Congestion Control for Machine Type Communications

Ahmed Amokrane*, Adlen Ksentini†, Yassine Hadjadj-Aoul†, Tarik Taleb‡
*IRISA/ENS Cachan, Rennes - Email: ahmed.amokrane@irisa.fr
†IRISA/University of Rennes 1, Rennes - Email: firstname.lastname@irisa.fr
‡ NEC Europe - Email: tarik.taleb@neclab.eu

Abstract—**One of the most important problems posed by cellular-based machine type communications is congestion. Congestion concerns both the radio access network and the mobile core network, impacting both the user data and the control planes. In this paper, we address the problem of congestion in machine type communications. We propose a congestion-aware admission control solution that selectively rejects signaling messages from MTC devices at the radio access network following a probability that is set based on a proportional integrative derivative controller reflecting the congestion level of a relevant core network node. We evaluate the performance of our proposed solution using computer simulations. The obtained results are encouraging. In fact, we succeed in reducing the amount of signaling, to reach a target utilization ratio of resources in the core network.**

I. Introduction

Machine Type Communications (MTC) in the 3GPP terminology (also called Machine-to-Machine (M2M)) refer to automated applications which involve machines or devices communication through a network without any human intervention. The MTC devices can be embedded in different environments such as cars, cell towers, and vending machines. They are generally spread in a wide area and should communicate through widely deployed networks. While some M2M deployments use short-range or proprietary radio links, mobile cellular-based M2M solutions are preferred. Mobile cellular-based M2M communications have the advantage of easier installation and provisioning, especially for short-term deployments. Cellular mobile networks offer different network technologies for M2M communications. On the other hand, M2M communications are seen promising to leverage revenue for mobile operators. Indeed, several forecasts state a significant market growth, over the next years, for both the MTC device and the MTC connectivity segments. The growth is expected at a Compound Annual Growth Rate (CAGR) exceeding 25% [3], [4], [5]. According to these forecasts, billions of machines or industrial devices will be potentially able to benefit from MTC.

However, to support M2M communications, a mobile operator has to accommodate its network to support a large number of MTC devices, which can overload its network resources and introduce congestion in the network at both the data and control planes. In fact, congestion may occur due to simultaneous signaling messages from MTC devices.For instance, if many devices detect an event at the same time, they send their alerts toward a central server at the same time, leading to congestion in the different nodes of the network, across the communication path.

In order to take advantage of the huge opportunities raised by a global M2M market over cellular networks, the 3GPP group , System Architecture - SA2, is in process of establishing requirements for 3GPP network system improvements that support MTC in the Evolved Packet System (EPS) [1]. Specifically, transport services for MTC as provided by the 3GPP system and the related optimizations are being considered as well as aspects needed to ensure that MTC devices and/or MTC servers and/or MTC applications do not cause network congestion or system overload.

In this paper, we focus on the problem of congestion when deploying MTC devices over mobile networks. To address this issue, we propose rejecting MTC signaling traffic at the eNodeBs when the Evolved Packet Core network (EPC) nodes such as Mobility Management Entity (MME) and Serving gateways (S-GWs) are congested. Our solution dubbed as Congestion-Aware Admission Control (CAAC), uses the classical control theory: (i) to model the impact of the MTC signaling traffic on the queue length at the MME/S-GW; (ii) to use the Proportional Integrative Derivative (PID) controller [12] to mitigate the MME/S-GW overload by rejecting the MTC traffic at the radio access network (eNodeBs). In other words, the PID controller will detect and control congestion defining the amount of MTC traffic to be admitted/rejected at the eNodeBs.

The remainder of this paper is organized as follows. Section II overviews the problem of congestion in LTE when deploying MTC devices and presents some related work. Section III presents a general and detailed description of our proposed solution CAAC. The proposed solution is, then, evaluated in Section IV through simulations. Finally, Section V highlights some future work and concludes this paper.

II. State of art

A. Problem statement

Cellular networks are mainly designed to handle Human to Human, Machine to Human and Human to Machine communications, where the proportion of the uplink traffic is low in comparison to the downlink traffic. In contrast, MTC applications communicate without any external intervention, and involve a huge number of cheap and low power devices that generate more signalization than the effective data. It is

978-1-4577-2053-6/12/$31.00 ©2012 IEEE

worth mentioning that congestion could arise when the MTC devices detect an event, and try to attach to the network and send data to the remote server at the same time. Figure 1 shows the potential locations of congestion in the Evolved Packet System (EPS) when MTC devices are trying to send data to a remote server. Indeed, the congestion could happen at different locations:

- The radio part as a lot of MTC devices are connected to the same eNodeB and consequently use the same channels leading to high contention.
- The Evolved Packet Core Network (EPC), mainly in: (i) the MME, which is responsible for managing the attachment of devices to the network; (ii) the S-GW in charge of carrying the traffic; (iii) the P-GW as a lot of MTC devices will send and receive data through it.

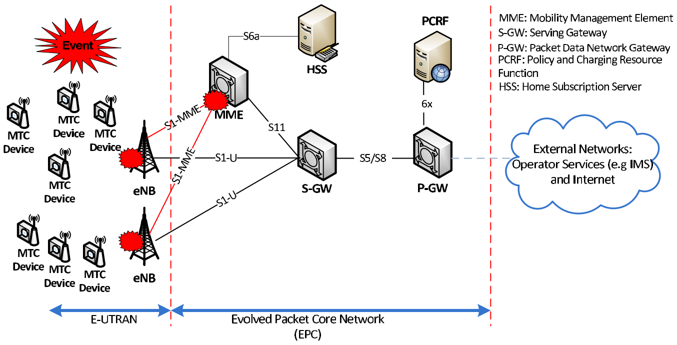
Event
MTC Device
MTC Device
MTC Device
MTC Device
MTC Device
MTC Device
MTC Device
eNB
eNB
S1-MME
S1-MME
S1-U
S1-U
S6a
HSS
MME
S11
S-GW
S5/S8
PCRF
6x
P-GW
MME: Mobility Management Element
S-GW: Serving Gateway
P-GW: Packet Data Network Gateway
PCRF: Policy and Charging Resource Function
HSS: Home Subscription Server
External Networks: Operator Services (e.g IMS) and Internet
E-UTRAN
Evolved Packet Core Network (EPC)

Fig. 1. Congestion location in the LTE architecture

Furthermore, regarding the nature of traffic, we can mainly separate the congestion location into two classes, at the user plane and at the control plane. The first class is caused by the data sent by the MTC devices on the uplink. Even if the amount of data per MTC device is small, the sum of all data can lead to congestion at the S-GWs and/or P-GWs. The second class of congestion is due mainly to the fact that MTC devices continuously generate signaling traffic to attach to the network, in response to paging, for tracking area updates, for the reason of mobility management, or when triggered by the remote server due to the occurrence of a specific event, etc. This high amount of signaling may cause congestion and overhead at the MME.

B. Related Work

Congestion that may occur due to simultaneous signaling messages from MTC devices can be significant as they may lead to peak load situations and may have a tremendous impact on the operations of a mobile network, mainly on the performance of vital nodes with scarce resources such as domain control plane nodes (e.g., SGSN in UMTS, MME in the EPS [6], [7]), gateways (e.g., GGSN in UMTS, PDN GWs in EPS), and Radio Access Network (RAN) (e.g., eNB, Base Stations and Radio Network Controller "RNC" in UMTS). Thanks to the wide bandwidth of LTE and other emerging radio technologies (e.g., LTE-advanced), congestion due to user data packets is, for the time being, of less importance. In the context of MTC, signaling congestion may happen due to a malfunction in the MTC server or application (e.g., synchronized recurrences of a particular MTC application) and/or due to massive attempts from a potential number of MTC devices to attach/connect to the network all at once [8]. A straightforward solution to signaling congestion can be in the form of designing MTC applications that are friendly to mobile network operators. However, mobile operators cannot risk the operations of their networks and the quality of the provided services by leaving the whole signaling congestion problem to MTC application developers. Existing mechanisms for MTC devices signaling congestion avoidance and overload control can be classified into two categories, namely soft mechanisms and rigid ones [1]. In the former, the mobile network operator takes "soft" measures to minimize the frequency of attempts (of MTC devices) to carry out a particular procedure without having to throttle them. In case of the latter, the mobile network operator takes rigid measures disallowing the concerned MTC devices from connecting to the network and/or executing the intended procedure. Solution examples that fall in the former category are:

- Reducing signaling due to TAUs from MTC devices by increasing the TAU period timer.
- Triggering MTC devices, following a pull model, to attach/connect to the network only based on explicit indications from the MTC server.

The vast majority of signaling congestion avoidance and overload control adopt a rigid strategy, implementing one or a set of the following approaches:

- First of all, MTC devices are grouped based on different metrics/features (e.g., low priority access, low mobility, online/offline small data transmission, etc)
- Forbidden/grant times are allocated for each MTC device based on its subscription in HSS.

MTC devices with time controlled MTC feature connect to the network only at certain time periods, predefined by the network operator, named "grant time intervals". The network also defines "forbidden time intervals" during which a MTC device is not allowed to connect to the network, be it the home network or a visited network. Over the grant time, assigned to a MTC device, the communication window is further limited. Indeed, in some scenarios and depending on the targeted application, MTC devices do not need to connect to the network during the total duration of the grant time interval.The access time of MTC devices is also randomized over the communication window/grant time. In case of multiple MTC devices attempting to connect to the network during a specific and short communication window/grant time, to avoid signaling congestion and to cope with possible network overload during communication windows, the communication windows of the different MTC devices can be distributed over the grant time interval, via for example, randomization of the start times of the individual communication windows. Another approach for tackling signaling congestion is by rejecting connection/attach requests, from MTC devices, by specific network nodes. This operation should take place targeting only MTC traffic, particularly those of MTC applications that are causing congestion.Rejection of connection/attach requests shall be done while ensuring a rejected MTC device does not immediately reinitiate the same request (e.g., only until after a predetermined back off time) and the rejected MTC devices do not attempt connecting to the network all at the same time, but rather at randomized times. Network attachment requests can be rejected either at (or nearby) RAN (e.g., eNB, RNC) or by SGSN/MME. In case the MTC access is controlled by RAN, whenever the network is about to get congested by MTC applications, i.e., following a congestion status feedback from PDN GWs/Serving Gateways or GGSN, MME/SGSN sends a notification message to RAN nodes triggering MTC access control indicating MTC barring information (e.g., barring factor, MTC group to block, barring time, etc). In case the MTC access is controlled by SGSN/MME, HSS may provision MME/SGSNs with information on grant time and forbidden time intervals in MTC subscription. Based on this feedback and also on local operator policies, SGSN/MME then determines authorized times for each MTC device and communicates them to the respective MTC devices via MTC server or by NAS (Non-Access-Stratum) signaling directly from MME/SGSN [9]. In case of congestion occurrence during the authorized times, SGSN/MME may reject connections from concerned MTC devices and provide them with back off times for later accesses, or simply send them a congestion notification message triggering them to reduce their data transmission rate. It should be noted that the latter incurs major impact on the MME/SGSN implementation.

In summary, one of the advantages of the RAN-based solution consists in the fact that there is no wastage in signaling from MTC devices that need to be blocked at first place. It also assists in controlling overload of both RAN nodes and core network nodes. One aspect which is missing so far in MTC signaling congestion control consists in the lack of a mechanism that handles a bulk of similar signaling messages from MTCs in a single shot (i.e., bulk MTC signaling handling). In [10], [11], the authors show the potential of handling signaling messages common to a group of MTC devices in bulk, as a complementary or alternative approach to the above-mentioned solutions.

III. CAAC: A Congestion-Aware Admission Controller

A. General Overview

One of the most efficient approaches to tackle signaling congestion consists in rejecting ongoing connections/attach requests of MTC devices by a specific network's node. This operation should only target the traffic originating from MTC devices, and particularly those of MTC applications causing the congestion. This means that the admission controller shall have no impact on non-MTC traffic.

To ensure a robust behavior of the controller, one needs to avoid the immediate reinitialization of the attach requests (i.e., the reinitialization process should be performed only after a predetermined back off time). Moreover, the rejected MTC devices should not attempt to connect to the network at the same time, but at randomized times. In other words, back off times should be indicated to the concerned MTC devices in a way to ensure a good distribution of future incoming attach requests. Such randomization of MTC devices' access can be performed at MTC devices level or by the network, including MTC servers. When it is triggered by the network, the operation of a MTC device starts after receiving paging request message from the network or from an application level data originating from the server. When it is computed by a particular device, its operation can start immediately after being paged or by adding some random delay after being paged.

Note that, the attach requests rejection can be done based on specific MTC group identifiers and/or by considering the MTC devices' traffic class, similar in spirit to the 3GPP's Access Class Barring solution. The solution proposed in CAAC, which combines Admission Control and Congestion Control, follows this approach. Indeed, each node in the EPC "Evolved Packet Core Network" (MME, S-GW and P-GW) monitors its state and detects congestion. Then, it sends notifications to all its downstream nodes (i.e. eNodeB) in order to reduce the uplink traffic. As stated above, the MTC traffic is rejected depending on its priority, which means that MTC traffic with high priority will be rejected with a lower probability than other lower priority traffic.

B. Congestion monitoring

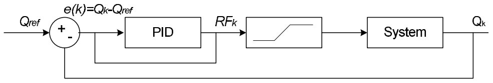

Fig. 2. The feedback loop for congestion monitoring

In order to handle and control the congestion, each node in the EPC has to monitor its queue length. According to this length it will send a signaling message, which specifies, to the upstream nodes, if the eNodeBs have to apply Admission Control for MTC traffic. The signaling message, to be ultimately received by the eNodeBs, include the congestion level which is represented by the *Reject Probability* parameter that specifies the level of Admission Control to be applied in order to reduce the amount of MTC applications' traffic. To derive the *Reject Probability*, we propose to use, in this paper, control theory, which consists in controlling a system that comprises one EPC node (likely to be overloaded) and the list of eNodeBs connected to it (e.g., 1 MME or 1 SGW and n eNodeBs).

Note that we focus, in the following, on the congestion which may happen at the MME. Such congestion is directly induced by the delay in processing the incoming packets at the application layer. This delay may lead to high and variable latencies at the input buffer, which may induce buffer overflow (i.e. packets loss). In contrast, the congestion in the other elements of the EPC, and particularly the P-GW, concerns the link layer, which may affect the outgoing buffer length.

Figure 2 depicts a feedback loop used to control the ingoing buffer length of the MME. The system will represent, then, all the elements impacting its length. The PID controller is used to derive the *Reject Factor* RF, which will help for converging the queue length to the reference or the desired queue length. The *Reject Factor* reflects, then, the system action to be done in order to reduce the congestion level .

According to the PID controller the $RejectFactor$ for each class of traffic C_i is, then, given by:

$$\begin{aligned} RF_i(k) &= RF_i(k-1) + K_p(e(k) - e(k-1) + K_i T_s e(k) \\ &\quad + \frac{K_d}{T_s}(e(k) - 2e(k-1) + e(k-2)) \end{aligned}$$

where : $Q(k)$, Q_{ref}, $e(k)$, and T_s represent, respectively, the queue length at the k^{th} sampling instant ($t = k * T_s$), the targeted queue length, the difference between the queue length at the instant k and the reference queue length ($e(k) = Q(k) - Q_{ref}$), and the sampling time. The constants K_p, K_i and K_d represent, respectively, the proportional gain, the integral gain, and the derivative gain.

The *Reject Factor* represents the amount of traffic to reject to reach the reference queue length. The higher the level of congestion is, the higher the $RejectFactor$'s value is. In contrast, lower values mean that the system is not congested.

Now, the RF has to be translated into a *Reject Probability* in order to be signaled to the eNodeBs participating in the congestion. To handle this, we propose to heuristically derive the *Reject probability* for each traffic class i of an eNodeB j as follows :

$$P_{i,j}(k) = \begin{cases} min(\frac{traffic_j}{\sum_{l=0}^{n} traffic_l}.R_i(k), 1) & \text{if} \quad R_i(k) \geq 0 \\ 0 & \text{else} \end{cases}$$

where $traffic_l$ is the amount of traffic coming from the eNodeB l.

The *Reject Probability* represents the proportion of traffic to reject to reach the reference value of the queue length. To ensure fairness between eNodeBs, this probability depends on the amount of traffic generated by an eNodeB, which participate to the congestion. Thus, the higher the traffic generated by an eNodeB is, the higher the *Reject Probability* (to be applied by this eNodeB to reduce the MTC traffic) will be.

The Admission Control is implemented at the eNodeBs level. At each reception of an attach request coming from an MTC device, the eNodeB applies Admission Control based on the *Reject Probability* coming from the MME. The incoming request of the class C_i is, thus, rejected with the probability $P_{i,j}$, by choosing randomly an uniform value between 0 and 1. If this value is greater than the $P_{i,j}$, then the MTC attach request is accepted. Otherwise, the request is rejected, and a back off time value is indicated to the MTC device in order to ensure a good distribution of future incoming attach requests over time.

IV. Performance Evaluation

A. Simulation settings

Parameter	Value	Details
nMME	1	The number of MMEs
nSGW	1	The number of S-GWs
neNB	10	The number of eNodeBs
nUEperCell	150	Number of MTC devices per cell
K_p	0.032	Proportional Gain
K_i	0.00051	Integrative Gain
K_d	0.000312	Derivative Gain
T_s	0.01 s	Resampling period
$Buffer\ size$	100	Buffer size (packets)
$Reference$	50	MME's Reference length (packets)
S	276	Service rate in the MME (requests/s)

TABLE I
Simulation Parameters

We implemented the CAAC solution in the ns3 simulator [2]. We simulated a system of 1 MME, 1 S-GW and n eNodeBs. The MTC traffic is modelled as a bursty traffic (see Fig 3), which represents the connection requests originating from the MTC devices. In fact, bursts are what characterize MTC applications since MTC devices are more likely to send data at synchronized periods due to time expiration or the occurrence of an event. Each cell contains the same number of MTC devices, which implies that each eNodeB will generate the same amount of traffic to the MME. In the following, we considered only one class of traffic. Other simulation parameters are shown in Table I.

B. Results

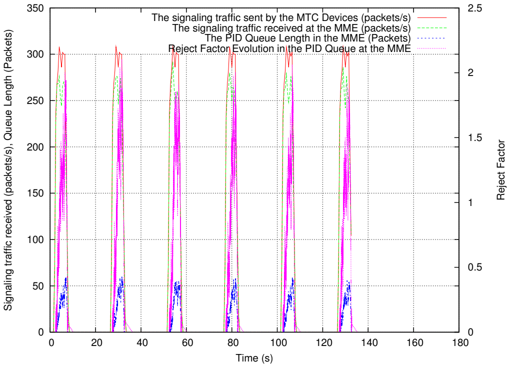

Fig. 3. The Reject Factor evolution depending on the received MTC signaling traffic and the queue length in the MME

The evolution of the *Reject Factor* regarding the received MTC signaling traffic and the MME queue length is given in Figure 3. We can, clearly, notice that the system reacts by increasing the *Reject Factor* value (i.e. the amount of MTC traffics to reject) each time the MME is overloaded by the MTC signaling traffic. Moreover, we note that the queue length is maintained around the reference value (50 attach requests) by the proposed control approach, which demonstrates the effectiveness of the proposed approach. Despite of receiving highly bursty traffic from MTC devices, neither overflow nor underutilization are observed. This clearly shows the efficiency of the proposed association, which considers a PID-based controller for measuring the congestion level and the eNodeB for rejecting MTC traffic.

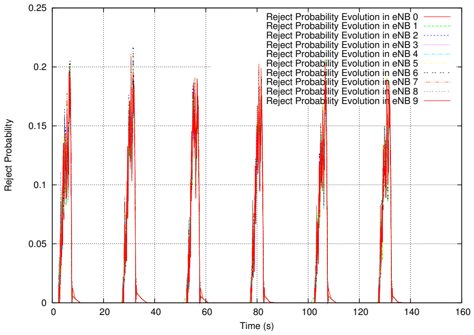

Fig. 4. The calculated Reject Probability in the eNodeBs

As the eNodeBs forward the same amount of uplink traffic to the MME, the *Reject Probability* value is practically the same for all the eNodeBs. Here, the *Reject Probability* is around 0.15, which represents $\frac{RF}{neNB}$ $(neNB = 10)$. Furthermore, we can, clearly, see that the Reject Probability value follows the evolution of the *Reject Factor*, which reflects the need to reduce the uplink traffic of the eNodeBs when the MME is overloaded.

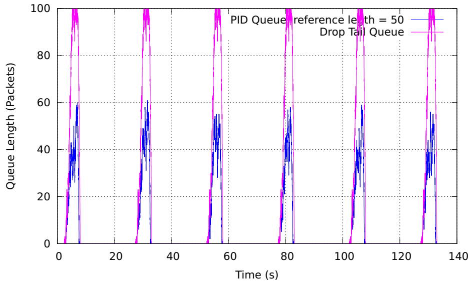

Fig. 5. The Queue Length in the MME

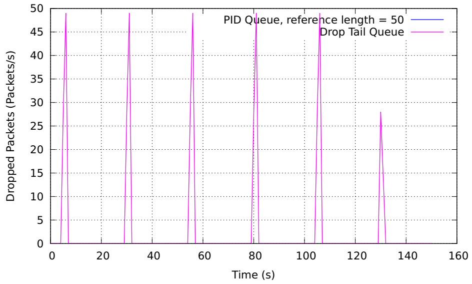

Fig. 6. Dropped packets in the MME

Now, we compare the CAAC solution to the case when no admission control is applied at the eNodeB and when a drop tail queue is used at the MME. Figures 5 and 6 show the MME queue length evolution and the dropped packets for the two cases. We note that when using both the PID-based queue in the MME and the Admission Control at the eNodeBs level, the queue length is maintained around the reference (50 packets or attach requests) and no packets are dropped (lost). Whereas, with a drop tail queue, bursts affect the queue length (i.e. overutilization) and packets are dropped (i.e. lost). At some points, 49 packets/s (15% of the the received traffic in the MME) are lost.

Despite the rejection of the MTC signaling traffic at the eNodeBs, the CAAC solution allows to transmit the same amount of MTC data as the case when no admission is applied. We argue this by the fact that CAAC solution reduces the signaling traffic at the Radio part while the case of no admission is used, the traffic is dropped at the MME. Thus, we succeed to reduce useless signalization between the eNodeBs and the MME , which allows optimizing network resources.

V. Conclusion

In this paper, we addressed the problem of congestion paused by cellular-based machine type communication. We introduced a novel solution, which reject MTC traffic at the radio part by using core network feedbacks regarding the network congestion state. The overloaded system was modeled and controlled with a PID controller that reflects the congestion level, and help to derive the amount of MTC traffic to be rejected by the eNodeBs. Simulation results show clearly that our solution avoids congestion and maintains the system at optimum level.

References

[1] 3GPP, "System Improvements for Machine Type Communications", TS 22.368 V10.1.0, Jun. 2010.
[2] ns3, http://www.nsnam.org.
[3] "Worldwide Cellular M2M Modules Forecast Market Brief", Beecham Research, Aug. 2010.
[4] S. Lucero, "Maximizing Mobile Operator Opportunities in M2M: The Benefits of an M2M-Optimized Network", ABI research, 1Q 2010.
[5] A. Malm and T. Ryberg, "Wireless M2M and Mobile Broadband Services", Berg Insight, Feb. 2007.
[6] 3GPP, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access", TS 23.401, Jun. 2010.
[7] 3GPP, "Architecture Enhancements for non-3GPP Accesses", TS 23.402, Jun. 2010.
[8] 3GPP, "Service Requirements for Machine-Type Communications", TS 22.368 V10.1.0, Jun. 2010.
[9] 3rd Generation Partnership Project, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3", 3GPP TS 24.301 V9.3.0, Jun. 2010
[10] T. Taleb and A. Kunz, "Machine Type Communications in 3GPP Networks: Potential, Challenges, and Solutions", to appear in IEEE Communications Magazine.
[11] T. Taleb and K. Samdanis, "Ensuring Service Resilience in the EPS: MME Failure Restoration Case", in Proc. IEEE Globecom 2011, Houston, USA, Dec. 2010.
[12] A. Visioli, "Practical PID control". Springer-Verlag Editor, ISBN 1-84628-585-2, 2006.