Noname manuscript No.
(will be inserted by the editor)

Overload Control in the Network Domain of LTE/LTE-A Based Machine Type Communications

Zahra Alavikia · **Abdorasoul Ghasemi**

Received: / Accepted:

Abstract It is expected that the LTE network, which includes the radio access network (RAN) and the core network (CN) in 3GPP LTE systems, will be overloaded due to the huge number of Machine-Type Communication (MTC) devices in the near future. Overload in the RAN and CN of the LTE may result in congestion occurrence, resource waste, Quality of Service (QoS) degradation and in the worst-case, it will cause service unavailability. In this paper, we have proposed an adaptive mechanism to manage a large number of MTC devices in both RAN and CN of the LTE network. We use Access Class Barring (ACB) scheme to regulate the MTC traffic according to the congestions level in the RAN and CN. We consider a scenario in which two-priority-based classes of MTC devices are contending for the RAN resources. At first, the overload problem in the RAN is formulated to find the number of allowable contending MTC devices of each class taking into account their required QoS. Then, an active load management policy based on additive increase multiplicative decrease rule is proposed to control the incoming load from multiple cells to the CN. To effectively limit the number of MTC devices in both RAN and CN, in the proposed approach, each Evolved Node B (eNB) updates the ACB factor upon overload detection in the RAN or CN in an adaptive manner. Simulation results show that the proposed mechanism is able to manage overload in the CN and RAN simultaneously.

Keywords Machine Type Communications (MTC) · Overload Control · Radio Access Network (RAN) · Core Network (CN) · Quality Of Service (QoS) · Access Class Barring (ACB)

Zahra Alavikia

Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran

Abdorasoul Ghasemi

Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran

1 Introduction

Machine-to-Machine communications (M2M) is considered to be the fundamental part of the Internet of Things (IoT), which aims to extend the communication to all things via the Internet. M2M communications is referred as Machine-Type Communications (MTC) by 3rd Generation Partnership Project (3GPP) as a key standardization body. MTC supports various automated systems with minimal human interaction. It brings several attentions in a wide range of applications such as monitoring systems, smart metering, healthcare and transportation systems [1, 2].

Long-Term Evolution (LTE)/LTE-Advanced (LTE-A) network is a good candidate for MTC infrastructure which attracts numerous attentions in literature due to its extensive coverage, mobility support, and high-speed data rate. However, the LTE network is optimized basically for light and mostly downlink human-type communications (HTC). The deployment of huge numbers of MTC devices with mostly uplink and batch arrival traffic raises various challenges in the LTE [3, 4]. Since MTC connections are predicted to reach up to 50 billion by 2020 [2], handling massive number of machines is one of the fundamental challenges for MTC in the LTE.

3GPP MTC related working groups determine several releases and studies to mitigate massive access of MTC devices which named as “overload control” [5]. Overload in a network is a condition in which the incoming load to a node is greater than the node’s available resources to afford that load. Overload leads to congestion which increases the access delay and the packet loss and hence severely decreases the throughput and the quality of service (QoS) of connections. Overload in the LTE could take place at both the Radio Access Network (RAN) and the Core Network (CN). When the number of MTC devices is more than the available radio resources, the RAN overload occurs. The overload in the CN occurs when the incoming access load from multiple cells is greater than the available resources at the CN’s nodes [6].

To alleviate MTC overload in the RAN and CN separately, 3GPP addresses some basic improvements. In the RAN, some approaches such as dynamic allocation of Random Access Channel (RACH) resources, separate RACH resources for MTC and HTC, Access Class Barring (ACB), class dependent back off time assignment, and pull based methods have been introduces by 3GPP [5]. Among the mentioned approaches, the ACB is considered to be a main solution in 3GPP because of its simplicity [5]. To overcome MTC overload in the CN, a rejection-based approach has been introduced by 3GPP [7]; where the requests from low priority MTC devices are rejected/discarded by the CN/RAN nodes upon overload detection. Several studies have been

conducted to improve the introduced solutions by 3GPP to control congestion in the RAN and CN separately in recent years [8–12].

Mostly focused on the RAN or CN overload control, these works does not consider the RAN and CN in congestion management simultaneously. RAN-base overload control mechanism which does not consider the CN load may lead to buffer overflow in the CN nodes at the presence of concurrent transmission of multiple cells. On the other hand the CN overload control mechanisms are reactive approaches which are not efficient from resource utilization perspective. This paper proposes an efficient mechanism to control the large number of connection requests from MTC devices in both radio and core of the LTE networks. The proposed method offers a proper ACB calculation method to specify the access probability for MTC devices in accordance with overload state in the RAN and CN nodes. The main contributions of this paper include:

- The random access problem in the RAN is formulated for two relevant MTC traffic, coordinated and uncoordinated event triggered data traffic, to satisfy the QoS of MTC devices.
- To avoid excessive load condition in the CN nodes as a result of concurrent transmissions of multiple cells, an additive increase multiplicative decrease (AIMD) based active queue management (AQM) method is proposed to inform each eNB about its admissible offered load.
- The proposed approaches in the RAN and CN are integrated to simultaneously control the excessive load of MTC devices in the RAN and CN through a proper ACB calculation method.

In the rest of this paper, a concise description of the MTC architecture model over the LTE network is presented and the related works in MTC overload control are reviewed in Section 2. The system model is introduced in Section 3. Section 4 is dedicated to the proposed solutions to overload control in the RAN, the CN and the RAN-CN mechanisms. Section 5 demonstrates evaluation of the proposed mechanisms. Finally, conclusion are presented in Section 6.

2 Backgrounds and Related Works

In this section, a simplified model of the MTC architecture in the LTE network and the overload concept in this model is introduced, and then some related works considering the congestion problem of MTC in the LTE are reviewed.

2.1 MTC Architecture Model in EPS

To provide IP connectivity between MTC devices and the Public Data Network (PDN), 3GPP System Architecture working group 2 (SA2) aims at using the Evolved Packet System (EPS) for MTC. In the context of the cellular system, EPS describes jointly LTE and System Architecture Evolution (SAE) and includes three important parts: RAN, CN, and service parts. The RAN and CN are denoted as Evolved Universal Terrestrial Access Network (E-UTRAN) and Evolved Packet Core (EPC) respectively. Service parts include remote services and their entities, e.g., application server [6]. From another perspective [13], MTC architecture can be seen as the interaction between three MTC domains, i.e., MTC device domain, MTC network domain, and MTC application domain. Fig. 1 shows a simplified model of the MTC architecture in EPS with different MTC domains where network domain includes the RAN and the CN sides. In the RAN side, the eNB acts as the base station which manages radio resources and handles device mobility in the cell. The main entities in the CN side include Mobility Management Entity (MME), Service GateWay (S-GW), and PDN GateWay (P-GW). The MME controls the radio connectivity, authentication, paging and bearer adjustment in the CN. The S-GW is the entity that handles local mobility for intra-3GPP handoffs and the P-GW provides the required interface with the PDN [11, 13].

When a large number of MTC devices simultaneously trigger to send data, they set up a connection through the uplink channel with the eNB. Due to numerous MTC devices and their event-trigger traffic, contention-based random access procedure is proposed to be an appropriate solution for MTC in the LTE. Random access procedure in LTE/LTE-A uses preamble transmission on Physical Random Access Channel (PRACH). In the random access process, several machines may select the same preamble in one PRACH. Therefore, the eNB might not detect the requesting machine correctly due to possibility of collision between them. If a proper load control scheme is not applied in the RAN/E-UTAN, an overload situation may be inevitable [3, 4]. After the successful transmission of an access request, the request from MTC device relays to the EPC/CN nodes through RAN/E-UTRANT node. Since multiple requests may be sent to the EPC/CN nodes simultaneously, the nodes of the EPC/CN may experience excessive load condition [6].

2.2 Related Works

Random access due to the simultaneous transmission by the huge number of MTC devices is susceptible to collision in an overload condition. As mentioned previously, overload can occurs at both the RAN and the CN

of LTE networks. To relieve this problem, several studies have attempted to control the massive number of requests in the RAN or the CN.

A classification of existing solutions to control overload in the domain of RAN is introduced in [3, 14]. Among existing solutions, the ACB scheme attracts numerous attentions. In the ACB, 16 different MTC classes [15] are barred according to their priorities. The eNB broadcasts barring parameters, i.e., barring probability, barring time, for each class to inform MTC devices. In the case of the overload, each MTC device can access the network if its unit uniformly distributed random drawing number is less than the barring probability; otherwise, it postpones its access according to the barring time. In [8] a heuristic algorithm has been proposed for updating barring probability in an adaptive manner. Authors of [16] obtained the adaptive access probability by designing a maximum likelihood estimator to estimate the number of MTC devices in a bursty MTC traffic condition. However, [8] and [16] did not take into account the priority of MTC devices in the ACB calculation. To provide prioritized access control in the RAN, different classes of MTC traffic have been introduced in [5]. In the overload condition, the requests from low-priority devices are rejected/barred/dropped due to their delay tolerant properties. To meet the QoS of different MTC classes, a dynamic ACB and class based back-off scheme is introduced in [9] to control the access of MTC devices to the radio resources. In [17], the joint access probability and resource allocation have been taken into account in the RAN. To find the ACB factor in adjacent cells, a cooperative scheme between eNBs is proposed in [10]. Although, congestion management techniques in each cell decrease the overload in the CN, but they do not guarantee the overload control in the CN since the relayed load from other cells do not taken into account, and the CN nodes may experience simultaneous traffic from multiple cells.

3GPP addresses some solutions to detect and minimize overload in the CN. As discussed in the 3GPP specification [7], upon the overload occurrence in the CN nodes, any new connection request may be rejected in the MME or eNBs. This can be issued by sending an overload indicator message to the MME/eNBs. Some pull-based overload management methods in the CN are also introduced by 3GPP and are extended in some works like [18, 19], where downstream mechanisms are applied to trigger MTC devices. This type of overload control strategies is not in the scope of this paper. From rejection-based perspective, authors in [11] investigated a congestion situation in the CN to avoid buffer overflow in the MME node. In their scheme, each eNB accept/reject connection requests from different MTC classes according to their reject factors, which has been announced by the CN node. A similar approach has been proposed in [12]. Although, in these methods each eNB determines the rejection probability from the receiving feedback of the MME, but they are reactive methods which result in the resource waste and may jeopardize the QoS of the high priority traffic.

In order to satisfy a given delay requirement of high priority traffic and utilize the available resources in the RAN and the CN effectively, we have offered a proper CN overload-aware method to specify access probability adaptively, associated with different priority of MTC groups.

3 System Model and Problem Statement

Fig. 2 shows a simplified system model of the LTE based MTC in which the RAN and the CN parts are separated by dash line. In the RAN, the eNB in each cell relays the requests of MTC devices to the CN through S1 interface. In each cell, we have considered a scenario in which there are two types of MTC devices: uncoordinated and coordinated event triggered devices which we named as *ClassI* and *ClassII* respectively. *ClassI* includes applications such as consumer electronics, fleet management and E-care. In *ClassI*, each device monitors the environment and send data to the eNB occasionally. *ClassII* includes a large number of devices such as smart meters; which have been scheduled to send their data to eNB periodically. To meet the QoS of the *ClassI* devices, these devices have higher priority to use network resources in comparison with *ClassII*.

In the traffic model of *CalssI*, each device is triggered with probability ϕ to send data upon event occurrence. Due to extremely low access delay and high successful transmission probability requirement of *ClassI* traffic, it is assumed that each device of *ClassI* has a queue of size one. Therefore, the system is consisted of N_T^1 individual queues, one for each device of *ClassI*. Since the new arriving packet contains the latest monitoring information, the waiting packet in the queue will be ignored. The traffic model of the *ClassI* devices can be represented as a two state Markov chain (see Fig. 3). Each device of *ClassI* is said to be in an active state, π_1, if it has one packet to transmit. When a device is in the active state, it sends a request to the eNB with probability p_{ACB}^1 to set up a connection through PRACH. We assume that the QoS of the *ClassI* devices is met if the success probability of these devices is greater than a predefined threshold, P_0.

The Beta distribution is used to model the correlated traffic of scheduled arrivals of the huge number of *ClassII* devices [5]. This distribution models the aggregated traffic of N_T^2 *ClassII* devices which are activated within a determined time, T_A, in a correlated manner. The parameters of the Beta probability density function are considered as $\alpha = 3$, $\beta = 4$, $T_A = 10s$ [5] [5]. The Beta distribution, g(t), with parameters α and β is defined by (1).

$$g(t) = \frac{t^{\alpha-1}(T_A - t)^{\beta-1}}{T_A^{\alpha+\beta-1} Beta(\alpha, \beta)}, \tag{1}$$

where $Beta(\alpha, \beta)$ indicates the Beta function [5]. In the activation time, each device of *ClassII* contends with other devices with probability p_{ACB}^2 to send data. There is not any barring time; which means collided

devices will retransmit their requests in the next opportunities. The number of retransmission attempts is not limited for the *ClassII* MTC devices.

The contention based random access procedure is done in a time slotted scheme in which time is divided into radio frames consists of ten sub-frames with *1ms* duration each. In each frame, some random access opportunities, in accordance with what has been defined by configurations in the LTE, is determined. More details about configurations in the LTE can be found in [20]. In each random access procedure, the eNB broadcasts the PRACH configuration index and the free preambles, which have not been assigned to HTC, to MTC devices. In what follows, it is assumed that there are M free resources which include frequency, time, and preamble in each frame.

Since both classes share the same resources in the RAN to maximize resource utilization, the eNB adaptively assigns the ACB factors to the devices of both classes to satisfy their QoSs. Whenever, an active MTC device passes the ACB procedure successfully, it uniformly selects one preamble from the dedicated preambles. When the eNB correctly detects an access request, it forwards the request to the CN node. Since the MME is more susceptible to congestion among various entities in the CN [7], in this paper just the overload condition in the MME is considered. Due to the sensitivity of the *ClassI* MTC devices to delay, it is assumed that there are reserved resources in the MME for these devices.

Fig. 2 also depicts a simplified message exchange process between MTC devices and the CN node through the eNB. In Fig. 2, the incoming messages which pass the ACB scheme, demonstrate the random access procedure.

The RAN and CN of the LTE are subject to the excessive load due to possible simultaneous access of *ClassII* devices. The objective of this paper is to control the massive requests of non-delay sensitive devices of *ClassII* in order to guarantee the expected number of successful transmissions of *ClassI* devices per frame. Also, this paper is an attempt to reduce the number of dropped requests of the *ClassII* devices in the CN while avoiding the resource underutilization conditions in the RAN and the CN.

4 Proposed Solution for Overload Control in the RAN and CN

To distinguish between the advantages and disadvantages of separately and simultaneously MTC overload control problem in the RAN and the CN parts of LTE, in this section we have proposed three different mechanisms: *RAN overload control*, *CN overload control* and simultaneous *RAN* & *CN overload control*. Fig. 4 shows a simplified message passing procedure in each mechanism.

In the *RAN overload control* mechanism, the eNB just considers the constraint of the RAN to calculate the ACB probability for both MTC classes. As shown in Fig. 4(a), in this case the eNB broadcasts the ACB factor in an adaptive manner to inform the MTC devices about their access probability. In this mechanism, if the number of requests from multiple cells is more than the service capacity of the MME, the requests of MTC devices will be dropped in the CN.

In the *CN overload control* mechanism, the eNB determines the offered load to the CN through the notification messages which have been received from the MME, see Fig. 4(b). To prevent the overload condition in the CN, the eNB rejects the excessive received requests of MTC devices in the RAN if the number of received requests are greater than the allowable offered capacity to the CN. Since in this mechanism the constraint of the RAN is not taken into account, it may jeopardize the QoS of high priority traffic in the RAN.

Fig. 4(c) shows the simultaneous *RAN* & *CN overload control* mechanism. In the *RAN* & *CN overload control* mechanism, the eNB considers both RAN and CN constraints to decide about the number of contending MTC devices. In this mechanism, eNB calculates and broadcasts the dynamic ACB probability according to the number of active MTC devices, available resources in the RAN, the required QoS of *ClassI* and also the notification messages which have been received from the MME. Although in the *RAN* & *CN overload control* mechanism the number of contending MTC devices can be controlled according to the constraints in the RAN and the CN simultaneously at the price of transmitting more signalling messages, see Fig. 4(c).

In what follows, the overload control mechanism in each mechanism is formulated in order to deal with congestion situation. The formulated mechanisms are compared with each other to emphasize the advantages and disadvantages of each mechanism.

4.1 Overload Control in the RAN

In the *RAN overload control* mechanism, overload in the RAN which leads to the QoS degradation and resource wasting is controlled through broadcasting the ACB probability by the eNB. To show this, in this section it is discussed how to obtain the ACB probability for two MTC classes in order to meet their QoS demands and also, to maximize resource utilization in the RAN. Since the traffic models of *ClassI* and *ClassII* in both cells of Fig. 2 are the same, we continue to calculate the ACB factors for cell 1 which can be applied in a similar way to cell 2. For simplicity of presentation, the index of cell number, i.e., 1, is omitted in this subsection. In what follows, we assume that the eNB knows the total number of MTC devices of both MTC classes.

Due to the limited number of the available resources in the RAN, it is necessary to control the number of contending MTC devices to maximize efficient resource usage and hence, to maximize the expected number of successful transmissions. Let $\bar{N}^1_{ACB}$ denotes the accepted number of *ClassI* devices which are allowed to participate in the contention. With the given M available resources in the RAN, the probability that one resource is selected by $\bar{N}^1_{ACB}$ MTC devices of *ClassI* which is denoted by P_M is computed by (2).

$$P_M = \frac{M}{\bar{N}^1_{ACB}}. \tag{2}$$

To maximize efficient resource usage, each resource must be selected by only one MTC device in average, which means $P_M = 1$ and hence, $\bar{N}^1_{ACB} = M$. If the number of contending MTC devices is greater/less than M, an overload/underutilization condition occurs respectively; this means that each resource is selected by greater/less than one user in average. In order to avoid this problem, the ACB probability is used to sustain the number of contending MTC devices near to M. Because the arrival of the devices of *ClassI* is time independent, the available resources are allocated to MTC devices of *ClassI* in order to maximize resource utilization. The devices of *ClassII* can benefit from the available resources provided that the QoS of *ClassI* devices is guaranteed. The optimum value of the ACB probability of *ClassI* can be acquired by dividing $\bar{N}^1_{ACB}$ to the expected number of active devices of *ClassI* as given in (3),

$$p^1_{ACB} = \frac{\bar{N}^1_{ACB}}{N_1} = \frac{\bar{N}^1_{ACB}}{N^1_T \Pi_1}, \tag{3}$$

where N_1 denotes the expected number of active devices of *ClassI* and Π_1 is the probability that the corresponding Markov chain is in active state in the steady state (see Fig. 3) as given by (4).

$$\Pi_1 = \frac{\phi}{\phi + p^1_{ACB}(1-\phi)}. \tag{4}$$

To find the value of p^1_{ACB} from (3), $\bar{N}^1_{ACB}$ is replaced by M as explained before, which results in (5).

$$p^1_{ACB} = \begin{cases} \frac{M\phi}{\phi(N^1_T+M)-M}, & M < N^1_T\phi \\ 1, & M \geq N^1_T\phi \end{cases} \tag{5}$$

When the number of active devices is less than M, p^1_{ACB} equals one; which means that there is no limitation in initiating the random access procedure.

The success probability of active devices of *ClassI*, P^1_{sc}, is acquired by dividing the expected number of successful transmitted requests, denoted by N^1_{sc}, to N_1 as given in (6).

$$P^1_{sc} = \frac{N^1_{sc}}{N_1} \tag{6}$$

where N_{sc}^1 can be obtained from (7).

$$N_{sc}^1 = MP_{M,sc}^1 \tag{7}$$

$P_{M,sc}^1$ in (7) denotes the probability that a given resource is selected by only one device which can be calculated by (8) because each resource is uniformly selected by each MTC device of *ClassI* with probability $\frac{1}{M}$, and will be utilized successfully if none of the contending devices of both classes select that resource.

$$P_{M,sc}^1 = \binom{p_{ACB}^1 N_1}{1} \frac{1}{M} (1-\frac{1}{M})^{p_{ACB}^1 N_1 + \bar{N}_{ACB}^2 - 1} \tag{8}$$

In (8), $\bar{N}_{ACB}^2$ denotes the accepted number of *ClassII* devices which are allowed to participate in the contention. To meet the QoS of *ClassI* devices, the success probability of these devices should be greater than P_0. Therefore, $\bar{N}_{ACB}^2$ can be found by solving (6) by replacing P_{sc}^1 with P_0 which results in (9).

$$\bar{N}_{ACB}^2 = \Big[(1 - p_{ACB}^1 N_1 + \frac{Ln(\frac{P_0}{p_{ACB}^1})}{Ln(\frac{M-1}{M})}) \Big]^+, \tag{9}$$

where $x^+ = \max\{x, 0\}$. Due to the time dependent arrival nature of the *ClassII* devices, the expected number of new arrivals of *ClassII*, denoted by a_2, varies in each frame. Therefore, in order to control the massive number of requests of *ClassII*, p_{ACB}^2 must be calculated in frame k using (10) for efficient utilization of remaining resources which shall be equivalent to $\bar{N}_{ACB}^2$.

$$p_{ACB}^2(k) = min\Big\{1, \frac{\bar{N}_{ACB}^2}{N_2(k)}\Big\}, \tag{10}$$

where $N_2(k)$ refers to the expected number of active *ClassII* devices in frame k and it includes $a_2(k)$ and those who did not transmit their requests to the eNB successfully in frame $k-1$. Therefore, $N_2(k)$ can be obtained as (11),

$$N_2(k) = a_2(k) + N_2(k-1) - N_{sc}^2(k-1). \tag{11}$$

where $N_{sc}^2(k-1)$ can be calculated in a similar way to (7) as in (12).

$$N_{sc}^2(k-1) = \binom{p_{ACB}^2(k-1)N_2(k-1)}{1} (1-\frac{1}{M})^{p_{ACB}^1 N_1 + p_{ACB}^2(k-1)N_2(k-1) - 1}. \tag{12}$$

The maximum of N_{sc}^2 in (12), $N_{sc,max}^2$, can be obtained through replacing $p_{ACB}^2(k-1)N_2(k-1)$ with $\bar{N}_{ACB}^2$ which results in (13).

$$N_{sc,max}^2 = \binom{\bar{N}_{ACB}^2}{1} (1-\frac{1}{M})^{p_{ACB}^1 N_1 + \bar{N}_{ACB}^2 - 1}. \tag{13}$$

Algorithm 1 RAN overload control Mechanism

1: Input: M, ϕ, N_T^1, N_T^2, P_0
2: Output: p_{ACB}^1, $p_{ACB}^2(k)$
3: Compute p_{ACB}^1 using (5)
4: Compute $\bar{N}_{ACB}^2$ using (9)
5: **for** the kth frame **do**
6: Compute $N_2(k)$ by (11)
7: Update $p_{ACB}^2(k)$ by (10)
8: **end for**

Since all N_T^2 devices of *ClassII* are activated within $[0,T]$, the number of *ClassII* arrivals in the k^{th} frame can be calculated using (14) according to the Beta probability density function.

$$a_2(k) = N_T^2 \int_{t_{k-1}}^{t_k} g(t)dt. \tag{14}$$

where t_{k-1} and t_k are the start and end time of the frame k.

We propose an iterative method using (10) and (11) to update the value of p_{ACB}^2 in each frame by computing $\bar{N}_{ACB}^2$ and N_2. Alg. 1 summarizes the main steps that each eNB should do in order to calculate the ACB probabilities for both MTC classes.

In the proposed *RAN overload control* mechanism there is not any message passing between eNBs and MME, which causes the overload control mechanism to be simple, however, it may lead to the overload in the CN as we will show by simulations.

4.2 Overload Control in the CN

The overload in the CN can cause to the buffer overflow in the MME node. As introduced in 3GPP standard [7], to overcome the overload condition in the CN nodes, the connection requests of low priority MTC devices can be rejected by the eNB when the MME notifies the overload condition to the eNB. In this subsection, we propose an iterative method for calculating the number of *ClassII* devices requests that passed the ACB check and should be rejected in the eNB according to the received feedbacks of the MME. For simplicity of analysis, in what follows, we will ignore the delay and error of the MME feedback to the eNB. In this subsection, $i \in \{1,2\}$ refers to the index of the cells.

To avoid both overload condition and system underutilization in the CN, the MME monitors its queue length state, denoted by Q, to informs eNBs to decrease or increase the offered load. The corresponding

feedback diagram of the system with two cells is depicted in Fig. 5(a). To specify the congestion level in the MME, we have used the variation of the queue length, which is introduced in some work to be a good indicator of the congestion level [21, 22].

The key idea of the CN overload control is monitoring the queue length of the MME and comparing it with a predefined threshold like Q_0. The MME feedbacks a binary signal, y, to the related eNBs to adjust the number of offered requests which is denoted by N_L^2. At the RAN side, each eNB should increase N_L^2 if receives $y = 0$ and should decrease N_L^2 if $y = 1$. If the number of requests which passes the ACB check in the RAN is greater than N_L^2, the eNB will reject some requests of *ClassII* to avoid the overload in the CN.

In the *CN overload control* mechanism, *ClassII* MTC devices are blocked with a proper fixed ACB probability to avoid the congestion collapse during the random access procedure. Let N_2^{max} be the number of maximum simultaneous *ClassII* MTC active devices. The fixed ACB probability should be selected near to $\frac{M}{N_2^{max}}$ to ensure that sufficient load is offered by the RAN while no congestion collapse will happen in the RAN. Notice that the access probability of the *ClassI* devices is obtained using (5).

To determine the number of rejected requests which is denoted by $N_{R,i}^2$, $N_{L,i}^2$ can be subtracted from the number of successful transmitted requests, $N_{sc,i}^2$, in each frame as (15).

$$N_{R,i}^2(k) = \left[N_{sc,i}^2(k) - N_{L,i}^2(k)\right]^+, \tag{15}$$

where $N_{L,i}^2$ is calculated by the eNB using the received binary feedback of the MME. Here, the eNB action is based on Additive Increase Multiplicative Decrease (AIMD) rule when the congestion or resource underutilization condition in the CN is detected. AIMD rule is one of the most popular approaches for controlling the sending rates through the binary feedback control message [21, 23]. The AIMD based approach for adjusting $N_{L,i}^2$ is given by (16).

$$N_{L,i}^2(t+\Delta t) = N_{L,i}^2(t) + \theta \Delta t \mathbf{I}(y=0) - \sigma \Delta t N_{L,i}^2(t)\mathbf{I}(y=1), \tag{16}$$

where $\mathbf{I}$ is the indicator function and $\theta > 0$ and $\sigma \in (0,1)$ are AIMD parameters describing the amount of increase or decrease within Δt seconds.

The variation of Q can be represented as (17).

$$Q(t+\triangle t) = \left[Q(t) + \left(N_{L,1}^2(t) + N_{L,2}^2(t)\right)\triangle t - D\triangle t\right]^+, \tag{17}$$

where D refers to the fixed service rate of the MME node. To apply (16) and (17) in a frame based system model, we replace $x(t+\triangle t)$ by $x(k+1)$, $x(t)$ by $x(k)$, and Δt by the duration of one frame.

The switching behavior of AIMD rule can be approximated by the sigmoidal function, $\frac{1}{1+e^{-\nu x}}$ where ν is the approximation coefficient. Note that the sigmoidal function tends to one and zero if ν is selected appropriately according to the range of variable x. Now, if $\triangle t \rightarrow 0$, the system dynamics can be modeled by the following nonlinear differential equations.

$$\begin{aligned} \dot{N}^2_{L,i}(t) &= \frac{\theta}{1+e^{\nu(Q(t)-Q_0)}} - \frac{\sigma N^2_{L,i}(t)}{1+e^{\nu(Q_0-Q(t))}} \qquad i=1,2 \\ \dot{Q}(t) &= \left[N^2_{L,1}(t)+N^2_{L,2}(t)-D\right]^+, \end{aligned} \tag{18}$$

where $\dot{x}(t)$ denotes the time-derivative of $x(t)$. The first differential equation in (18) models the AIMD(θ,σ) rule which is used in each cell. The second differential equation in (18) reflects the MME's queue size dynamics as the difference between the incoming traffic of both cells and the service rate of the MME.

By linear approximation of these nonlinear equations around the equilibrium point $(N^{2*}_{L,i}, Q^*)$ and assuming that the same AIMD parameters are used in each cell, we can find simple linear differential equations that describes the system dynamics. This linear approximation is used to simply find the effect of AIMD parameters on the system transient response behavior.

The equilibrium point of (18) is obtained by $\dot{Q}=0$ and $\dot{N}^2_{L,i}=0$, where $i=1,2$. Since in the overload condition the incoming traffic from each cell to the MME is greater than $\frac{D}{2}$, then $N^{2*}_{L,1}=N^{2*}_{L,2}$ and therefore,

$$Q^* = Q_0 - \frac{\ln\frac{\sigma D}{n\theta}}{\nu}; \qquad N^{2*}_{L,i} = \frac{D}{n}, \tag{19}$$

where n denotes the number of cells. Therefore, with given Q^* and $N^{2*}_{L,i}$, the linear system of (18) at the equilibrium point is given by:

$$\begin{aligned} \delta\dot{N}^2_{L,i}(t) &= -I_N \delta N^2_{L,i}(t) - I_Q(\delta Q(t) - Q_0) \\ \delta\dot{Q}(t) &= n\delta N^2_{L,i}(t) \end{aligned} \tag{20}$$

where $\delta N^2_{L,i} \triangleq N^2_{L,i} - N^{2*}_{L,i}$, and $\delta Q \triangleq Q - Q^*$. In (20), I_N and I_Q are the values of the derivative of $\dot{N}^2_{L,i}$ respect to $N^2_{L,i}$ and Q at the equilibrium point respectively. I_N and I_Q can be obtained as (21),

$$I_N = \frac{n\sigma\theta}{\sigma D + n\theta}; \qquad I_Q = \frac{\nu\sigma\theta D}{n\theta+\sigma D}. \tag{21}$$

By taking Laplace transform of (20), the linear dynamics of the system can be illustrated as a block diagram in Fig. 5(b). This figure contains an inner and an outer feedback loops. To keep the queue length of the MME around its determined threshold in the outer loop, the difference between Q and Q_0, which is called error signal,

Algorithm 2 CN overload control Mechanism in eNB

1: Input: σ, θ, M, ϕ, N_T^1,
2: Output: $N_{R,i}^2(k)$, p_{ACB}^1
3: Compute p_{ACB}^1 using (5)
4: Set $N_{L,i}^2(1) := 0$
5: **for** kth frame **do**
6: Monitor $N_{sc,i}^2(k)$
7: Monitor y
8: Compute $N_{L,i}^2(k)$ by Eq. (16)
9: Update $N_{R,i}^2(k)$ by Eq. (15)
10: **end for**

is used as feedback to each cell to control its offered load. Then the inner feedback loop is applied to bring the actual value of $N_{L,i}^2$ closer to the desired value which is obtained through the error signal of the outer loop.

According to the block diagram in Fig. 5(b), the overall system transfer function is given by (22).

$$T(s) = \frac{nI_Q}{s^2 + sI_N + nI_Q}. \tag{22}$$

By considering T(s), we can find the optimal value of the AIMD parameters in order to satisfy the desired speed of the system and keep the equilibrium point around its determined threshold. To satisfy the desired speed of the system, the settling time of the system, $t_s = \frac{4(\sigma D + n\theta)}{\theta\sigma}$, is considered to be less than the determined threshold, γ. Therefore, the conditions of finding the desired value of AIMD parameters are $t_s \leq \gamma$, $\tilde{q} \simeq Q_0$ where $\theta > 0, \quad \sigma \in (0,1)$. With these conditions, the optimal value of AIMD parameters are $\sigma = \frac{8n}{\gamma}$ and $\theta = \frac{8D}{\gamma}$.

In order to compute the number of rejected requests in the eNB, $N_{R,i}^2$, we propose an iterative algorithm using (15) and (16). The main steps that each eNB performs to calculate $N_{R,i}^2$ are summarized in Alg. 2.

The proposed *CN overload control* mechanism is a queue-aware mechanism which can control the variation of queue length and hence prevents the abrupt dropping requests in the CN. However, the rejection policy in this mechanism leads to the inefficient resource utilization and also a decrease in the successful transmitted requests of *ClassI*.

4.3 Simultaneous RAN and CN overload control

In this subsection, in order to meet the constraints of the RAN and the CN simultaneously, the overload control solutions in sections *4.1* and *4.2* are combined together. We note that in the *RAN & CN overload control*

mechanism, the ACB probability of *ClassI* and the variation of the MME's queue length are obtained using (5) and (17) respectively. Each eNB calculates the ACB probability for the devices of *ClassII*, $p^2_{ACB,i}$, through the messages that are received directly from the MME while considering the QoS of *ClassI*. To consider the QoS of *ClassI*, each eNB sets $N^2_{sc,max}$ as the maximum of admissible offered load to the MME in frame k, that is,

$$N^2_{L,i}(k) = min\Big\{(N^2_{L,i}(k-1)+\theta), N^2_{sc,max}\Big\}\boldsymbol{I}(y=0) + \Big(N^2_{L,i}(k-1) - \sigma N^2_{L,i}(k-1)\Big)\boldsymbol{I}(y=1) \tag{23}$$

$N^2_{sc,max}$ in (23) represents the maximum transmitted requests of *ClassII* devices which is calculated according to the QoS of *ClassI* using (13).

Let $\bar{N}^2_{ACB,i}$ denotes the accepted number of *ClassII* devices which are allowed to participate in the contention according to the imposed constraints of the RAN and the CN. Therefore, by using the computed value of $N^2_{L,i}(k)$ in (23), (24) can be solved numerically to find $\bar{N}^2_{ACB,i}$ in each frame.

$$N^2_{L,i}(k) = \binom{\bar{N}^2_{ACB,i}(k)}{1}(1-\frac{1}{M_i})^{\bar{N}^2_{ACB,i}(k)+N_1P^1_{ACB}-1}. \tag{24}$$

By knowing $\bar{N}^2_{ACB,i}(k)$, $p^2_{ACB,i}$ in the k^{th} frame can be computed as given in (25).

$$p^2_{ACB,i}(k) = min(1, \frac{\bar{N}^2_{ACB,i}(k)}{N_{2,i}(k)}), \tag{25}$$

where $N_{2,i}(k)$ denotes the expected number of active *ClassII* devices in frame k which is calculated in a similar way to (11). Alg. 3 shows the ACB calculation process by the eNB in the *RAN & CN overload control* mechanism.

A comparison of the *RAN overload control*, the *CN overload control* and the *RAN & CN overload control* mechanisms in terms of the average number of successful *ClassI* transmissions, average number of dropped *ClassII* requests, the queue length of the MME, and the required number of signalling messages are summarized in Table 1. In the *RAN & CN overload control* mechanism, the accurate adjustment of contending MTC devices of *ClassII* prevents the request of MTC devices to be rejected/dropped in the RAN/CN. Due to the available interface between the eNB and the MME and also the broadcasting channel in the RAN, the message passing process in the *RAN & CN overload control* mechanism is applicable. However, this mechanism incurs more signalling messages. The rejection policy in the *CN overload control* mechanism wastes resources and degrades the QoS of *ClassI* devices. Also, the dropping policy in the *RAN overload control* mechanism degrade the QoS of *ClassII* devices.

Algorithm 3 RAN & CN overload control Mechanism

Input: M, ϕ, N_T^1, N_T^2, P_0, σ, θ
Output: p_{ACB}^1, $p_{ACB}^2(k)$
Set $N_{L,i}^2(1)$:=0
Compute p_{ACB}^1 by Eq. (5)
Compute $\bar{N}_{ACB}^2$ by Eq. (9)
Compute $N_{sc,max}^2$ by Eq. (13)
for kth frame **do**
    Monitor y
    Compute $N_{L,i}^2(k)$ by Eq. (23)
    Find $\bar{N}_{ACB,i}^2(k)$ by Eq. (24)
    Compute $N_2(k)$ by Eq. (11)
    Update $p_{ACB}^2(k)$ by Eq. (25)
end for

5 Performance Evaluation

In this section, we compare the performance of the proposed mechanisms for overload control of MTC devices in the RAN, CN, and RAN & CN by simulations. The traffic of *ClassI* and *ClassII* are generated according to the two state Markov chain model in Fig. 3 and the Beta(3,4) distribution respectively. The arrival times of *ClassII* traffic bursts in cell 1 and cell 2 are considered to be 0 and 3 seconds respectively, while the time-independent *ClassI* traffic is running at the background. In this system, each MTC device of *ClassI* can transmit its request with probability p_{ACB}^1 as in (5). Also, each MTC device of *ClassII* can participate in the random access procedure according to the probability p_{ACB}^2 which are derived from (10) and (25) for the *RAN overload control* and *RAN & CN overload control* mechanisms, respectively. We simulate a frame-based random access procedure; where different events including new arrival requests of each class, resource selection by MTC devices, and transmitting the connection requests to the MME by the eNB occur at the end of a frame time, i.e., we consider one random access procedure at the end of each frame time. In the following simulations, AIMD parameters are computed according to what has been discussed in section 4.2. It is considered that a cell is active when it forwards some requests of the *ClassII* devices to the MME node. At first we assume that the eNB can estimate the number of active MTC devices of *ClassII* in the next frame while there is no error in the received feedbacks from the MME. At the end of this section, the effects of the estimation and feedback errors on the performance of *RAN & CN overload control* mechanism is evaluated. The simulation results are

the average of 40 independent runs which are compared with the analytical results in each scenario. The values of the system parameters for the simulations are summarized in Table 2.

5.1 Comparing the Performance of the Proposed Mechanisms

Since in the proposed mechanisms it is assumed that there are reserved resources for the devices of *ClassI* in the CN, eNB just considers the number of available resources in the RAN to calculate p^1_{ACB}. In Fig. 6, the average number of successful transmission of the *ClassI* devices against different access probability for these devices for different values of resources, $M = 35 \sim 60$, for one cell is depicted. As it is expected the maximum number of successful transmission is happened at the calculated optimum value of p^1_{ACB} which is derived in (5).

Next we consider the average number of successful transmissions of *ClassI* and the average queue length of the MME as two main constraints in the RAN and the CN respectively. We show that these constraints can be met simultaneously in the *RAN & CN overload control* mechanism, while the separate overload control in the RAN or CN can just guarantee one constraint at each time interval. We should note that the simultaneous satisfaction of the *RAN & CN overload control* is achieved at the cost of exchanging the required signalling.

In Fig. 7, the average number of successful transmissions of *ClassI* in cell 1 and cell 2 in the presence of *ClassII* traffic are shown for each mechanism. As it is shown in Fig. 7(a) and Fig. 7(b), $N^1_{sc,1}$ and $N^1_{sc,2}$ in the *RAN overload control* and *RAN & CN overload control* mechanisms are above the determined threshold, 8.4. However, in the *CN overload control* mechanism, $N^1_{sc,1}$ and $N^1_{sc,2}$ are decreased severely due to the non-controlled arrivals of *ClassII*. In addition, $N^1_{sc,1}$ and $N^1_{sc,2}$ in the *RAN & CN overload control* mechanism are greater than *RAN overload control* mechanism for $3s < t < 7s$ in cell 1 and for $5s < t < 10s$ in cell 2. This is imposed by service capacity of MME which enforces the eNB in the *RAN & CN overload control* mechanism to decrease the number of contending MTC devices of *ClassII*.

Fig. 8 shows the average queue length of MME for each mechanism. Since both *CN overload control* and *RAN & CN overload control* mechanisms are queue aware mechanisms, the number of requests in the queue are sustained around the determined threshold in these mechanisms. This prevents from resource underutilization or buffer overflow conditions in the MME's queue. While in the *RAN overload control* mechanism the average queue length of MME increases extremely in the presence of massive arrival of MTC devices. In this case, the requests of MTC devices are dropped in the CN when the buffer of the MME is overflowed. In the *CN overload control* mechanism, the offered load by each cell to the MME node is based on the exact value of N^2_L however

in the *RAN & CN overload control* mechanism, the offered load is based on the ACB probability which itself comes from N_L^2. This causes the amplitude of oscillations in the *CN overload control* mechanism to be less than the *RAN & CN overload control* mechanism for the same number of simulations in Fig. 8. Since we have obtained AIMD parameters for the case in which both cells are active, the average queue length decreases for $t > 13s$ where only one cell is active.

The dynamics of the queue length in the *RAN* & *CN overload control* mechanism is presented in Fig. 9 where it is assumed that the beginning of the burst traffic for two cells is the same. In this figure the average and the corresponding 98% confidence interval of the MME queue length in each 40 frames for 20 independent runs are depicted and compared with the approximated expected behavior of differential equation in (18). The simulation results are consistent with the expected behavior after the building up phase of the MME's queue. That is due to the Beta distribution arrival model of the MTC devices, a few numbers of requests are generated initially and after some notification messages of the MME the offered load reach to the average capacity of the MME.

To show the performance of the *RAN* & *CN overload control* mechanism in controlling the number of contending MTC devices of *ClassII*, the average number of successful transmissions of *ClassII* in the RAN for cell 1 and cell 2 are shown in Fig. 10(a) and Fig. 10(b), respectively, in comparison with the RAN or CN overload control.

In the *RAN* & *CN overload control* mechanism, the maximum value of the $N_{sc,1}^2$ and $N_{sc,2}^2$ are adjusted according to the imposed constrains by the RAN and CN. Therefore, the maximum value of $N_{sc,1}^2$ in the *RAN* & *CN overload control* mechanism for $t < 3s$ when MME's queue is empty, reaches to the constrain of the RAN, i.e, 17.1. While, the maximum of $N_{sc,1}^2$ and $N_{sc,2}^2$ for $t > 3s$ reaches to the service capacity of the MME, i.e, 7 and 14 when both cells and one cell are active respectively. As we can see in Fig. 10, $N_{sc,1}^2$ and $N_{sc,2}^2$ in the *RAN overload control* and *CN overload control* mechanisms are more than the *RAN* & *CN overload control* mechanism which leads to the droped/rejected requests in the CN/RAN respectively. To obtain the number of the dropped/rejected requests, $N_{sc,i}^2$ in the *RAN overload control*/*CN overload control* mechanism can be subtracted from $N_{sc,i}^2$ in the *RAN* & *CN overload control* mechanism respectively.

Table 3 shows the total number of dropped, rejected, and retransmitted requests for three mechanisms. Since the value of ACB probability is updated according to the imposed constraints of the RAN and CN, the requests of the *ClassII* devices are not dropped or rejected in the *RAN & CN overload control* mechanism. However, in the *RAN overload control*/*CN overload control* mechanism the non-controlled behavior of the *ClassII* devices results some requests of this class to be dropped/rejected. Also, the number of retransmitted

requests can be controlled through the adaptive ACB scheme which is applied in the *RAN overload control* and *RAN & CN overload control* mechanisms. While, the non-adaptive ACB scheme in the *CN overload control* mechanism increases the number of retransmitted requests.

5.2 The Effect of Incomplete Information

In this subsection, at first a simulative study has been provided to investigate the sensitivity of the *RAN* & *CN overload control* mechanism against the possible estimation error of N_1 and N_2. Then the effect of feedback error on the proposed mechanism has been investigated. We consider a scenario in which we have: $N_T^2 = 15000$, $D = 20(request/frame)$, $t_s = 1s$, $T_1 = 0s$, and $T_2 = 5s$. The following simulations are the results of 40 independent runs and depicted the average of the interested parameter in every 10 frames.

Let the estimated values of N_1 and N_2 be υN_1 and υN_2 respectively, where $\upsilon > 1$ and $\upsilon < 1$ reflect over-and under-estimation of the number of devices. The number of successful transmissions of *ClassII*, *ClassI* devices and the queue length of the MME in cell 1 for different scenarios of over-estimation ,$\upsilon = 1.5$, and under-estimation, $\upsilon = 0.5$, are shown in Fig. 11 respectively. In Fig. 11(a) when $t < 6s$, p_{ACB}^2 is calculated using the imposed constraint of the RAN which causes every error in the estimation of N_1 and N_2 will result in either the resource underutilization or overload condition. However, due to the adaptive overload control mechanism in the MME, the maximum value of N_{sc}^2 is limited to the service capacity of the MME which controls the underestimated value of N_1 or N_2. When $t > 6s$, p_{ACB}^2 is calculated using the imposed constraint of the MME which causes N_{sc}^2 not to vary considerably. This happens due to the no erroneous in the received feedback of the MME.

As it is show in Fig. 11(a), the over-estimation of the number of devices decreases the efficient resource utilization and hence increases the total service time of devices of *ClassII*. While in the under-estimation scenario, the number of contending devices of both classes are increased which leads to a decrease in N_{sc}^1, as shown in Fig. 11(b). Finally, the effect of estimation error on the queue length of the MME is shown in Fig. 11(c). Since in the *RAN & CN overload control* mechanism the ACB parameter is calculated using the received feedback of the MME, the queue length is sustained near to its determined threshold for $\upsilon > 1$ or $\upsilon < 1$.

In Fig. 12, the effect of feedback error on N_{sc}^2, N_{sc}^1 and Q are shown respectively. We assume feedback error is occurred when the binary message of the MME is received with error. The feedback error increases the amplitude of oscillations in N_{sc}^2, N_{sc}^1 and Q. In Fig. 12(a) when $t < 6s$, the imposed constraint of the RAN

limits the maximum oscillation of N_{sc}^2 through controlling the number of contending MTC devices of *ClassII*. This causes a reduction in N_{sc}^2 compared to the case that the feedback is error free. When $t > 6s$, the imposed constraint of the CN determines the number of contending MTC devices so, there is no limit on the maximum number of contending MTC devices from the RAN point of view. Therefore, the value of N_{sc}^2 dose not differ noticeably from the error free feedback case. Also, the less value of N_{sc}^2 for $t < 6s$ leads to greater values of N_{sc}^1 as is shown in Fig. 12(b). In Fig.12(c), the variations of the MME's queue length in both cases are shown. As expected we have more oscillations in the erroneous feedback case in the comparison with the error free scenario.

6 Conclusion

In this paper, we have investigated the performance of the simultaneous and separate RAN and CN overload control problems for LTE/LTE-A based machine type communications. We have formulated the random access process in the RAN for two relevant MTC traffics to update the access probability by considering their QoS requirements. Then, the AIMD rule is used to avoid the overload in the CN node. The proposed approaches in the RAN and CN, can sustain the load level in the RAN and CN according to their service capacities separately. Also, an efficient simultaneous RAN and CN overload control mechanism is proposed which uses the MME' queue notification messages to find the ACB factor for the MTC devices of each class. Simulation results are provided to show the effectiveness of the proposed mechanisms in terms of meeting the QoS of each class traffic.

References

1. K.-C. Chen and S.-Y. Lien, "Machine-to-machine communications: Technologies and challenges," *Elsevier, Ad Hoc Networks*, vol. 18, pp. 3-23, July 2014.
2. O. Vermesan and P. Friess, "Internet Of Things-From Research and Innovation to Market Deployment", *(Aalborg, Danmark: River Publisher, 2014, 1st edn. 8-30.)*
3. A. Laya, L. Alonso, and J. Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communication? A survey of alternatives," *IEEE Communication Surveys & Tutorials*, vol. 16, no. 1, pp. 4-16, Feb. 2014.
4. F. Hussain, A. Anpalagan, and R. Vannithamby, "Medium access techniques in M2M communication:survey and critical review," *Transactions on Emerging Telecommunications Technologies*, DOI: 10.1002/ett.2869, Sep. 2014.
5. 3rd Generation Partnership (3GPP), "Study on RAN improvements for machine type communications; Release 11, v.11.0.0," Sophia-Antipolis Cedex, France, TR 37.868, sep. 2011.
6. C. A.Haro and M. Dohler, Machine-to-Machine(M2M) Communications: Architecture, Performance and Applications, *(Elsevier, 2015, 1st edn, pp. 158-164.)*

7. 3rd Generation Partnership (3GPP), “System improvements for machine-type communications; Release11, v.11.0.0,” Sophia-Antipolis Cedex, France, TR 23.888, sep. 2011.
8. S. Duan, V. Shah-Mansouri, and V. W. S. Wong, “Dynamic Access Class Barring for M2M Communications in LTE Networks,” *IEEE Global Telecommunications Conference (GLOBECOM 2013)*, pp. 4747-4752, Dec. 2013.
9. T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, “PRADA: Prioritized Random Access With Dynamic Access Barring for MTC in 3GPP LTEA Networks,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2467-2472, Jun. 2014.
10. S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, “Cooperative Access Class Barring for Machine-to-Machine Communications,” *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 27-32, Jan. 2012.
11. A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, “Cellular-Based Machine-to-Machine: Overload Control ,” *IEEE Network*, vol. 26, no. 6, pp. 54-60, Nov. 2012.
12. A. Amokrane, A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, “Congestion control for machine type communications,” *IEEE International Conference on Communications, (ICC 2012)*, pp. 778-782, June 2012.
13. *Machine-to-Machine Communications (M2M): M2M Service Requirements, v2.1.1*, ETSI TS 102 690, July 2013.
14. M. Tauhidullslam, A.-E. M.Taha, and S. Akl, “A Survey of Access Management Techniques in Machine Type Communications,” *IEEE Communication Magazine*, vol. 52, no. 4, pp. 74-81, Apr. 2014.
15. 3rd Generation Partnership (3GPP), “Evolved Universal Terrestrial Radio Access(E-UTRAN);Radio Resource Control (RRC); Release 10, v.10.5.0,” Sophia-Antipolis Cedex, France, TS 36.331, Mar. 2012.
16. M.Tavana, V.S.Mansouri, and V. W.S.Wong, “Congestion Control for Bursty M2M Traffic in LTE Networks,” *IEEE International Conference on Communications, (ICC 2015)*, pp. 5815-5820, June 2015.
17. C.-Y. Oh, D. Hwang, and T.-J. Lee, “Joint Access Control and Resource Allocation for Concurrent and Massive Access of M2M Devices,” *IEEE Transactions on Wireless Communications*, vol. PP, no. 99, pp. 1-11, Mar. 2015.
18. A. Ksentini and et.al, “Congestion-Aware MTC Device Triggering,” *IEEE Internation Conference on Communication,(ICC 2014)*, pp. 294-298, June 2014.
19. S.-I. Sou, and S.-M. Wang, “Performance Improvements of Batch Data Model for Machine-to-Machine Communication,” *IEEE Communication Letters*, vol. 18, no. 10, pp. 1775-1778, Aug. 2014.
20. 3rd Generation Partnership (3GPP), “Evolved Universal Terrestrial Radio Access(E-UTRAN); Physical channels and modulation; Release 11, v.11.1.0,” Sophia-Antipolis Cedex, France, TS 36.21, Sep. 2013.
21. V. Misra, W.-B. Gong, and D. Towsley, “Fluid-based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED,” *in Proc. ACM SIGCOMM*, vol. 30, no. 4, pp. 151-160, Aug. 2000.
22. M. Chen, and et al., “Normalized Queueing Delay: Congestion Control Jointly Utilizing Delay and Marking,” *IEEE/ACM Transactions on Networking*, vol. 17, no. 2, pp. 618-631, July 2008.
23. J. Li, E. Gong, Z. Sun, and H. Xie, “QoS-Based Rate Control Scheme for Non-Elastic Traffics in Distributed Networks,” *IEEE Communication Letters*, vol. 19, no. 6, pp. 1089-7798, June 2015.

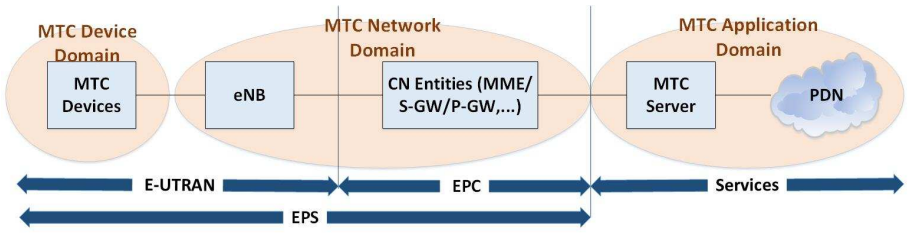

Fig. 1 MTC architecture model in EPS.

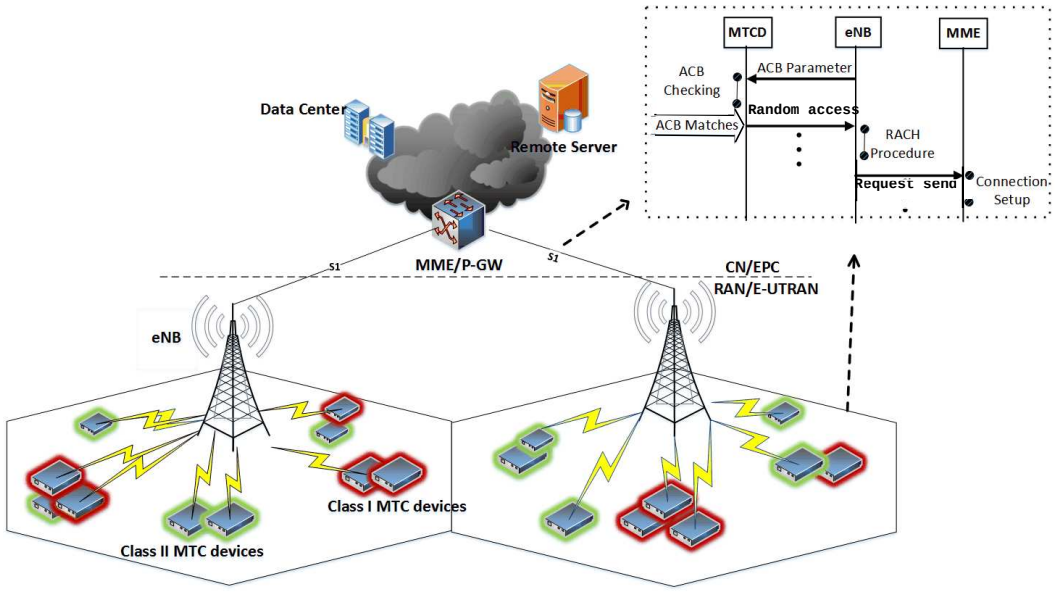

Fig. 2 System model of MTC in LTE and activation procedure in the ACB scheme.

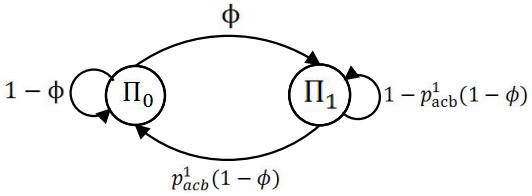

Fig. 3 Traffic model of the *ClassI* MTC.

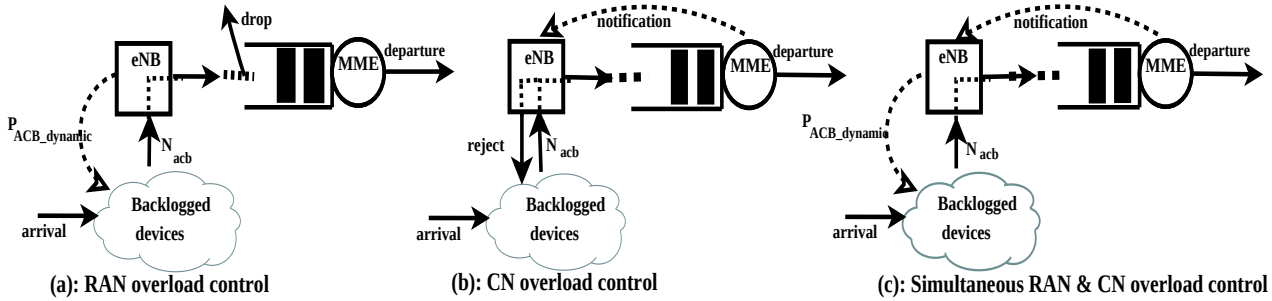

Fig. 4 Three overload control schemes in the network domain of MTC.

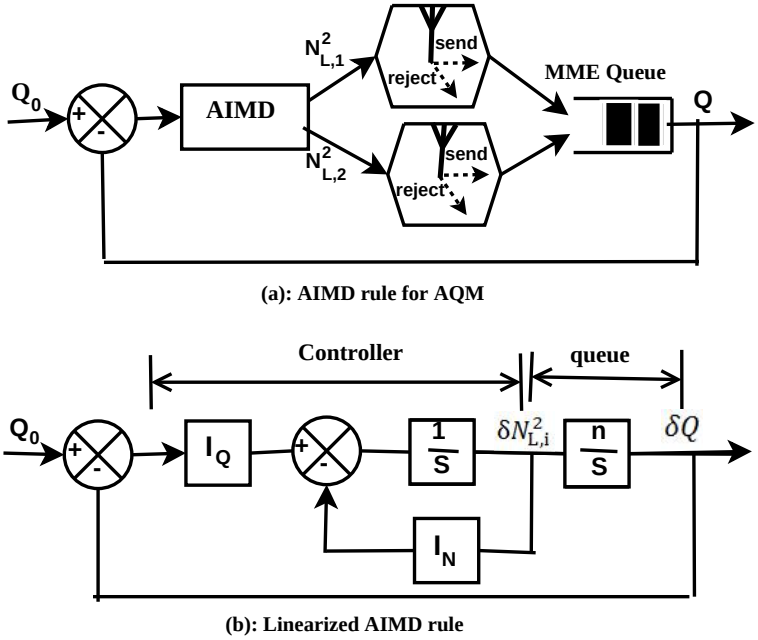

Fig. 5 Block-diagram of the AIMD based MME queue management.

Table 1 Comparison of the overload control for the RAN, the CN, and the RAN & CN mechanisms

	Mechanisms		
	RAN overload control	CN overload control	RAN & CN overload control
Avg. no. of successful transmissions of *ClassI*	Guaranteed	Not Guaranteed	Guaranteed
Avg. no of dropped requests of *ClassII* in the MME	Not Guaranteed	Guaranteed	Guaranteed
The queue length of MME	Not Guaranteed	Guaranteed	Guaranteed if $N_{sc}^2(k) > N_L^2(k)$
Number of signaling messages	Moderate	Moderate	High

Table 2 System parameters

Parameter	Details	Value
ϕ	Arrival probability of devices of *ClassI*	0.03
N_T^1	Total number of *ClassI* devices	400
N_T^2	Total number of *ClassII* devices	10000
B	Buffer size	160
Q_0	Threshold of the queue length (requests)	80
D	Service rate of the MME(request/frame)	14
P_0	Required success probability of *ClassI*	0.7
M_1, M_2	Number of resources in cell 1, cell2	100, 150
T_1, T_2	Burst arrival time of *ClassII* devices in cell 1, cell 2 (second)	0, 3
γ	defined settling time (second)	0.8

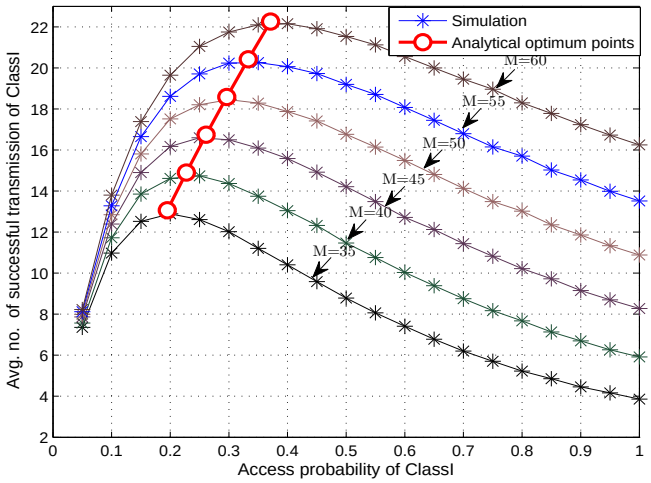

Fig. 6 Average number of successful transmission of *ClassI* MTC against p_{ACB}^1 for $\phi = 0.6$, $N_T^1 = 200$, and different values of M.

Table 3 Comparison of the main performance metrics of ClassII devices for cell 1 in the RAN, CN, and RAN & CN mechanisms

Performance metrics of ClassII devices	**Mechanisms**		
	RAN overload control	CN overload control	RAN & CN overload control
Total no. of dropped requests $(\times 10^3)$	3.5	-	-
Total no. of rejected requests $(\times 10^3)$	-	12.3	-
Total no. of retransmitted requests $(\times 10^3)$	13.9	206.9	12.8

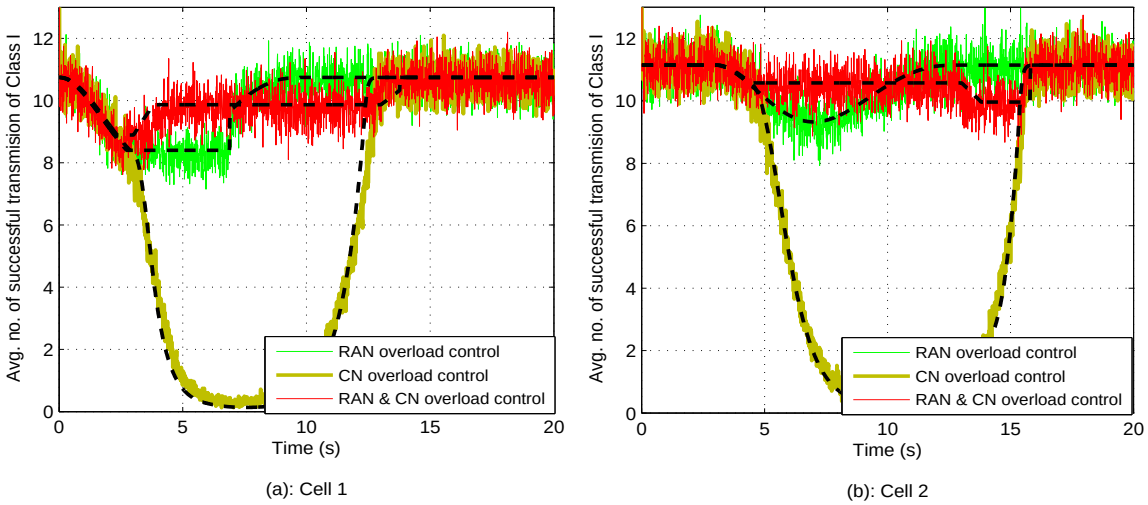

Fig. 7 The average number of successful transmissions of *ClassI* over the time in two cells (dash line:the expected trend by analysis; line:simulation).

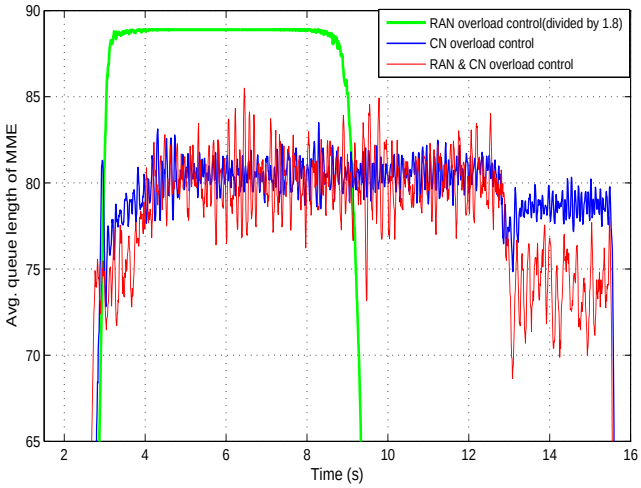

Fig. 8 The average queue length of the MME over the time.

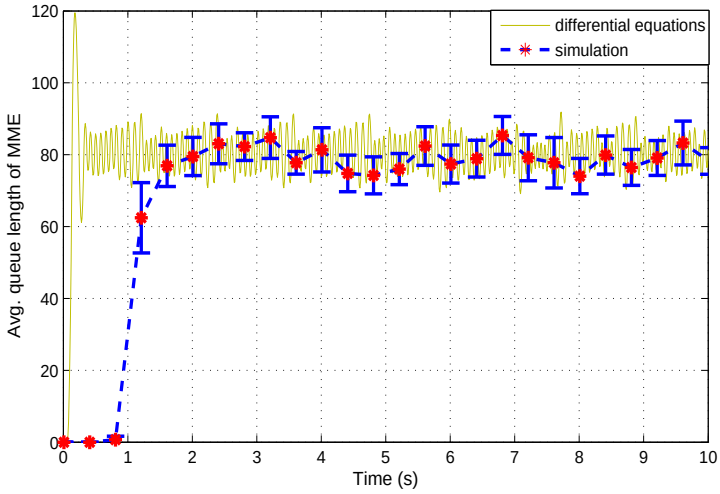

Fig. 9 Queue length variation of the MME over the time.

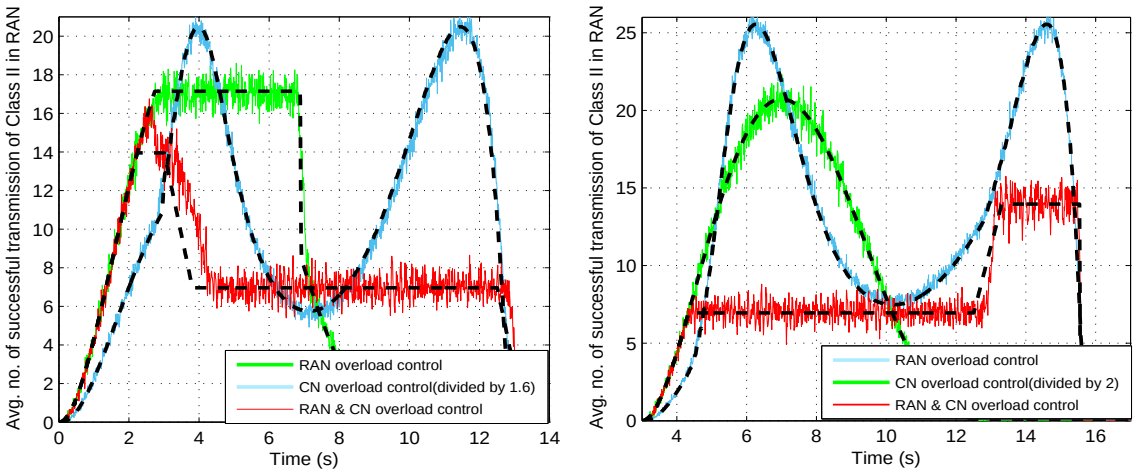

Fig. 10 Average number of successful transmissions of *ClassII* over the time in two cells (dash line: the expected trend by analysis; line:simulation).

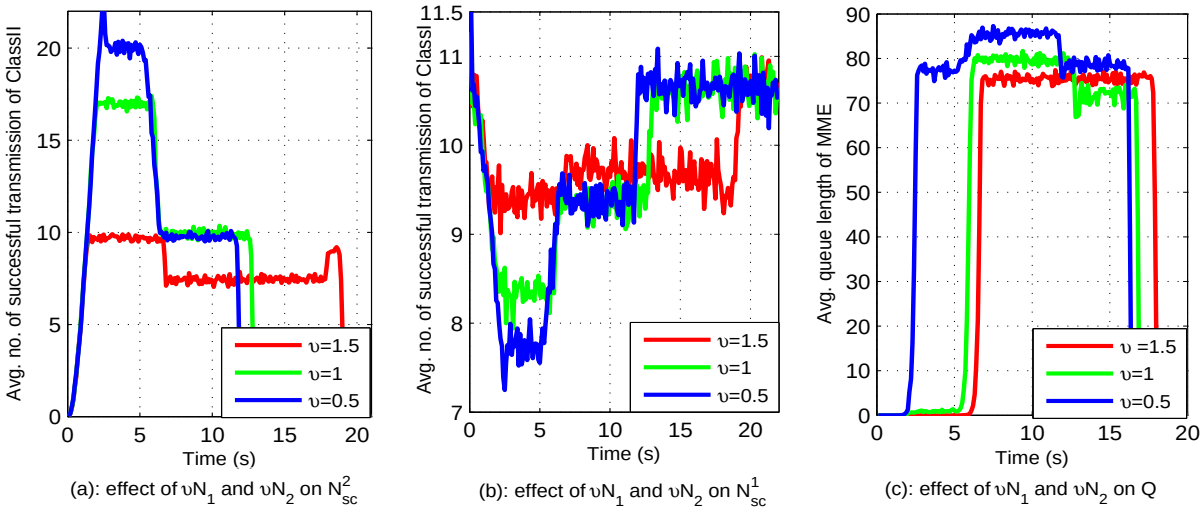

Fig. 11 The average number of successful transmissions of *ClassII* and *ClassI* and the queue length of the MME in cell 1 for υN_1 and υN_2 with values of $\upsilon = 0.5, 1, 1.5$.

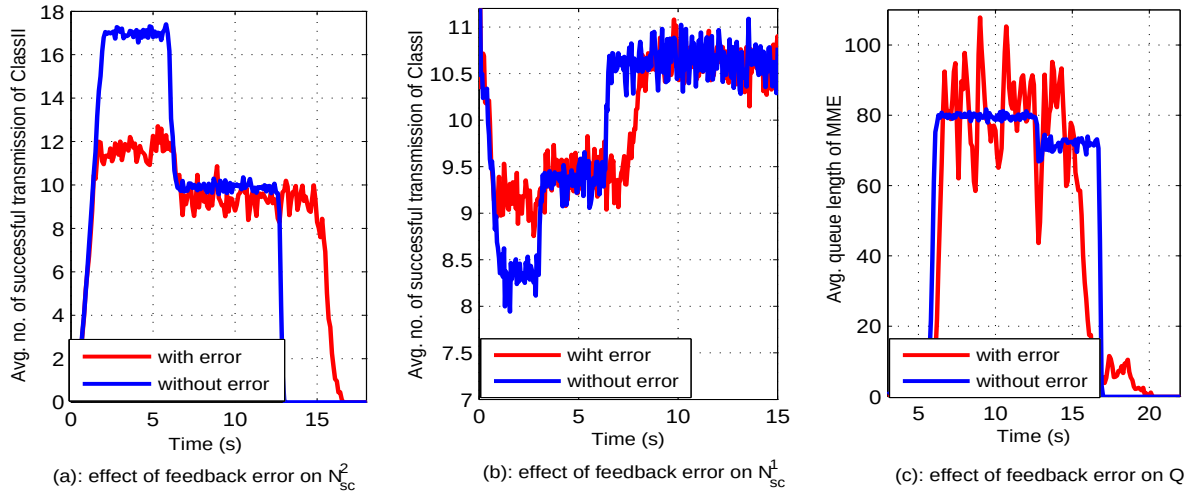

Fig. 12 The average number of successful transmissions of *ClassII* and *ClassI* and the queue length of the MME in cell 1 for feedback error with probability of 0.4.