \geq 1$ and $E[T_P] \geq 1$, let us first write the derivative of U with regard to T_W as

$$U' = \frac{f(T_W)}{2(\ln 2 - 2\lambda)(T_W - 1)^2 \ln 2} \tag{19}$$

where

$$f(T_W) = -c_{\text{MTD}}\lambda(4\theta\lambda + \ln 2)(T_W - 1)^2 \ln 2 + 2(\ln 2 - 2\lambda)^2(K - c_{\text{eNB}}). \tag{20}$$

It can be obtained from (20) that $f(T_W) = 0$ has a single nonzero root for $T_W \in (1, +\infty)$, which is given by

$$T'_W = 1 + (\ln 2 - 2\lambda)\beta \tag{21}$$

where $\beta = \sqrt{[(2(K - c_{\text{eNB}}))/(c_{\text{MTD}}\lambda(4\theta\lambda + \ln 2)\ln 2)]}$, and $f(T_W) > 0$ for $T_W \in (1, T'_W)$ and $f(T_W) < 0$ for $T_W \in (T'_W, +\infty)$, indicating that the network utility U increases with T_W for $T_W \in (1, T'_W)$ and decreases with T_W for $T_W \in (T'_W, +\infty)$. Thus, it is maximized at T'_W.

Moreover, to satisfy the constraints $E[n] \geq 1$ and $E[T_P] \geq 1$, we can obtain that the waiting period length T_W should be no smaller than $T_W^{\min}$, i.e., $T_W \geq T_W^{\min}$, according to (3), (4), and (9), where

$$T_W^{\min} = \frac{\ln 2 - 2\lambda + \lambda \ln 2}{\lambda \ln 2}. \tag{22}$$

It can be seen that the optimal waiting period length T_W^* should be T'_W if $T'_W > T_W^{\min}$, or $T_W^{\min}$ otherwise. To determine $T'_W > T_W^{\min}$ or not in terms of the input rate λ, let us define

$$g(\lambda) = T_W^{\min} - T'_W. \tag{23}$$

By combining (21)–(23), we can further have that $g(\lambda) = 0$ has a single nonzero root $\lambda_0 = [(c_{\text{MTD}} \ln 2)/(2(K \ln 2 - c_{\text{eNB}} \ln 2 - 2c_{\text{MTD}}\theta))]$, and $\lim_{\lambda \to 0} g(\lambda) = +\infty$ and $\lim_{\lambda \to \ln 2/2-} g(\lambda) = 0$. According to the intermediate value theorem, it can be concluded that if $\lambda < \lambda_0$, then $g(\lambda) > 0$, $T_W^{\min} > T'_W$ and the optimal waiting period length $T_W^* = T_W^{\min}$; otherwise, $T_W^* = T'_W$.

References

[1] V. B. Mišić and J. Mišić, *Machine-to-Machine Communications: Architectures, Standards and Applications*. New York, NY, USA: CRC Press, 2014.

[2] "Cisco visual networking index: Forecast and trends, 2017–2022," San Jose, CA, USA, Cisco, White Paper, Feb. 2019.

[3] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digit. Commun. Netw.*, vol. 1, no. 1, pp. 1–19, Feb. 2015.

[4] *Comparing Push and Pull Based Approaches for MTC*, document TSG-RAN WG2 #71 R2-104873, 3GPP, Sophia Antipolis, France, Aug. 2010.

[5] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in LTE networks: Modeling and throughput optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771–2785, Apr. 2018.

[6] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.

[7] C.-Y. Oh, D. Hwang, and T.-J. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4182–4192, Apr. 2015.

[8] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive congestion control algorithm for Bursty M2M traffic in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec. 2016.

[9] G.-Y. Lin, S.-R. Chang, and H.-Y. Wei, "Estimation and adaptation for Bursty LTE random access," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2560–2577, Apr. 2016.

[10] H. Jin, W. T. Toor, B. C. Jung, and J.-B. Seo, "Recursive pseudo-Bayesian access class barring for M2M communications in LTE systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8595–8599, Sep. 2017.

[11] C.-H. Wei, R.-G. Cheng, and S.-L. Tsao, "Performance analysis of group paging for machine-type communications in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3371–3382, Sep. 2013.

[12] T. Deng and X. Wang, "Performance analysis of a device-to-device communication-based random access scheme for machine-type communications," *Wireless Pers. Commun.*, vol. 83, no. 2, pp. 1251–1272, Mar. 2015.

[13] B. Han and H. D. Schotten, "Grouping-based random access collision control for massive machine-type communication," in *Proc. IEEE Globecom*, Singapore, Dec. 2017, pp. 1–7.

[14] G. Farhadi and A. Ito, "Group-based signaling and access control for cellular machine-to-machine communication," in *Proc. IEEE VTC*, Las Vegas, NV, USA, Sep. 2013, pp. 1–6.

[15] R. Harwahyu, R.-G. Cheng, and R. F. Sari, "Consecutive group paging for LTE networks supporting machine-type communications services," in *Proc. IEEE PIMRC*, London, U.K., Sep. 2013, pp. 1619–1623.

[16] O. Arouk, A. Ksentini, and T. Taleb, "Group paging-based energy saving for massive MTC accesses in LTE and beyond networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1086–1102, May 2016.

[17] R. Harwahyu, X. Wang, R. Sari, and R.-G. Cheng, "Analysis of group paging with pre-backoff," *EURASIP J. Wireless Commun. Netw.*, vol. 43, no. 1, pp. 1–9, Feb. 2015.

[18] J. Chen, Y.-T. Lin, and R.-G. Cheng, "A delayed random access speed-up scheme for group paging in machine-type communications," in *Proc. IEEE ICC*, London, U.K., Jun. 2015, pp. 623–627.

[19] W. Cao, A. Dytso, G. Feng, H. V. Poor, and Z. Chen, "Differentiated service-aware group paging for massive machine-type communication," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5444–5456, Nov. 2018.

[20] H. S. Jang, B. C. Jung, and D. K. Sung, "Dynamic access control with resource limitation for group paging-based cellular IoT systems," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5065–5075, Dec. 2018.

[21] O. Arouk, A. Ksentini, Y. H. Aoul, and T. Taleb, "On improving the group paging method for machine-type-communications," in *Proc. IEEE ICC*, Sydney, NSW, Australia, Jun. 2014, pp. 484–489.

[22] 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 Things J.*, vol. 2, no. 5, pp. 427–434, Oct. 2015.

[23] A. Laya, C. Kalalas, F. Vazquez-Gallego, L. Alonso, and J. Alonso-Zarate, "Goodbye, ALOHA!" *IEEE Access*, vol. 4, pp. 2029–2044, 2016.

[24] H. M. Gursu, M. Vilgelm, W. Kellerer, and M. Reisslein, "Hybrid collision avoidance-tree resolution for M2M random access," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 4, pp. 1974–1987, Aug. 2017.

[25] A. J. E. M. Janssen and M. J. D. Jong, "Analysis of contention tree algorithms," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2163–2172, Sep. 2000.

[26] J. I. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 5, pp. 505–515, Sep. 1979.

[27] G. C. Madueño, Č. Stefanović, and P. Popovski, "Efficient LTE access with collision resolution for massive M2M communications," in *Proc. IEEE Globecom Workshops*, Austin, TX, USA, Dec. 2014, pp. 1433–1438.

[28] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A random access for massive machine-type communications in smart cities," *IEEE Commun. Mag*. vol. 55, no. 1, pp. 76–83, Jan. 2017.

[29] J. Yuan, H. Shan, A. Huang, T. Q. S. Quek, and Y.-D. Yao, "Massive machine-to-machine communications in cellular network: Distributed queueing random access meets MIMO," *IEEE Access*, vol. 5, pp. 2981–2993, 2017.

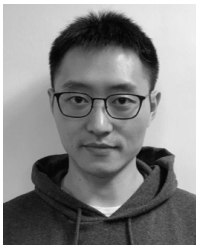
[30] R.-G. Cheng, Z. Becvar, and P.-H. Yang, "Modeling of distributed queueing-based random access for machine type communications in mobile networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 129–132, Jan. 2018.

[31] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) Protocol Specification, V13.3.0*, 3GPP Standard TS 36.331, Jan. 2017.

[32] M. Koseoglu, "Lower bounds on the LTE-A average random access delay under massive M2M arrivals," *IEEE Trans. Wireless Commun.*, vol. 64, no. 5, pp. 2104–2115, May 2016.

[33] "Study on RAN improvements for machine-type communications, V11.0.0," 3GPP, Sophia Antipolis, France, Rep. TR 37.868, Oct. 2011.

[34] M. L. Molle and G. C. Polyzos, "Conflict resolution algorithms and their performance analysis," Comput. Syst. Res. Inst., Univ. Toronto, Rep. CS93-300, Toronto, ON, Canada, Jul. 1993.



Wen Zhan (S'17) received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2012 and 2015, respectively, and the Ph.D. degree from the City University of Hong Kong, Hong Kong, China, in 2019.

His current research interests include Internet-of-Things, machine-to-machine communications, and wireless random access networks.

Xinghua Sun (M'13) received the B.S. degree from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2008, and the Ph.D. degree from the City University of Hong Kong (CityU), Hong Kong, in 2013.

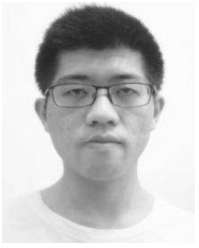
In 2010, he was a visiting student with INRIA, Rocquencourt, France. In 2013, he was a Post-Doctoral Fellow with CityU. From 2015 to 2016, he was a Visiting Scholar with the University of British Columbia, Vancouver, BC, Canada. From 2014 to 2018, he was an Associate Professor with NJUPT. Since 2018, he has been with Sun Yat-sen University, Guangzhou, China. His current research interests include wireless networking and Internet of Things.

Yitong Li received the B.Eng. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 2011 and 2016, respectively.

He is currently an Assistant Professor with the School of Information Engineering, Zhengzhou University, Zhengzhou, China. His current research interests include the performance evaluation and optimization of wireless random access networks.

Feng Tian (M'13) received the Ph.D. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2008.

From 2013 to 2015, he was a Visiting Scholar with Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. He is currently a Professor with the Nanjing University of Posts and Telecommunications. His current research interests include performance optimization and algorithm design for wireless networks, cognitive radio networks, and big data driven optimization for wireless networks.



Hong Wang received the B.S. degree in communication engineering from Jiangsu University, Zhenjiang, China, in 2011, and the Ph.D. degree in information and communication engineering from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 2016.

From 2014 to 2015, he was a Research Assistant with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, where he was a Senior Research Associate with the State Key Laboratory of Millimeter Waves and Department of Electronic Engineering from 2016 to 2018. Since 2016, he has been an Instructor with the Department of Communication Engineering, NUPT. His current research interests include broadband wireless communications, particularly in interference management in HetNets, nonorthogonal multiple access, and nonorthogonal waveforms.