Temporal Load Balancing of Time-driven Machine Type Communications in Mobile Core Networks

Go Hasegawa
Osaka University
Osaka 560–0043, JAPAN
Email: hasegawa@cmc.osaka-u.ac.jp

Takanori Iwai
NEC Corporation
Kanagawa 211–8666, JAPAN
Email: t-iwai@hx.jp.nec.com

Naoki Wakamiya
Osaka University
Osaka 565–0871, JAPAN
Email: wakamiya@ist.osaka-u.ac.jp

***Abstract*—Machine Type Communications (MTC) has been paid much attention as a new communication paradigm to increase mobile network traffic. Most of MTC terminals are time-driven, that is, they send and receive data periodically. Therefore, network access requests on mobile core networks are concentrated at a specific timing, which results in instantaneous increase in network load. Considering the fact that such time-driven MTC would accept a certain amount of latency in their cyclic communication, in this paper, we propose a scheduling method of communication timings of time-driven MTC terminals to mitigate traffic concentration. We extend the standardized back-off mechanism of 3GPP to configure the back-off time length for each terminal to decrease the number of concurrent bearers in the network, while satisfying requirements on communication latency. We compare proposed methods by simulation experiments and reveal that we can achieve almost zero access rejections at reasonable communication quality by a simple timeslot selection algorithm when the core network maintain the timeslot assignment status for accommodated User Equipments. To the best of our knowledge, this is the first proposal to alleviate short-term congestion of mobile core networks by MTC with TDMA-like network control.**

I. Introduction

The recent growth of mobile networks and proliferation of mobile devices such as smart-phones and tablets result in rapid increase in the amount of mobile network traffic, while the penetration rate of cellular phones into developed countries reaches nearly 100% [1]. In [2], it is estimated that the amount of mobile network traffic will increase 12-fold from 2012 to 2018 [2]. Furthermore, Machine Type Communication (MTC) for Machine-to-Machine (M2M) applications [3] has attracted considerable attention as a new communication paradigm which further increases mobile network traffic.

One of characteristics of MTC is that the number of terminals is much larger than the number of traditional mobile phones. In fact, according to [4], the number of MTC devices will soon reach 10 times the number of mobile phones. In addition, patterns of M2M communication are rather different from mobile phones. In general the frequency and the amount of traffic between MTC terminals are lower [5]. However, mainly due to its wide coverage, mobile network operators are requested to accommodate such MTC terminals for M2M applications in their networks. In the 3rd Generation Partnership Project (3GPP), various types of MTC services are considered [6] and required network functions for such services are now under considerations [7]. However, due to different characteristics of MTC terminals described above, such MTC traffic may burden the control plane of the core networks.

In existing mobile cellular networks such as 3G and LTE systems, the core network establishes tunnels (identical to bearers in 3GPP terminology) together with related information for each User Equipment (UE), and maintains them permanently until a UE is disconnected from the network. This is required to provide full IP-reachability as well as small paging delay. Therefore, when MTC terminals are accommodated into the mobile network, the amount of resources required for maintaining UEs increases drastically beyond the capacity. In existing works various approaches ave been considered to alleviate congestion caused by M2M traffic. However, these researches mainly focus on decreasing the long-term load on the control plane of the mobile core networks.

On the other hand, it is considered that the Average Revenue Per User (ARPU) of MTC terminals would be substantially smaller compared with traditional mobile phone terminals [8]. Consequently, we expect that we cannot recover the cost for accommodating MTC terminals to existing mobile cellular networks with the current system and cost structure.

In this paper we focus on a different problem in accommodating M2M traffic into mobile networks. It is the concentration of access timing to the mobile network by *time-driven* MTC terminals. Since in time-driven M2M applications massive terminals such as sensors and smart meters access the mobile network at regular intervals, the core network suffers from instantaneous high load while the long-term average is small enough. It is apparently inefficient and costly to prepare resources for the peak traffic especially when we consider small ARPU of M2M applications. We can find some existing rmesearches on accommodating MTC traffic to mobile core networks, such as protocol simplification [9], call admission control [10, 11], load balancing at core nodes [12, 13], context-aware approach [14, 15], and grouped communication [16, 17] However, most of These researches mainly focus on decreasing the long-term load on the control plane of the mobile core networks and there is no proposal for short-term network congestion by M2M communications.

To tackle the above-mentioned problem, in this paper, we introduce a novel approach of temporal load balancing of time-driven M2M traffic. To mitigate the concentration of access from many MTC terminals, the core network schedules and controls their access timing. To minimize modification to the 3GPP standards while most of existing work described above

involve substantial modifications, we extend the standardized back-off mechanism to configure the back-off time length for each terminal. Our main objective is to decrease the number of concurrent tunnels while satisfying requirements on communication latency of MTC terminals. We present the overall architecture four alternatives in access timing selection by UE and access rejection policy by the core network. We conduct simple numerical evaluation to confirm the effectiveness of the proposed method and obtain some insights on mobile network control for M2M applications. To the best of our knowledge, this is the first proposal to alleviate short-term congestion of mobile core networks by M2M communication with temporal access control. As far as we know, our method is a novel approach for temporal load balancing for time-driven M2M communication.

The rest of this paper is organized as follows. Section II describes the current mechanism for temporal load balancing and back-off mechanisms in 3GPP and their problems for accommodating M2M communications. Section III proposes a control mechanism of access timings of MTC terminals. We evaluate the proposed mechanism by numerical evaluations in Section IV. Finally we conclude this paper and discuss future work in Section V.

II. Temporal load balancing for M2M terminals

A. Back-off mechanism in 3GPP

When a large number of UEs in the mobile cellular network tries to make communications via base stations, called as evolved Node Bs (eNodeBs) in LTE networks, simultaneously, the mobile core network that aggregates the eNodeBs suffer from high load to manage the communications. To resolve such congestion of the core network, the back-off mechanism is standardized in 3GPP as a control method of communication timings of UEs by the core network [18, 19]. The back-off mechanism sends a message to a UE to notify a time length for which the UE stops to access the core network. The UE waits for the informed time length when receiving the back-off message. There are two kinds of back-off mechanisms in 3GPP standard, which are Radio Resource Control (RRC)-level and Non-Access-Stratum (NAS)-level.

1) RRC connection reject: When the core network is congested, the Mobility Management Entity (MME) in the core network can send an OVERLOAD START message to eNodeBs to resolve the congestion. The eNodeB receiving the message denies the connection request from UEs by sending RRC Connection Reject message to incoming UEs with extended wait time. The extended wait time is configured randomly between 1 to 1,800 seconds [20]. This back-off mechanism is mainly used for alleviating the congestion of the radio network between UEs and eNodeBs.

2) NAS-level congestion control: When a UE sends a Service Request message to the core network after establishing the RRC connection with NAS protocol [21], MME can send a Service Reject message to the UE with back-off timer value. In 3GPP, the back-off timer value is chosen randomly from a default value range of 15 and 30 minutes [22]. This back-off mechanism is mainly utilized to alleviate the overload of MME and to resolve the congestion at Serving Gateway (S-GW) and Packet Data Gateway (P-GW), called as Mobility Management back-off and Session Management back-off, respectively.

B. Temporal load balancing of M2M communication

Some MTC terminals such as smart meters and environmental sensors are *time-driven*, that is, they send their data to the server at regular intervals [23]. When the number of such MTC terminals accommodated into the mobile cellular networks increases, some of communication requests are synchronized at the same time, that increases the load of the core network temporarily, while the long-term average load remain bearable. This may degrade the stability of the core network and increase the network cost when accommodating such temporal access concentration.

On the other hand, such M2M communication applications can allow the delay in their communication to some extent, while the network access by mobile phones requires quite a short latency in paging and making outgoing calls. This gives us a chance to consider a method to control the communication timings of UEs according to the application requirements in terms of communication latency, towards the temporal load balancing of the core networks. In 3GPP, such function is called as "Time controlled feature," [6], which is selected as one of thirteen important issues for realizing M2M communications in mobile cellular networks [7]. However, in the current standardization phase, the discussion has just started and there is no detailed mechanism to realize such functions.

C. Problems in existing mechanisms for temporal load balancing

When we apply the existing back-off mechanisms in 3GPP to realize the temporal load balancing of M2M communication, we will face the following two major problems.

1) Inflexibility in generating back-off messages: Since the main objective of the existing back-off mechanisms is to resolve the congestion in the core network, the back-off messages can be generated when the core network is congested or the core nodes are overloaded. Furthermore, in 3GPP standard, eNodeBs or MMEs can inform the back-off timer length to UEs only when the access request from the UEs are rejected. This means that we must reject the access request from UEs when we want to set the back-off timer for temporal load balancing. This significantly degrades the communication quality of UEs, as well as increasing the core network load for processing the retransmissions for rejected access requests.

2) Incapability of handling heterogeneous quality requirements: The existing back-off mechanisms have limited ranges for timer length as explained in Subsection II-A. In addition, we can not intentionally set different back-off timer lengths to UEs that have different requirements for communication latency. Therefore, we can not realize adaptive control of access timing from UEs. Furthermore, we can not handle the heterogeneous requirements for M2M application quality with the existing back-off mechanisms. For example, we cannot maintain the quality of M2M applications which require a small communication latency.

III. Proposed method

In this section, we propose a novel back-off mechanism to achieve temporal load balancing of M2M communications. We focus on M2M applications in which MTC terminals send data to servers on the external network (e.g. the Internet) at regular intervals, and assume that the mobile core network know the detailed communication requirements of MTC terminals.

A. Design goals

1) Decreasing the number of concurrent tunnels in the core networks: As explained in Section I, a tunnel is established and maintained persistently for each UE in the core network. After attaching to the core network, a UE keeps its status as CONNECTED or IDLE according to the UE's communication. The core network maintain the status of each UE as well as the tunnel itself. It means that the number of accommodated UEs in the core network is limited according to the amount of network and core node resources. Therefore, decreasing the number of concurrent tunnels by temporal load balancing has positive effect on the management of mobile core networks. In what follows, we consider the situation where the number of concurrent tunnels are strictly limited and the proposed mechanism controls the communication timings of MTC terminals to keep the number of concurrent tunnels being under the limitation.

2) Satisfying the allowable latency of MTC terminals: We define the *allowable latency* of a MTC terminal as the upper limit of the time length from when the terminal generates data to be sent to the server to when it successfully access the core network. We assume the allowable latency would be determined by M2M service providers such as electric power companies for smart meters. The proposed mechanism determines the communication timing of each MTC terminal considering its allowable latency.

3) Small modifications to 3GPP standard: In order to minimize the effect on the existing mobile core networks, we extend the existing back-off mechanisms with small modifications to realize the proposed mechanism. In detail, we add functions to existing nodes in the mobile core network.

B. Concept of proposed method

To achieve the above-mentioned design goals, we introduce a TDMA-like temporal access control from MTC terminals. In detail, to avoid the simultaneous access from many MTC terminals, the core network schedules their access timing. We first model the requirement of M2M communications in terms of access timing as communication policy. We then construct the mechanism for informing the MTC terminals of their appropriate access timing to avoid the core network comgestion while preserving the communication policy of MTC terminals. For this purpose, we extend the format of back-off timer information and propose the method to caclulate and notify the appropirate access timing for MTC terminals.

C. Communication model of M2M terminals

In the existing mobile core network, MME and HSS maintains UE's information such as phone number, Access Point Name (APN), and QoS level. In the proposed method, we model the communication of MTC terminals with the following three parameters.

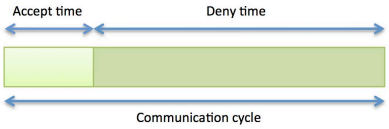


Fig. 1. Extended back-off timer information

- Communication cycle
  This means the length of the cycle at which a MTC terminal send data to the server.
- Communication length
  This is the required time for each communication, which is the duration from when the UE begins to access the core network to when the UE finish sending its data and close the connection. In general, this is determined by data size, network bandwidth, and the time for UE to attach to and detach from the core network.
- Allowable latency
  As explained in Subsection III-A, this represents the maximum latency from when the terminal generates data to be sent to the server to when it successfully access the core network.

We call the combination of the above three values as *communication policy* of MTC terminals to be maintained in the mobile core network.. We assume that the communication policy is determined by M2M service providers. For example, consider the case where each sensor terminal sends the data in every 60 minutes, it takes one minute to send the data to the server, and the server needs to collect all data from the sensor terminals within ten minutes. In this case, we use 60, 1, and 10 minutes for the communication cycle, the communication length, and the allowable latency, respectively. Whereas the different communication policy can be set to each MTC terminal, we assume that the M2M service providers set the same communication policy to a significant number of MTC terminals.

D. Functions

1) Extended back-off timer information: In the proposed mechanism, a MME sends the *extended back-off timer information* to a UE, and the UE determines the communication timing according to the information. We define the extended back-off timer information as a combination of communication cycle, accept time, and deny time. Figure 1 depicts the relationships among these values. The proposed method determine the accept time and deny time to control the access timing of the MTC terminal. In detail, in each communication cycle the MTC terminal access the core network to send the data only in the accept time. On the other hand, in deny time, the MTC terminal is prohibited to access the core network. In general case, the length of accept time is identical to the communication length defined above.

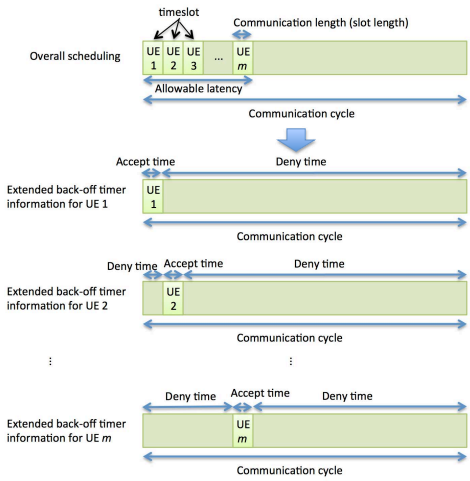


Fig. 2. Communication schedule for all UEs and extended back-off timer information for each UE

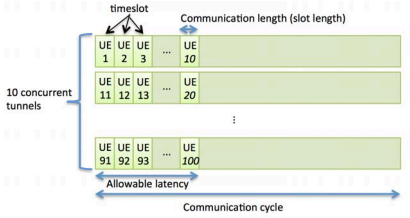


Fig. 3. Communication schedule for multiple concurrent tunnels

2) Calculating extended back-off timer information: Here we assume that all UEs have identical communication policy. Note that we can easily extend the proposed method to allow heterogeneous communication policies.

A communication cycle of the UEs, is divided into *timeslots*, whose length is identical to the communication length. The proposed mechanism assigns a timeslot to each UE within the allowable latency, as depicted at the topmost picture in Figure 2, meaning that it determines the *communication schedule* of all UEs, while satisfying the allowable latency. The remaining pictures in Figure 2 explain the extended back-off timer information for UEs in the the communication schedule. For the example in SubsectionIII-C, the communication cycle is 60 minutes, which is divided into sixty timeslots with one minute. Since the allowable latency is ten minutes, the proposed method assigns ten timeslots from the beginning of the communication cycle to UEs. By configuring the values of communicaion cycle and allowable latency, we can accommodate various type of M2M communications with the proposed method.

Figure 2 shows the simplest case where we allow only one concurrent tunnel in the network. In Figure 3, we show an example in which we allow ten concurrent tunnels and assign timeslots for 100 UEs. In the figure the same timeslot is assigned to ten UEs (UE 1, UE 11, UE 21, ..., UE 91). Therefore, the core network send the identical extended back-off timer information for these UEs.

Note that the detailed algorithm for assigning timeslots to UEs is out of scope of this paper. In Section IV, we introduce some simple algorithm for timeslot assignment to evaluate the fundamental performance of the proposed mechanism.

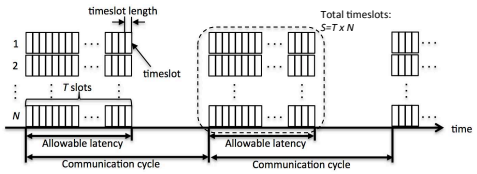


Fig. 4. Network model

3) Notification of extended back-off timer information: In the proposed method, the core network sends the extended back-off timer information for a UE when the UE tries to access the core network, regardless the access is accepted or rejected. In detail, when the access from UE is accepted, the core network determines the extended back-off timer information for the next communication cycle and notify it to the UE.

On the other hand, when the access is rejected, we have two options. The one is that the core network determines the communication timing for the UE in the current communication cycle so that the UE can retry the access in the same communication cycle. This can be implemented by adding a field for such information in RRC OVERLOAD START messages or NAS-level Service Reject messages. The other is that the communication timing for the UE in the next communication cycle so that the UE gives up sending data at the current communication cycle and wait for the next communication cycle. This can be implemented easier compared with the former method since it does not need the modification of signalling messages in 3GPP.

E. Accommodating heterogeneous MTC devices

In the explanation of the proposed method above assumes the identical communication policies. However, we can easily accommodate the heterogeneous MTC devices which have various communication policies. The MTC devices which has various communication length can be accommodated by assigning multiple successive time slots in each communication cycle. For different communication cycle, we make groups of MTC devices based on the communication cycle values, and assign tunnel(s) for each group.

IV. Performance Evaluation

In this Section we first introduce the models and assumptions for performance evaluation, including the detailed algorithm for timeslot selection for UEs and access rejection policy, that are out of scope of the proposed mechanism in Section III. We then show some simulation results to confirm the fundamental performance of the proposed mechanism and present some insights for mobile network control for M2M application traffic.

A. UE and core network models

We assume that the communication policy of all UEs are identical, meaning that all UEs have the same values for

communication cycle, communication length, and allowable latency. The number of UEs is denoted by U. We divide the time from the beginning of the communication cycle to the end of the allowable latency into T timeslots. Figure 4 depicts the core network resource model with T timeslots for the allowable latency and N maximum concurrent tunnels. The total number of timeslots for each communication cycle, denoted by S, becomes $T \times N$. The proposed mechanism assigns S timeslots for U UEs.

When the number of UEs that access the core network at a certain timeslot becomes larger than N, the access of excess UEs are rejected due to *core network congestion*. We also introduce the wireless access rejection probability, p, that represents the probability at which a UE's access to the wireless network is rejected due to *wireless network congestion*. Therefore, a UE fails to send data at a certain timeslot in two cases. The first case is that the UE tries to access the wireless network but the access is rejected due to wireless network congestion. The other case is that after the successful access to the wireless network without wireless network congestion, the access to the core network is rejected due to core network congestion.

In the following performance evaluation, when the access fails by either of the above two reasons, the UE retries the access at the next timeslot. Otherwise, the UE successfully accesses the core network and sends data to the server and the core network sends the extended back-off timer information to the UE and the UE waits for the next communication cycle according to the extended back-off timer information. Furthermore, when a UE fail to access the core network within the allowable latency, the UE gives up sending data at the current communication cycle and wait for the next communication cycle.

B. Timeslot selection of UEs without extended back-off timer information

When a UE does not have the extended back-off timer information from the core network, for example the first access to the core network, the UE determines the access timing to the core network by itself. This means the timeslot selection by the UE within its allowable latency. The algorithm in the timeslot selection affects the communication quality of UEs and the core network performance significantly.

For example, when the access timing of UEs is distributed equally within the allowable latency, the probability of occurring core network congestion becomes small since the average number of UEs that access at the same timeslot is minimized. However, the UEs that access the core network at around the end of the allowable latency may fail sending data within the allowable latency due to wireless network congestion or core network congestion. On the other hand, when all UEs select timeslots from the beginning of the communication cycle, the number of UEs sending data within the allowable latency increases. However, since the core network congestion may happen frequently at the beginning of the communication cycle, the core network load increases by the processing overhead for access rejection procedure.

Therefore, we introduce the following two algorithms for UE's timeslot selection.

(1) Equal timeslot selection: A UE without extended back-off timer information selects the timeslot equally within the allowable latency.

(2) Greedy timeslot selection: A UE without extended back-off timer information selects the timeslot from the beginning of the communication cycle.

C. Access rejection policy

The fact that a UE obtains the extended back-off timer information from that core network means that the core network assigns a certain timeslot for the UE. However, even in this case, the UE may fail to access the core network due to core network congestion since new UEs without extended back-off timer information may access the core network at the same timeslot. If the core network maintains the timeslot assignment status for all UEs, such rejection can be avoided by prioritizing the UEs' accesses with extended back-off timer information. However, handling the timeslot assignment status may increase the load of core network nodes.

Therefore, we introduce the following two algorithms for access rejection policy.

(a) Prioritized rejection: When the number of UEs access the core network at a certain timeslot exceeds the limitation of the concurrent tunnels N, the core network selects UEs without extended back-off timer information for access rejection. This algorithm needs the maintenance of the timeslot assignment status by the core network.

(b) Non-prioritized rejection: When the number of UEs access the core network at a certain timeslot exceeds the limitation of the concurrent tunnels N, the core network selects UEs at equal probability for access rejection regardless of the extended back-off timer information. When a UE's access is accepted at the certain timeslot, the core network generates the extended back-off timer information for the UE so that they use the same timeslot in the next communication cycle. This algorithm does not need the maintenance of the timeslot assignment status by the core network. On the other hand, when a UE's access with the extended back-off timer information is rejected, the UE discard the timer information.

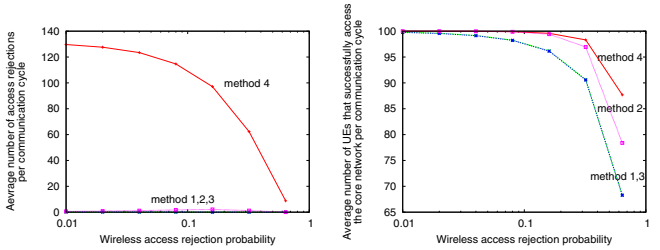
D. Simulation settings and evaluation metrics

We conduct the simulation experiments of the proposed mechanism in Section III with detailed algorithms in Subsections IV-B and IV-C. We assume that the number of UEs accommodated in the network, U is 100, and they have identical communication policy. The communication length is identical to the length of a single timeslot, and communication cycle and allowable latency equal to 1,000 timeslot and 5 timeslots, respectively. Therefore, when the upper limit of the concurrent tunnels, N, is equal to or larger than 20, we can assign timeslots within the allowable latency for all UEs in ideal case. Note that we have confirmed that the similar performance evaluation results are obtained with larger networks with several million users per one MME, fitting to the actual network environment [24].

We have the following two evaluation metric. The one is the average number of access rejections per communication cycle due to core network congestion, which is called as *access*

TABLE I. FOUR METHODS FOR PERFORMANCE COMPARISON

	Equal timeslot selection	Greedy timeslot selection
Prioritized rejection	method 1	method 2
Non-prioritized rejection	method 3	method 4



(a) Access rejections by core network congestion (b) Successful UEs within communication latency

Fig. 5. Effect of wireless network quality

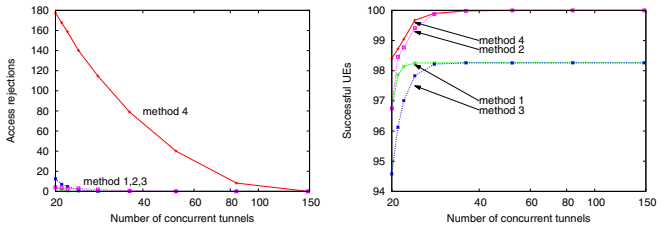
rejections in what follows. This metric is closely related to the load of the core network for access rejection procedure. The other is the average number of UEs per communication cycle that successfully access the core network and send the data to the server within the allowable latency. We denote this by *successful UEs* hereafter. This value directly affects the communication quality of M2M applications.

We combine the timeslot selection algorithm in Subsection IV-B and rejection algorithm in Subsection IV-C to make four candidates presented in Table I for comparative performance evaluation. For each parameter setting we conduct 1,000 times simulation experiments, each of which has 1,000 communication cycles, and take average values of evaluation metrics.

E. Evaluation results and discussions

Figure 5 presents the simulation results when we set N to 28. We observe the access rejections and successful UEs in Figures 5(a) and 5(b), respectively, as a function of the wireless access rejection probability (p). From these figures we can observe that method 4, which utilizes greedy timeslot selection and non-prioritized rejection algorithms, has the largest access rejections while the largest number of successful UEs. This means that method 4 can provide the best quality of UEs communication, that affects the quality of M2M applications, at the sacrifice of core network load for access rejection procedures. On the other hand, method 2 can obtain reasonable performance in terms of access rejections and successful UEs. Considering the difference between methods 2 and 4, the reason of this is that since method 2 preserves the timeslot assignment for UEs by the prioritized rejection algorithm, the access rejection occurs at the beginning several communication cycles.

In Figure 5(a), we can see that the number of access rejections in method 4 decreases when the wireless access rejection probability becomes large. This is because most UEs cannot send the access request to the core network due to frequent wireless access rejections. This is also the reason for Figure 5(b) where the number of successful UEs decreases when the wireless access rejection probability becomes large. Furthermore, method 1 and method 3 degrade the number of successful UEs significantly as compared with method 2 and 4. This is because UEs that select timeslots around the end of the allowable latency frequently fail to access the core network within the allowable latency.



(a) Access rejections by core network congestion (b) Successful UEs within communication latency

Fig. 6. Effect of upper limit of concurrent tunnels

Figure 6 present the results when p is set to 0.08 and vary N from 20 to 150 to observe the effect of the upper limit of the concurrent tunnels in the core network. From Figure 6(a), we can observe that for methods 1, 2, and 3, we can decrease the access rejections significantly by relaxing the upper limit slightly. This is due to the effect of prioritized rejections for method 1 and 2. For method 3, when N increases the average number of UEs that access at the same timeslot decreases rapidly, that decreases the access rejections effectively.

From the results in Figure 6(b), we can also confirm that in method 2 and method 4 almost all UEs can send data to the server within the allowable latency when N is larger than around 30. This is because UEs greedily selects timeslots from the beginning of the communication cycle. On the other hand, in method 1 and method 3, we can not completely alleviate the UEs that cannot send data within the allowable latency even when N is enough large. This is due to the effect of wireless access rejections that occur regardless of the core network condition.

V. CONCLUSION

In this paper, we proposed a temporal load balancing mechanism in mobile core networks to decrease the access concentration by MTC terminals communicating at regular intervals. Although our proposal in the present paper is at the conceptual level, simulation results gave us the insight that when we can take the cost for maintaining the timeslot assignment status for accommodated UEs, we can achieve almost zero access rejections and reasonable UE's communication quality by a simple timeslot selection algorithm. If we should avoid managing the timeslot assignment status, the access rejections and UE's communication quality have the trade-off relationships. We also discussed implementation issues of the proposed method.

For future work, we plan to evaluate the performance of the proposed mechanism when UEs have heterogeneous communication policies. We also should compare the system capacity, in terms of the number of accommodated UEs with reasonable delay. Furthermore, we need to consider detailed protocol design and implementation issues, including the odifications required for association of a tunnel to multiple UEs for mobility support, paging, and similar functions at gateway nodes, i.e. S-GW and P-GW.

REFERENCES

[1] The World Bank Group, "Mobile-cellular subscriptions," Mar. 2014. [Online]. Available: http://data.worldbank.org/indicator/IT.CEL.SETS.P2

[2] Ericsson, "Ericsson mobility report: On the pulse of the networked society," Feb. 2014. [Online]. Available: http://www.ericsson.com/res/docs/2014/ericsson-mobility-report-february-2014-interim.pdf

[3] G. Lawton, "Machine-to-Machine technology gears up for growth," *Computer*, vol. 37, no. 9, Sep. 2004.

[4] W. Sun and M. Song, "A general M2M device model," in *Proceedings of SWS 2010*, Aug. 2010, pp. 578–581.

[5] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A first look at cellular machine-to-machine traffic - large scale measurement and characterization," in *Proceedings of ACM SIGMETRICS 2012*, Jun. 2012.

[6] 3GPP, "Service requirements for Machine-Type Communications (MTC); Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.368, Dec. 2013. [Online]. Available: http://www.3gpp.org/DynaReport/22368.htm

[7] ——, "System improvements for Machine-Type Communications (MTC)," 3rd Generation Partnership Project (3GPP), TS 23.888, Sep. 2012. [Online]. Available: http://www.3gpp.org/DynaReport/23888.htm

[8] A. Daj, C. Samoila, and D. Ursutiu, "Digital marketing and regulatory challenges of Machine-to-Machine (M2M) communications," in *Proceedings of REV 2012*, Jul. 2012, pp. 1–5.

[9] Y. Chen and W. Wang, "Machine-to-machine communication in LTE-A," in *Proceedings of VTC2010-Fall*, Sep. 2010, pp. 1–4.

[10] K. Jun, "Enabling massive machine-to-machine communications in LTE-Advanced," in *Proceedings of GPC 2013*, May 2013, pp. 563–569.

[11] A. Amokrane, A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Congestion control for machine type communications," in *Proceedings of ICC 2012*, Jun. 2012.

[12] D. Bouallouche, "Congestion control in the context of machine type communications in 3GPP LTE networks," *Master thesis internship report, University of Rennes*, Aug. 2012.

[13] R. Vaidya, C. Yadav, J. Kunkumath, and P. Yadati, "Network congestion control: Mechanisms for congestion avoidance and recovery," in *Proceedings of ACWR 2011*, Dec. 2011.

[14] P. Makris, D. N. Skoutas, N. Nomikos, D. Vouyioukas, and C. Skianis, "A context-aware backhaul management solution for combined H2H and M2M traffic," in *Proceedings of CITS 2013*, May 2013.

[15] X. Jian, Y. Jia, X. Zeng, and J. Yang, "A novel class-dependent back-off scheme for machine type communication in LTE systems," in *Proceedings of WOCC 2013*, May 2013.

[16] G. Farhadi and A. Ito, "Group-based signaling and access control for cellular machine-to-machine communication," in *Proceedings of VTC Fall 2013*, Sep. 2013.

[17] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communication Magazine*, vol. 49, pp. 66–74, Apr. 2011.

[18] 3GPP, "General Packet Radio Service (GPRS); Service description; Stage 2," 3rd Generation Partnership Project (3GPP), TS 23.060, Mar. 2014. [Online]. Available: http://www.3gpp.org/DynaReport/23060.htm

[19] ——, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," 3rd Generation Partnership Project (3GPP), TS 23.401, Mar. 2014. [Online]. Available: http://www.3gpp.org/DynaReport/23401.htm

[20] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification," 3rd Generation Partnership Project (3GPP), TS 36.331, Mar. 2014. [Online]. Available: http://www.3gpp.org/DynaReport/36331.htm

[21] ——, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3," 3rd Generation Partnership Project (3GPP), TS 23.301, Mar. 2014. [Online]. Available: http://www.3gpp.org/DynaReport/23301.htm

[22] ——, "Mobile radio interface Layer 3 specification; Core network protocols; Stage 3," 3rd Generation Partnership Project (3GPP), TS 24.008, Mar. 2014. [Online]. Available: http://www.3gpp.org/DynaReport/24008.htm

[23] S. H. Hung, C. H. Chen, and C. H. Tu, "Performance evaluation of Machine-to-Machine (M2M) systems with virtual machines," in *Proceedings of WPMC 2012*, Sep. 2012, pp. 159–163.

[24] Ericsson, "Ericsson SGSN-MME," Sep. 2014. [Online]. Available: http://www.ericsson.com/ourportfolio/products/sgsn-mme