IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 12, 2018, accepted October 22, 2018, date of publication October 29, 2018,
date of current version November 30, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2878424*

# Cluster-Based Group Paging for Massive Machine Type Communications Under 5G Networks

**QI PAN[1,2], XIANGMING WEN[1,2], (Senior Member, IEEE), ZHAOMING LU[1,2], WENPENG JING[1,2], AND LINPEI LI[1,2]**

[1]Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Beijing Laboratory of Advanced Information Networks, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Qi Pan (panqiouc@sina.com)

**ABSTRACT** Machine-type communications (MTCs) have been envisaged to play a key role within the future 5G cellular network. To handle the massive MTCs (mMTCs) and alleviate the congestion of the radio access network, the group paging (GP) scheme was proposed by the third-generation partnership project. However, its performance quickly decreases in the face of massive simultaneous channel accesses. In this paper, we propose a two-phase cluster-based GP (CBGP) scheme. First, owing to the advantages of low-cost, high-access capacity and handy deployment, IEEE 802.11ah is introduced to increase the capability of coping with massive access attempts. The separation of inner cluster data collection and header-based data transmission phases greatly alleviates the access congestion of cellular networks, reducing the access delay and increasing the successful access probability for mMTC devices. Besides, we also derive mathematical models of the CBGP scheme in terms of the successful access probability and average access delay. Moreover, effects from different numbers of clusters on the performance of the CBGP scheme are investigated and the optimal number of clusters is also derived, adaptive to different access scales. At last, numerical results are presented to validate the accuracy of our analytical models, demonstrate the effectiveness of the proposed CBGP scheme, and verify the optimal number of clusters, providing insights for the coming 5G cellular system design.

**INDEX TERMS** 5G, 802.11ah, access control, cellular networks, group paging, machine type communications, radio access network, Internet of Things.

## I. INTRODUCTION

Machine type communications (MTC) have become a new paradigm where devices can interact autonomously without or with few human interventions, which are also named machine-to-machine (M2M) communications. Ubiquitous sensing and connected MTC devices will provide enhanced situational awareness, data-driven decision analysis and automated response [1]. These devices are capable of measuring, collecting, delivering and analyzing the outside environmental information for various vertical applications, such as smart metering, fleet management, healthcare monitoring, intelligent transportation system (ITS) and so on, forming the so-called Internet of Things (IoT) [2].

The number of MTC devices is kept growing and will reach 26 billion by 2020 [3]. Furthermore, the IoT product and service suppliers will generate incremental revenue exceeding 300 billion by 2020 [4]. While massive MTC bring enormous opportunities, it also poses significant challenges for their distinct network requirements. Depending on the major demands and the specific application scenarios, M2M communications can be classified into two types: massive machine type communication (mMTC) and ultra-reliable machine type communication (uMTC), which are key use cases of coming 5G network [5]. The mMTC devices are always numerous with low complexity and constrained power. And uMTC devices have critical requirements in terms of latency and reliability. In this paper, we mainly focus on congestion alleviation enhancements under mMTC scenarios.

To achieve an ubiquitous connectivity for mMTC devices, the cellular network, like Long Term Evolution (LTE) and LTE-Advanced (LTE-A) network, has become one of the most promising solutions for support of M2M communications due to its wide coverage, easy installation and flexible resource management methods with high maturity. Ericsson forecast that a significant portion of the IoT applications

would be served by the cellular network in the future [6]. However, they are originally designed for human-to-human (H2H) communications which are always large data volume and high data transmission rate. Significantly different from the conventional H2H communications, the characteristics of M2M traffic are always massive access attempts, infrequent short payload transmission, power-constrained. If the existing cellular infrastructures are used to support the M2M applications, it would have been likely to suffer heavy congestion and overload both in the radio access network (RAN) and core network (CN).

Before data transmission in LTE/LTE-A networks, unconnected M2M devices need to execute the random access procedure (RAP) for connection establishments with the evolved-Node B (eNodeB). Compared to the large amount of access requests, the finite available radio access resources are not sufficient. This would lead to severe congestion and system overload in the RAN. In addition, due to the defect of slotted ALOHA-based RAPs, collisions become severer, especially for the event-driven M2M communications. For example, if devices encounter earthquakes, hurricanes or reboot after a huge scale of power outages, thousands of adjacent devices would send the access request messages almost at the same time. Consequently, massive simultaneous access attempts lead to heavy congestion and even make the network collapse, posing a giant impact on H2H communications. What is worse, collisions also bring about the ceaseless back-off process and retransmission, prolonging the access delay, wasting finite resources and increasing the power consumption of power-constrained M2M devices. Therefore, proper congestion control mechanisms are urgently required to handle the massive IoT access attempts [7].

### A. RELATED WORK

To alleviate congestion and overload from mMTC devices, there has been lots of research focusing on the random access improvements over cellular networks. The Third Generation Partnership Project (3GPP) has carried out a series of feasible solutions [3], [8]. According to the way MTC traffic is generated, current RAN-overload control schemes can be classified into two classes: push-based and pull-based approaches [9].

In push-based overload control mechanisms, the network access procedures are initiated by devices [10], [11]. Request messages are pushed by devices to the eNodeB. Typical push-based solutions include Access Class Barring (ACB), Extended Access Barring (EAB), Resource Separation Scheme (RSS), back-off scheme, slotted access scheme, etc. The key idea of the above solutions is to scatter the massive simultaneous access attempts into a longer time period to decrease the chance of collisions. For example, in the ACB, devices are categorized into several classes according to different application requirements and each class is assigned with a barring factor, $p$, and a barring time, $T$. Before executing the RAP, a device would randomly generate a value, $m$, varying from 0 to 1. Only when $m < p$, then the device can be allowed to access the network. Otherwise, it needs to perform a random back-off algorithm with a window size of $T$ for retransmission. In a similar approach, the back-off scheme assigns a back-off procedure for M2M devices, distributing the access attempts randomly in a back-off window. Soltanmohammadi *et al.* [3] gave a detailed summary of advantages and disadvantages of push-based congestion control schemes.

In the pull-based solutions, the access procedures are initiated by the eNodeB [10]. The eNodeB informs specific M2M devices for network connection and data transmission. One of the most typical pull-based solutions is paging. In the LTE/LTE-A network, a downlink paging channel is assigned to transmit paging messages. Devices would listen to their predetermined paging channel at paging occasions to obtain paging messages. One paging message can page at most 16 devices and two consecutive paging messages are available per 10 *ms* radio frame. If there are 1600 M2M devices waiting to be paged, this will cost 1000 *ms* with 100 paging messages, which results in long delay and low resource efficiency. Hence, a group paging (GP) mechanism was proposed to use one paging message for one group of M2M devices to address the above issues [12]. Via dividing into multiple groups, devices in one paging group share with the same group identity (GID). Once their GIDs are found in the broadcast paging message, devices within the paged groups should execute the RAP for connection establishments.

Extensive research has also focused on the validation, analysis and improvement of the GP scheme. In [13], the performance of GP was primarily presented via computer simulation. The detailed analytical model of the GP scheme was proposed and evaluated in terms of collision probability, access successful probability, statistics of preamble transmission, utilization of random access opportunities (RAOs), etc. [9]. For improving resource efficiency in smart metering, a dynamic radio resource allocation algorithm for GP was proposed [14]. However, the dynamically allocated preambles are not sufficient and could not work when the number of pending smart meter devices is larger than limited preambles. In [15], another dynamic resource allocation (DRA) scheme was proposed for GP. The reserved RAOs could be dynamically adapted to the predicted number of contending devices in each random access slot (RAS). However, it only focused on the resource efficiency and ignored the key problem of access congestion resulted from massive MTC devices.

However, when the amount of devices in a paging group becomes large, there is still severe congestion for finite random access resources [16]. Devices need to spend more time delay and energy for network connection establishments, even be dropped due to excessive retransmission. In order to solve this issue, plenty of proposals to distribute centralized massive access to longer time slots has been put forward. Harwahyu *et al.* [17] proposed to repeat the GP process until all the data transmission was completed, which was named Consecutive Group Paging (CGP). Devices failed in the first GP interval would continue to access the network in subsequent paging intervals. However, when there are more

groups, the performance of CGP would be worse with low resource efficiency. In [18], a traffic scattering for group paging (TSFGP) mechanism was proposed. Devices are enabled to start the RAP separately in the time domain. The number of access attempts in each RAS would be significantly decreased to alleviate the congestion. Similar to the spirit of [18] and [19], the authors proposed a further improved version of TSFGP (FI-TSFGP). The FI-TSFGP leveraged a better estimation of both the total number of arrivals and the amount of successful MTC devices in a stable state [20]. In addition, another improvement for GP is the pre-backoff scheme [16]. After receiving the paging message, devices would execute a back-off algorithm before starting the RAP, realizing the dispersion of access attempts in the time domain. There is no doubt the above mechanisms could alleviate the congestion of GP mechanism to some extent. However, devices still need to spend more time on the random distribution of access attempts in the time domain, causing long access delay. And the long time of channel occupancy also means low resource utilization and waste of access resources.

Besides, cluster-based research for random access enhancement is summarized as below. Wang *et al.* [21] proposed a cluster-based random access scheme for M2M communications. The inter-cluster devices reuse the dedicated random access resources which are reserved by eNodeB. The predetermined cluster headers aggregate data from inter-cluster devices and send them to the eNodeB. However, M2M devices would still suffer severe congestion when there are so many clusters that the resources allocated to each cluster are insufficient. In [22], an improved random access (I-RA) scheme for GP was proposed. Device-to-device (D2D) technology was utilized by cluster headers to collect data within the clusters. Mathematical models were also presented to evaluate the performance of the proposed scheme. However, there are no details about the exact description nor the mathematical model of the inner-cluster data collection with D2D. In [23], a reliable and real-time uplink access and downlink machine-type multiple service (MtMS) was designed and analyzed. Taking advantages of small-cells, home-evolved NodeBs (HeNBs) were used to aggregate control and data traffic for the sake of overload reduction toward the core network. However, the subgroups are paged by HeNBs one by one. The delay for paged devices would be long due to the paging dispersion of subgroups in time domain. In [7], a two-layered group-based access mechanism was proposed. Within the same paging group, devices would be further divided into multiple access groups where a designated device delegates the RAPs of other devices in the access groups. However, no mathematical model was illustrated to characteristic the performance of the proposed scheme and there are also no details about data transmission within the access groups. Furthermore, the effect of access groups on the performance is not investigated and the optimal access group size for maximizing the performances is also not discussed. In addition, employing traditional cellular network to support the inner-cluster data collection still faces the same congestion problem and also leads to higher cost due to the expensive price of LTE/LTE-A communication modules.

### B. CONTRIBUTIONS & ORGANIZATION

In this paper, we propose a cluster-based group paging (CBGP) mechanism to enhance the GP mechanism. Firstly, paging groups are usually classified by the communication features, i.e., quality of service (QoS) requirements, service types or deployment in a specific location [8]. Dividing the communication into two phases could significantly decrease the number of access attempts for cellular network, alleviate the congestion, decrease the energy consumption of MTC devices and ensure high QoS guarantees for various M2M applications. Furthermore, unlike the above literature, IEEE 802.11ah is assumed to be adopted for the inner-cluster data collection phase for power-constrained M2M devices to enhance the capability of handling massive MTC access attempts. IEEE 802.11ah is designed to support ubiquitous sensors and various IoT application, which can support up to 8000 node devices' energy-efficient communications [24]. Different from the LTE/LTE-A cellular network, low power wide area networks (LPWAN) are designed for the ubiquitous and massive power-constrained M2M devices, including IEEE 802.11 ah, SigFox, LoRaWAN, Narrowband IoT (NB-IoT), etc. Due to the mature technology and flexible deployment, IEEE 802.11 ah, as an IoT-oriented solution, is one of the most promising technologies for support of the inner-cluster data collection for massive power-constrained devices. Therefore, congestion of the inner-cluster data collection phase can be significantly alleviated and less number of header devices establishes the network connection without or with a few collisions. Hence, congestion for cellular networks from massive access attempts could be greatly mitigated. Afterwards, mathematical models would also be derived and analyzed to characterize the performance of proposed CBGP scheme. Besides, the effect from different numbers of clusters on the performance of CBGP would be investigated based on the proposed analytic models. Finally, the optimal number of clusters would also be derived, adapting to the dynamic M2M traffic. The main contributions of this paper can be summarized as below:

- The CBGP scheme is proposed to alleviate the congestion and overload in RAN of cellular networks. Through dividing paging groups into more clusters, the number of access attempts for cellular networks would decrease. Simultaneously, introducing the IEEE 802.11ah into the inner-cluster data collection phase can cope better with the massive access attempts within the clusters. LPWAN and cellular networks are cooperatively utilized to enhance the conventional GP scheme for the 5G mMTC scenarios.
- Mathematical analytical models for the CBGP scheme are presented. Both the inner-cluster data collection phase and header-based data transmission phase to cellular networks have been modeled and analyzed from the

perspective of successful access probability and average access delay.
- The effect of different numbers of clusters on the performance of the proposed CBGP scheme is investigated and the optimal number of clusters is derived, adapting to the dynamic amount of access attempts.
- The computer simulations are utilized to evaluate the accuracy of mathematical models for the CBGP scheme. Besides, the results also verify the optimal number of clusters for the proposed CBGP scheme with a better performance compared to CBGP schemes with different fixed numbers of clusters, conventional GP scheme and other GP enhancements.

The reminder of this paper is structured as follows. In section II, the system model is presented, including the specific access procedures and corresponding analytical models for the inner-cluster data collection and header-based data transmission phases. We give the analytic model and derive the optimal number of clusters for our proposed CBGP scheme in section III. The simulation results are provided in section IV, verifying the effectiveness of the proposed CBGP scheme, its mathematical models and optimization. Section V concludes this paper at last.

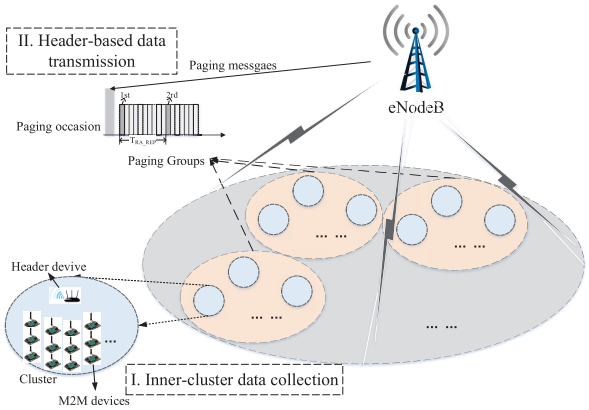



**FIGURE 1.** System architecture of proposed CBGP scheme.

## II. SYSTEM MODEL

Consider there are $M$ M2M devices under the coverage of $W$ cellular base stations, running various MTC applications. Massive M2M devices would execute the RAPs for network connection establishments once paged. Devices within the paging groups are further divided into more clusters. Sensed data are collected firstly within the clusters and then forwarded to base stations for particular applications. Let the number of paging groups be $N_p$ and each paging group is separated into $N_c$ clusters according to the locations for the convenience of data collections as shown in Fig. 1. Header devices would be elected in each cluster with consideration of location, residual energy capacity, communication ability, etc. This header can be regarded as an agent or gateway to gather sensed data within the clusters and assume there are $M_c$ devices in a cluster. Hence, data communication between sensor devices and eNodeB can be divided into two phases as shown in Fig. 1:

Inner-cluster
- data collection phase,

Header-based
- data transmission phase.

For simplicity, we assume that number of devices in each paging group and cluster is equivalent in this paper. In addition, Table 1 summarize the definitions of variables used in the analytical models for clarity.

**TABLE 1.** Variables used in the analytical modules.

| | |
|---|---|
| $CW$ | size of contention window under Halow |
| $k$ | the back-off stage under Halow |
| $S_i$ | index of the RAW slot |
| $M$ | number of M2M devices in the system |
| $N_p$ | number of paging groups |
| $N_c$ | number of clusters in one paging group |
| $M_c$ | number of M2M devices in one clusters |
| $L$ | number of access slots in one RAW slot |
| $P(j)$ | probability that $j$ devices choose one same RAW slot |
| $P_1, P_2$ | successful access probability for first/second case of no conflicts under Halow |
| $T_1, T_2$ | time delay for first/seconds case of no conflicts under Halow |
| $H(k)$ | number of devices executing the $k$th back-off stage |
| $\sigma$ | the time duration of time unit under Halow |
| $L_{CW}$ | the access opportunities during the $k$th back-off stage |
| $P_{s,dcf}(H)$ | the successful transmission probability for $H$ devices |
| $T_{total}$ | the total time for the devices of pervious $k$ stages |
| $D(k)$ | time for data transmission within the $k$th back-off stage |
| $T_{dcf}$ | average access delay to complete DCF procedure |
| $P_{p_1,s}, P_{p_2,s}$ | successful access probability for inner-cluster data collection, header-based data transmission phase. |
| $T_{p_1}, T_{p_2}$ | average access delay for inner-cluster data collection, header-based data transmission phase |
| $h$ | number of paging groups paged in one message |
| $R$ | number of available preambles |
| $T_{RA\_REP}$ | time interval between two continuous RASs |
| $W_{RAR}$ | RAR window size |
| $N_{RAR}$ | maximum number of RARs in one Msg2 |
| $N_{ACK}$ | maximum number of successful MTC devices within the RAR window |
| $N_{PT_{max}}$ | the maximum number of preamble transmission |
| $I(j)$ | number of access attempts in the $j$th RAS |
| $I_s(j)$ | number of successful access attempts in the $j$th RAS |
| $K_{min}, K_{max}$ | minimal and maximal values of RAS when the failed devices could restart the RAP in the current RAS. |
| $I_{k,F}[n-1]$ | number of devices which failed their $n1$th retransmission in the $k$th RAS |
| $\alpha_{k,j}$ | the portion of back-off interval of the $k$th RAS that overlaps with the $j$th RAS |
| $T_{GP}$ | time interval between two consecutive paging occasions in the CBGP scheme |
| $T_{p1,max}$, $T_{p2,max}$ | maximum time for inner-cluster data collection, header-based data transmission phase |
| $T_{avg}$ | average access delay for the CBGP scheme |
| $P_{CBGP}$ | successful access probability for the CBGP scheme |
| $U_{N_c,M}$ | utility function of the proposed CBGP scheme for optimal number of clusters |
| $T_{N_c,M}$ | average access delay of CBGP scheme with $M$ devices and $N_c$ clusters |
| $P_{s,N_c,M}$ | successful access probability for the CBGP scheme with $M$ devices and $N_c$ clusters |

### A. INNER-CLUSTER DATA COLLECTION PHASE

The header of clusters aggregates data from devices via the IEEE 802.11ah for its high access capacity, low cost, flexible deployment and high technical maturity.

Once devices learn that their groups have been paged in the paging message, header devices would begin the data collection process and others within clusters would be informed by the clusters and prepare for the inner-cluster data transmission. The access capacity of header devices is vital for massive devices within the clusters. Hence, we introduce the new standardized IEEE 802.11ah, which is also named Wi-Fi Halow in our proposed CBGP scheme [25]. It can be seen as a simplified version of traditional IEEE 802.11ac and specially revised to accommodate to characteristics of M2M communications.

In order to evaluate the performance of inner-cluster data collection phase for massive MTC devices, we derive the expressions of the average access delay and successful access probability for the inner-cluster data collection phase.

#### 1) RESTRICTED ACCESS WINDOW UNDER Wi-Fi HALOW

In order to deal with massive access attempts, Wi-Fi Halow introduces the restricted access window (RAW) mechanism. It limits a set of devices that would access channels at the same time and spreads their network requests over a longer period of time. For that, header devices (Access Point, AP) distribute a large amount of devices into several specific RAW slots via the RAW parameters set (RPS) included in the periodic beacon frames. Devices are forbidden to transmit data in alien RAW slots. Therefore, severe congestion can be alleviated via this dispersion in the time domain. In Fig. 2, we can easily figure out how RAW mechanism works.

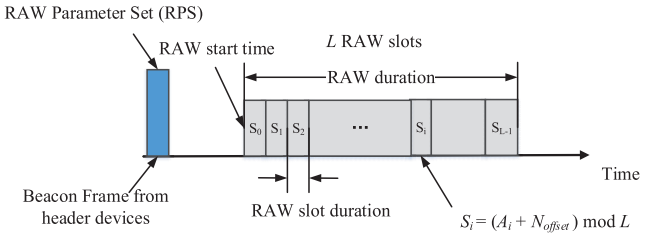



**FIGURE 2.** Slot assignment procedure in RAW mechanism.

Header devices of clusters would broadcast the RPS periodically in beacon frames, informing devices when they are permitted for data transmission. For the sake of energy saving, devices can always remain in the sleep mode except for their RAW slots. Inside their RAW slots, devices would follow the traditional distributed coordination function (DCF) for data transmission. The DCF procedure implements carrier sense multiple access with collision avoidance (CSMA/CA) method. Devices could start data transmission only when the medium is sensed idle for a short time period. In addition, a truncated binary exponential back-off mechanism would be utilized here for contention alleviations.

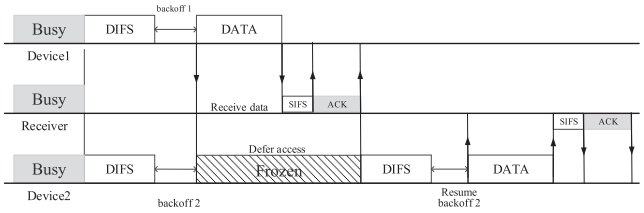



**FIGURE 3.** Overview of DCF MAC access process.

When their RAW slots come, devices would implement the DCF process as shown in Fig. 3. Devices always keep sensing the transmission medium. If the channel is sensed idle for a time period of DCF Inter-frame space (DIFS), devices would generate a random integer value from a uniform distribution over the contention window, $[0, CW_0 - 1]$, and execute the back-off procedure. Then, the counter would decrease one by one if the channel is sensed idle for a unit time slot $\sigma$. The back-off counter would be frozen when the channel is sensed busy. It would not resume the back-off counting until the channel is sensed idle again for a time period of DIFS. When the counter decreases to zero, data transmission will be initiated. If the channel is sensed idle and only one device start to transmit data in this time slot, the data would be received correctly and an ACK from header devices would be returned after a short inter-frame space (SIFS). Otherwise, devices need to retransmit this data and a new contention window, $CW_k$, is provided. The contention window can be expressed as below:

$$CW_k = min\{CW_{max}, 2^k CW_0\}, \tag{1}$$

where $k$ indicates the back-off stage and increases one by one in case of contentions until it reaches the maximum stage, $m$, i.e., $CW_{max} = 2^m CW_0$. In this paper, we assume there are limitations on the retransmission, i.e., $RL_{max}$. Once retransmission times exceeds the upper limits, this would be dropped as a failed data transmission.

Once collided, devices would restart the back-off counter with the contention window of $CW_k$ and continue the above processes until the data is successfully transmitted or dropped.

#### 2) MODELING OF RAW FOR INNER-CLUSTER DATA COLLECTION

Sensed data from ubiquitous M2M devices is collected via Wi-Fi Halow by the headers and then sent to the Internet via LTE. Here, details about the mathematical model of inner-cluster data collection phase would be elaborated. The explicit analytic model of average access delay and successful access probability would be presented.

Devices within clusters can obtain their allocated RAW slots by receiving the periodic beacon frames, including the traffic indication map (TIM). Then they would fall into sleep mode again until their RAW slots for energy saving. The RAW slot index for the device $i$ is $S_i$ and the association ID (AID) is $A_i$. There are $L$ access time slots in RAW and

$M_c$ devices limited by the lowest and highest AID indicating the location, traffic, type, energy saving mode, etc. [26]. The mapping function between the slot index and the AID can be expressed as below [27]:

$$S_i = (A_i + N_{offset}) mod L. \tag{2}$$

where $N_{offset}$ is defined as the offset parameter for time slot allocations. As described above, the allocation of slot indexes can be regarded as randomly selecting an RAW slot for inner-cluster devices within the RAW time period.

And the number of devices in one cluster which would transmit data via Wi-Fi Halow is $M_c = M/N_p * N_c - 1$, where there is one M2M device selected as the header for data aggregation.

Sensed data can be successfully transmitted if no conflicts take place. The scenarios where no conflicts happened can be approximately split into two cases [27]. The first one is the case that only one device is assigned for one RAW slot. With no other devices contending, this device can solely occupy the RAW slot for its data transmission. On the other hand, when the amount of access is large, there is more than one device assigned to the same RAW slot. These devices need to execute the DCF process as mentioned above for data transmission to headers. For simplicity and operability, one cluster of devices is assigned into one same RAW duration for data collections.

As devices would be uniformly distributed into the RAW access slots within the RAW duration, the probability that $j$ devices choose one same RAW slot can be easily derived as

$$P(j) = \binom{M_c}{j} (\frac{1}{L})^j (1 - \frac{1}{L})^{M_c - j}, \tag{3}$$

where $\binom{m}{k}$ is k-combinations.

For the first case of no conflicts, only one device is distributed into one RAW slot. The probability, $P_1$, can be expressed as

$$P_1 = (1 - \frac{1}{L})^{M_c - 1}. \tag{4}$$

The time delay needed for the first case is the sum of random selections of RAW slots, $L/2$, and the basic transmission delay without retransmission, $T_0$. No conflicts take place in the selected RAW slot. This can be indicated as

$$T_1 = L/2 + T_0 = L/2 + DIFS + 2 * SIFS + T_{DATA} + T_{ACK}, \tag{5}$$

where $T_{DATA}$ and $T_{ACK}$ represent the time duration for transmission of data and ACK separately.

In the second case, assume that there are $H$ devices choosing and contending one same RAW slot. The back-off procedure would be executed after sensing the idle channel for a time period of DIFS and there could be at most $CW_{min} = CW_0$ devices which can complete the data transmission at the first back-off stage. We assume that the number of devices executing the $k$th stage of back-off process is $H(k)$. As devices start the first back-off stage at the same time after waiting for the idle channel of a DIFS period, for simplicity, we assume the devices begin their next back-off stage at the same time. Then we can derive the analytical model as

$$\begin{aligned}
&H(1) = H;\\
&H_s(1) = H(1) \times (1 - \frac{1}{L_{CW_1}})^{H(1)-1};\\
&H_f(1) = H(1) - H_s(1);\\
&\quad \dots \quad \dots \quad \dots\\
&\quad \dots \quad \dots \quad \dots\\
&H(k) = H_f(k-1);\\
&H_s(k) = H(k) \times (1 - \frac{1}{L_{CW_k}})^{H(k)-1};\\
&H_f(k) = H(k) - H_s(k); \quad 2 \le k \le m\\
&\quad \dots \quad \dots \quad \dots\\
&H(m) = H_f(m-1);\\
&H_s(m) = H(m) \times (1 - \frac{1}{L_{CW_{max}}})^{H(m)-1};
\end{aligned} \tag{6}$$

where $H_s(k)$ and $H_f(k)$ separately represent the number of successful and collided devices in the $k$th back-off stage. Obviously, the number of devices in the first back-off stage is $H$. There are $L_{CW_k} = CW_k/\sigma$ access opportunities during the $k$th back-off stage. Successful transmission means that only one device finish its $k$th back-off procedure in one time slot which is occupied by this successful device itself. Hence, the probability for devices successfully completing the data transmission at the $k$th back-off stage can be expressed as $(1 - 1/L_{CW_k})^{H(k)-1}$ as shown in equation (6). In addition, data transmission requests of M2M devices will be dropped once they collided again in their $m$th back-off stages.

Hence, we can get the successful transmission probability for M2M devices completing the DCF procedure as

$$\begin{aligned}
P_{s,dcf}(H) &= \frac{\sum_{k=1}^{m} H(k) \times (1 - \frac{1}{L_{CW_k}})^{H(k)-1}}{H}\\
&= \frac{\sum_{k=1}^{m} H_s(k)}{H}.
\end{aligned} \tag{7}$$

Devices under the DCF process would always freeze their DIFS timers or back-off counters once the transmission medium is sensed busy. The successful transmission must be separate in the time domain and others have to be waiting until the successful transmission is completed one by one. Assume the time for devices to finish the data transmission within the $k$th back-off stage is $D(k)$. Then the expression of transmission delay can be obtained as below:

$$\begin{aligned}
&T_{total}(0) = 0;\\
&T_{total}(k) = T_{total}(k-1) + T_0 * H_s(k) + \frac{CW_k}{2};\\
&D(k) = T_{total}(k-1) + T_1 + \frac{CW_k}{2};
\end{aligned} \tag{8}$$

where $T_{total}(k)$ denotes the total time delay needed for the previous devices which completed their data transmission within the $k$ times of back-off processes. The access delay for devices which complete their data transmission in the $k$th

stage of back-off process can be expressed as the time needed for the previous $k-1$ back-off processes and the average time needed for the current stage of back-off process. Hence, the average access delay for M2M devices to complete DCF procedure can be derived as

$$T_{dcf}(H) = \frac{\sum_{k=1}^{m} D(k)}{\sum_{k=1}^{m} H_s(k)}. \tag{9}$$

The probability that $H$ devices choose the same RAW slot and succeed in finishing data transmission to the header devices can be expressed as

$$P_2(H) = P(H) \cdot P_{s,dcf}, \tag{10}$$

where $H$ is from 2 to $M_c$.

Hence, for the second case with no collisions, the probability that one device succeeds in accessing the channel can be derived as

$$P_2 = \sum_{H=2}^{M_c} P_2(H). \tag{11}$$

In summary, we can get the successful access probability to complete the data transmission from devices to headers within the clusters, $P_{p1,s}$, as following:

$$P_{p1,s} = P_1 + P_2. \tag{12}$$

Similarly, the average access delay for the first phase of CBGP scheme can be obtained as:

$$T_{p1} = P_1 \cdot T_1 + \sum_{H=2}^{M_c} P_2(H) \cdot T_{dcf}(H). \tag{13}$$

Until now, the detailed mathematical models for the inner-cluster data collection phase have been derived from the perspective of successful access probability and average access delay. The analytical models would be evaluated in the latter simulations.

### B. HEADER-BASED DATA TRANSMISSION PHASE

Data aggregated within the clusters would be sent to the LTE/LTE-A cellular networks for subsequent applications. All the paged header devices would perform the random access channel (RACH) procedure for connection establishment. Here, details about the channel access process under the LTE/LTE-A network would be elaborated, also including the corresponding mathematical models in terms of the successful access probability and average access delay.

#### 1) RACH PROCEDURE UNDER LTE NETWORK

When the device is in *RRC_IDLE* state with a network request, not synchronized, about to handover or after a radio link failure, a separate physical random access channel (PRACH) is provided to handle these above situations. There are two categories of RACH processes defined in cellular network: contention-free RACH procedure and contention-based RACH procedure. The contention-free RACH procedure is mainly applied when connection to the network has existed, e.g., for handover among different base stations. The access resources and related configuration would be pre-assigned by eNodeB. In addition, when devices initiate the network access for the first time or have lost the synchronization, contention-based RACH procedure would be performed. Due to the constrained power characteristic of M2M devices, it is not proper to keep them always active and attached to the cellular network all the time. Hence, we only consider the contention-based RACH procedure and the RACH procedure represents the contention-based RACH procedure in this following part of this paper.

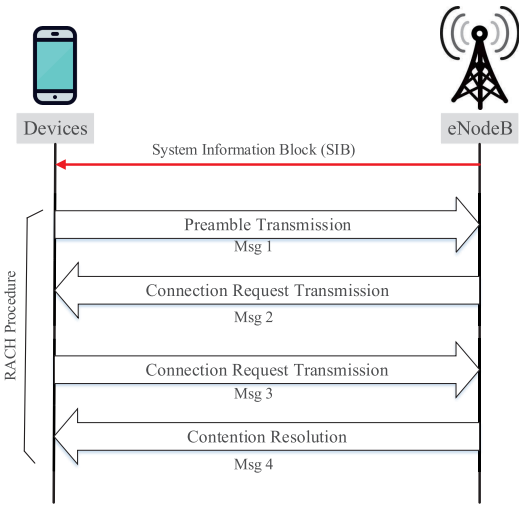



**FIGURE 4.** The random access procedure under cellular network.

Before starting the RACH procedure, devices would firstly learn about the details of access channels via the System Information Block (SIB) messages from eNodeB, including the preamble index, the PRACH configuration index, PRACH frequency offset, etc. Details of the contention-based RACH procedure can be shown in Fig. 4. Owing to the small-data characteristic of IoT traffic, we assume that once the connection is established, the data would also be transmitted at once.

1) Preamble Transmission: The Msg 1 is from terminals to the eNodeB as a random access request. Devices randomly choose one available preamble sequence and transmit it to eNodeB. The preamble sequences contain the random access radio network temporary identity (RA-RNTI) and preamble configuration index information. If devices choose different preambles, the eNodeB can recover their signals. The reason is that preambles are orthogonal Zadoff-Chu (ZC) sequences so that multiple access interference (MAI) is negligible [28]. Once one preamble is chosen by two MTC devices, collisions would take place in the PRACH. After transmitting the preambles, devices would wait for the random access response (RAR).
2) Random Access Response: After receiving the preamble sequences from M2M devices, the eNodeB would

decode RA-RNTI to obtain the spent transmitting time and calculate the propagation delay as an offset for subsequent Msg 3 transmission. The eNodeB returns RAR messages in Msg 2 via the physical downlink control channel (PDCCH). A RAR message includes a timing advance (TA) instruction compensating for the time difference between transmission and reception, an uplink grant for Msg 3 and a cell radio network temporary identifier (C-RNTI). However, if multiple devices choose the same preamble sequence and eNodeB cannot detect the collisions, they would receive same RAR message.

3) Connection Request Transmission: If devices can receive RAR message within the RAR window, they would send Msg 3 via the physical uplink sharing channel (PUSCH). Otherwise, this random access would be regarded as a failure and devices would proceed a back-off algorithm for retransmission. The Msg 3 conveys actual access requirements, such as RRC connection requests, tracking area update (TAU) or scheduling requests. Collided devices would receive the same RAR and transmit their connection request message via the same uplink resources. As a result, eNodeB cannot decode their RRC connection requests due to the mutual interference.
4) Contention Resolution: After receiving Msg 3, eNodeB would return acknowledgements to successful devices with the identification (ID) information in contention resolution messages. Devices which could find their IDs in the Msg 4 would continue the data transmission. Otherwise, devices realize that they have encountered collisions. Then they would go back to sleep mode, perform the back-off algorithm and re-initiate a new RAP until the limited retransmission times are reached.

### 2) MODELING OF RACH FOR HEADER-BASED DATA TRANSMISSION

The data transmission under LTE/LTE-A network has been described in detail above. Here, the corresponding analytical model would be derived.

Assume that $h(1 \leq h \leq min(N_p, 16))$ groups are paged in one paging message and there are $R$ available preamble sequences. Let the time interval between two continuous RASs be $T_{RA\_REP}$ and the RAR window size be $W_{RAR}$. Therefore, the number of access attempts from header devices to eNodeB is $I = h \times N_c$. These header devices randomly choose preambles for connection establishments and this is the typical balls-into-bins problem in the probability theory [29]. The probability that a ball falls into a bin is $1/R$. Then the probability that $w$ balls fall in one bin can be easily derived as

$$P(w) = \binom{I}{w}(\frac{1}{R})^w(1-\frac{1}{R})^{I-w}. \tag{14}$$

The network connections can be successfully established only when all 4 messages of RACH procedure are successfully finished. Hence, the successful probability for completing RACH procedures can be represented as

$$P_s(j) = P(1) = \frac{1}{R}(1-\frac{1}{R})^{I(j)-1} \approx \frac{e^{-I(j)/R}}{R}, \tag{15}$$

where $I(j)$ means the whole number of access attempts in the $j$th RAS [12]. Hence, the number of successful M2M access attempts in the $j$th RAS, $I_s(j)$, can be easily derived as

$$I_s(j) = I(j)P_s(j) = \frac{I(j)}{R}e^{-I(j)/R}. \tag{16}$$

In addition, the total number of access attempts in the $j$th RAS, $I(j)$, may include various devices with different retransmission after the back-off procedures. Therefore, it can be expressed as

$$I(j) = \sum_{n=1}^{N_{PT_{max}}} I_j[n], \tag{17}$$

where $n$ means the number of retransmission for header devices and $N_{PT_{max}}$ indicates the maximum retransmission limit. If the number of retransmission exceeds $N_{PT_{max}}$, this access attempt would be dropped. In addition, the maximum number of RARs that can be sent in one Msg2 message is $N_{RAR}$. Hence the maximum number of successful access devices within one RAR window is $N_{ACK} = N_{RAR} \times W_{RAR}$. Similarly, the successful number of devices in $j$th RAS, $I_s(j)$, can also be obtained as

$$I_s(j) = \begin{cases} \sum_{n=1}^{N_{PT_{max}}} I_j[n]e^{-I_j/R}, \\ \sum_{n=1}^{N_{PT_{max}}} I_j[n]e^{-I_j/R} \leq N_{ACK}. \\ \frac{\sum_{n=1}^{N_{PT_{max}}} I_j[n]e^{-I_j/R}}{\sum_{n=1}^{N_{PT_{max}}} I_j[n]e^{-I_j/R}} * N_{ACK}, \quad otherwise. \end{cases} \tag{18}$$

Obviously, the number of devices transmitting their first preambles is $I_j[1] = I = h \times N_c$. The retransmitted devices in the current RAS were collided before and completed their random back-off waiting time periods in latest RAS. They would start the RAP in the current RAS. Unlike the common access model under LTE/LTE-A network, there are no new arrivals except the first RAS. The number of access attempts in each RAS is monotonous decreasing. The number of devices which would transmit their $n$th preambles in the $j$th RAS, $I_j[n]$, can be expressed as

$$I_j[n] = \sum_{k=K_{min}}^{K_{max}} \alpha_{k,j} I_{k,F}[n-1], \tag{19}$$

where $K_{min}$ and $K_{max}$ are the minimal and maximal values of RAS when the devices found their $n-1$th retransmission failed and they could complete their back-off algorithms and start the RAP in the current RAS. $I_{k,F}[n-1]$ denotes the number of devices which failed their $n-1$th retransmission in the $k$th RAS.

Based on the RACH procedure described above, $\alpha_{k,j}$, $K_{min}$, $K_{max}$ can be derived according the temporal relations among the failed RA time slots, random back-off time period and the retransmitted RASs [9], [20]. The devices failing in $k$th RAS restart the RACH procedure in $j$th RA time slot after a random

back-off time with contention window of $W_{BO}$. Therefore, the expressions can be derived as

$$K_{min} = \lceil (j-1) + \frac{1-(W_{RAR}+W_{BO})}{T_{RA\_REP}} \rceil;$$
$$K_{max} = \lfloor j - \frac{W_{RAR}+1}{T_{RA\_REP}} \rfloor;$$
$$\alpha_{k,j} = \begin{cases} \frac{(k-1)T_{RA\_REP}+W_{RAR}+W_{BO}-(j-2)T_{RA\_REP}}{W_BO}, \\ \quad \text{if } K_{min} \le k \le j - \frac{W_{RAR}+W_{BO}}{T_{RA\_REP}} \\ \frac{T_{RA\_REP}}{W_{BO}}, \quad \text{if } j - \frac{W_{RAR}+W_{BO}}{T_{RA\_REP}} \le k \\ \quad \le j - 1 - \frac{T_{RAR}+W_{RAR}}{T_{RA\_REP}} \\ \frac{(j-1)T_{RA\_REP} - ((k-1)T_{RA\_REP}+W_{RAR})}{W_{BO}}, \\ \quad \text{if } (j-1) - \frac{W_{RAR}}{T_{RA\_REP}} \le k \le K_{max} \\ 0, \quad otherwise. \end{cases} \tag{20}$$

Hence, the successful access probability of header-based data transmission phase, the ratio between the number of successful M2M devices and the overall access attempts, can be derived as

$$P_{p2,s} = \frac{\sum_{j=1}^{T_R} I_j e^{-I_j/R}}{N_c \cdot h}, \tag{21}$$

where $T_R$ denotes the time duration of the paging interval.

Let $T_{p2}$ be the average access delay for devices to complete the RACH procedure under LTE/LTE-A networks. Hence, the average time needed for the header-based data transmission phase can be derived as

$$T_{p2} = \frac{\sum_{i=1}^{T_R} T_i I_i e^{-I_i/R}}{\sum_{i=1}^{T_R} I_i e^{-I_i/R}}, \tag{22}$$

where $T_i$ is the time consumption for devices successfully completing the RACH procedure in the $i$th RA time slot. As there are no new arrivals during the $T_R$ time period, the time delay of $T_i$ is $i$, i.e., $T_i = i$.

## III. MODELING AND OPTIMIZATION OF CBGP

Based on the analytical models for two phases of CBGP, the mathematical analysis of communications between massive M2M devices and the eNodeB would be given in this section.

Firstly, the time interval between two consecutive paging occasions in the CBGP scheme, $T_{GP}$, can be derived as below:

$$T_{GP} = T_{p_1,max} + T_{p_2,max};$$
$$T_{p_1,max} = (\sum_{k=0}^{m} 2^k CW_0 + m * T_0) * M_c;$$
$$T_{p_2,max} = 1 + (N_{PT_{max}} - 1)\lceil \frac{T_{RAR}+W_{RAR}+W_{BO}}{T_{RA\_REP}} \rceil. \tag{23}$$

where $T_{p_1,max}$ and $T_{p_2,max}$ respectively denote the maximum time for inner-cluster data collection period and header-based data transmission phase in CBGP scheme.

The average access delay of the proposed CBGP scheme is the sum of average delay spent on the inner-cluster data collection and data transmission. Hence, the expression of the average access delay, $T_{avg}$, could be obtained as

$$T_{avg} = T_{p1} + T_{p2}. \tag{24}$$

Similarly, the probability of transmitting data to the eNodeB successfully should also be considered. The overall successful data transmission implies that transmission of two phases should both be successful. Hence, the overall successful access probability of the proposed CBGP scheme, $P_{CBGP}$, could be derived as

$$P_{CBGP} = P_{p1,s} \times P_{p2,s}. \tag{25}$$

According to the above analysis, the proposed CBGP scheme can significantly alleviate the congestion. However, there is still a trade-off problem. Given the fixed number of M2M devices and paging groups, different numbers of clusters poses distinct influences on the overall performance of the proposed CBGP scheme. More clusters in the paging group would decrease the number of uplink attempts for the inner-cluster data collection phase. Then access conflicts during the inner-cluster data collection phase would be highly alleviated. However, it also increases the access attempts to perform the RAPs, leading to more collisions and longer time delay in the header-based data transmission phase. Due to the different performances of the two phases, it is necessary to figure out the effect from different numbers of clusters on the overall performance and derive the optimal number of clusters in order to dynamically suit massive access attempts from ubiquitous M2M devices.

As shown in Fig. 5(a), the average access delay and successful access probability for different numbers of MTC devices and clusters are presented. Each point in this figure means the performance of proposed CBGP scheme with a fixed number of clusters handling fixed amount of network requests.

For given amount of access attempts, different numbers of clusters leads to totally different performances. As we can see in Fig. 5(a), we assume that one paging message can page 10 paging groups. When the number of access attempts in one paging group is 2400, the performance can be maximized when the number of clusters, $N_c$, equals to 5. The average access delay is low and successful access probability is high. When access attempts increase to 24000, the performance of proposed CBGP scheme deteriorates for both the average access delay and successful access probability. However, there is still an optimal number of clusters for the given amount of access attempts, i.e., $N_c = 12$. Hence, there is always an optimal number of clusters for different amount of access attempts. The optimal number of clusters can be derived by minimizing the distance to point (0, 1) in Fig. 5(a),

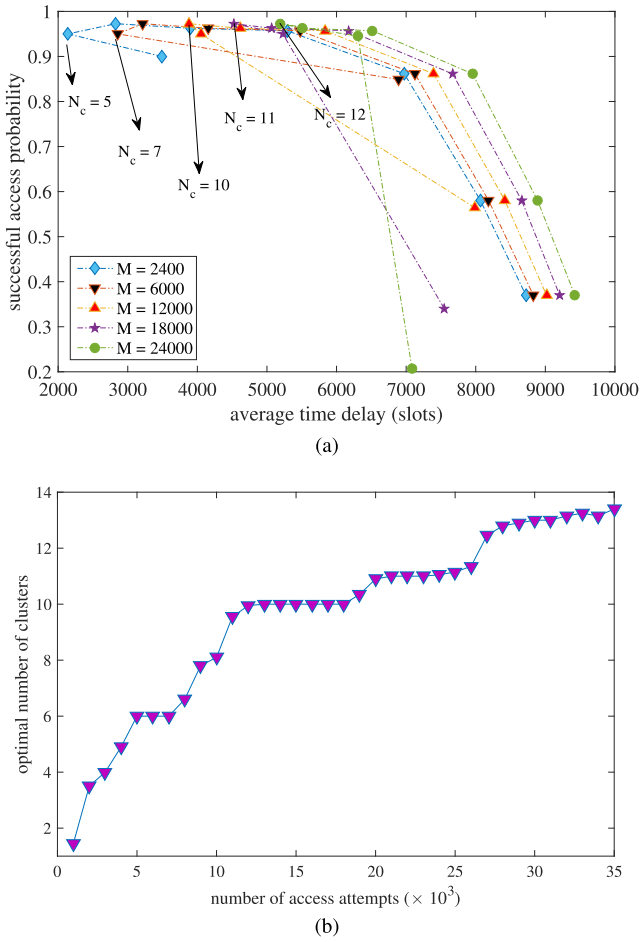



**FIGURE 5.** Effect of different numbers of clusters on the system performance. (a) Performance for different numbers of clusters and access attempts. (b) Optimal number of clusters for different numbers of access attempts.

maximizing the successful access probability and minimizing the average access delay.

Hence, we can define the utility function of the proposed CBGP scheme for the optimal number of clusters as

$$U_{N_c,M} = T^2_{N_c,M} + (1 - P_{s,N_c,M})^2, \tag{26}$$

where $T_{N_c,M}$ and $P_{s,N_c,M}$ separately represent the average access delay and successful access probability for the CBGP schemes with $N_c$ clusters when there are $M$ M2M devices. It reveals the squared distance between the specific node to the target point $(0, 1)$. Based on the analytical model of the proposed CBGP scheme, we could derive the optimal number of clusters that maximizing the utility function, i.e., $max\{U_{N_c,M}\}$. The derived optimal number of clusters is as shown in Fig. 5(b) and would be verified in the subsequent simulations.

## IV. SIMULATION AND ANALYSIS

Computer simulations are conducted on top of Matlab software to verify the effectiveness of proposed CBGP models, which are based on a Monte-Carlo approach. Each point represents the average value of 3000 samples for high accuracy.

In the proposed CBGP scheme, we may adjust the amount of access attempts in the inner-cluster data collection and header-based data transmission phases, number of clusters within a paging group. The parameter settings used in the simulations are summarized in Table 2 [12], [27]. For validation of the access performances of the proposed CBGP scheme, the channel fading, interference and hidden stations are not taken into consideration. The failed transmission is only the results of collisions.

**TABLE 2.** Simulation parameter settings.

| Parameters | Values |
|---|---|
| PRACH configuration Index | 6 |
| Total number of preambles | 54 |
| Time duration of LTE subframe | $1ms$ |
| Maximum retransmission times under LTE | 10 |
| Number of paging groups | 10 |
| Number of clusters | 2, 5, 8, 15 |
| Back-off indicator of LTE | 20 |
| Size of RAW | 10 |
| Minimum size of contention window | 3 |
| Maximum size of contention window | 10 |
| Number of MTC devices in each cluster | 0 to 35000 |
| Size of ACK | 120 $bits$ |
| Time duration of a back-off slot for DCF | 13 $\mu s$ |
| Maximum retransmission times under DCF | 8 |
| Size of RAW slot | $2ms$ |

Firstly, the accuracy of analytical models for the two phases of CBGP scheme would be validated. For inner-cluster data collection phase, the size of RAW in Wi-Fi Halow is fixed as 10 and size of RAW slot is $2ms$ [25]. In addition, for the header-based data transmission phase, an eNodeB would reserve 54 RAOs in each RAS ($R = 54$) to page a group size of $0 - 35000$ MTC devices without H2H traffic [12]. And the results are shown in Fig. 7, demonstrating that the proposed analytical models can accurately match two phases of the proposed CBGP schemes.

Afterwards, we would study the effect of different numbers of clusters on the proposed CBGP scheme ($N_c = 2, 5, 8, 15$) and verify the optimal number of clusters for different amount of access attempts. The proposed CBGP does perform better than the conventional GP and FI-TSFGP. Furthermore, the optimal number of clusters can be adaptively adjusted to maximize the successful access probability and minimize the average access delay according to the number of access attempts.

### A. EVALUATION OF PROPOSED ANALYTICAL MODEL FOR CBGP

The accuracy of proposed analytical models is evaluated in this subsection. The average access delay and successful access probability for massive M2M devices would both be simulated and presented as following.

Fig. 6 presents the evaluation of proposed analytical models for the inner-cluster data collection phase. The number of access attempts within the clusters is ranging from 0 to 10000. As we can see, when the number of access attempts is small at the beginning, the successful access probability is high

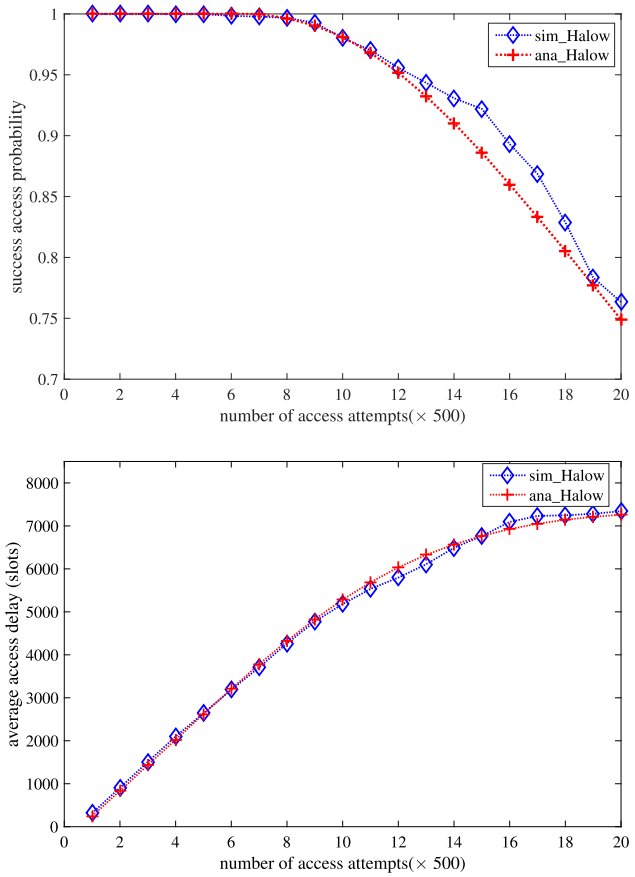



**FIGURE 6.** Evaluation of the proposed analytic model for inner-cluster data collection.

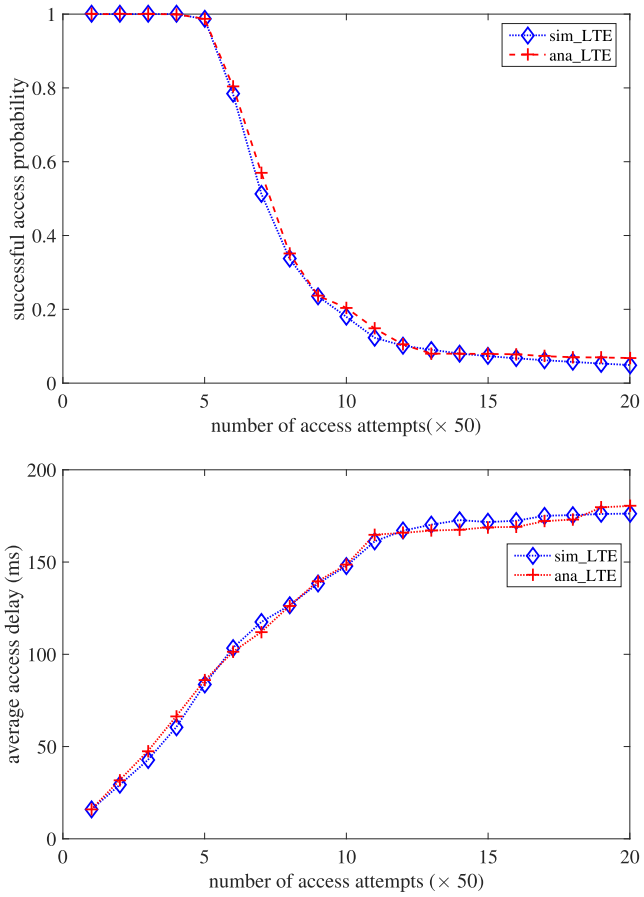



**FIGURE 7.** Evaluation of the proposed analytic model for data transmission.

and the average access delay is low. It is because that small amount of access attempts are separated into several RAW slots in the time domain. Less number of devices assigned to one RAW slot can hardly encounter conflicts. In addition, there is still a truncated binary exponential back-off mechanism for collision avoidance. Afterwards, with more access arriving, the average access delay gradually increases and the successful access probability starts to decrease when the number of access attempts is about 4500 as shown in Fig. 6. One RAW slot is likely to be distributed into more than two devices, which may cause conflicts during the DCF process. Longer back-off periods would be required for collided M2M devices. The retransmission would even expire the limitations and these network attempts would be dropped as real ''failure,'' decreasing the successful access probability. However, when more network requests arrive, the successful access probability approaches to zero and the average access delay becomes very high and remains steady as shown in Fig. 6. This is because that collisions become more severe and the access scale has expired the access capacity for Wi-Fi Halow. More network requests are dropped as ''failure,'' resulting in the sharp decrease of successful access probability. Successful devices mostly spend maximum retransmission constrain and the average access delay begins to keep steady. In conclusion, the proposed analytical model of inner-cluster data collection phase accurately matches the simulation results with tiny deviations.

For header-based data transmission under LTE cellular networks, the evaluations of proposed mathematical models are presented as shown in Fig. 7. The number of access attempts per RAS for header-based data transmission is ranging from 0 to 1000. Firstly, the successful access probability keeps as high as 1 and the average access delay increases slowly from a very low value. It is because that radio access resources of LTE/LTE-A cellular network are sufficient with respect to the small number of access attempts. All the required connections can be easily established without or with a few collisions. Only a small number of retransmission and back-off procedures are needed. Hence, the successful access probability keeps as high as 1 and the average access delay keeps low. Gradually with more access arriving, the average access delay begins to increase significantly and the successful access probability also starts to decrease sharply. This is due to that the access resources become deficient, leading to more collisions, more retransmission and more back-off processes. What is worse, expiration of devices' retransmission causes the transmission failure and decrease the successful access probability. More back-off processes and retransmission mean higher access delay for M2M devices. Afterwards, the successful access probability even decreases

to 0 and the average access delay also increases to about 170*ms* and remains steady as shown in Fig. 7. The access collisions become more severe. The number of network requests is so high that the back-off algorithm could not work for the congestion alleviations. Devices cannot complete the RACH procedure after limited retransmission, leading to more dropped devices. The small number of successful access almost spends all their retransmission times, making their access delays keep steady. To sum up, the successful access probability and average access delay from the proposed analytical models are consistent with the simulation results.

### B. EVALUATION OF PROPOSED CBGP

In this subsection, to evaluate the validity of the proposed mechanism, the CBGP scheme with different numbers of clusters, conventional GP mechanism and FI-TSFGP scheme are simulated. Moreover, the effects from different numbers of clusters on the performances of CBGP schemes would also be investigated. Finally, the optimal number of clusters would be verified as derived from the mathematical models shown in Fig. 5(b). We assume there are 10 paging groups in this area and the number of access attempts in one paging group varies from 0 to 35000 as shown in Table 2. As the average access delay of FI-TSFGP scheme is too large with respective to other CBGP schemes and conventional GP mechanism, we give the detailed presentation for the CBGP schemes with different clusters and conventional GP scheme in Fig. 8(c).

Firstly for the CBGP schemes with a small number of clusters and conventional GP scheme, when the amount of access attempts, $M$, is small, the successful access probability keeps as high as 1 and the average access delay is low as shown in Fig. 8. The limited resources are sufficient for the inner-cluster data collection phase via Wi-Fi Halow and the number of access attempts for the header-based data transmission phase is also small. There are no dropped or blocked attempts and the average access delays increase slowly due to a few back-off processes and retransmission. Then when the access scale increases to $M = 3000$, the conventional GP scheme experiences sharp deteriorations on the successful access probability and growth on the average access delay. The access resources for the GP scheme are relatively insufficient with respect to the number of access attempts, leading to severe collision and congestion. After the constrained retransmission, the failed devices still encounter collisions and would be dropped at last. Devices require more retransmission times to complete the connection establishments, resulting in longer access time delay.

Gradually, when there are more network requests, the performances of CBGP schemes with a small number of clusters also begin to deteriorate. Both the successful access probability and average access delay start to worsen as shown in the Fig. 8(a) and (c). It is due to the conflicts happened during the inner-cluster data collection phase. As shown in Fig. 6, the increased access attempts for inner-cluster data collection phase would cause severe conflicts. The small and fixed number of clusters means that there are only a small

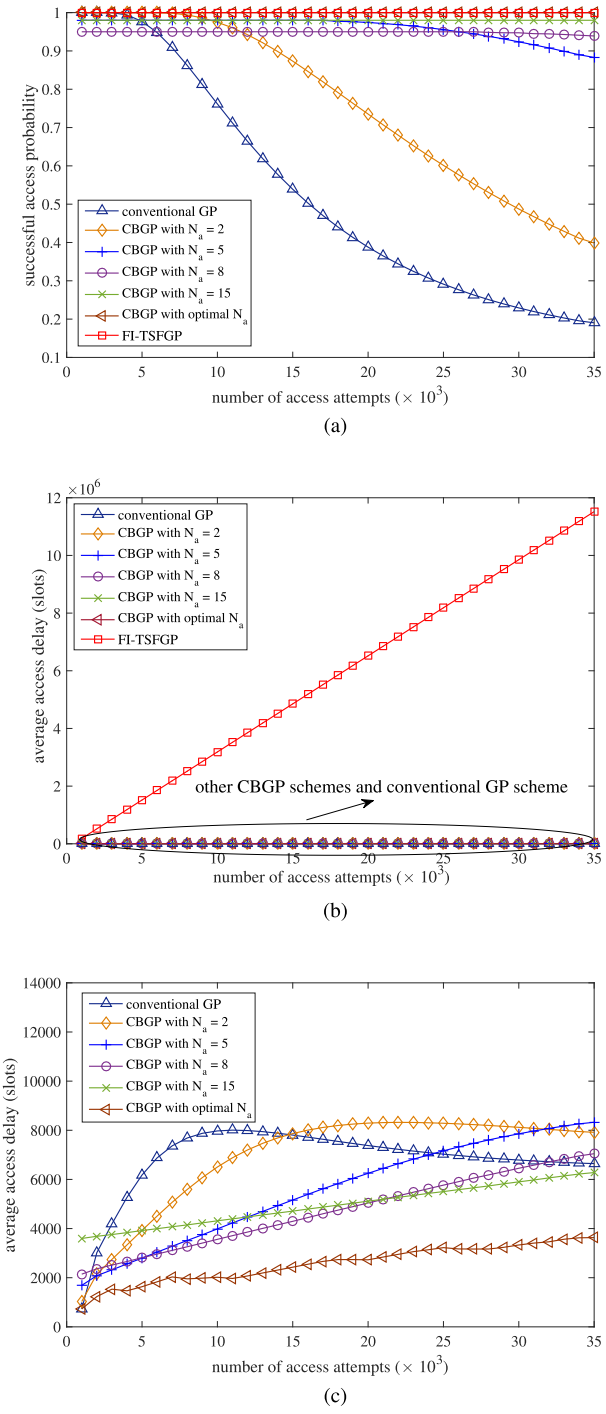



**FIGURE 8.** Evaluation of the proposed analytic model for data transmission. (a) Comparison of successful access probability. (b) Comparison of average access delay (1). (c) Comparison of average access delay (2).

amount of access attempts for the cellular network. Hence, the congestion of header-based data transmission phase is negligible.

Afterwards, when the access scale becomes higher than about 10000, the average access delay of conventional GP scheme begins to decrease gradually and is even lower than

the proposed CBGP mechanism. This is because that congestion becomes severer and finite retransmission times cannot ensure the complements of connection establishments. Only a few MTC devices accomplish the RAPs occasionally. Then the overall time delay of the successful access attempts decrease and average access delay also degrades, which is lower than the proposed CBGP scheme. At $M = 20000$, the CBGP with 2 clusters also demonstrates the same changes to the average access delay as shown in Fig. 8(c).

Besides, for more clusters in the paging group, the successful access probabilities always keep constantly high but lower than 1 for a long time with the increasing access attempts. And the average access delays keep higher than others in the face of a few access attempts as shown in Fig. 8, e.g., $N_c = 8$ and $N_c = 15$. This is because that there would be more devices contending for the RACH procedure under LTE network. No matter how many access attempts there are, the congestion from RACH procedures always exists. Hence, the successful access probability cannot be as high as 1 at the beginning. Fewer devices transmit data in the inner-cluster data collection phase with few conflicts and the successful access probability would maintain the same until there are severe collisions happened in the inner-cluster data collection phase. For the same reason token, the average access delay is higher for the larger number of clusters at the beginning as shown in Fig. 8(c). However, as more and more access attempts come, the average access delays of CBGP schemes with a larger number of clusters become lower than that with a smaller number of clusters, just as shown in Fig. 8(c). It is due to the severe collisions also happened in the inner-cluster data collection phase. Massive access attempts arrive, leading to severer conflicts in the inner-cluster data collection phase. Dividing paging groups into more clusters means that there are fewer access attempts for the first phase of the CBGP scheme. Hence, when the number of access attempts is small, the proposed CBGP schemes with a small number of clusters perform better. And the CBGP schemes with large number of clusters achieve a better performance in the face of massive access attempts.

In addition, the FI-TSFGP scheme can achieve high successful access probability as high as 1 as shown in Fig. 8(a). However, this is at the cost of the access delay, which is far higher than the CBGP schemes as shown in Fig. 8(c). The number of arrived devices per RAS is fixed to keep stable state. When there are massive access attempts arriving, the number of RASs needed will be large, leading to long queuing delay. MTC devices need to wait for their RASs to access the network. As the number of access attempts per RAS is stable, causing the delay linearly increases with the number of access attempts as shown in Fig. 8(b).

Finally as shown in Fig. 8, the CBGP scheme with optimal number of clusters keeps highest successful access probability and lowest average access delay compared to other CBGP schemes with different fixed numbers of clusters and the FI-TSFGP scheme. In order to obtain the goal of best performance, the number of clusters is dynamically adjusted according to the amount of access attempts. Through the pre-derived optimal value of clusters, the CBGP scheme would keep adaptive to dynamic number of access attempts from mMTC devices, maximizing the successful access probability and minimizing the average access delay.

In conclusion, the simulation results show that the proposed CBGP scheme performs better than the traditional GP in terms of successful access probability and average access delay. In addition, the optimal number of clusters for CBGP schemes is also verified, dynamically adaptive to the number of access attempts.

## V. CONCLUSION

In this paper, we propose a CBGP scheme for congestion and overload control under the mMTC scenarios. Based on the current GP mechanism, paging groups are further divided into clusters for data collection. Due to the advantages of low cost, high access capacity and handy deployment, IEEE 802.11ah is utilized to gather the sensed data from devices in the clusters and headers upload the collected data to the LTE/LTE-A cellular network. Additionally, the corresponding mathematical models are derived, which can characterize the performance of proposed CBGP scheme. Furthermore, the effect from different numbers of clusters on the performance of CBGP is also investigated and the optimal number of clusters can be derived, aiming to obtain the best performance in the face of different access scales. Finally, numerical results illustrate that the analytical model matches well with the simulation results and effectiveness of proposed CBGP scheme is validated. The optimal number of clusters for proposed CBGP is also verified, adaptively adjusted according to the number of access attempts.

## REFERENCES

[1] P. Annamalai, J. Bapat, and D. Das, "Constellation constraining-based coverage enhancement technique for MTC devices in LTE-A," *IEEE Wireless Commun. Lett.*, vol. 5, no. 6, pp. 596–599, Dec. 2016.

[2] Z. Alavikia and A. Ghasemi, "A multiple power level random access method for M2M communications in LTE-A network," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 6, p. e3137, 2017.

[3] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A survey of traffic issues in machine-to-machine communications over LTE," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 865–884, Dec. 2016.

[4] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[5] J. Guo, S. Durrani, X. Zhou, and H. Yanikomeroglu, "Massive machine type communication with data aggregation and resource scheduling," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 4012–4026, Sep. 2017.

[6] "Cellular networks for massive IoT," Ericsson, Stockholm, Sweden, Tech. Rep. Uen 284 23-3278, Jan. 2016. [Online]. Available: https://www.ericsson.com/res/docs/whitepapers/wp iot.pdf

[7] G. Farhadi and A. Ito, "Group-based signaling and access control for cellular machine-to-machine communication," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2013, pp. 1–6.

[8] *System Improvements for Machine Type Communications*, document Std. TR 23.888, 3GPP, Dec. 2011.

[9] 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.

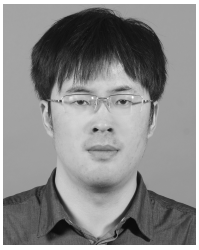
[10] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A. C.-C. Hsu, "Overload control for machine-type-communications in LTE-advanced system," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 38–45, Jun. 2012.

[11] *Comparing Push Pull Based Approaches for MTC*, document RAN2#71, Std. R2-104 873, 3GPP, 2010.
[12] *Study RAN Improvements for Machine-Type Communications*, document TR Std. TR 37.868, 3GPP, Sep. 2011.
[13] *Further Analysis of Group Paging for MTC*, document Std. R2-113198, 3GPP, May 2011.
[14] C.-H. Wei, R.-G. Cheng, and F. M. Al-Taee, "Dynamic radio resource allocation for group paging supporting smart meter communications," in *Proc. IEEE 3rd Int. Conf. Smart Grid Commun. (SmartGridComm)*, Nov. 2012, pp. 659–663.
[15] 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.
[16] R. Harwahyu, X. Wang, R. F. Sari, and R.-G. Cheng, "Analysis of group paging with pre-backoff," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 34, Feb. 2015.
[17] R. Harwahyu, R.-G. Cheng, and R. F. Sari, "Consecutive group paging for LTE networks supporting machine-type communications services," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 1619–1623.
[18] O. Arouk, A. Ksentini, and T. Taleb, "Group paging optimization for machine-type-communications," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 6500–6505.
[19] O. Arouk, A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "On improving the group paging method for machine-type-communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 484–489.
[20] 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.
[21] S.-H. Wang, H.-J. Su, H.-Y. Hsieh, S.-P. Yeh, and M. Ho, "Random access design for clustered wireless machine to machine networks," in *Proc. 1st Int. Black Sea Conf. Commun. Netw., (BlackSeaCom)*, Jul. 2013, pp. 107–111.
[22] 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, 2015.
[23] 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.
[24] M. Park, "IEEE 802.11ah: Sub-1-GHz license-exempt operation for the Internet of Things," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 145–151, Sep. 2015.
[25] L. Tian, E. Khorov, and S. Latré, and J. Famaey, "Real-time station grouping under dynamic traffic for IEEE 802.11ah," *Sensors*, vol. 17, no. 7, p. 1559, 2017.
[26] E. Khorov, A. Lyakhov, A. Krotov, and A. Guschin, "A survey on IEEE 802.11ah: An enabling networking technology for smart cities," *Comput. Commun.*, vol. 58, pp. 53–69, Mar. 2015.
[27] C. W. Park, D. Hwang, and T.-J. Lee, "Enhancement of IEEE 802.11ah MAC for M2M communications," *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1151–1154, Jul. 2014.
[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] R. Motwani and P. Raghavan, "Randomized Algorithms," *Algorithms and theory of computation handbook*. M. J. Atallah M. Blanton, Eds. London, U.K.: Chapman & Hall, 2010, p. 12.

**QI PAN** was born in Linyi, China, in 1990. He received the B.S. degree in electronic and information engineering from the Ocean University of China, Qingdao, China, in 2014. He is currently pursuing the Ph.D. degree in information and communication engineering with the Beijing University of Posts and Telecommunications, Beijing, China. His research interests mainly focus on machine-type communications, Internet of Things, random access mechanism, and offloading in heterogeneous networks and future networks.

**XIANGMING WEN** received the B.E., M.S., and Ph.D. degrees in electrical engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China. He is currently the Vice President of BUPT, where he is also a Professor with the Communication Network Center and the Director of the Beijing Key Laboratory of Network System Architecture and Convergence. In the last five years, he has authored over 100 papers. His current research interests include broadband mobile communication theory, multimedia communications, and information processing. He is the Vice Director of the Organization Committee of the China Telecommunication Association. He is the Principle Investigator of over 18 projects, including the National Key Project of Hi-Tech Research and Development Program of China (863 Program) and the National Natural Science Foundation of China.

**ZHAOMING LU** received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2012. He joined the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, in 2012. His research interests include open wireless networks, QoE management in wireless networks, software-defined wireless networks, and cross-layer design for mobile video applications.



**WENPENG JING** received the B.S. degree in communication engineering from Shandong University in 2012 and the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications in 2017. He is currently a Post-Doctoral Researcher with the Beijing University of Posts and Telecommunications. His research interests include radio resource allocation, energy efficiency optimization, and interference management in heterogeneous networks.

**LINPEI LI** received the B.S. degree in electronic information engineering from Northeastern University at Qinhuangdao in 2015. She is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. Her research interests include Internet of Things, wireless communications with unmanned aerial vehicles, and energy-efficient problems in UAV-aided communications.