Call Admission Control of Machine-to-Machine Communications for satisfying Delay Constraint in LTE-Advanced

Kyungkoo Jun

Dept. of Embedded Systems Engineering,
University of Incheon,Incheon, Korea
kjun@incheon.ac.kr

Abstract

As concerns about energy efficiency and conservation grow, smart grid and intelligent transportation systems that consist of a large number of sensors, actuators and controllers have positioned as a crucial infrastructure. In such systems, machine-to-machine (M2M) communications between components make it possible the full automatic control of the systems. M2M communications is different from existing systems in the sense that it involves an enormous number of machine-type-communication (MTC) devices. Also traffic patterns and as well as QoS requirements vary widely depending on applications. LTE-Advanced of 3GPP provides a list of supportive functions to facilitate the M2M communications. To this end, a massive admission control scheme for the M2M communications was proposed. However, it is neither scalable nor adaptive particularly when the transmission interval of a device is relatively longer than its delay constraint. As a result, the call blocking probability under such scenario is much higher than normal cases. In this paper, we propose a method that is free from such limitation. Furthermore it can decrease the computational overhead under the condition that the transmission interval and the delay meet certain conditions. Through a set of simulations, we show the improvement in the call blocking probability when using the proposed method. We also provide the theoretical proofs that the proposed method can satisfy the delay constraint.

Keywords: *M2M communications, LTE-Advanced, call admission control, delay constraint*

1. Introduction

Green technologies such as smart grid systems [1] and intelligent transportation systems [2] draw interest as energy efficiency becomes of the first importance. Such systems are designed to collect data from a huge number of sensors and then determine actions after analyzing the data, which are then carried out by an enormous number of actuators. It is, however, almost impossible to manage the sensors and actuators manually because of their excessively large quantity. Hence, the provision of full automatic control mechanisms for such systems is essential. Also, M2M communications in the state of the art has expanded from a one-to-one connection to a complex and diverse system of networks that exchange data with personal appliances. The expansion of IP networks across the world has made it far easier for M2M communication to take place and has reduced the amount of power and time necessary for information to be communicated between machines. It is considered as one of the tipping point to ignite the proliferation of the adoption of the M2M communication based applications. To this end, machine-to-machine (M2M) communications is considered as one of the enabling technologies for automatic control.

Even though the M2M communications play a significant role in a large span of applications, the following common characteristics can be easily identified in most cases. The first notable characteristic is enormity in quantify. That is, the number of involved machine-type-communication (MTC) devices such as sensors and actuators is very large particularly compared with existing applications in which the involved components can be numbered less than dozens. Secondly, the data exchange pattern is diverse. Mostly, the communications occur between MTC devices and controllers. The traffic volume, burstness and interval between packet exchanges are different depending on the applications. For example, the intervals between data transmissions range from 10 ms to several minutes. Thirdly, QoS requirements such as maximum delay, jitter and throughput are also diverse depending on applications. According to the purpose of applications, the QoS parameters are determined and enforced. Considering the usefulness of M2M based applications, the variety in the QoS perspective can be easily beyond expectation.

To support M2M communications, existing infrastructure of cellular mobile networks need to be modified because of the following limitations. Since the cellular systems were developed mainly for voice communications, it is obvious that the M2M communications are not a good fit for it. There are dozens of issues to be addressed regarding the support of the M2M communications in the cellular systems. However, due to the limitation of the space, we discuss a few of them here. One of the issues is how to efficiently handle the massive number of the call admission control requests from a large number of MTC devices. Furthermore, such requests occur randomly and simultaneously in the time domain. Such massiveness overburdens the radio resource management (RRM) system of the cellular system. The other issue stems from the broad range of the QoS requirements. It particularly complicates the RRM such that the data to be kept which calls are in progress increases dramatically in quantity. Typically, the guarantee of time--related constraints is important in control systems because it determines the reliability and safety of whole systems. Such criticalness of the satisfaction of the QoS requirements induces lots of difficulties and complexities to the implementation of the RRM.

LTE-Advanced [3] of 3GPP proposes a list of required functions [4][5] that are designed to facilitate the M2M communications in cellular network environment. We discuss three of them which are closely related with the topic that this paper addresses. The first function is to be able to allocate a small number of physical resource blocks (PRB) to MTC devices. Without such function, the allocated blocks to MTC devices are too many to use all of them, leaving unused blocks wasted. It is because MTC devices transmit a tiny amount of data at a time. The second function provides time-controlled access. It forces MTC devices to access the radio resources only during the granted time interval (GTI). It facilitates to manage the radio access of MTC devices. Otherwise it would be complex and hard because it needs to manage the resource allocation each for a large number of MTC devices. The third functions is the group-based management. A set of MTC devices is grouped together according to categories. Then the radio access is controlled in the unit of group rather than individual MTC devices. The grouping function is essential for the scalability required to manage a large quantity of MTC devices efficiently.

With the help of these three functions, call admission control schemes [6, 7] for the scenario where a massive number of MTC devices involve were proposed. The schemes assume that MTC devices transmit data periodically and the QoS requirements are imposed with a constraint of maximum jitter. The schemes have drawn a sufficient condition that can determine whether a new request for the call admission be granted without harming the jitter constraints of existing ongoing calls. They have used a cluster-based approach; MTC devices are grouped into clusters according to QoS requirements. It also enforces that, once a new MTC device is admitted, it must belong to one of the clusters. It means that at least one cluster should exist or new cluster is created that has the same QoS requirements with the newly admitted one. Once the MTC devices are managed in the unit of clusters, the radio resource allocation is also performed in the level of clusters. For example, for the radio resource allocation, a base station which is eNodeB in LTE systems allocates GTIs to clusters. Then all the MTC devices belonging to a cluster should compete or share the time slots in GTIs to send or receive data from the base station.

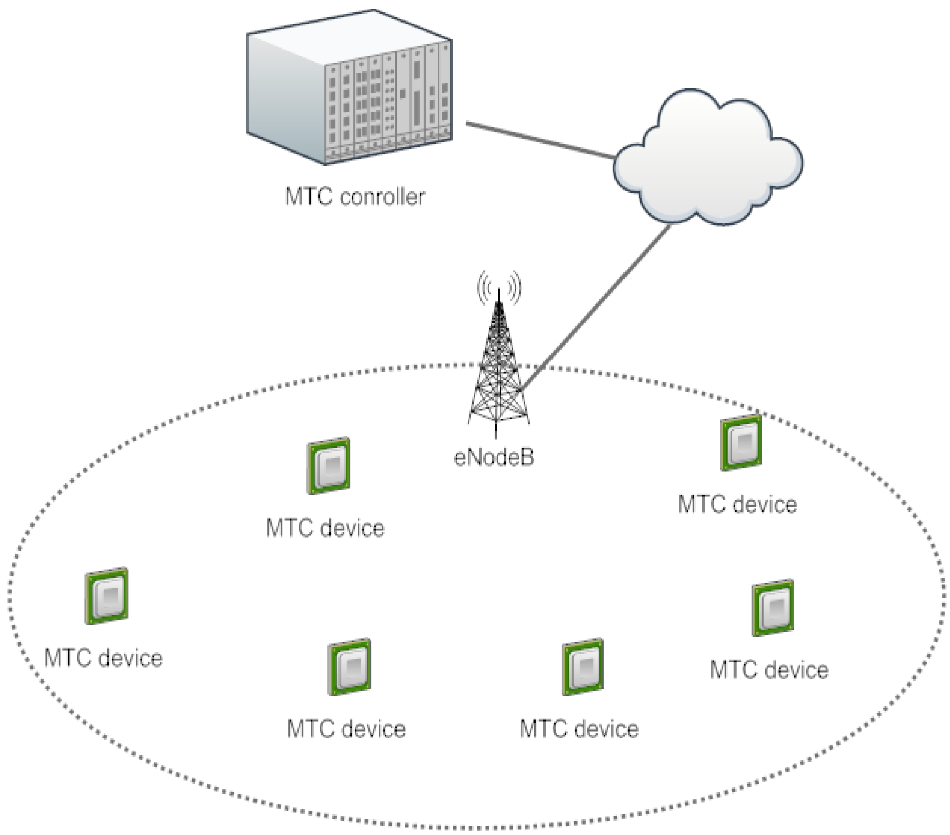

Figure 1. A Cellular Mobile System Model using M2M Communications as Infrastructure

However, the schemes of [6, 7] have the limitation that the clusters of which transmission interval is long but its delay constraint is relatively short have high call rejection probability . It means that the MTC devices that would belong to the clusters once they are admitted are more likely to be rejected than those of other clusters. It is because of the priority based algorithm of the schemes. The algorithm determines the order of the radio resource allocation according to the priority. But higher priority is given to the clusters with short transmission interval and long delay constraint. Such way of priority decision is advantageous in satisfying QoS requirements. However, it has the side effect to increase the call block probability.

This paper proposes a call admission method that employs a different sufficient condition than the aforementioned schemes. We prioritize the resource allocation according to delay expiration. It has the effect to lower the call block probability while meeting the required QoS parameters. We also propose an optimized sufficient condition that can reduce the computational overhead of the call admission procedure. It is only applicable for the situation where the interval and the delay meet certain conditions. Even though such limitation narrows the applicability of our proposed scheme, it is able to provide insight to understand the call admission problem in the M2M communications. This paper is organized as follows. Section 2 describes the proposed method in detail and proves its validity. Section 3 presents the simulation results and Section 4 concludes the paper.

2. Call Admission Control for the MTC Devices

We present a new novel call admission control scheme for the MTC devices in LTE advanced systems. As shown in Figure 1, we assume the case where all the MTC devices that request call admission belong to one eNodeB. The MTC devices generate data periodically and are required to deliver to MTC controllers located on external networks. Thus the traffic from the MTC devices should pass through eNodeB. It requires that the MTC devices ask eNodeB for the call admission to obtain radio resources to convey data to the MTC controllers. Once admitted, eNodeB is responsible for allocating resources to the MTC devices.

We also assume that delay constraints are imposed. The delay is the elapsed time from the moment of the data creation at an MTC device until it finishes its transmission to an eNodeB. It is not allowed to exceed a maximum. Therefore, we represent the characteristics of an MTC device regarding communications as (p, d) where p is the data creation interval and d is the maximum delay.

We manage the MTC devices by grouping them into clusters according to QoS requirements. The MTC devices with the same QoS parameters belong to a same cluster. With cluster $i = 1, \ldots, N$, the QoS requirements can be represented as (p_i, d_i) to mean the interval and the maximum delay. Also, n_i is the number of MTC devices that belong to cluster i.

The eNodeB allocates to the clusters a set of PRBs which is called GTI. One PRB has both time and frequency domains. The duration of the GTI is τ. Therefore, the maximum number of the MTC devices that can transmit during one GTI, L is

$$\underset{L}{argmax} \sum_{j=1}^{L} t_j \leq \tau \quad (1)$$

where t_j is the transmission time of an MTC device j belonging to a cluster with $1 \leq j \leq n_i$.

If the GTIs of different cluster i and j are overlapped each other and the priorities assigned to cluster i and j are pr_i and pr_j, then

$$pr_i < pr_j \quad if \quad d_i < d_j \quad (2)$$

where the lower pr_i the higher the associated priority is. The clusters k with lower pr_k than other clusters l are allocated the GTIs at subsequent available times after all of clusters l are assigned with resources. In the case that $d_i = d_j$, then the priorities are

$$pr_i < pr_j \quad if \quad p_i > p_j, \;\; d_i = d\_j \quad (3)$$

It is because such prioritization can avoid the situation in which the resource allocation to the longer-period clusters is repeatedly postponed by the shorter-period ones because a shorter period can occur multiple times within a longer period. Throughout the paper, it is assumed that the cluster number $i = 1, \ldots, N$ is used interchangeably to mean priority too; the lower the number, the higher the priority.

To inform the MTC devices belonging to a same cluster that GTI is allocated, physical downlink control channel (PDCCH) is used by eNodeB. The allocation information also carries the slot locations within GTI. Each MTC device is able to determine which slot it should use to transmit data. Besides GTI frames, Non-GTI time frames and the non-occupied slots in the GTIs are allocated to user equipments (UE) if there exists demand for it.

1. Receives a request for call admission from an MTC device with (p_i, d_i)
2. If there exists a cluster with (p_i, d_i)
   - 2.1 If there exists a slot?
     - 2.1.1 Accept the request and finish.
3. Else
   - 3.1 If (p_i, d_i) can satisfy required sufficient condition for the admission
     - 3.1.1 If the delay constraint of all the clusters can be satisfied
       - 3.1.1.1 create a new cluster with (p_i, d_i) and accpet the request
   - 3.2 Else
     - 3.2.1 Reject the request

Figure 2. The Proposed Algorithm for the Call Admission Control

When the MTC devices terminate the calls, eNodeB is informed from them. It then reclaims corresponding slots and reallocate them for UEs or other MTC devices. However, the case that such notification is not made due to some reasons such as abrupt breakdown of the MTC devices can be handled, too. For this, eNodeB needs to check the status of the slot occupancy by monitoring physical uplink control channel (PUCCH). The channel is used to carry the information about slot usage statistics, that is, which devices use which slots. If eNodeB finds that a slot is detected as being not used more than n during a specified threshold period φ without prior notification, the corresponding call is assumed as being terminated and the slot allocation is canceled. The parameters of n and φ should be determined carefully because they affect the utilization of resource. The GTI allocation to the clusters continues as long as there exist more than one working MTC device. However, once all the MTC devices in a cluster terminate calls, the cluster is removed from the list that eNodeB manages and the GTI is not allocated any more.

Figure 2 shows the proposed algorithm for the call admission control. An MTC device asks the call admission by sending a request of (p_i, d_i) to eNodeB through random access channel (RACH). If there already exists a cluster satisfying (p_i, d_i) and the associated GTI has at least one available slot, the request is accepted. Then the MTC device joins the cluster. Otherwise, it is required to create a new cluster to admit new call. However it should be checked in advance that the creation of the new cluster does not harm the delay satisfaction of the existing clusters. The check procedure is discussed in detail shortly. The new cluster is created only if the delays of existing clusters are not violated. If the created cluster happens to have the same (p_i, d_i) that at least one of existing clusters has, then the new cluster is assigned a lower priority than the existing ones.

We discuss the condition that the proposed algorithm uses. It is the sufficient condition that must be satisfied before creating a new cluster is as follows. It analyzes whether the delay constraint can be satisfied. The condition is checked against to all clusters which are active as well as including the one that is considered to be created. Let $w(i)$ the worst case time that cluster i takes to finish its transmission, then

$$w(i) = \tau \sum_{k=1}^{n} y_k \cdot \left\lceil \frac{p_i}{p_k} \right\rceil \qquad (4)$$

$$\begin{cases} y_k = 0 \ if \ k \le i \\ y_k = 1 \ if \ k > i \end{cases}$$

Eq. 4 shows that the transmission of cluster i is delayed until the clusters which have higher priorities than cluster i. Thus $w(i)$ becomes the sum of the transmission times of all the clusters $k = 1, \dots, i-1$ that have the higher priorities than cluster i plus its transmission time.

Regarding Eq. 4 when the transmission period of more than one clusters overlap, the clusters with lower priorities must postpone their transmission until the higher priority ones finish. Thus, the maximum transmission time of cluster i should include delay. Since the duration of GTI per cluster is τ, cluster i which is delayed by cluster k should wait until all the MTC devices of cluster k finish transmission.

Based on $w(i)$, the sufficient condition that guarantees the satisfaction of d_i of all the clusters is as follows.

$$w(i) \le d_i, \forall\ i \in \{1, \dots, N\} \qquad (5)$$

Then, the admission decision based on Eq. (5) is very straightforward and simple. Its computational complexity is $O(n^2)$ because the condition should be compared with all the clusters. However, in some case, the complexity can be reduced to $O(n \cdot logn)$ under certain conditions. For example, if the delay and the data generation interval is in a relationship of arithmetic multiplication, part of the condition can be eased without having to check its satisfaction. However, such requirement is hard to be enforced when considering the variety of M2M applications.

3. Performance Evaluation

This section evaluates the proposed method by the simulations. The simulations assume there is one eNodeB and its bandwidth is 20 MHz. One GTI consists of 100 PRBs and its duration τ is 1 ms. Within one GTI, maximum 20 MTC devices can transmit, allocating 5

PRBs for each device. It is also assumed that 5 PRBs are enough resources for one MTC device to transmit data.

The simulations create MTC devices one by one with the intervals that follow the exponential distribution of the average of 1 ms. Once created, the MTC devices choose randomly one of (p_i, d_i) shown in Table 1 as the QoS requirement. And the MTC devices that are admitted by the eNodeB have the call duration that follows the exponential distribution with the average of 15 ms. only uplink data transmission is assumed.

Once the calls are terminated, the MTC devices stop their operation, not requesting any more calls. All the MTC devices are the same distance away from the eNodeB and their locations are fixed. The simulations each of which lasts 1000 ms are repeated several times and the results are averaged.

Table 1. The Data Generation Interval and Delay Constraints of Clusters in the Simulations

Priority	Interval (ms)	Delay (ms)
1	500	2
2	450	4
3	400	6
4	350	8
5	300	10
6	250	12
7	200	14
8	150	16
9	100	18
10	50	20

From the simulation results, we can see that the proposed algorithm is adequate to provide performance as well as the satisfaction of the QoS requirements. Figure 3 shows the average throughput of the MTC devices per cluster. Figure 4 shows the average transmission delay of the MTC devices per cluster. We can see that the delays are less than the constraints presented in Table 1. These results confirm that the proposed algorithm is able to satisfy the QoS requirements. Figure 5 presents the call block probability of clusters. We compare the results with the case when no sufficient condition is enforced. As shown, the blocking probabilities with the sufficient condition are lower than without the condition in all of the clusters. It implies that the sufficient condition plays an additional role to improve blocking as well as the satisfaction of the QoS requirements. Figure 6 shows the average GTI utilization of the clusters. All the results are higher than 90% implying that the proposed algorithm does not harm the resource utilization for the QoS guarantee.

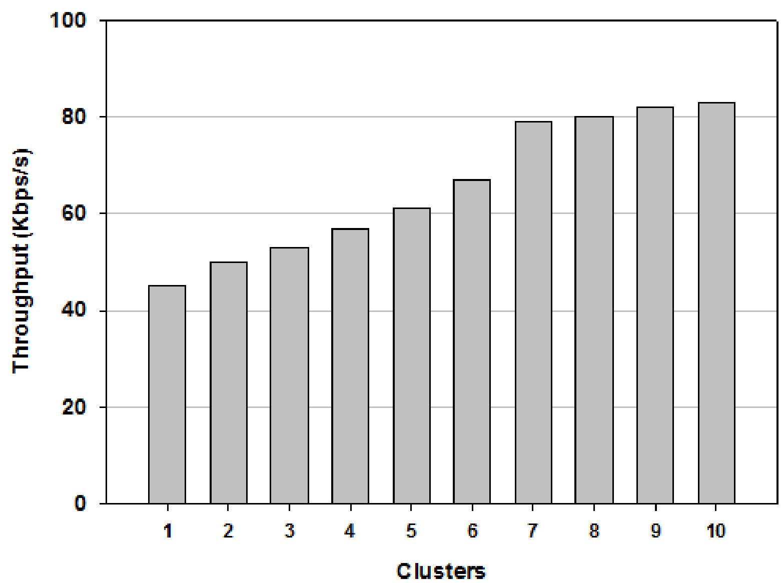

Figure 3. The Throughput of Clusters

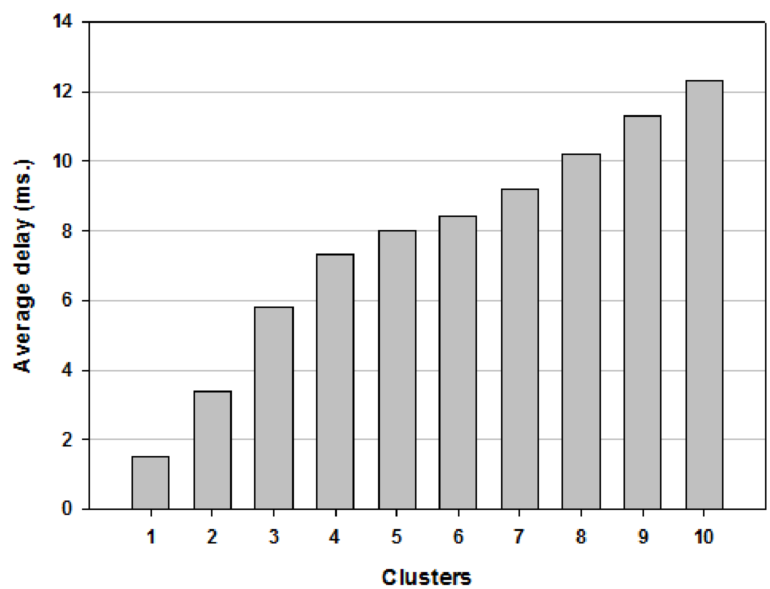

Figure 4. Average Delay of Clusters

4. Conclusions

This paper proposed a call admission control method for the M2M communications in LTE-Advanced systems. It improves the call admission probability compared with the conventional method by using the delay based priority assignment scheme. Furthermore it suggests that the computation overhead required for the call admission can be reduced if the transmission interval and the required delay meet the certain conditions. The proofs that the proposed method can satisfy the delay constraint were provided and the simulation results confirmed it, showing also the comparison of the blocking probability.

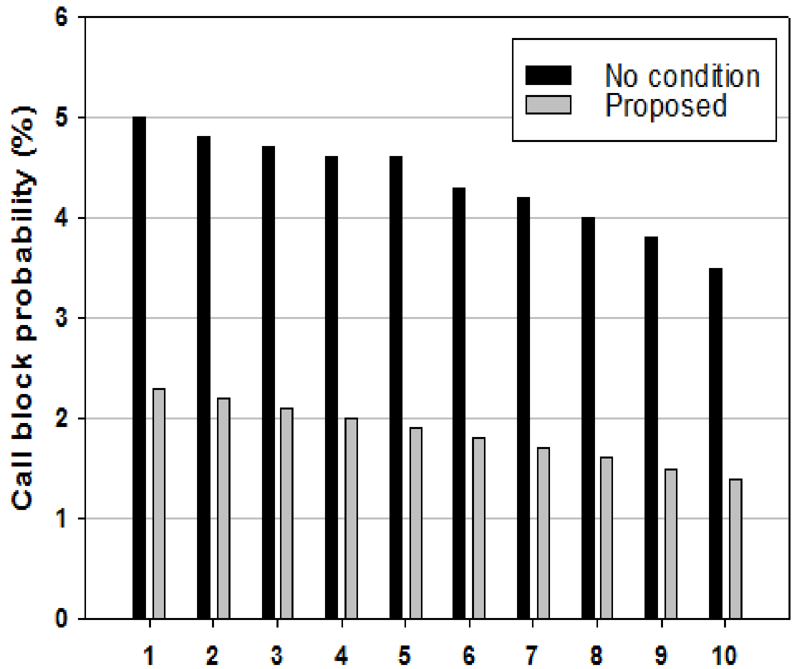

Figure 5. Call Blocking Probabilities of Clusters

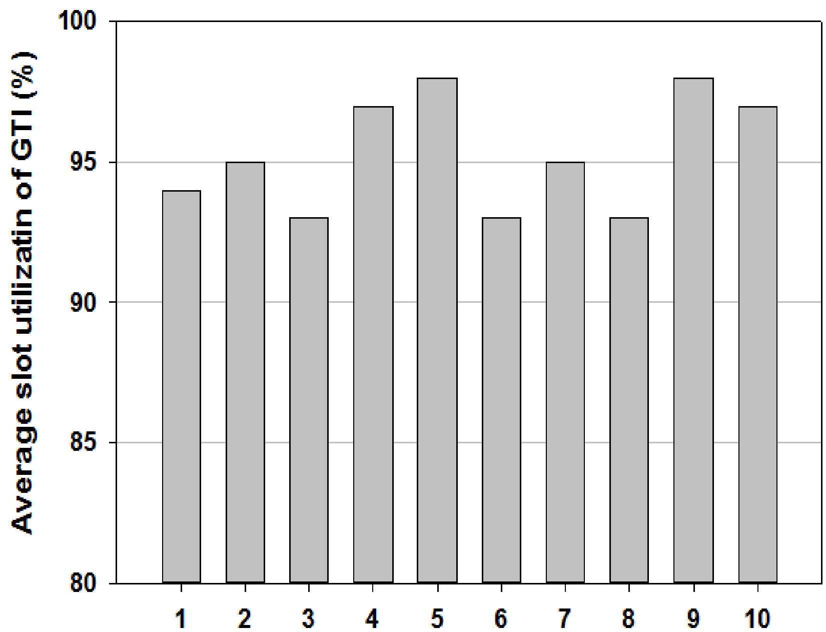

Figure 6. Average Slot Utilization of GTI of Clusters

References

[1] C. Gellings, "The Smart Grid: Enabling Energy Efficiency and Demand Response", CRC Press, **(2009)**.
[2] M. Chowdhury and A. Sadek, "Fundamentals of Intelligent Transportation Systems Planning", Artech House, **(2003)**.
[3] http://www.3gpp.org.
[4] 3GPP TS 22.368 V11.1.0, "Service Requirements for Machine Type Communications", **(2011)** March.
[5] 3GPP TS 23.888 V1.1.1, "System Improvements for Machine Type Communications", **(2011)** April.
[6] S. Lien and K. Chen, "Massive Access Management for QoS Guarantees in 3GPP Machine-t—Machine Communications", IEEE Comm. Letters, vol. 15, no. 3, **(2011)** March.
[7] S. Lien, K. Chen and Y. Lin, "Toward Ubiquitous Massive Accesses in 3GPP Machine to Machine Communications", IEEE Comm. Mag., **(2011)** April.

Authors

Kyungkoo Jun is currently an associate professor of Embedded Systems Engineering, Incheon National University. His research interests are cellular mobile network systems and artificial intelligence and embedded systems.