MME Support for M2M Communications using Network Function Virtualization

Pilar Andres-Maldonado, Pablo Ameigeiras, Jonathan Prados-Garzon, Juan Jose Ramos-Munoz and Juan Manuel Lopez-Soler

Wireless and Multimedia Networking Lab
Dept. of Signal Processing, Telematics, and Communications, UGR
Granada, Spain
pam91@correo.ugr.es, {pameigeiras, jpg, jjramos, juanma}@ugr.es

***Abstract*—The use of massive Machine to Machine (M2M) communications on future mobile networks may lead to a signaling traffic explosion. Small Data Transmission (SDT) procedure appears as an efficient option for M2M small data transfer in Long Term Evolution (LTE). However, this procedure entails more processing load in the Mobility Management Entity (MME). Moreover, the fixed capacity in current LTE core hardware-based infrastructure can limit the scalability of this solution. To overcome this, we propose to: i) virtualize hardware dedicated MME (vMME) using Network Function Virtualization (NFV), ii) prioritize the vMME processing of H2H signaling messages by means of priority queues, and iii) use the DSCP field to identify priorities. The results show that, by increasing the number of NFV instances, the vMME capacity can be raised to manage the massive M2M SDT requests. Additionally, they show that the delay increase of H2H control plane procedures, caused by M2M communications, can be mitigated. Therefore, we conclude that our solution eases the deployment of massive M2M communications in future mobile networks.**

***Index Terms*—NFV; 5G; LTE; Machine-to-Machine.**

I. Introduction

The foreseen increase of M2M communications brings a new signaling and data burden to mobile networks. In LTE, the transmission of data from an idle User Equipment (UE) requires the use of the Service Request procedure to allocate UE's network resources. This procedure implies the download of the UE's context to the eNodeB (eNB) and the establishment of the bearers. Unfortunately, most of M2M communications involve small and occasional data transmissions. This leads to numerous release and reallocation resource procedures which create an excessive increase of signaling load. In the present paper we concentrate on massive and delay tolerant M2M communications that transmit infrequent and small data packets.

One efficient option to convey this type of data packets is SDT, a dedicated procedure with an optimized sequence of LTE messages [1]. SDT uses the pre-established Non Access Stratum (NAS) security context to transfer one IP packet as NAS signaling without establishing Radio Resource Connection (RRC) security. At first, the UE and the eNB establish the RRC connection to send the small uplink data onto the initial NAS uplink message to the MME. Then, the MME uses the UE security context previously stored to authenticate and decrypt the message, and forms the GTP-U (GPRS Tunneling Protocol - User data) packet with the information obtained, to send it to the Serving Gateway (S-GW), as shown in Figure 1.

The adoption of the SDT procedure to convey packet data transmissions from M2M communications would imply a massive increase of the signaling load processing. The Radio Access Network (RAN) will experiment a lack of radio resources due to the large number of simultaneous UEs trying to establish the RRC connection with the eNB [2]. In the core, where we focus on this paper, the MME's capacity will need to be increased to handle new functionalities imposed [1]. This, combined with the current high exposition of the MME to signaling in LTE [3], and the fixed capacity of current core LTE hardware-based infrastructure, can limit the scalability of the SDT solution.

To overcome this limitation, NFV provides a novel framework to deploy network services onto virtualized servers. NFV

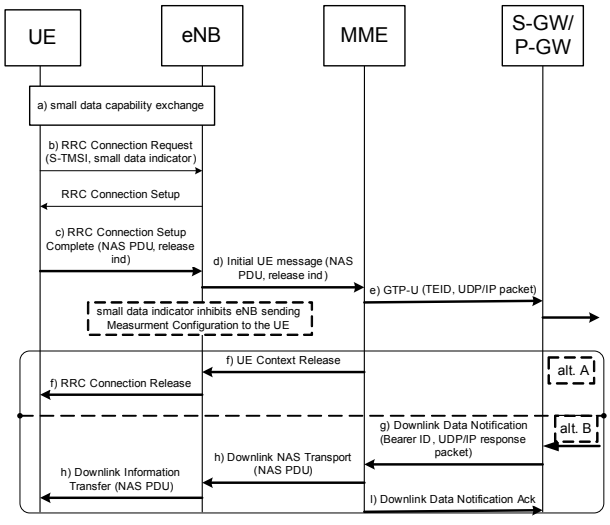

Figure 1. SDT procedure sequence [1].

benefits include, among others, reduced CAPEX and OPEX investments, openness of platforms, scalability and flexibility or shorter development cycles [4].

In this paper, we propose a new solution to mitigating the incurred signaling overload on the MME. The solution is composed of three points. The first point consists on replacing conventional hardware dedicated MME entities by NFV instances, called virtualized MME (vMME). Our results show that, by increasing the number of NFV instances, the vMME capacity can be raised to manage the massive M2M SDT requests. However, our results also show that, as the vMME response time is equal for H2H and M2M, the addition of more vMME NFV instances cannot always avoid the rise of the vMME response time for H2H procedures.

To address this issue, we also propose the following two points. The second point consists on prioritizing the vMME processing of H2H signaling messages over signaling messages of delay tolerant M2M communications by means of priority queues in the NFV instances. The third point consists of using the Differentiated Services Code Point (DSCP) classes to identify the priority of the signaling packets in the control plane. Our results show that the increase delay experienced by H2H signaling traffic, when M2M communications are included, can be alleviated by adding priorities in the control plane, at the expense of decrease M2M signaling priority, which does not imply a critical penalty for the M2M delay tolerant applications considered here.

The paper is organized as follows. Section II presents the system model. Section III describes the proposed signaling management approach. In Section IV, we show the results of the simulations. Finally, Section V draws the main conclusions of the paper.

II. System Model

We consider a LTE network, with a MME, which handles UEs control procedures requests. We assume two types of communications: H2H and M2M. The H2H UEs have sessions, which consist of activity periods separated by readings time periods. During activity periods, the H2H UE generates traffic, according to the UE's application running. For M2M communications, we consider low cost/low power consumption massive M2M communications, which we assume that send occasional and small data transmissions, and that are delay tolerant [5]. For simplicity, we consider only two types of M2M UEs: M2M high priority (HP) devices and M2M low priority (LP). This could be generalized for more types of M2M UE devices.

The H2H and M2M UEs data transmissions trigger control procedures in the network. Each control procedure involves several signaling messages between different control plane entities. From all control procedures of LTE, we focus on the ones which generate more signaling load on MME entities [6], see Table I. For each procedure and message, we model the processing tasks to be performed by the MME. We assume that M2M UEs small data transmissions are handled by SDT procedure, as shown in Figure 1. For simplicity, we focus on M2M uplink small data transmissions, since SDT procedure is similar in downlink transmissions. We assume H2H UEs move following a fluid-flow mobility model, while M2M UEs are stationary devices.

TABLE I. Considered control procedures

Com. Type	Control procedure	MME pkts processed	Used to
H2H	UE Triggered Service request	3	Send new data from the idle UE to the network and the UE does not have available resources.
	eNB Triggered S1 Release	3	Release UE's resources due to its inactivity. UE's state changes from connected to idle.
	X2-Based Handover	2	Switch the bearers end point from the source to the target eNB due to UE's mobility.
M2M	SDT procedure	1	Send small data packets from the idle M2M UE to the network and the M2M UE does not have available resources.

III. Proposal

Our solution is composed by three main points, explained in the following subsections.

A. Virtualized MME

The first point of our proposal consists of replacing hardware dedicated MME entity by virtualized NFV instances of MME and scale the number of instances according to the MME load. Our solution is based on the 1:N architecture extracted from [7], represented in Figure 2. This mapping option is based on the web services paradigm and decomposes each LTE core entity into multiple elements, which combined form a virtual component pool. These elements are classified in three types: i) the *front end* (FE), which is responsible of the communication between entities, ii) a stateless virtual component (W), which implements the virtualized network functions, and iii) the *state database* (SDB), which stores all UE's session state and allows a stateless design. External entities will see the virtual component pool as a single node. This enables scale out/in of elements of the pool without impacting other nodes. However, synchronization issues appears due to the communication between the SDB and the different virtual elements inside the entity, which can be solved serializing the access to the SDB, or between different nodes of the core to perform the control procedure, which could increase processing delay [7].

We model the architecture of the virtualized MME as shown in Figure 3. This model is based on [8][9] and it is composed of the following entities:

- Arrival process of signaling messages: H2H or M2M devices which generate traffic that triggers control procedures requests in the network. The signaling messages

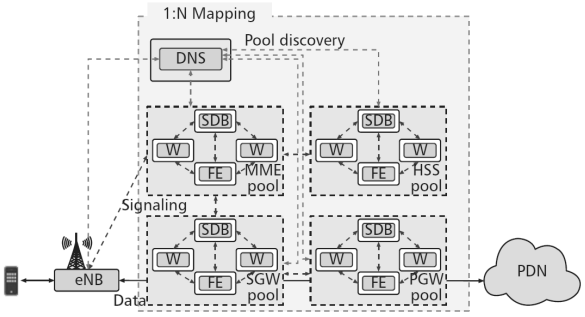

Figure 2. Architecture reference model for 1:N mapping [7].

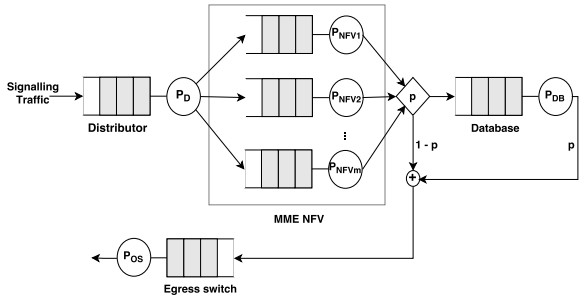

Figure 3. Virtualized MME model [8].

needed to perform these procedures are processed by vMME NFV instances.

- Distributor: Acts as a load balancer between vMME NFV instances. It distributes signaling messages depending on the average workload of each instance.
- Database: Shared database for vMME NFV instances which is accessed during each transaction. The database stores protocol and UE's state.
- vMME NFV instances: NFV instances which virtualize MME functionalities. We suppose that vMME NFV instances are identical. Each control procedure needs a different number of messages, which can involve other core entities not considered here to perform it. To improve NFV processing, the control procedures are splitted into request and response transactions. The protocol and UE's context is kept in the shared database. This allows the vMME NFV instances to retake the state of a procedure after the reception of a new signaling message and continue with it.
- Egress switch: Signaling messages output switch.

We model the distributor, the shared database and the egress switch as single processor queues, and the vMME NFV instances as a $M/G/m$ queueing systems.

We denote S the service time needed for each vMME NFV instance to process the signaling message. S is a random variable that depends on the transactions needed to process the message. The average service time for the messages of the procedures in Table I are extracted from [8]. For SDT procedure, we assume an average service time of $1.05 \cdot 10^{-4} s$. The shared database is accessed during each transaction with a probability p, as we consider every request processed by the MME will need an access to the shared database, $p = 1.0$.

Let us define the mean vMME response time $\overline{T}$ as the time required by the vMME to process a message and generate the corresponding reply. The mean vMME response time is composed of several factors: $\overline{T}_D$ denote the mean response time of the distributor node, $\overline{T}_{NFV}$ denote the mean response time of vMME NFV instances, $\overline{T}_{DB}$ denote the processing time of the shared database and $\overline{T}_{OS}$ denote the egress switch node processing time. So, $\overline{T}$ can be calculated as

$$\overline{T} = \overline{T}_D + \overline{T}_{NFV} + \overline{T}_{DB} + \overline{T}_{OS} \quad (1)$$

In order to scale the capacity of the vMME according to the load it has to process, we assume that the number of vMME NFV instances m, used as a dimensioning criterion in our results, is selected as expressed in (2), where $\overline{T}_{max}$ represents the maximum permitted mean vMME response time

$$m = \min\{M : \overline{T} \leq \overline{T}_{max}, M \in \mathbb{N}\} \quad (2)$$

An increase of signaling load on the MME caused by M2M traffic is to be compensated with an increase number of vMME NFV instances. However, $\overline{T}$ is equal for H2H and M2M, which implies that in certain situations, the addition of more vMME NFV instances cannot avoid the rise of $\overline{T}$ compared to scenarios without M2M traffic involved.

B. Priority Queue Discipline

We propose to prioritize H2H signaling messages over M2M signaling messages. The goal is mitigating the rise of the mean vMME response time suffered by H2H procedures due to the signaling overload generated by massive M2M communications. For this purpose, we propose to organize the signaling messages received by the vMME through non-preemptive priority queues inside vMME NFV instances. Messages belonging to same priority obey the first-come first-served discipline. Then, signaling messages with higher priority are served in the vMME NFV instance before others with low priority. The corresponding vMME model is represented in Figure 4.

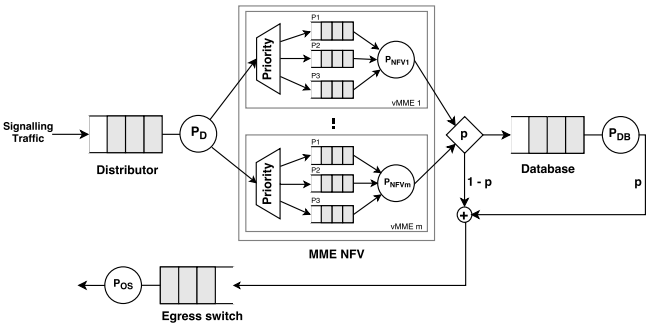

Figure 4. Proposed vMME system model.

C. Priority Management

In LTE, signaling messages between a UE and a MME are secured with NAS security context. To transfer these signaling messages over the radio interface, the RRC protocol is used between the UE and the eNB. When a UE wants to send a NAS signaling message to the MME, the message is delivered to the eNB as included in a RRC signaling message. Then, the eNB sends the NAS signaling message contained in a S1AP signaling message to the MME. Figure 5 shows the control plane protocol stacks for mentioned LTE entities. As the eNB cannot know the content of a NAS message, which holds useful information to sort signaling messages sent to the MME, we propose to use RRC Establishment cause in the eNB to discern signaling messages priorities.

Current signaling traffic over eNB-MME interface is marked as high strict priority [11]. Therefore, it is mapped to the Expedited Forwarding (EF) class in the DSCP field of the IP packet transporting the signaling message. As all signaling traffic is marked equally, the vMME cannot apply prioritized queuing of the signaling messages before being processed by the vMME NFV instances. We propose to use the DSCP field of the IP packet transporting the signaling message to discern signaling traffic from different types of communications. This IP header field is easier to analyze by FE elements due to it is not required a deep packet inspection. By adding priorities to the signaling traffic, the vMME distributor can schedule the control messages taking into account their priority. The DSCP classes used in this paper are summarized in Table II.

Since the eNB is responsible for uplink packet marking, the eNB will mark the IP datagram of the signaling messages according to the UE's RRC Establishment cause. Specially for M2M communications, which use SDT procedure, the RRC Establishment cause reported by the M2M UE when the RRC connection is established in the SDT procedure will be analyzed by the eNB to determine the DSCP class for the M2M UE SDT signaling. For this, it will take advantage of the possible values "small data" or "low priority small data", as described in [1]. Other possible RRC Establishment cause value to differentiate priorities in the signaling messages can be "delay tolerant access", introduced within the release 10 version of the 3GPP specifications [12], and currently used if the UE has been configured for "low priority NAS signalling".

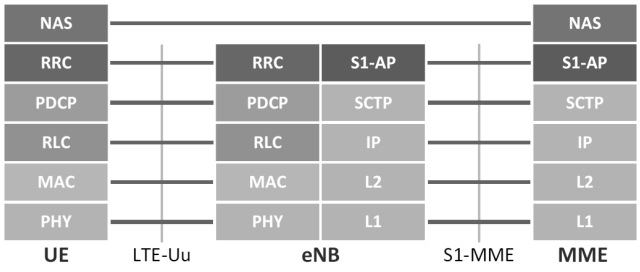

Figure 5. LTE control plane protocol stacks [10].

IV. Evaluation

In this section we evaluate the impact of using the SDT procedure on the vMME mean response time. As authors of [8], we generate procedure requests using NS-3 simulator [13]. The queue model presented in Section III is simulated using the Matlab Simulink framework.

A. Experiment Setup

We evaluate three scenarios:

- Scenario 1: M2M data traffic is not conveyed by the SDT procedure. The vMME processes signaling messages generated only by H2H UEs.
- Scenario 2: M2M data traffic is conveyed by the SDT procedure with no priorities. The vMME processes signaling messages generated by H2H UEs and by M2M UEs.
- Scenario 3: Similar to Scenario 2 but with priorities. The vMME applies the prioritization scheme presented in Section III.

1) H2H traffic models: H2H communications use three possible applications along their sessions: web browsing [14], HTTP progressive video [15] and video calling [16]. At the beginning of the session, one of these applications is selected. Web browsing application download time of a session depends on the web page size, the link data rate, and the time needed for the web browser to parse the embedded objects of the web page. HTTP progressive video application follows the Youtube traffic model, in which the download rate ranges from a initial period of high downloading rate, to a constant limited rate after this initial period. The number of downloaded video clips per session is set to follow a geometric distribution [17]. Video calling application generates a constant bit rate traffic at 1.5 Mbps during the activity period duration.

2) M2M traffic models: The M2M HP devices follow a traffic model extracted from [18], which is modeled as a Markov Modulated Poisson Process, but without taking into consideration the coordinated behavior for M2M devices. The M2M LP devices follow a traffic model based on [19], which sends infrequent report transmissions.

Scenarios 2 and 3 have three M2M devices per each H2H UE. We assume $\overline{T}_{max} = 3\ ms$. The main vMME characteristics, and details of the traffic models shown in Table III, are extracted from [8].

TABLE II. Priority treatment

Type of signaling traffic	RRC Establishment Cause	DSCP class	Priority
H2H	Mo-signaling	EF	1
M2M HP devices	Small data	AF41	2
M2M LP devices	Low priority small data	AF31	3

TABLE III. Traffic models characterization

Com. Type	Traffic Type	Parameters	Statistical Characterization
H2H ($\overline{IAST}$ = 1200 s [20])	Web browsing (HTTP) P_{app} = 0.74	Main Object Size	Truncated Lognormal Distribution: μ=15.098 σ=4.390E-5 min=100Bytes max=6MBytes
		Embedded Object Size	Truncated Lognormal Distribution: μ=6.17 σ=2.36 min=50Bytes max=2MBytes
		Number of Embedded Objects per Page	Truncated Pareto Distribution: mean=22 shape=1.1
		Parsing Time	Exponential Distribution: mean=0.13seconds
		Reading Time	Exponential Distribution: mean=30seconds
		Number of pageviews per session	Geometric Distribution: p=0.893 mean=9.312
	HTTP progressive video P_{app} = 0.03	Video Encoding Rate	Uniform distribution with ranges: (2.5, 3.0)Mbps / (4.0,4.5)Mbps / (12.5, 16.0)Mbps / (20.0, 25.0)Mbps, for equiprobable itags: 137 / 264 / 266 / 315 respectively.
		Video Duration	Distribution extracted from [15]
		Reading Time	Exponential Distribution: mean=30seconds
		Number of videoviews per session	Geometric Distribution: p=0.6 mean=2.5
	Video calling P_{app} = 0.23	Call Holding Time	Pareto Distribution: k=-0.39 s=69.33 m=0
		Number of calls per session	Constant = 1
M2M	M2M HP	Discretization time interval	Δ_T = 1 sec
		Markov chain state transition matrix	$P = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix}$ where $p = 6.75 \times 10^{-5}$ and $q = 1.47 \times 10^{-4}$
		Markov chain state rates	λ_1 = 0.0015 packets/s; λ_2 = 0.065 packets/s
		Packet Size	100 b
	M2M LP	Mean arrival rate	Poisson Distribution: λ = 0.0167 packets/s
		Packet Size	8 b

B. Experimental Results

To show the impact of the inclusion of M2M communications, Figure 6 depicts the mean vMME response time versus the number of H2H UEs for Scenarios 1 and 2. According to Figure 6, the mean vMME response time increases exponentially with the number of H2H UEs. When $\overline{T} = \overline{T}_{max}$, a new vMME NFV instance is added to the system, represented as a new curve. The results for Scenario 2 show that, by increasing m, the vMME's capacity rises to manage the massive M2M SDT requests. However, as $\overline{T}$ is equal for H2H and M2M UEs, there are some ranges where the addition of more vMME NFV instances cannot avoid the rise of $\overline{T}$ compared to Scenario 1 in which the M2M traffic is not involved.

Figure 7 depicts the mean vMME response time versus the number of H2H UEs for Scenarios 1 and 3. For almost the entire considered range of the number of H2H UEs, the mean vMME response time of H2H signaling messages in Scenario 3 is lower than in Scenario 1. That is, for almost the entire considered range of the number of H2H UEs, the proposed prioritized treatment of the signaling messages manages to prevent the increase of the mean vMME response time in H2H signaling traffic caused by the processing of the M2M traffic. Furthermore, this prioritized treatment allow H2H UEs and M2M HP devices signaling traffic to reduce their exponential signaling delay growth, at the expense of increase M2M LP devices signaling traffic delay, which reach a mean value of 9.65 ms. For delay tolerant M2M applications, this assumed increase of the mean vMME response time for M2M LP devices signaling traffic does not imply a critical penalty.

V. Conclusion and Future Work

In this paper we propose a new approach to handle the foreseen increase of signaling traffic in MME entities due

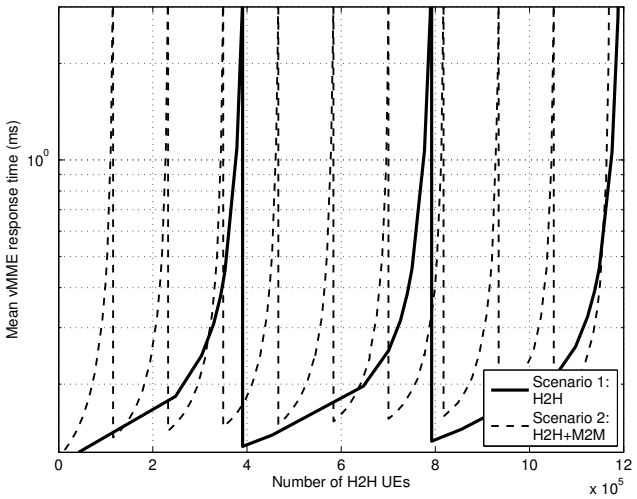

Figure 6. vMME response time in Scenarios 1 and 2 (three M2M devices per each H2H UE).

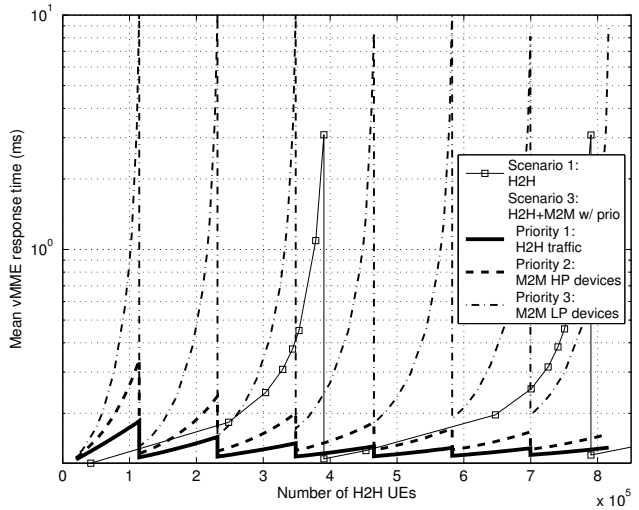

Figure 7. vMME response time in Scenarios 1 and 3 (three M2M devices per each H2H UE).

to massive M2M communications deployment, with no significant penalty in H2H. Particularly, we propose to replace conventional hardware dedicated MME entities by NFV (vMME) instances, as well as to prioritize the control plane signaling traffic with different DSCP classes. The reported results have shown that giving priority to H2H traffic can mitigate the increase delay experienced by H2H signaling traffic in H2H and M2M scenarios when delay tolerant M2M communications are included. Therefore, we can conclude that the proposed solution facilitates the massive deployment of M2M communications in future mobile networks.

For the future work, we intend to incorporate further LTE entities to the model. Apart from that, it could be interesting to analyze priorities with bound queues, or possible NFV overheads in the vMME instances proposed.

ACKNOWLEDGMENT

This work is partially supported by the Spanish Ministry of Economy and Competitiveness (project TIN2013-46223-P).

REFERENCES

[1] *Study on Machine-Type Communications (MTC) and other mobile data applications communications enhancements*, 3GPP TR 23.887 Rel 12 v12.0.0, 2013.

[2] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "Reduced M2M Signaling Communications in 3GPP LTE and Future 5G Cellular Networks," Accepted at Wireless Days Conference 2016, March 2016.

[3] "Signaling is growing 50% faster than data traffic," White Paper, Nokia Siemens Networks, 2012.

[4] ETSI, "Network Function Virtualization: An Introduction, Benefits, Enablers, Challenges, & Call for Action," 2012. [Online]. Available: portal.etsi.org/NFV/NFV_White_Paper.pdf

[5] *Service requirements for Machine-Type Communications (MTC)*, 3GPP TS 22.368 Rel 13 v13.1.0, 2014.

[6] Alcatel-Lucent, "Managing lte core network signaling," Application note, 2014.

[7] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, "EASE: EPC as a service to ease mobile core network deployment over cloud," *Network, IEEE*, vol. 29, no. 2, pp. 78–88, 2015.

[8] J. Prados-Garzon, J. J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado, and J. M. Lopez-Soler, "Latency Evaluation of a Virtualized MME," Accepted at Wireless Days Conference 2016, March 2016. [Online]. Available: http://arxiv.org/pdf/1512.02910v1.pdf

[9] J. Vilaplana, F. Solsona, I. Teixidó, J. Mateo, F. Abella, and J. Rius, "A queuing theory model for cloud computing," *The Journal of Supercomputing*, vol. 69, no. 1, pp. 492–507, July 2014.

[10] N. C. Group, "LTE Network Architecture: Basic," Netmanias Technical Document, July 2013.

[11] E. M. Metsala and J. Salmelin, Eds., *LTE Backhaul: Planning and Optimization*. Wiley, 2015.

[12] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification*, 3GPP TS 36.331 Rel12, 2015.

[13] G. Riley and T. Henderson, *The ns-3 Network Simulator*. Springer Berlin Heidelberg, 2010, pp. 15–34.

[14] NGMN, "NGMN Radio Access Performance Evaluation Methodology," NGMN Alliance, Tech. Rep., 2008.

[15] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Analysis and modelling of youtube traffic," *Transactions on Emerging Telecommunications Technologies*, vol. 23, pp. 360–377, June 2012.

[16] T. D. Dang, B. Sonkoly, and S. Molnar, "Fractal analysis and modeling of voip traffic," in *11th International Telecommunications Network Strategy and Planning Symposium. NETWORKS 2004*. IEEE, June 2004, pp. 123–130.

[17] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Characterizing user sessions on Youtube," *SPIE 6818, Multimedia Computing and Networking 2008, IEEE*, vol. 6818.

[18] C. Anton-Haro and M. Dohler, *Machine-to-Machine (M2M) Communications: Architecture, Performance and Applications*. Woodhead Publishing, 2015.

[19] *Machine-to-Machine (M2M) Evaluation Methodology Document (EMD)*, IEEE 802.16p-11/0014, May 2011.

[20] I. Tsompanidis, A. H. Zahran, and C. J. Sreenan, "Mobile network traffic: a user behaviour model," in *Wireless and Mobile Networking Conference (WMNC), 2014 7th IFIP*. IEEE, May 2014, pp. 1–8.