Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2015, Article ID 703486, 10 pages
http://dx.doi.org/10.1155/2015/703486

Research Article

Congestion-Aware Signaling Aggregation and Barring Scheme for LTE Based Machine-to-Machine Network

Fangmin Xu, Chao Qiu, Pengbiao Wang, and Xiaokai Liu

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Fangmin Xu; xufm@bupt.edu.cn

Received 27 March 2015; Accepted 8 June 2015

Academic Editor: Sherwood W. Samn

With the recently progress of Machine-to-Machine (M2M) communication technology, especially the enormous M2M devices and unique service of M2M, some challenges are emerging to the traditional wireless access and core networks, especially the congestion problem due to simultaneously bursty M2M service. Following this paradigm, the purpose of this paper is to support and optimize the signaling aggregation and barring of M2M services based on cellular network. With LTE network being the example access network, a congestion-aware signaling aggregation and barring scheme is designed considering the various requirements of M2M services and the congestion situation in the network entity. Theoretical analysis and experimental simulations show that this scheme can improve the system efficiency and greatly alleviate the signaling congestion, especially for the bursty M2M service.

1. Introduction

Machine-to-Machine (M2M) communication (or Machine Type Communication (MTC) or Internet of Things) is a revolutionary innovation for information society in the last decade. It is a dynamic global network infrastructure based on standard and interoperable communication protocols where physical and virtual “things” are seamlessly integrated into the information network through radio-frequency identification (RFID), short range wireless communication devices, and various sensors [1].

With the cooperation of enormous smart M2M devices (e.g., the smart metering device), tremendous data traffic from massive devices would be gathered and transmitted simultaneously to access network entities which are designed for traditional Human-to-Human (H2H) communications [2]. Firstly the tremendous data and signaling may lead to traffic and signaling overload situations and may have a great impact on the operations of the cellular network; secondly, Quality of Service (QoS) guarantee schemes of cellular network which was originally designed for H2H communication may be challenged by unique services of M2M, such as small traffic, real-time transmission, burst, and low mobility [3].

The majority of signaling congestion avoidance approaches could be classified into the grouping mechanism and the access class barring scheme. Grouping mechanism reduces the signaling overload by grouping the M2M devices into one group. The M2M group delegates implement the signaling aggregation or compression. Access class barring scheme permits only parts of M2M devices to access the network in the case of congestion.

Taking the LTE network as the example access network of M2M network, to address the signaling congestion problem in LTE based M2M network, we propose a novel solution, congestion-aware signaling aggregation and barring (CASAB). The approach could alleviate the signaling congestion automatically by the simple cooperation of network entity and M2M devices. The network entity informs the M2M gateway about the possible occurrence of congestion and the estimate signaling arrival rate. Consequently, M2M gateway adjusts the signaling aggregation level and the barring probability for different M2M services accordingly considering both the delay requirement of different services and the congestion situation.

The rest of this paper is organized as follows. Firstly, the architecture of LTE based M2M network, especially the possible congestion factors, and existing solutions are introduced.

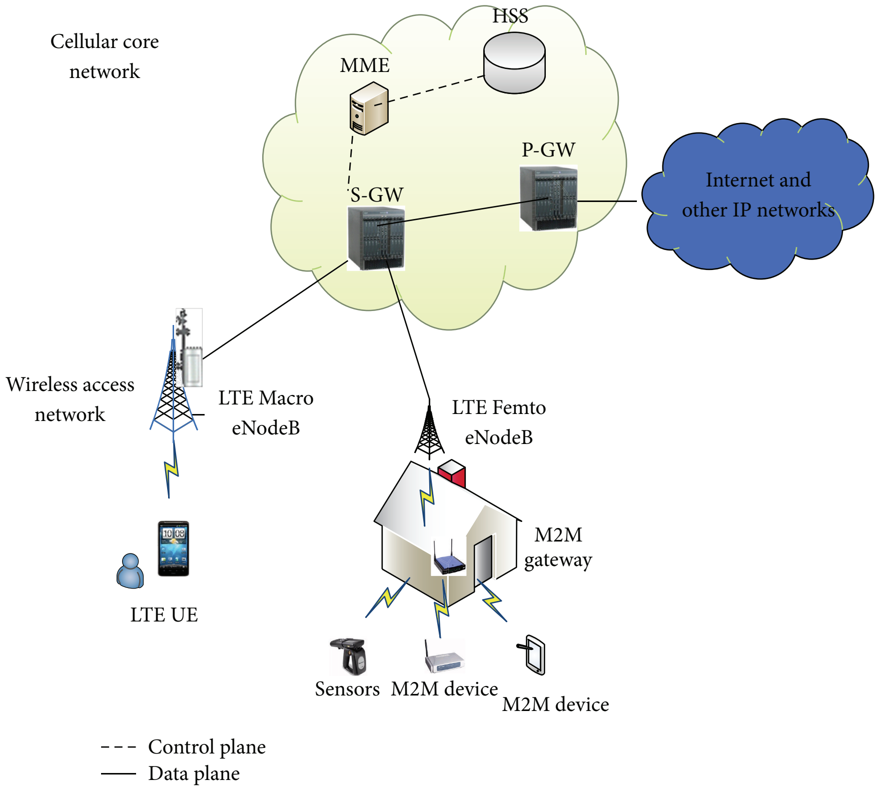

Figure 1: Architecture of LTE network supporting M2M Service.

In Section 3, the signaling aggregation mechanism and the analytical model are discussed. We present a general and detailed description of our proposed solution, CASAB, and the key function modules of CASAB. The proposed solution is then evaluated through simulations. Finally, we highlight some future research work and conclude this paper.

2. M2M Network Architecture and Related Work

As illustrated in Figure 1, M2M devices that belong to the same M2M domain access the LTE network through the M2M gateway and connect to the M2M servers via the LTE access and core network. The M2M gateway acts as an entry point in the access network [4].

LTE eNodeB (including Marco and Femto eNodeB) is the base station for LTE network which sends and receives radio transmissions to all mobile terminals via air interface. Serving gateway (S-GW) acts as a router and forwards data between the base station and the Packet Data Network (PDN) Gateway. The PDN Gateway (P-GW) is responsible for the communication with the internet, IP address allocation for the UE, QoS enforcement, and flow-based charging. The Mobility Management Entity (MME) controls the high-level operation of the mobile by means of signaling messages and Home Subscriber Server (HSS) [5, 6].

Congestion may appear in both the access and core network, possible in the form of data congestion and signaling congestion. The congestion of data traffic in M2M happens rarely since M2M devices send and receive small amounts of data. But it may frequently happen that lots of devices send their data simultaneously, leading to congestion mainly in the core part, especially in the S-GW/P-GW [7].

The congestion of signaling appears in all the architecture. It is due to the fact that the devices continuously generate signaling message to attach to the network, when triggering from initial attachment, transmitting data and alerts, bearer management, and so forth, which causes signaling overhead and congestion in MME mainly [8]. In this paper, we focus on the signaling congestion in MME.

Congestion problem in LTE based M2M network has been addressed in previous literatures. The majority of signaling congestion avoidance approaches could be summarized into following two categories.

(1) Grouping Based Mechanism. Group parsing and aggregation of SIP message in LTE based M2M network to reduce the message overload are proposed in [9]. A designated device (group delegate) is chosen to implement the random access procedure of the devices in an access group [10]. The devices are grouped based on the similarity of their mobility patterns, and only the leader performs mobility management on behalf of the other devices in the same group [11]. The overload of the signaling message content for a group of MTC devices sharing redundant information elements is compressed in [12]. However, the group parsing and aggregation procedure will introduce inevitable process delay and complexity.

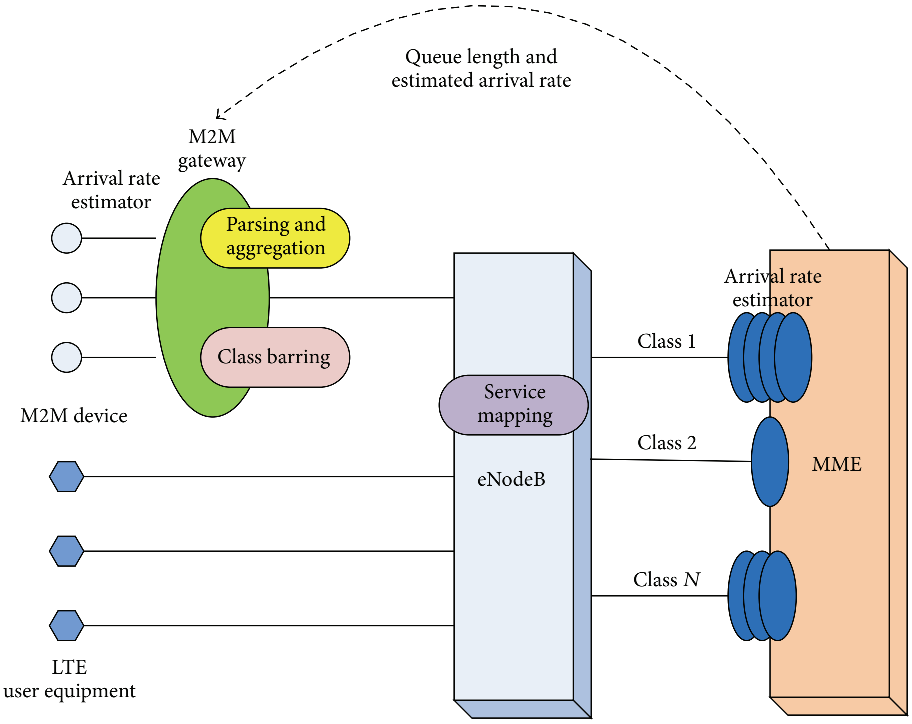

Figure 2: Function modules of proposed CASAB scheme.

(2) Allocation for Each M2M Device Class of Forbidden/Grant Times. Access class barring (ACB) [13] is one of the typical schemes discussed in 3GPP to reduce the collision probabilities of random access resource. LTE eNodeB broadcasts a probability p called ACB factor. Each M2M device determines whether it is temporarily barred from accessing the cell with probability $1-p$. However, the reject probability for each class is difficult to dynamically calculate and adjust.

In summary, CASA scheme proposed in our paper employs the advantage of group aggregation and class barring. Based on the queue statistical information feedback from the network entity, M2M gateway adaptively adjusts the signaling aggregation parameters of various M2M services.

3. Signaling Aggregation Mechanism for M2M

The system model and function modules of CASAB mechanism are given in Figure 2. There are four basic function modules: parsing, aggregation, class barring, and service mapping.

Signaling parsing module reads out the key fields of incoming message, including the source address, destination address, and signaling type. Signaling aggregation implements the regeneration of signaling message from several signaling messages according to predefined rules in a predefined period. There are two key parameters for parsing and aggregation: aggregation level (number of aggregated signaling messages) and length of buffer period. The calculation of the two parameters is determined by the feedback information (including the queue length and estimated signaling arrival rate of each class) from the MME. The detail of the design will be discussed later.

Similar to ACB scheme in LTE, class barring function rejects the access request of signaling randomly when severe congestion happens.

Service mapping function classifies the M2M service and LTE service based on the QoS requirement of various services (such as tolerate delay, delay jitter, and transmission rate) into several classes. The design of mapping criteria is out of the scope of our paper. Literature [14] has discussed the approach of mapping service of different network extensively.

The MME will count the queue length and estimate the arrival rate of each class periodically. The estimation methods have been investigated in prior works [15]. A simple Maximum Likelihood (ML) estimation approach is based on the raw cumulative sum of the arrival signaling.

3.1. Signaling Parsing and Aggregation. Considering the bursty feature of M2M service (the signaling is generated in short period with the same destination and service type), the signaling parsing and aggregation scheme for M2M work as follows:

(1) M2M device initiates signaling message (such as *RRCConnectionRequest*), in an example format as shown in Algorithm 1.

(2) When M2M gateway receives signaling from other M2M devices, it stores the signaling in the buffer. There are several buffers corresponding to different service types.

(3) If the buffer is not empty, the timer is started (denote the length of the timer by T_{buff}, as shown in Figure 3(a)).

(4) If the timer is expired or the number of signaling messages in the buffer is equal to the maximum number of aggregated messages, which is represented by the parameter aggregation level, denoted by L_A ($L_A \geq 1$), M2M gateway

```
struct signaling{
uint32_t src_id;    // Unique Source ID, example: china_bupt_sice_ax2154
uint32_t dst_address;    // Destination Address, example: 10.2.1.3
uint4_t serv_type; // M2M Service Type, example: Smart_Meter
uint4_t sig_type; // M2M Signaling Type, example: Connection_Request
. . .        // Other fields    };
```

Algorithm 1

```
<?xmlversion="1.0", encoding="UTF-8",
xmlns:xsi="http://www.w3.org/M2Mparsing-instance">
<complexType>
   <sequence>
   <element name="dst_address"> // Collect the destination address
   <element name ="serv_type"> // Collect the service type
   <element name ="sig_type"> // Collect the signaling type
   </sequence>
   <element name="src_id"> // Collect the source ID
</complexType>
```

Algorithm 2

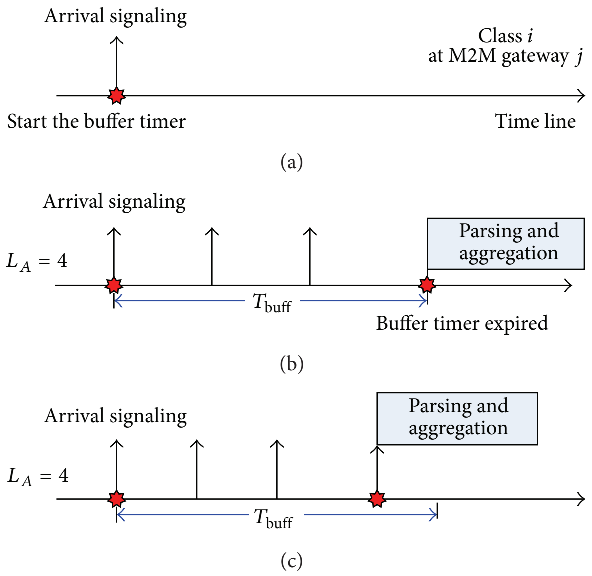

Figure 3: Illustration of signaling aggregation and buffer.

parses and aggregates the signaling messages in the buffer based on the aggregation level. The parsing and aggregation can be simply depicted by configured XML document as shown in Algorithm 2.

The principle is parsing the signaling depending on the *dst_address*, *serv_type*, and *sig_type* fields. Signaling with the same *dst_address*, *serv_type*, and *sig_type* fields could be aggregated by listing the element in *src_id* into the aggregated signaling.

The example of aggregated signaling is shown in Algorithm 3.

In case $L_A = 1$, the parsing and aggregation procedure is skipped.

(5) M2M gateway sends the aggregated messages to LTE eNodeB.

In summary, there are two key parameters that could be adjusted according to congestion status: T_{buff} and L_A. The procedure to determine the two parameters will be discussed later in detail.

The procedure is given by some examples in Figure 3. There are two possible scenarios: scenario (b) corresponds to the case in which the timer of buffer is expired, while scenario (c) illustrates the case in which a maximum number of aggregated messages have been collected before timer expired.

3.2. Service Mapping. As in [16], M2M services could be classified into four classes based on the QoS requirements and mapping with traditional H2H services defined in LTE as follows:

(i) Conversational (Class 1): these services requires a given delay constraint, such as traditional voice service and emergency alarm application in M2M.

(ii) Streaming (Class 2): applications like audio and video services are delay jitter sensitive rather than delay sensitive. Typical M2M streaming application includes remote monitoring in e-Health/e-Home services.

(iii) Interactive (Class 3): interactive traffic is broadly characterized by the request response pattern of the user. Round trip moderate delay is one of the key features. Typical interactive application includes web browsing.

(iv) Best efforts/background (Class 4): data traffic of applications such as e-mail and file download could be

```
struct signaling{
uint32_t src_id;    // Unique Source ID, example: china_bupt_sice_ax2154, china_bupt_sice_dx246x
uint32_t dst_address;    // Destination Address, example: 10.2.1.3
uint4_t serv_type; // M2M Service Type, example: Smart_Meter
uint4_t sig_type; // M2M Signaling Type, example: Connection_Request
. . .        // Other fields    };
```

Algorithm 3

Table 1: Mapping relationship between M2M applications and LTE services.

	Class 1 (conversational)	Class 2 (streaming)	Class 3 (interactive)	Class 4 (background)
M2M services	Security, payment	Health, consumer device	Tracking, remote control	Metering
LTE services	Voice	Video	Web browsing	FTP
Priority	High	Moderate	Moderate	Low
Features	Low latency	Low delay jitter	Moderate latency	No grantee

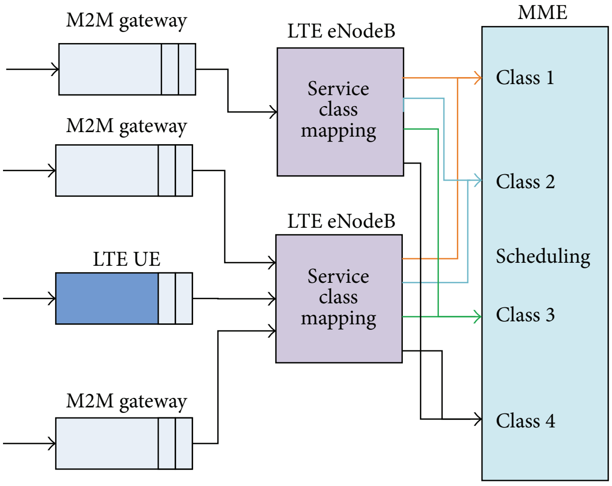

Figure 4: System model for signaling scheduling at MME.

delivered in the background since such applications do not require tight delay.

The mapping relation of M2M services and LTE services is given in Table 1. For each class, there is a signaling queue converged at the MME from attached LTE eNodeBs. Furthermore, there is a maximum tolerate delay for signaling of each class, denoted by $D_m(i)$ for class i ($i \in \{1, 2, 3, 4\}$).

3.3. Congestion Status Feedback and Parameter Adjustment. As shown in Figure 4, the converged signaling at the MME could be modeled as several individual queue processes.

Since we are considering aggregate flows originated at multiple sources (LTE eNodeBs), it is reasonable to assume that the arrival process is locally Poissonian at least at small-time scales. Denote the signaling queue length of class i at the MME at time instance $k * T$ by $L_Q(i, k)$, where T is the length of estimation period. A simple ML based traffic arrival rate estimator would be used at the MME.

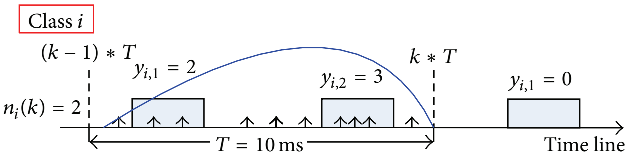

Figure 5: Illustration of arrival rate estimation.

Each flow is observed by an independent monitor that measures the number of arrivals in n_i time slots of fixed duration T and reports the raw measurements to the estimator. The estimation of arrival rates is carried out over an observation window of length t wherein the arrival processes can be assumed to be stationary and Poisson, for example, 1 second or 1 minute depending on the application. For each class $i \in \{1, \ldots, N = 4\}$ we introduce the following variables:

(i) $\lambda(i, k)$: the (unknown) arrival rate of class i to be estimated at time instance $k * T$;

(ii) $n_i(k) \geq 1$: the counting process representing the number of measurements during the estimation period, while each measurement requires unit time, such as 1 ms;

(iii) $y_{i,l}$ (with $l = 1, \ldots, n_i$): the lth measurement of the number of arrivals in the lth observation window during $[(k-1) * T, k * T]$, that is, $y_{i,l} \sim \text{Poisson}(\lambda(i, k))$ i.i.d. Note that the measurement should count the arrival signaling before aggregation.

As shown in Figure 5, $\lambda(i, k)$ is estimated at time $k * T$ based on the sample $y_{i,l}$ ($l = 1, \ldots, n_i(k)$) by simply computing the empirical mean of the available measurements by the cumulative sum of the measurements. This yields the Maximum Likelihood estimator of $\lambda(i, k)$; that is,

$$\tilde{\lambda}(i, k) = \frac{1}{n_i(k)} \sum_{l=1}^{n_i(k)} y_{i,l}. \quad (1)$$

As the M2M signaling arrival has autocorrelation feature with last period, through the period unknown to the estimator, time serial models could be used to reduce the estimation error of ML estimator. Considering the difference service feature of M2M application and LTE services, the smooth of average arrival rate is treated differently for difference service class.

(1) For normal M2M service or LTE services at M2M gateway or MME, the arrival rate of class i at time instance $k * T$ could be estimated by

$$\lambda(i,k) = \theta * \widetilde{\lambda}(i,k) + (1-\theta) * \widetilde{\lambda}(i,k-1), \quad 0 \leq \theta \leq 1, \tag{2}$$

where θ is the forget factor. By (1) and (2), at MME, the arrival rate for M2M signaling (class i) and LTE signaling (class i) could be obtained separately by checking the *serv_type* field of signaling, represented by $\lambda_{\mathrm{M2M}}(i,k)$ and $\lambda_{\mathrm{LTE}}(i,k)$, respectively.

(2) For bursty M2M service at M2M gateway, since for most of the time there is no bursty traffic, the estimated arrival rate is quite low toward 0. By setting a threshold Th_b, the bursty period and normal period could be classified. The threshold Th_b could be adjusted automatically. The detailed procedures are given in the following.

Initial setting $\mathrm{Th}_b = 0.01$, $\lambda_{\mathrm{hist}}(i) = 0$.

At time instance $k * T$, estimate the average arrival rate and update the historical rate $\lambda_{\mathrm{hist}}(i)$ and the threshold Th_b by

$$\begin{gathered}
\lambda(i,k) = \begin{cases} \theta * \widetilde{\lambda}(i,k) + (1-\theta) * \lambda_{\mathrm{hist}}(i) & \text{if } \widetilde{\lambda}(i,k) \geq \mathrm{Th}_b \\ \widetilde{\lambda}(i,k) & \text{else,} \end{cases} \\
\lambda_{\mathrm{hist}}(i) = \begin{cases} \lambda(i,k) & \text{if } \widetilde{\lambda}(i,k) \geq \mathrm{Th}_b \\ \lambda_{\mathrm{hist}}(i) & \text{else,} \end{cases} \\
\mathrm{Th}_b = \begin{cases} \lambda(i,k) * \eta & \text{if } \widetilde{\lambda}(i,k) \geq \mathrm{Th}_b,\ \widetilde{\lambda}(i,k) \leq \lambda(i,k) \\ \mathrm{Th}_b & \text{else,} \end{cases}
\end{gathered} \tag{3}$$

where η is the trigger factor based on the required false alarm rate (the larger η, the smaller the false alarm rate). Note that, in above estimation procedure, we ignore the notation in subscript to identify the estimation location (at MME or M2M gateway) and the service type (LTE service or M2M service) for the sake of simplification. However, it must be kept in mind that the rate is estimated whether in MME or M2M gateway and whether for LTE service or M2M service.

In a word, the MME will feedback the estimated arrival rates $\lambda_{\mathrm{M2M}}(i,k)$ and $\lambda_{\mathrm{LTE}}(i,k)$ and length of queue $L_Q(i,k)$ to M2M gateway in following two modes:

Table 2: Periodic and aperiodic feedback message example.

Periodic	Estimated arrival rate	Queue length
Class 1	λ_1	$L_Q(1)$
Class 2	λ_2	$L_Q(2)$
$\vdots$	$\vdots$	$\vdots$
Class N	λ_N	$L_Q(N)$
Aperiodic	Estimated arrival rate	Queue length
Class i	λ_i	$L_Q(i)$

(i) Periodic feedback: MME feedback the estimated arrival rate and queue length periodically; the feedback period is the same as the estimation period T. For instance, the estimation time is $k * T$ ms and the feedback time is $k * T + 4$ ms, where 4 ms is the processing delay. The example of feedback message format is given in Table 2.

(ii) Aperiodic feedback: since the period of periodic feedback is relatively long, to feedback latest congestion status, aperiodic feedback could be triggered anytime. When the queue length is larger than a certain number, or the estimated signaling rate is higher than a predefined threshold, the MME will generate the feedback message and send it to the M2M gateway.

3.4. Congestion Control Related Parameters Calculation.

The M2M gateway decides T_{buff} and L_A for each class based on the feedback of estimated arrival rates $\lambda_{\mathrm{M2M}}(i,k)$ and $\lambda_{\mathrm{LTE}}(i,k)$ and length of queue $L_Q(i,k)$. In following paragraph, we drop the time index k for simplification.

By neglecting the propagation delay, the total delay (start at the M2M gateway receives the signaling and end after the MME process receives the signaling) of M2M signaling could be calculated as follows:

$$\begin{aligned} D_{\mathrm{total}}(i) = {} & T_{\mathrm{buff}}(i) - T_{\mathrm{Arrival}}(i) + D_{\mathrm{Queue}}(i) \\ & + D_{\mathrm{Process}}(i), \end{aligned} \tag{4}$$

where T_{buff} is the timer length for class i in the M2M gateway. T_{Arrival} is the time gap between the arrival time and the timer start time; therefore, T_{Arrival} is uniform distributed between $[0, T_{\mathrm{buff}}]$. D_{Queue} is the wait time in the MME due to congestion, and it depends on the queue length and the arrival rate at the MME. D_{Process} is the delay due to parsing and aggregation of the signaling, and it is determined by the aggregation level L_A.

Our objective is to guarantee the total delay to be no larger than the maximum tolerate delay of class i; that is,

$$D_{\mathrm{total}}(i) \leq D_m(i). \tag{5}$$

Among the four factors in right side of (4), T_{buff} and T_{Arrival} are neglectable compared with queue delay and process delay. Based on queuing theory, the queue in MME could be modeled as a $M/M/1$ model. Denote the signaling process rate at the MME by $\mu(i)$ (assume that $\mu(i)$ is known at M2M gateway), and assume that the M2M service distribution of

```
Initiate set $k := 0, x(0) := 1, \varepsilon := \infty$
Start:
for $i = 1, i < 5$ do
  While $\varepsilon > T_h$
    $x(k+1) = x(k) - \dfrac{-\left(\lambda_{\text{M2M}}(i,k)/x(k)^2\mu(i)\right) + D_{\text{Single}} * \exp(x(k)-1)}{2\left(\lambda_{\text{M2M}}(i,k)/x(k)^3\mu(i)\right) + D_{\text{Single}} * \exp(x(k)-1)}$
    $\varepsilon = \text{abs}(x(k+1) - x(k))$
    $k = k + 1$
  end while
$L_A^*(i) = \text{floor}(x(k))$
$i = i + 1$
end for
End
```

Algorithm 4

different M2M networks is identical. The effective arrival rate after aggregation would be roughly rewritten as

$$\lambda_{\text{eff}}(i) = \frac{\lambda_{\text{M2M}}(i)}{L_A(i)} + \lambda_{\text{LTE}}(i). \tag{6}$$

Therefore, the average waiting time for the aggregated signaling is

$$D_{\text{Queue}}(i) = \frac{L_Q(i) + \lambda_{\text{eff}}(i) + 1}{\mu(i)}. \tag{7}$$

The parsing and aggregation process delay of the signaling D_{Process} grows exponentially with the aggregation level L_A; that is,

$$D_{\text{Process}}(i) = D_{\text{Single}}(i) * \exp\left(L_A(i) - 1\right), \tag{8}$$

where $D_{\text{Single}}(i)$ is the process delay for single signaling message. It is quite easy to prove that the total delay (4) is a concave function of parameter L_A:

$$\frac{\partial^2 D_{\text{total}}(i)}{\partial L_A(i)^2} = 2 * \frac{\lambda_{\text{M2M}}(i)}{L_A(i)^3 * \mu(i)} + D_{\text{Single}}(i) * \exp\left(L_A(i) - 1\right). \tag{9}$$

Since $L_A(i) \geqslant 1$, $\lambda_{\text{M2M}}(k)$, $\mu(i)$, and $D_{\text{Single}}(i)$ are generally positive. Therefore, $\partial^2 D_{\text{total}}(i)/\partial L_A(i)^2 > 0$.

To minimize the total delay (4), let the derivation of T with respect to L_A be 0, and we will get

$$\frac{\partial D_{\text{total}}(i)}{\partial L_A(i)} = -\frac{\lambda_{\text{M2M}}(i)}{L_A(i)^2 * \mu(i)} + D_{\text{Single}}(i) * \exp\left(L_A(i) - 1\right) = 0. \tag{10}$$

The explicit solution for (10) is unable to get. However, by numerical computation (Newton's method), iteration algorithm (see Algorithm 4) could be used to get the approximate solution for (10), where Th_h is the threshold to control the convergence of iteration algorithm, L_A is an integral, and $\text{Th}_h = 0.1$.

After obtaining the approximate optimal value for aggregation level L_A^*, denote the estimated signaling (M2M class i) arrival rate in M2M gateway j at time instance $k * T$ by $\overline{\lambda}_j(i,k)$. We could simply determine the buffer length of M2M gateway j for class i at time instance $k * T$ by the following relationship:

$$T_{\text{buff}}(i,j,k) = \frac{L_A^*(i)}{\overline{\lambda}_j(i,k)}. \tag{11}$$

3.5. Class Barring Scheme. It is assumed that the M2M signaling arrival pattern at different M2M gateway is identical. For the final step, substituting the optimal L_A^* into (4), we should compare the minimum total delay with the maximum delay of class i. If condition (5) is not fulfilled, class barring mechanism should be employed here. A factor p for class i is adaptively updated at M2M gateway j at time instance $k * T$. Our objective is to provide maximum permitted opportunity under the constraint of delay requirement. The problem can be formulated by

$$\max_{p_i} \quad \{p_i \times N_i\} \tag{12}$$

$$\text{s.t.} \quad \overline{D}_{\text{total}}(i) \leq D_m(i), \tag{13}$$

where we drop the indexes j and k, N_i is the number of M2M signaling messages (Class i) after aggregation at time instance $k * T$, and $\overline{D}_{\text{total}}(i)$ is the estimated average total delay of class i at M2M gateway j at time instance $k * T$ according to (4). Since N_i is independent of p_i, (12) equals maximum p_i. In constraint condition (13), among the four items in total delay, T_{buff}, D_{Process}, and T_{Arrival} are independent of p_i. By employing the barring mechanism, the average queue delay after barring can be estimated roughly by

$$D_{\text{Queue}}(i) = \frac{L_Q(i) + \left(\lambda_{\text{M2M}}(i) \times p_i\right)/L_A(i) + \lambda_{\text{LTE}}(i) + 1}{\mu(i)}. \tag{14}$$

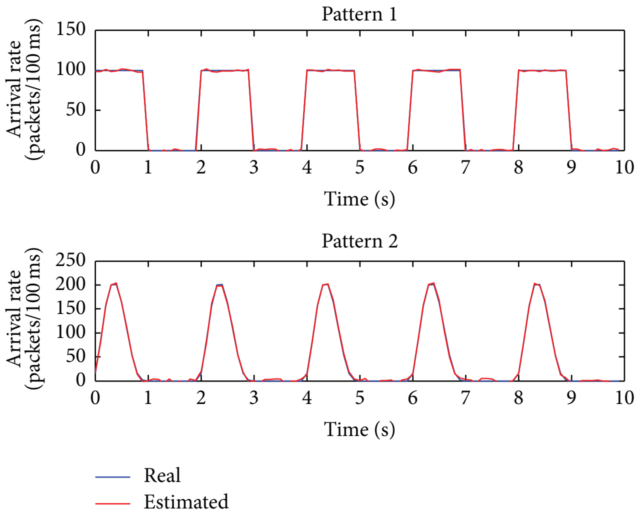

Figure 6: Estimated arrival rate for M2M traffic patterns.

Based on constraint condition (13), barring factor p_i should satisfy

$$\begin{aligned}\overline{D}_{\text{total}}(i) &= \frac{1}{2}T_{\text{buff}}(i) + D_{\text{process}}(i) \\ &\quad + \frac{L_Q(i) + (\lambda_{\text{M2M}}(i) \times p_i)/L_A(i) + \lambda_{\text{LTE}}(i) + 1}{\mu(i)} \\ &\le D_m(i),\end{aligned} \tag{15}$$

where the fact that the mean value of T_{Arrival} equals $(1/2)T_{\text{buff}}$ is utilized in above formula. After simplification, the factor p_i should meet the following requirement:

$$p_i \le \frac{L_A(i) \times \left[D_m(i) - (1/2)\,T_{\text{buff}}(i) - D_{\text{process}}(i)\right] \times \mu(i) - 1 - L_Q(i) - \lambda_{\text{LTE}}(i)}{\lambda_{\text{M2M}}(i)}. \tag{16}$$

In summary, the congestion-aware barring mechanism could be described by

$$p_i = \min\left\{\frac{L_A(i) \times \left[D_m(i) - (1/2)\,T_{\text{buff}}(i) - D_{\text{process}}(i)\right] \times \mu(i) - 1 - L_Q(i) - \lambda_{\text{LTE}}(i)}{\lambda_{\text{M2M}}(i)}, 1\right\}. \tag{17}$$

4. Performance Evaluation and Analysis

In this section, we implement simulations to demonstrate the effectiveness of proposed CASAB scheme by Matlab. A system consists of 1 MME and 10 LTE eNodeBs is considered. One M2M network and normal LTE signaling traffic coexisted in each LTE eNodeB. The normal LTE signaling arrives following the Poisson process with a mean arrival rate λ_{LTE}.

The traffic models for M2M are considered the same as those in [15]. M2M signaling traffics in different eNodeBs are identically distributed. We only consider two types of traffic here. Traffic pattern 1 (uniform distribution) represents the scenario in which M2M devices access the network uncorrelated over a period of time (1 s). Traffic pattern 2 (beta distribution over 1 s, $\alpha = 3$, and $\beta = 4$) can be considered as a practical M2M scenario in which a large amount of M2M devices access the network in a correlated manner, for example, sensors reporting an emergency event. In fact, bursts are what characterize M2M applications since M2M devices are more likely to send data in a short period due to the occurrence of an event, as shown in Figure 6. The number of M2M devices in the simulated 10 eNodeBs is equal. For the sake of simplicity, in the simulations, we considered only one class of M2M services and corresponding one type of LTE signaling.

Firstly, some experiments are implemented to verify the accuracy of arrival rate estimation scheme proposed in the paper. Figure 6 shows the real arrival rate of traffic and the estimated arrival rate for patterns 1 and 2 in one M2M gateway. In order to make the figure clearer, the number of M2M devices in the gateway is equal to 3000 and the trigger factor η is equal to 0.02. The estimation accuracy is getting better after the initial period. In short, the average relative estimation error for both pattern is less than 10%, and the estimation performance for pattern 1 (uniform) is slightly better than pattern 2 (beta).

Congestion and failure rate of generated M2M signaling traffic are considered as the two metrics to measure the ability of congestion alleviation. To calculate these probabilities numerically, we apply the $M/M/1/K$ queue system at MME for each class. Monte Carlo simulations with different seeds are employed in the paper, and performances for each run are averaged to estimate the average performance of proposed scheme. Indeed, the system has a finite buffer (maximum K waiting positions) in the queue with FIFO (First In First Out)

Table 3: Parameters setting in the simulation.

Parameter	Value	Description
λ_{Normal}	1/ms	Arrival rate of LTE signaling (single eNodeB)
μ	5/ms	Serving rate at MME
D_{single}	0.05 ms	Delay of process single signaling
θ	0.6	Forget factor
n_i	2	Measurement windows per period
T	100 ms	Update period
K	5	Buffer length at MME
D_m	10 ms	Maximum tolerate delay
Arrival rate (pattern 2)	1 packet/ms	Arrival rate of M2M pattern 1/2 (single M2M device)

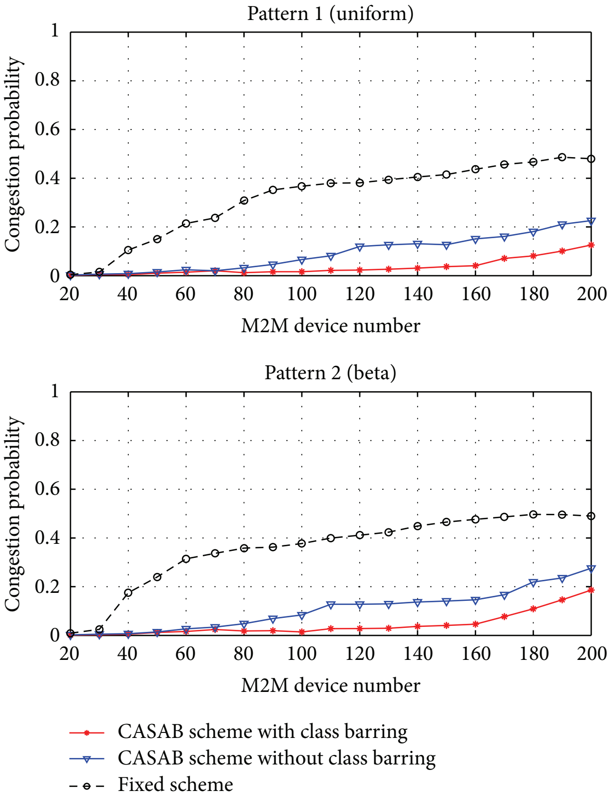

Figure 7: Congestion probability versus number of M2M devices.

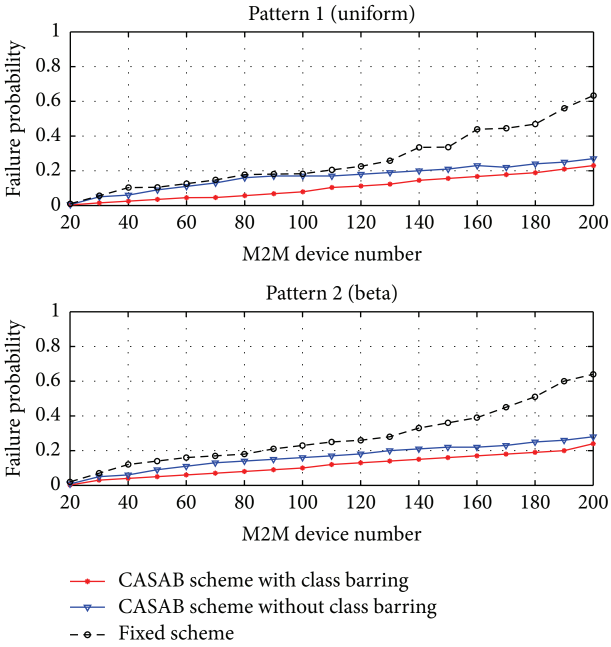

Figure 8: Failure probability versus number of M2M devices.

queue principle. The values for all parameters in following simulations are summarized in Table 3.

The congestion probability is defined as the probability of having the number of signaling messages in the queue greater than or equal to the predefined maximum queue length K. For comparison, we also evaluate the performance with traditional scheme by fixed $L_A = 1$, $T_{\text{buff}} = 0$ (denoted by fixed scheme in the figure legend). Total 100 simulations with different random seeds are evaluated and the average congestion probability is collected. In each simulation, the simulation time length is 10 s. The relationship between congestion probability and the number of M2M devices is given in Figure 7. Obviously, CASAB scheme has better performance in high traffic load region due to the fact that larger aggregation level is adopted. In addition, because of the correlated feature of pattern 2, the congestion probability for M2M traffic pattern 2 is slightly higher than pattern 1. Meanwhile, class barring mechanism effectively reduces the traffic load especially in high load region; hence CASAB with class barring mechanism achieves better performance compared with CASAB scheme without class barring.

The failure probability is defined as the probability that the delay of signaling is larger than the maximum tolerate delay or the signaling being rejected at MME. Figure 8 shows the average failure probability with different M2M device number. The CASAB scheme with class barring mechanism shows better performance, especially in high load region.

Compared to the fixed scheme without adaptive signaling aggregation, above results show that our CASAB scheme achieves much better performance in reducing signaling congestion and failure. This proves the benefit of our algorithm.

5. Conclusion

In this paper, based on the convergence architecture of M2M network and LTE cellular network, congestion-aware signaling aggregation and barring scheme is proposed considering the features of M2M service. The signaling parsing and aggregation parameters and class barring factors are adaptively adjusted based on the congestion status of LTE network entity and traffic estimator. Evaluation results show that the proposed solution avoids signaling congestion and maintains good performance of the system. In further work, more precise estimation of arrival rate and more realistic simulation are needed.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The project is supported by the Key Program of the National Natural Science Foundation of China (Grant nos. 61431008 and 61271180) and National Special Fund of the Internet of things of China, BUPT Youth Innovation Project (BUPT-2015RC01).

References

[1] J. P. Conti, "The internet of things," *IET Communications Engineer*, vol. 4, no. 6, pp. 20–25, 2006.

[2] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, "From today's INTRAnet of things to a future INTERnet of things: a wireless- and mobility-related view," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 44–51, 2010.

[3] Y. Zhang, R. Yu, S. Xie, W. Yao, Y. Xiao, and M. Guizani, "Home M2M networks: architectures, standards, and QoS improvement," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 44–52, 2011.

[4] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 60–65, 2011.

[5] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, 2011.

[6] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: potential, challenges, and solutions," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 178–184, 2012.

[7] A. Amokrane, *Congestion control in the context of machine type communications in long term evolution networks [M.S. thesis]*, ENS Cachan, University of Rennes, Rennes, France, 2011.

[8] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Cellular-based machine-to-machine: overload control," *IEEE Network*, vol. 26, no. 6, pp. 54–60, 2012.

[9] S. Yang, X. Wen, W. Zheng, and Z. Lu, "Grouping aggregation and on-demand parsing mechanism for congestion mitigation in 3GPP machine-to-machine communications," *Information Technology Journal*, vol. 13, no. 13, pp. 2196–2203, 2014.

[10] G. Farhadi and A. Ito, "Group-based signaling and access control for cellular machine-to-machine communication," in *Proceedings of the IEEE 78th Vehicular Technology Conference (VTC-Fall '13)*, pp. 1–6, September 2013.

[11] H.-L. Fu, P. Lin, H. Yue, G.-M. Huang, and C.-P. Lee, "Group mobility management for large-scale machine-to-machine mobile networking," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1296–1305, 2014.

[12] T. Taleb and A. Ksentini, "An efficient scheme for MTC overload control based on signaling message compression," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '13)*, pp. 342–346, 2013.

[13] S. Duan, V. Shah-Mansouri, and V. W. S. Wong, "Dynamic access class barring for M2M communications in LTE networks," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '13)*, pp. 4747–4752, IEEE, Atlanta, Ga, USA, December 2013.

[14] R. Liu, W. Wu, H. Zhu, and D. Yang, "M2M-oriented QoS categorization in cellular network," in *Proceedings of the 7th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '11)*, pp. 1–5, IEEE, Wuhan, China, September 2011.

[15] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, John Wiley & Sons, New York, NY, USA, 2001.

[16] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 184–192, 2012.