Received July 14, 2014, accepted August 23, 2014, date of publication September 22, 2014, date of current version October 2, 2014.

Digital Object Identifier 10.1109/ACCESS.2014.2359649

Lightweight Mobile Core Networks for Machine Type Communications

TARIK TALEB[1], (Senior Member, IEEE), ADLEN KSENTINI[2], (Senior Member, IEEE), AND ABDELLATIF KOBBANE[3]
[1]Aalto University, Espoo 02150, Finland
[2]University of Rennes 1, Rennes 35065, France
[3]École Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Mohammed V University, Rabat 11000, Morocco

Corresponding author: A. Ksentini (aksentin@irisa.fr)

ABSTRACT Machine type communications (MTCs) enable the communications of machines (devices) to machines over mobile networks. Besides simplifying our daily lives, the MTC business represents a promising market for mobile operators to increase their revenues. However, before a complete deployment of MTC over mobile networks, there is need to update the specifications of mobile networks in order to cope with the expected high number (massive deployment) of MTC devices. Indeed, large scale deployment of MTC devices represents an important challenge as a high number of MTC devices, simultaneously connecting to the mobile network, may cause congestion and system overload, which can degrade the network performance and even result in network node failures. Several activities have been led by 3GPP to alleviate system overload introduced by MTC. Most of the devised approaches represent only incremental solutions. Unlike these solutions, we devise a complete new architectural vision to support MTC in mobile networks. This vision relies on the marriage of mobile networks and cloud computing, specifically exploiting recent advances in network function virtualization (NFV). The aim of the proposed vision, namely LightEPC, is: 1) to orchestrate the on-demand creation of cloud-based lightweight mobile core networks dedicated for MTC and 2) to simplify the network attach procedure for MTC devices by creating only one NFV MTC function that groups all the usual procedures. By doing so, LightEPC is able to create and scale instances of NFV MTC functions on demand and in an elastic manner to cope with any sudden increase in traffic generated by MTC devices. To evaluate LightEPC, some preliminary analysis were conducted and the obtained analytical results indicate the ability of LightEPC in alleviating congestion and scaling up fast with massive numbers of MTC devices in mobile networks. Finally, a real-life implementation of LightEPC on top of cloud platform is discussed.

INDEX TERMS Machine type communications (MTC), network function virtualization (NFV), mobile cloud networking, carrier cloud.

I. INTRODUCTION

Recent forecasts predict a dramatic increase in mobile data traffic in the future [1]. Such traffic growth may become partially driven by the connections of potential numbers of communicating objects (i.e., sensors and actuators), known as Machine to Machine (M2M) or Machine Type Communication (MTC) devices, to the Internet through mobile networks. MTC devices support a broad variety of applications (e.g., eHealth and Intelligent Transport Systems – ITS), in a wide variety of domains. MTC is expected to connect a potential number of devices, exceeding 60 billions of M2M connections by 2020 [2]. Meanwhile, launching new services based on MTC over mobile networks represents an excellent business opportunity for mobile operators to increase their revenues and to cope with the relatively stagnant Average Revenue Per User (ARPU). However, deploying a massive number of MTC devices over mobile networks will put pressure not only on the network infrastructure (i.e., at both the Radio Access Network - RAN and the Core Network - CN) but also on the interconnecting nodes (e.g., routers and gateways), whereby system overload may occur when a large number of MTC devices attempt simultaneous connections to the mobile networks. Such situation could happen when a similar event is detected by a high number of MTC devices or after a network outage. Besides negatively impacting regular mobile user traffic and possibly downgrading the Quality of Experience (QoE) of mobile users, system overload caused by MTC traffic can lead to failure of network equipment [6], [15]. For instance, mobile gateway failure can occur due to a sudden increase in traffic, due in turn to MTC devices.

There have been several activities, led by 3GPP groups, around adapting the specifications of next-generation mobile networks (e.g., Long Term Evolution – LTE) to cope with mobile connections from a large number of MTC devices, ultimately aiming at reducing the impact of MTC on mobile networks and regular mobile user traffic [6]. The main idea is to separate MTC traffic from regular mobile user traffic using techniques at the RAN level, such as Extended Access Class Barring (EAB) and Extended Wait Timer (EWT) or granting dedicated times for MTC traffic [6]. However, it is generally agreed that MTC traffic is characterized by being infrequent and using small packet sizes. Therefore, it is mandatory to adapt and reduce the network attach procedure for MTC devices taking into account their traffic patterns. Such adaptation could be implemented in the form of an ''elastic and on-demand'' network attach procedure. One solution could be found in the convergence of mobile networks and the cloud through the Network Function Virtualization (NFV) concept [16], [17]. Indeed, NFV is a new paradigm that allows running network functions, as software, on Virtual Machines (VMs) instantiated on general-purpose hardware rather than on standalone dedicated hardware. The usage of NFV would help in creating and scaling network resources for MTC devices on demand and when needed. Such flexibility could be exploited for the case of time-controlled MTC accesses whereby resources are created only during specific periods [17].

In this paper, we devise a new architectural vision based on NFV for supporting MTC in mobile networks. Stemmed from the fact that MTC traffic is infrequent with small packet sizes, we propose a new and lightweight MTC control plane management, namely Lightweight Evolved Packet Core (EPC - LightEPC). The aim of LightEPC is to simplify the control plane procedures (e.g., network attach and bearer establishment) for MTC devices by orchestrating the creation and lifecycle management of one single NFV function, hosted in a Data Center, which groups all usual procedures for attaching MTC devices and subsequently establishing communication with remote MTC servers.

This paper is organized as follows. Section II presents some research work related to MTC and NFV. Section III describes in details the LightEPC framework. Its evaluation, based on preliminary analysis, is shown in Section IV. Section V presents an implementation recommendation for LightEPC. Finally, the paper concludes in Section VI.

II. RELATED WORK

Several specifications, at both architecture and system levels, have been amended to the 3G/4G standards to efficiently support MTC in cellular networks. One relevant output is the new MTC-aware 3GPP network architecture introduced in Release 12 and shown in Fig. 1.

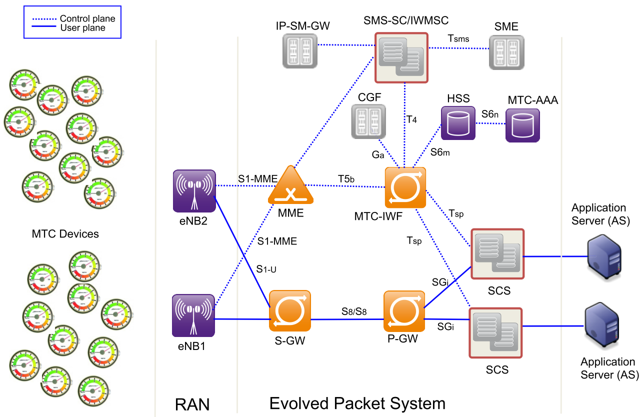

FIGURE 1. Envisioned MTC architecture as studied in 3GPP.

The proposed architecture distinguishes three domains: (i) the MTC device domain, (ii) the transport network domain, and (iii) the MTC application domain. The MTC application domain consists of MTC servers, under the control of the mobile network operator or a third-party MTC provider. Besides the usual CN entities (e.g. Mobility Management Entity – MME, Packet Data Network Gateway – P-GW, and Serving Gateway – S-GW), two new entities have been recently added to the Evolved Packet System (EPS) architecture. One entity is the MTC- Inter Working Function (MTC-IWF) that hides the internal PLMN (Public Land Mobile Network) topology and relays/translates signaling messages/protocols used over the Tsp interface to invoke specific functionalities in PLMN. It is also responsible for relaying trigger requests from the Service Capability Server (SCS) after checking authorization and reporting the acceptation or denial of these requests. The SCS entity connects the MTC application servers to the 3GPP network. SCS entities are connected to one or more MTC-IWF, depending on the locations of MTC devices. Table 1 lists most relevant nodes that form EPS.

Although MTC defines promising business opportunities for mobile operators, a large scale deployment of MTC devices, without an adequate engineering of their associated traffic and signaling, could result in potential congestion of mobile networks. As stated before, system overload or congestion may occur due to simultaneous signaling messages from MTC devices resulting from: (i) a malfunction in a MTC server; e.g., MTC devices trying to reconnect multiple times to the server which is down; (ii) multiple attempts from a large number of MTC devices to attach/connect to the network all at once (e.g., in order to report the detection of a particular event); and (iii) multiple device trigger requests from different MTC servers. Aiming for alleviating congestion at mobile core networks, the Core Network Overload (CNO) Study Item was initiated in 3GPP [3], providing system requirements and introducing a few solutions. Existing solutions for MTC signaling congestion avoidance and overload control may be classified into two classes, namely proactive and reactive. Proactive solutions are categorized, in turn, into two

TABLE 1. EPS's most important nodes.

Node	Description
eNB	Evolved Node B; LTE's base station.
MME	Mobility Management Entity, a control plane entity for all mobility related functions, paging, authentication, and bearer management in EPS.
MTC-IWF	MTC Interworking Function hides the internal Public Land Mobile Network (PLMN) topology and relays/translates signaling messages/protocols used over *Tsp* to invoke specific functionality in the PLMN.
HSS	Home Subscriber Server, main database containing subscription-related information.
P-GW	Packet Data Network Gateway, interfaces with the Packet Data Network (e.g., Internet).
S-GW	Local mobility anchor for intra-3GPP handoffs.
SCS	Services Capability Server is the entity that connects the MTC application domain to the network domain.

other categories. The first category solutions differentiate MTC traffic from the non-MTC traffic using techniques, such as the grouping and clustering of MTC devices based on different metrics/features (e.g., low mobility, QoS requirement [4], belonging to macro or femtocell [5]). Afterwards, network access is differentiated for each MTC group by allocating different "grant time periods" for each MTC group or setting up different RAN access parameters in order to prioritize the non-MTC traffic. Approaches of the second category reduce the amount of signaling traffic by aggregating the MTC requests either at RAN (i.e., Bulk signaling) [6] or by creating profile ID [7] that replaces common Information Elements (IEs) used by a group of MTC devices, which allow reducing the size of signaling messages and accordingly associated processing load.

On the other hand, reactive solutions react to congestion only after its occurrence, rejecting or delaying MTC attach/connect requests at RAN. Indeed, in some situations, there is need to reduce the MTC traffic by a specific amount implementing admission control at eNBs or even at MTC devices [8]. Admission control could be activated at eNBs upon receiving a congestion signal from the EPS nodes (e.g., MME and HSS). Alternatively, it could be communicated to the MTC devices as in the 3GPP Access Class Baring (ACB) solution. Effectively, ACB reduces the collision probability of transmitting the bulk of preambles at the same Resources Access Channel (RACH) resource. Based on the parameters broadcasted by eNBs, a User Equipment (UE) determines whether it is temporarily barred from accessing the cell. An access class barring factor or access probability (p) determines the probability that access is allowed. If a UE-generated random number n is equal to or greater than p, then access is barred for a mean access barring time duration. In legacy ACB scheme, there are 16 access classes. AC 0-9 represents normal UEs, AC 10 represents an emergency call, and AC 11-15 represents specific high priority services, such as security services and public utilities (e.g., water/gas suppliers). A UE may be assigned one or more access classes depending on the particular cell access restriction scheme. Similar in spirit to this concept, the authors in [8] proposed a congestion-aware admission control solution. The proposed solution selectively rejects signaling messages from MTC devices at RAN following a probability that is set based on Proportional Integrative Derivative (PID) controller, and is derived at a particular CN node (e.g., MME). To further reduce signaling generated by MTC devices, the authors also proposed in [7] an optimized triggering solution for low mobility MTC devices. This solution renders the triggering operation less costly in case of triggering low mobility MTC devices, by limiting the triggering operation to a specific network area and also by reducing the number of involved network nodes, mainly avoiding MME. Other solutions that cope with MTC signaling congestion are discussed in [6].

Whilst several research work have addressed network congestion caused by MTC, most presented approaches represent incremental solutions to engineer the MTC traffic. To the best knowledge of the authors, no prior research work has considered the use of cloud computing technologies and NFV to build a lightweight EPC to specifically handle MTC traffic in an elastic way and on demand, two important features of cloud computing. Devising such lightweight EPC and tailoring it for MTC defines the focus of this paper as detailed in the next section.

Thanks to the numerous advantages it offers in terms of network configuration flexibility, scalability, and elasticity, NFV has emerged as an important topic of inquiry among different stakeholders in the telecommunications arena. Several pioneering research work have been conducted to enable the creation and runtime management of mobile networks over the cloud, studying different implementation options [16] and devising an entire framework for the creation of end-to-end mobile services, including mobile transport networks, on the cloud [17]. Software Defined Networking (SDN) has been also considered in virtualization of mobile network functions over OpenFlow-based networks [18]. Other research work also considered the usage of SDN to virtualize the control plane of a mobile network on the cloud [19]. The concept of NFV is also explored in [10] and [20], focusing on the virtualization of the control plane; separately or jointly with the user data plane. Some other research work have also investigated the problem of network function placement aiming for the optimization of different metrics relevant to the mobile network performance [9], [11] and considering the features of provided services [27].

III. LIGHTWEIGHT EPC FOR MTC

As stated earlier, MTC communications impose new constraints on the mobile network architecture, consequently creating a need for solutions to cope with new and unusual traffic types. So far the devised relevant solutions rely on modifying either the RAN entities or CN's, which constitute only incremental solutions. In this paper, we propose a new architectural vision to support MTC communications in 3GPP-based mobile networks. This vision exploits the progress achieved so far in NFV, required for carrier cloud whereby software is decoupled from hardware [16], [17]. Indeed, NFV aims at running network functions in virtualized environments on VMs on top of virtualized platforms, rather than on dedicated expensive hardware.

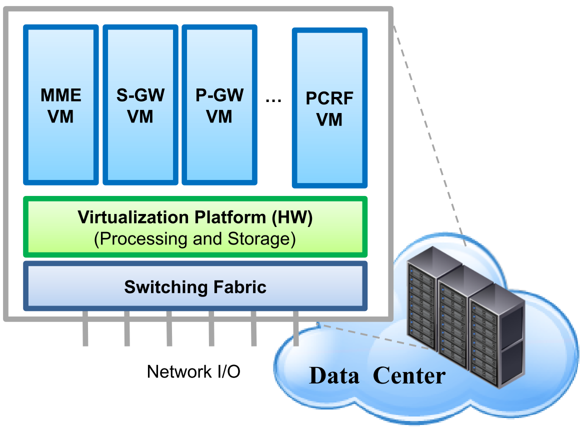

FIGURE 2. EPS as a service.

Fig. 2 depicts a schematic illustration of the so-called EPS as a Service (EPSaaS) architecture [16], [17], wherein some EPC network functions (e.g., MME, P-GW, and S-GW) are instantiated on VMs on top of a virtualized platform, running in a Data Center (DC) or across multiple DCs, and are interworked by the use of a suitable SDN technology. Such new vision enables network operators to cope with the growing mobile traffic while reducing CAPEX/OPEX (Capital/Operational Expenditures) [16], [17]. Indeed, the flexibility of this design allows network operators to flatten the architectures of their mobile networks by deploying instances of EPC entities across a federated cloud of DCs and hence geographically distribute/decentralize the mobile core network, which (i) enables the offload of all or a selection of user traffic nearby users and accordingly avoid wasting resources of the core network infrastructure [12] and (ii) increases user's QoE since the user traffic gets routed through the nearest NFV instance of the EPC entities.

However, MTC traffic is known for its unique characteristics: it is of short duration and involves small and frequent data transmissions. For instance, when a measurement has to be reported by a MTC device to a remote server or the remote server triggers the MTC device, the device has to attach to the mobile network following the whole signaling procedures (e.g., attach request and bearer creation request) to only send a few bytes in small data packets. The established bearers (i.e., radio, S1 and S5) are maintained for certain duration and then released. Such procedures may consume significant resources, particularly in case of a high number of MTC devices trying to simultaneously attach to the mobile network. To cope with this issue, the 3GPP System Architecture group 2 (SA2) has introduced several new connection setup procedures for small data transmissions [13]. These procedures are dedicated to optimize both RAN and CN. Regarding the latter, most proposed solutions reduce the signaling procedures related to the establishment and management of bearers. In some of these solutions, MTC small data traffic is handled as Short Message Service (SMS). In fact, SMS transmission over LTE does not require the establishment of S5/S8 bearers; it is directly encapsulated into a NAS (Non-Access Stratum) message and handled by the MME and forwarded to a special entity SMS Service-Center (SMS-SC). The latter is responsible for transferring the SMS to its destination. However, a SMS message size does not typically exceed 140 bytes, which limit the MTC small data size to be carried by the mobile network. To avoid the creation of S1 and S5/S8 bearers, another solution considered the use of a dedicated protocol, Small Data Transmission (SDT), with a specific SDT-Packet Data Unit (PDU) (up to 1 kilo bytes), which are encapsulated into signaling messages such as NAS, T5 and Tsp. Similar to the SMS solution, MME handles these messages and finds the appropriate MTC-IWF, which will transfer this data to the remote MTC server. In this solution, all MTC small data are routed through the MTC-IWF entity without involving the S-GW/P-GW. Fig. 3 shows the case of small data transfer for uplink and downlink transfer.

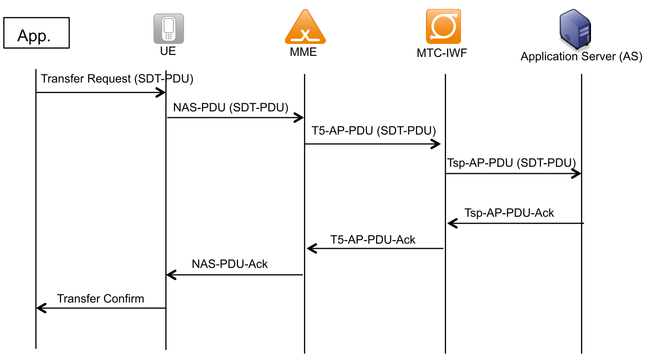

FIGURE 3. Small data transfer in 3GPP [14].

Another alternative for small data transmission is to use a combined gateway (P-GW/S-GW), namely CGW, without involving MTC-IWF entities. Such solution maintains the S1 bearer between RAN and MME, but does not require the S5/S8 bearer.

Stemmed from the fact that the 3GPP group has proposed several mechanisms to efficiently handle small data transmission procedures, we propose a new control plane management, dubbed LightEPC, for MTC devices exploiting NFV. The purpose of LightEPC is to simplify the MTC control plane management procedure, while being as flexible as possible to handle MTC traffic. Rather than having the usual MTC-IWF, MME, P-GW and S-GW, we envision a single virtualized network function (VNF) hosted in DCs. Thus, network operators may build and scale LightEPC instances on demand, creating the concept of MTC as a Service (MTCaaS). Besides avoiding the risk of MTC traffic overloading the system by separating MTC control plane from regular control, LightEPC aims at reducing the cost for mobile network operators to rapidly enter into the MTC market with low investment. LightEPC instances may work in parallel to the usual EPC entities, which may be used to handle regular traffic. No modifications are foreseen for eNBs, apart of their abilities to recognize MTC traffic and to redirect it to the LightEPC instances. To do so, eNBs may use the internal identifiers of MTC devices, such as International Mobile Subscriber Identity- IMSI. Here, we assume that the internal identifier is specific and could be used to distinguish MTC devices from classical UEs as well as to which group they belong. Indeed, such identifier could be formulated like the Ethernet MAC address: a part to detect MTC devices and their group and another part representing any other relevant information.

As mentioned earlier, the network elements that suffer most from MTC massive attach attempts are MME followed by data anchor and mobility gateways (i.e., P/S-GWs). Separating MTC control plane from regular traffic control plane constitutes a good solution to reduce the load on these elements. Indeed, the flexibility that both cloud and LightEPC VNF offer, facilitates the on-demand creation of instances of MTCaaS for each MTC group or application. Thus, it is possible to support several MTC applications and services in parallel (ensuring MTC service scalability) without impacting the regular traffic of other UEs. Such flexibility could be also exploited to create dynamic instances of LightEPC. For example, time-controlled MTC devices, like electricity/gas meters, require resources during only specific time periods. Consequently, relevant utility providers may request the creation of specific LightEPC instances to gather data during the specific time periods and release resources after the measurements are successfully transmitted [17]. The group ID of MTC devices and their associated LightEPC instances (IP address) are then communicated to RAN nodes (i.e., eNBs), in order that eNBs know where to redirect the MTC traffic. Depending on the underlying MTC application, cloud resources dedicated for the relevant LightEPC instance may be released and reassigned periodically. To ensure service continuity, recreated LightEPC instances must be reachable by eNBs through the same Fully Qualified Domain Name (FQDN), the same IP address, or through the use of adequate SDN technologies (e.g., OpenFlow) in case of a change in the IP address. The envisioned MTC LightEPC instances include all necessary functionalities to transport data gathered from MTC devices, e.g., using SDT protocol as well as Combined Gateway. Indeed, LightEPC goes one step further by integrating not only S-GW and P-GW, but also the MTC-IWF and SCS functionalities. Fig. 4 portrays the envisioned architecture of LightEPC. Fig. 5 illustrates the sequence of the main steps to follow for the creation and management of a LightEPC instance dedicated for an identified MTC service.

In the envisioned architecture, eNBs and SCSs are assumed to be equipped with a new function dedicated to detect and identify MTC service types, namely MTC Service Type Detection Function (MTC-STDF). Indeed, upon a trigger or the occurrence of an event, MTC devices and/or MTC servers issue signaling messages to connect to the network and/or to trigger MTC devices to attach to the network. These signaling messages are intercepted and analyzed by the MTC-STDF function. Note that MTC service type detection and identification could be carried out at any point on the path between MTC devices and MTC servers. For instance, it could be carried out at a function collocated with MTC-IWF, P-GW, or SMS-SC depending on the underlying MTC data delivery protocol (i.e., SMS-, T5-, or Tsp-based). In practice, the MTC-STDF function monitors the size of the MTC signaling messages and the frequency at which they are transmitted, i.e., measuring the inter-arrival time between two consecutive signaling messages from the same MTC device, from different MTC devices but belonging to the same group identified by a unique ID, or from the same or different SCSs but triggering the same group of MTC devices, etc. If this inter-arrival time is shorter than a predetermined/configurable threshold, and/or the number of MTC signaling messages, from/targeting the same MTC device/group, issued during a particular time interval exceeds a certain threshold, the MTC-STDF may qualify/identify the MTC application relevant to the MTC signaling messages as "an MTC application with frequent and dynamic requested content", similar in spirit to [24]. MTC-STDF may employ any suitable learning technique (e.g., Neuronal Networks, Fuzzy techniques, etc.) to MTC traffic analysis to improve the accuracy of MTC type detection, similar in concept to [25] and [26].

Once the MTC service type is identified or the features of its traffic (i.e., MTC message size, frequency of MTC messages,, MTC group size, etc) are extracted, MTC-STDF informs and forwards the MTC service type and traffic features to the Policy Enforcement Entity (PEE) as part of the LightEPC (for MTC) service orchestrator), residing in the cloud. The LightEPC service orchestrator also contains a repository of images of the network functions that may be forming a LightEPC instance (e.g., P-GW, S-GW, and MME). A high-level diagram architecture of PEE is shown in Fig. 4. Shown are also the interactions of PEE with MTC-STDF and the Cloud Controller that is in charge of instantiating required resources/VMs on the virtual infrastructure platform of the cloud, using a suitable cloud management tool (e.g., OpenStack). PEE principally consists of four units, namely Policy Decision Making (PDM), Cloud Resource Assessor (CRA), Individual Policy Enforcer (IPE), and Run-Time Policy Orchestrator (RPO). Upon detection and identification of an MTC service, MTC-SDTF (at both RAN and SCS) reports this event to the PDM unit. Depending on the characteristics and features of the identified MTC service, PDM

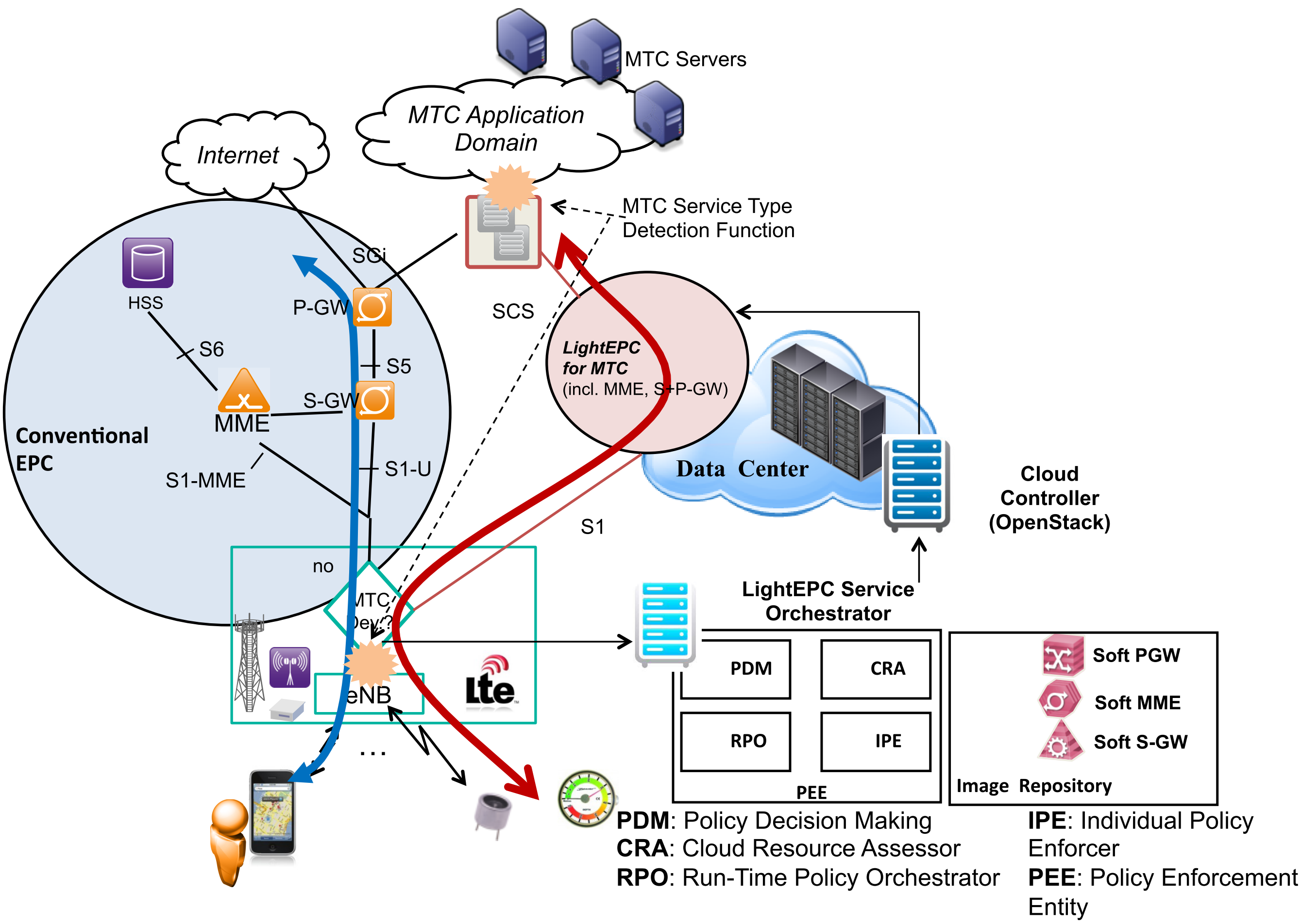

FIGURE 4. Envisioned LightEPC architecture for the support of MTC services.

decides whether the traffic of such MTC service shall be routed over the existing mobile core network or whether an instance of LightEPC shall be instantiated and dedicated for the MTC traffic. In case of the latter, PDM decides also on the size, location [9], [27], and features of the VM(s) that will be hosting LightEPC. PDM's decision is then communicated to CRA that assesses the required cloud resources to enforce it on the traffic of the relevant MTC service. For instance, in case PDM decides to create a LightEPC for the MTC service, resources for instantiating images of LightEPC or a subset of its virtualized network functions (e.g., S-GW, P-GW, and MME) become therefore required. Once the needed resources are identified, they are communicated to the cloud controller that deploys them, e.g., using OpenStack. IPO then instantiates images of adequate VNFs (e.g., S-GW, P-GW, MME, etc) on deployed VMs as per the decision of PDM. RPO then orchestrates the underlying LightEPC during its run-time and per changes in the MTC service detection and based on internal as well as external triggers (e.g., cloud resource monitoring, service level agreement controller, etc). Exploiting cloud computing technologies, a policy orchestration could indicate the instantiation of an additional LightEPC to scale up, turning down another to scale down, replacing VNF running on a VM with another one as per changes in the MTC service and the behavior of its devices. With its four units, PEE follows a common lifecycle policy management model whereby decision and policy design is conducted by PDM and CRA, decision implementation and deployment are carried out by IPE, and policy provisioning, runtime and operation, and disposal are conducted by RPO. Once the creation of a LightEPC for a specific MTC service is completed, PEE, namely PDM, informs the concerned SCSs and RAN nodes of the created LightEPC instructing them to route all MTC traffic relevant to the MTC service via the LightEPC. With the flexibility that cloud computing and the PEE architecture offer, a network operator may instantiate VMs and run on them suitable VNFs to specifically handle traffic of any MTC application.

The instantiated LightEPC instances also provide communication interface (i.e., mimicking the standardized SGi interface connecting P-GWs with Packet Data Networks in EPS) to remote MTC servers in order to allow them trigger MTC devices or receive MTC data. LightEPC may also incorporate a small-scale HSS-like database for authorized MTC devices

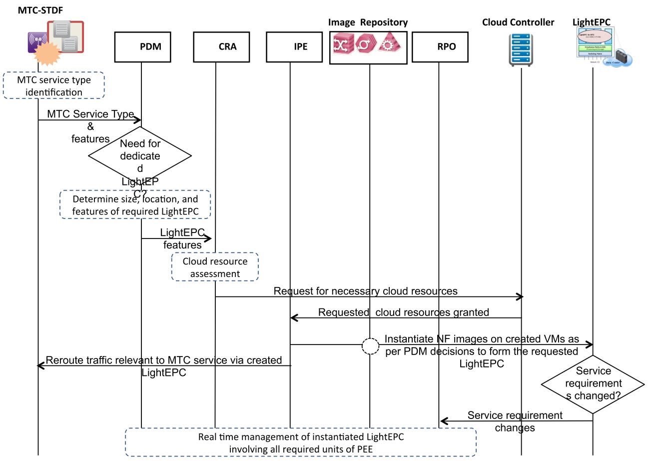

FIGURE 5. **Important steps for the creation and management of a LightEPC instance dedicated for an identified MTC service.**

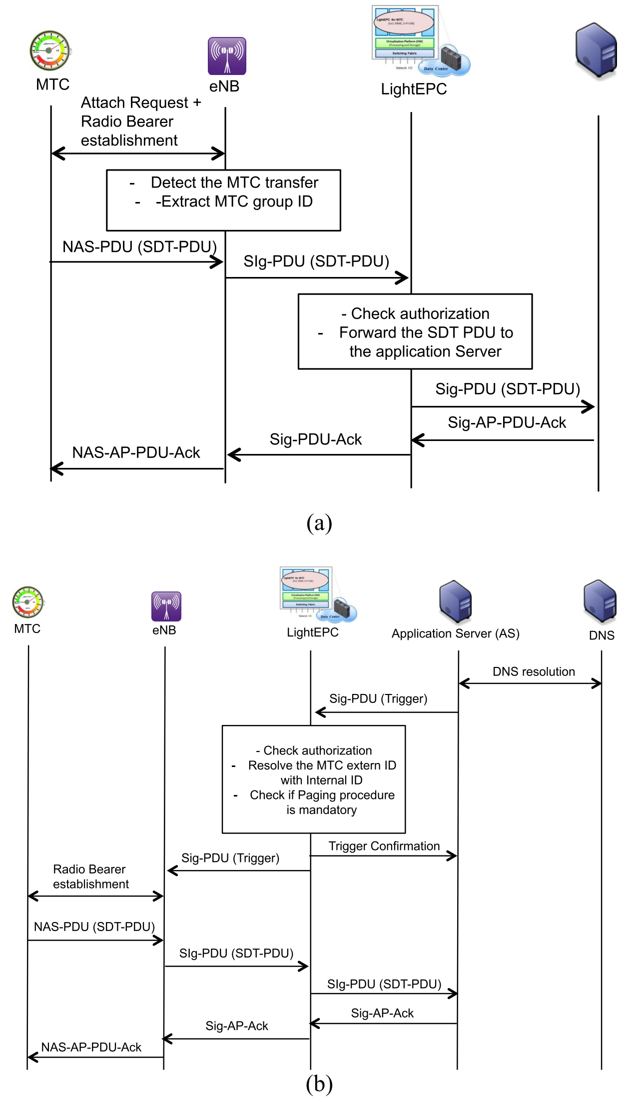

FIGURE 6. **Connection set up in the envisioned LightEPC architecture. (a) MTC device-initiated attach. (b) Trigger-based connection set up.**

and their location over the network. Fixed MTC devices, such as electricity/gas meters, are labeled as non-mobile UEs and will be associated to the IDs of their corresponding eNBs, wherein no paging procedure is needed. Regarding mobile MTC, used for instance by Intelligent Transport Systems, their locations can be retrieved at different granularity (e.g., cell, tracking area, and service area) depending on the current state of the device (i.e., active or idle). In LightEPC, if a majority of MTC devices (e.g., based on configurable thresholds) are detected, by MTC-STDF, roaming to another location where a geographically nearby DC is available, a ''hot'' migration of LightEPC VM can be then made to the new optimal DC similar in spirit to the Follow Me Cloud concept [14] and as per decision of the RPO unit. This means that VM migration must be done in real-time without impacting service connectivity. In fact, several technologies exist which allow ''hot'' VM migration, depending on the network level where they operate, i.e. layer 2 or layer 3. For instance, solutions such as TRILL (Transparent Interconnection of Lots of Links) and SPB (Shortest Path Bridging) can be used at layer 2, while VXLAN (Virtual Extensible LAN) and NVGRE (Network Virtualization using Generic Routing Encapsulation) can be employed as mixed layer2/layer3 solutions [15]. LightEPC is also responsible for assigning IP addresses to MTC devices as it is assumed to incorporate a subset of functions of a data anchor gateway, namely P-GW. It is provisioned with non-overlapping ranges of IPv4 address or an IPv6 global unicast. Such simplification is indeed inline with SDT protocol proposed by 3GPP in [14] whereby MME has to assign IP addresses to MTC.

Fig. 6 illustrates the MTC communication flow with the remote server for both MTC device-initiated connection setup and trigger-based connection setup. As stated before, we rely on a combination of SDT protocol and CGW mechanisms. Once a MTC device has a small data packet to send, it attaches as usual to RAN, namely eNB, and establishes a radio bearer. Referring to the device's internal ID, eNB detects that the attaching UE is an MTC device of an MTC group for which a LightEPC has been created, and directs its SDT PDU (Protocol Data Unit) message to the associated LightEPC. LightEPC then verifies the authorization rights of the MTC device to transmit data (either by consulting the mobile network's HSS or an internal HSS-like database) and forwards the data packet to the destined MTC-Server. On the other hand, when the MTC server desires triggering one or a group of MTC devices, it sends a connection trigger request to LightEPC in charge of the MTC device group, indicating the group ID or the MTC device ID. Alternatively, the connection trigger request is transmitted to an available SCS that recognizes the MTC device, its group and whether it is associated with a running LightEPC. In case the latter is affirmative, SCS forwards the connection trigger request to the corresponding LightEPC. After a successful authorization checkup, LightEPC checks the type of MTC device (i.e., mobile or fix). The MTC device is not paged if it is a UE with limited mobility feature. LightEPC then directly forwards the connection trigger request to the eNB managing the cell where the MTC device resides. Otherwise, the paging procedure takes place, triggering the MTC device to attach to the network (i.e., to establish a radio bearer and to get an IP address from the LightEPC). After successful attach, LightEPC proceeds with small data transmission using the same encapsulation procedure as described before.

IV. PERFORMANCE ANALYSIS

Besides the trivial gain in reducing signaling messages used for establishing bearers, in this section we focus on the capability of the overall LightEPC architecture in scaling up with the number of MTC devices. Given the fact that mobile core networks are usually over-dimensioned and hence no link congestion is expected to occur within mobile core networks, we specifically focus on the analysis of the number of attach requests handled by a mobile core network entity, such as MME, in both the classical dedicated hardware-based LTE architecture and when the proposed LightEPC concept is used. For this purpose, we adopt an analysis similar in spirit to that of [7]. We assume that MTC devices are grouped, as per their services, in M groups, wherein each group contains k devices. Therefore, the network supports a total of M.k MTC devices. We assume that each MTC device sends an attach request following a Poisson distribution with intensity λ. The expected numbers of MTC attach requests during a period T for both solutions, namely the classical LTE procedure and the proposed LightEPC are obtained respectively as follows:

$$E[attach - request] = \sum_{i=1}^{m*k} i * p(i)$$

$$E[attach - request] = \sum_{i=1}^{k} i * p(i)$$

where $p(i)$ denotes the probability that i MTC devices send an attach request. The main difference consists in the fact that in case of the proposed LightEPC concept, each MTC group is associated with one LightEPC instance, and hence no more than k devices can connect to LightEPC. Whereas, in the classical solution, all $M * k$ devices have to connect to the same MME.

Fig. 7. shows the expected number of MTC device-initiated attach requests for two values of λ and for both mechanisms. $\lambda = 1$ means that the traffic intensity is low, while $\lambda = 10$ shows a case where traffic intensity is high. The period T is fixed to one hour, and k is set to 5000. As our aim is to evaluate the scalability of each solution, we varied the number of MTC groups (M) to increase the number of MTC devices in the network. Clearly, we observe that increasing the number of MTC devices in the network increases the number of attach requests to be handled by MME. This number reaches high values when the traffic intensity increases. In this case, high numbers of attach requests result in overloading MME

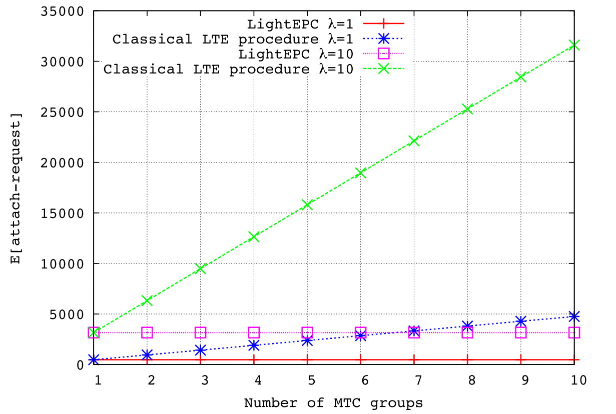

FIGURE 7. Expected number of MTC attach requests for different numbers of MTC groups.

increasing its queue size [8], which may ultimately lead to dropping some attach requests, including those that may come from regular UEs. This may increase the connections setup delays, which would consequently impact users' QoE. Using one LightEPC instance per MTC group maintains the number of MTC attach requests to be handle at a constant value. Of course, the number of requests increases by increasing the traffic intensity, but it is still manageable by the LightEPC instances as more VMs are flexibly instantiated to accommodate the traffic increase.

V. IMPLEMENTATION VISION

The technologies enabling the virtualization of network functions are currently in their infancy, wherein new architectures and systems are still needed. To this aim, recent standardization activities have been launched by ETSI and IRTF [21], [22] establishing new working groups on network function virtualization. It is worth noting that ETSI NFV group is supported by leading telecom operators and equipment vendors, and has already published different documents to build the basis of the NFV architecture and systems that should definitely benefit the envisioned LightEPC architecture. On the other hand, the ClickOS initiative [23] aims at building a technology enabler for NFV based on open source tools. ClickOS is a minimal OS (Operating System) based on XEN software platform optimized for middlebox processing. By middlebox, we mean all hardware-based network appliances used to run a specific network function (e.g., firewall, Intrusion Detection System – IDS, and Network Address Translation –NAT). ClickOS includes the software modular router, Click, in order to process packets and acts as router or firewall. As one of the challenges of NFV is the ability to process packets as fast as hardware-based solutions, ClickOS leverages the XEN I/O subsystem by changing the back-end switch, virtual net devices, and back/front end-drivers. Results presented in [23] show that ClickOS is capable to forward packets at around 30Gbps, proving that NFV could achieve the same performance as hardware-based solutions. Furthermore, ClickOS is able to boot in only few seconds. Therefore, putting all these features together, ClickOS represents a highly relevant platform to implement LightEPC. Indeed, all mentioned functionalities required by LightEPC, such as assigning IP addresses, managing mobility, and establishing connections to remote MTC servers, can be easily implemented and ran on top of ClickOS, thanks to its ability to manage and forward high number of packets. Moreover, in order to scale with the number of MTC devices, LightEPC has to be instantiated instantaneously, which could be achieved by ClickOS VMs as it boots in only few seconds.

VI. CONCLUSION

One important requirement on the upcoming 5G mobile system is to support high number of user equipment and devices in need of connections to the mobile network. Among these devices, MTC devices will represent the lion share. However, a high number of MTC devices attaching simultaneously to the mobile network may introduce system overload and negatively impact the overall network performance. Most existing solutions to this issue present incremental approaches to engineer the MTC traffic, solving only partially the problem (at RAN or CN). In this paper, we proposed a new vision to support different MTC services in mobile networks, "marrying" between cloud computing and mobile networking, exploiting recent advances in NFV. The proposed solution, dubbed as LighEPC, simplifies the network attach procedure for MTC devices by creating on-demand only one NFV MTC function that groups all the usual procedures. Thus, LightEPC is able to create and scale instances of NFV MTC functions on demand and in an elastic manner to cope with any sudden increase in MTC traffic. Whilst the results presented herein demonstrate the efficiency and scalability of LightEPC, they are admittedly preliminary. It is therefore the intention of the authors to develop a more complete analysis, e.g., using optimization theories, that models the system and recommends how many instances of LightEPC to instantiate to support a predetermined MTC service with particular characteristics, such as number of MTC devices, MTC traffic rate, patterns of MTC flows, etc. This defines one of the future research directions of the authors in relevance to the topic of the paper. The authors would also use the ClickOS framework to implement LightEPC and build its Proof of Concept.

REFERENCES

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018*, Cisco, San Francisco, CA, USA, Feb. 2014.

[2] *The Internet of Everything Global Private Sector Economic Analysis*. [Online]. Available: http://www.cisco.com/web/about/ac79/doc/innov/IoE-Economy.pdf, accessed Sep. 24, 2014.

[3] *Architecture Enhancements to Facilitate Communications With Packet Data Networks and Applications*, document 3GPP, TS 23.682, V12.2.0, Jun. 2014.

[4] S.-Y. Lien and K.-C. Chen, "Massive access management for QoS guarantees in 3GPP machine-to-machine communications," *IEEE Commun. Lett.*, vol. 15, no. 3, pp. 311–313, Mar. 2011.

[5] A.-H. Tsai, L.-C. Wang, J.-H. Huang, and T.-M. Lin, "Overload control for machine type communications with femtocells," in *Proc. IEEE VTC Fall*, Québec, QC, Canada, Sep. 2012, pp. 1–5.

[6] T. Taleb and A. Ksentini, "Follow me cloud: Interworking federated clouds and distributed mobile networks," *IEEE Netw.*, vol. 27, no. 5, pp. 12–19, Sep./Oct. 2013.

[7] T. Taleb and A. Ksentini, "On alleviating MTC overload in EPS," *Ad Hoc Netw.*, vol. 18, pp. 24–39, Jul. 2014.

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

[9] T. Taleb and A. Ksentini, "Gateway relocation avoidance-aware network function placement in carrier cloud," in *Proc. ACM MSWiM*, Barcelona, Spain, Nov. 2013, pp. 341–346.

[10] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, and E.-D. Schmidt, "A virtual SDN-enabled LTE EPC architecture: A case study for S-/P-gateways functions," in *Proc. IEEE SDN Future Netw. Services (SDN4FNS)*, Trento, Italy, Nov. 2013, pp. 1–7.

[11] F. Z. Yousaf, J. Lessmann, P. Loureiro, and S. Schmid, "SoftEPC—Dynamic instantiation of mobile core network entities for efficient resource utilization," in *Proc. IEEE ICC*, Budapest, Hungary, Jun. 2013, pp. 3602–3606.

[12] T. Taleb, K. Samdanis, and A. Ksentini, "Supporting highly mobile users in cost-effective decentralized mobile operator networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 7, pp. 3381–3396, Sep. 2014.

[13] *Study on RAN Improvements for Machine-type Communications*, document 3GPP, TR 37.868, V11.0.0, Sep. 2011.

[14] T. Taleb and A. Ksentini, "Follow me cloud: Interworking federated clouds and distributed mobile networks," *IEEE Netw.*, vol. 27, no. 5, pp. 12–19, Sep./Oct. 2013.

[15] T. Taleb and K. Samdanis, "Ensuring service resilience in the EPS: MME failure restoration case," in *Proc. IEEE GLOBECOM*, Houston, TX, USA, Nov. 2011, pp. 1–5.

[16] T. Taleb *et al.*, "EASE: EPC as a service to ease mobile core network," *IEEE Netw.*, to be published.

[17] T. Taleb, "Toward carrier cloud: Potential, challenges, and solutions," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 80–91, Jun. 2014.

[18] N. McKeown *et al.*, "OpenFlow: Enabling innovation in campus networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, Apr. 2008.

[19] J. Kempf, B. Johansson, S. Pettersson, H. Luning, and T. Nilsson, "Moving the mobile evolved packet core to the cloud," in *Proc. IEEE 8th Int. Conf. WiMob*, Barcelona, Spain, Oct. 2012, pp. 784–791.

[20] J. Batalle, J. Ferrer Riera, E. Escalona, and J. A. Garcia-Espin, "On the implementation of NFV over an openflow infrastructure: Routing function virtualization," in *Proc. IEEE SDN Future Netw. Services (SDN4FNS)*, Trento, Italy, Nov. 2013, pp. 1–6.

[21] *ETSI NFV Group*. [Online]. Available: http://www.etsi.org/technologies-clusters/technologies/nfv, accessed Sep. 24, 2014.

[22] *IRTF NFV Group*. [Online]. Available: http://trac.tools.ietf.org/group/irtf/trac/wiki/nfvrg, accessed Sep. 24, 2014.

[23] J. Martins *et al.*, "ClickOS and the art of network function virtualization," in *Proc. USENIX NSDI*, Seattle, WA, USA, Apr. 2014, pp. 459–473.

[24] T. Taleb and A. Ksentini, "Impact of emerging social media applications on mobile networks," in *Proc. IEEE ICC*, Budapest, Hungary, Jun. 2013, pp. 5934–5938.

[25] T. Taleb and A. Ksentini, "QoS/QoE predictions-based admission control for femto communications," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 5146–5150.

[26] Z. M. Fadlullah, T. Taleb, A. V. Vasilakos, M. Guizani, and N. Kato, "DTRAB: Combating against attacks on encrypted protocols through traffic-feature analysis," *IEEE/ACM Trans. Netw.*, vol. 18, no. 4, pp. 1234–1247, Aug. 2010.

[27] M. Bagaa, T. Taleb, and A. Ksentini, "Service-aware network function placement for efficient traffic handling in carrier cloud," in *Proc. IEEE WCNC*, Istanbul, Turkey, Apr. 2014.

TARIK TALEB (S'04–M'05–SM'10) received the B.E. (Hons.) degree in information engineering and the M.Sc. and Ph.D. degrees in information science from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a faculty staff with the School of Engineering, Aalto University, Espoo, Finland. He has been a Senior Researcher and 3G Partnership Project Standardization Expert with NEC Europe Ltd., Heidelberg, Germany, where he is leading the NEC Europe Labs Team, working on research and development projects on carrier cloud platforms. Prior to his work at NEC until 2009, he was an Assistant Professor with the Graduate School of Information Sciences, Tohoku University. His research interests include architectural enhancements to mobile-core networks (in particular 3GPP), mobile cloud networking, mobile multimedia streaming, congestion control protocols, handoff and mobility management, inter-vehicular communications, and social media networking. He has also been directly involved in the development and standardization of the Evolved Packet System as a member of 3GPP's System Architecture Working Group.

Dr. Taleb is a Distinguished Lecturer of the IEEE Communications Society (ComSoc), and a Board Member of the IEEE ComSoc Standardization Program Development Board. He is serving as the Vice Chair of the Wireless Communications Technical Committee: the largest in the IEEE ComSoC. He also served as a Secretary and the Vice Chair of the Satellite and Space Communications Technical Committee of the IEEE ComSoc (2006–2010). As an attempt to bridge the gap between academia and industry, he founded and has been the General Chair of the IEEE Workshop on Telecommunications Standards: From Research to Standards, which is a successful event that received the Best Workshop Award from the IEEE ComSoC. He has been on the Technical Program Committee of different IEEE conferences, including the IEEE Global Communications Conference, the IEEE International Conference on Communications, and the IEEE Wireless Communications and Networking Conference, and has chaired some of their symposia. He is/was on the Editorial Board of the *IEEE Wireless Communications Magazine*, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, and a number of Wiley journals. He was a recipient of the IEEE ComSoc Asia-Pacific Best Young Researcher Award in 2009, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2008, the Funai Foundation Science Promotion Award in 2007, the IEEE Computer Society Japan Chapter Young Author Award in 2006, the Niwa Yasujirou Memorial Award in 2005, and the Young Researcher's Encouragement Award from the Japan Chapter of the IEEE Vehicular Technology Society in 2003. Some of his research works have received best paper awards at prestigious conferences.

ADLEN KSENTINI (SM'14) is currently an Associate Professor with the University of Rennes 1, Rennes, France. He is a member of the INRIA Rennes team Dionysos. He received the M.Sc. degree in telecommunication and multimedia networking from the University of Versailles, Versailles, France, and the Ph.D. degree in computer science from the University of Cergy-Pontoise, Cergy-Pontoise, France, in 2005, with a dissertation on QoS provisioning in IEEE 802.11-based networks. His other interests include future Internet networks, mobile networks, QoS, QoE, performance evaluation, and multimedia transmission. He is involved in several national and European projects on QoS and QoE support in future wireless and mobile networks. He has co-authored over 60 technical journal and international conference papers. He was a recipient of the Best Paper Award from the 2012 IEEE International Conference on Communications and 2005 Association for Computing Machinery (ACM) Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems. He is the Technical Program Committee (TPC) Chair of the IEEE Third Workshop on Standards on Telecommunication (collocated with Globecom 2014), and the Workshop Chair of the ACM/IEEE QShine 2014. He was the Guest Editor of the *IEEE Wireless Communication Magazine* and the *IEEE Communication Magazine*. He has been on TPC of major IEEE ComSoc Conferences, International Conference on Communications/Global Communications Conference, International Conference on Multimedia and Expo, Wireless Communications and Networking Conference, and International Symposium on Personal, Indoor and Mobile Radio Communications.

ABDELLATIF KOBBANE has been an Associate Professor with the École Nationale Supérieure d'Informatique et d'Analyse des Systemes, Mohammed V University, Rabat, Morocco, since 2009. He received the M.S. degree in computer science, telecommunication, and multimedia, and the Ph.D. degree in computer science from Mohammed V University at Agdal, Rabat, in 2003 and 2008, respectively. His research interests lie with the field of wireless networking, performance evaluation in wireless network and NGN, ad hoc networks, DTNs, mesh networks, cognitive radio, mobile computing, mobile social networks, and future networks. He is a member of the Laboratoire des Systèmes d'Information Mobiles and MIS Research Group Leader.