Paper

Network Function Virtualization: Mitigating the Impact of VoLTE on the Policy and the Charging System

Youness Jouihri[1], Zouhair Guennoun[1], Youssef Chagh[1], and Driss Zahi[2]

[1] *Electronics and Communications Laboratory (LEC), Mohammadia School of Engineering (EMI), Rabat Mohammed V University (UM5R), Rabat, Morocco*
[2] *Gregor Research Lab, Pantheon-Sorbonne University, Paris, France*

Abstract—Needless to say that telecommunications' operators are showing increasing interests toward solving the dilemma of optimizing network resources while offering state-of-the-art quality of service. Recently, operators have shown an increasing interest to investigate solutions for better control on network utilization, service usage and monetization. They also noticed a significant growth in Diameter signaling and more specifically in signaling related to policy management. A massive introduction of Voice over LTE (VoLTE) service will have a significant impact on the systems handling policy signaling, as VoLTE will reshape the landscape of Long Term Evolution (LTE) policies and completely change the way policy management works. However, this massive approach is meant to provide significant competitive advantages for operators offering LTE services and still require circuit-switched network to provide voice service. The biggest challenge for those operators is to find an appropriate solution, scalable enough to handle the unpredictable growth of Diameter signaling. In this paper, a model, based on Network Function Virtualization (NFV) technology is proposed, able to address the challenges of massively introducing VoLTE, without impacting existing services and without jeopardizing current revenues. In presented approach, the standard VoLTE call flows, referenced user's behavior and latest experiments' results on NFV technology are used.

Keywords—NFV, policy diameter signaling, policy management, VoLTE.

1. Introduction

Maintaining the profitability of services and keeping a high level of customer satisfaction is becoming a challenging situation after the drastic growth in data services. The intensive deployments of LTE networks and the high penetration of smart phones are the major drivers for such growth.

Many research studies have shown that telcos will soon reach an "end-of-profit" point if they do not find appropriate solutions to sustain their profit margin [1]. As the cost of the required network expansions to handle this growth increases faster than the generated profit. Consequently, the focus of operators is put on policy management signaling to sustain the profitability of their business and efficiently control the infrastructure.

A wide range of research studies treated the topic of handling the growth of data services [2], [3]. However, literature addressing the scalability issue of the systems handling the signaling associated with those services is difficult to find. In contrast to existing literature, the research aims to investigate the challenges faced by LTE operators while managing the LTE Diameter signaling, more specifically those related to the policy management signaling after introducing VoLTE.

This paper is structured as follows. Section 2 highlights the growth in Diameter signaling policies versus current policy management systems' limitations. In Section 3, the benefits and impacts of massive VoLTE introduction are shown. In Section 4, the challenges of using Network Function Virtualization (NFV) technology to mitigate the VoLTE introduction impact are studied. Finally, Section 5 concludes the paper.

2. Growth in Diameter Signaling Policies Versus Current Policy Management System Limitation

2.1. Growth of Diameter Signaling

Over the past few years, mobile operators reported a demand shift from voice to data services, and started investing massively to support the impressive growth in mobile data traffic. However, most of them have given less interest to the signaling growth. When comparing the global LTE Diameter signaling growth and the global IP traffic (including mobile data traffic) growth, authors find that the Diameter signaling grows at a Compound Annual Growth Rate (CAGR) of 68%, and will surpass the total global IP traffic by 2019 growing at a CAGR of 23% [4], [5] (Fig. 1).

The Diameter signaling has considerably increased. The high LTE network deployment and smartphone penetration, along with the changing subscribers' behavior were the key players in this drastic increase.

Special care should be given to the Diameter signaling volume growth, as it may overwhelm the mobile operators existing infrastructures, and jeopardize both the cost forecasted for future expansions and the end users' Quality of Service (QoS). Mobile operators should study the Diameter signaling and make their infrastructure ready to deliver future mobile services, with similar or better quality than the current one.

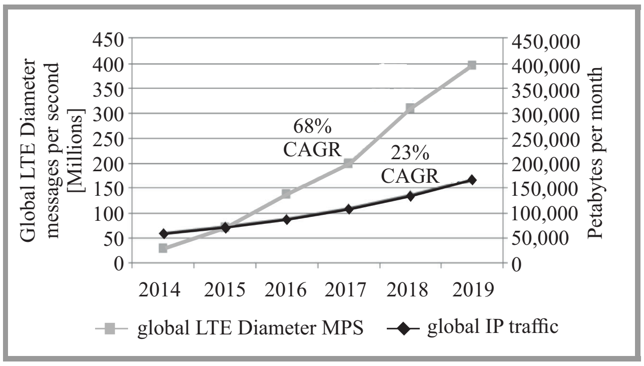


Fig. 1. LTE Diameter signaling and mobile data growth.

In an LTE network, the Diameter signaling traffic is needed to handle Mobility Management, LTE Broadcast and other signaling activities such policy management signaling, required to provide and to handle the LTE services. The policy management signaling is becoming the dominant Diameter traffic and will continue its growth (Fig. 2) [4]. As it is gaining more importance and will significantly influence future services. Operators usually rely on the use of policies for fair usage and traffic management cases. Meanwhile, they started implementing complex use cases to generate more revenues, and providing a wider range of service packages to their end users.

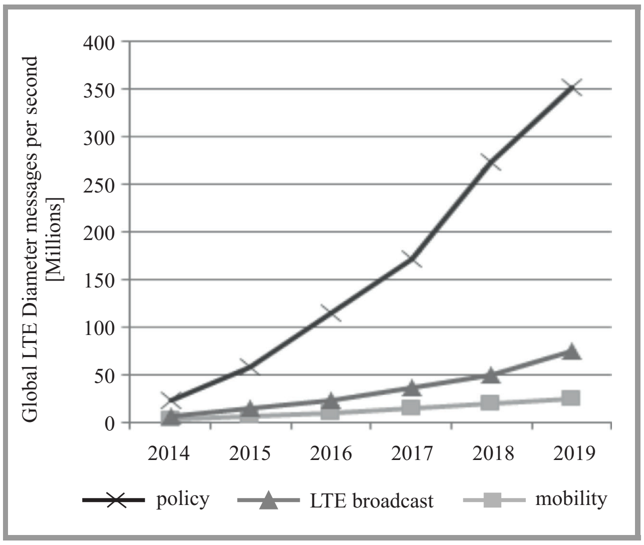


Fig. 2. Global LTE Diameter signaling by use case.

The policy management signaling growth is impacted by the mobile plans concerning service policies. Such plans are the consequence of a significant migration to smartphones, which generate approximately 20 times higher signaling and traffic [6]. Nevertheless, authors still believe that the launch of Voice over LTE service (VoLTE) will have a crucial impact on the policy management processes and growth.

2.2. Reasons and Management Challenges of Policy Signaling Growth

As highlighted earlier, mobile operators are shifting their investment from a voice-dominant towards a data-dominant infrastructure. During the voice-dominant period, the revenue growth was proportional to the traffic volume [7]. However, after moving to the data-dominant network, the data traffic is growing faster than the associated revenues (Fig. 3) [7]. This results in a gap between traffic and revenue, mainly caused by inadequate pricing for some services (e.g. the unlimited data packages), as well as the fierce competition between operators. Moreover, if operators continue to equally treat the data packet's content and enable all the services by default, Over The Top (OTT) players will take the lion's share of the revenue generated by OTT services, which causes a loss of potential revenues. If no remedy action is taken by operators, they might reach an "end of profit" point, as highlighted in many studies [1], [8].

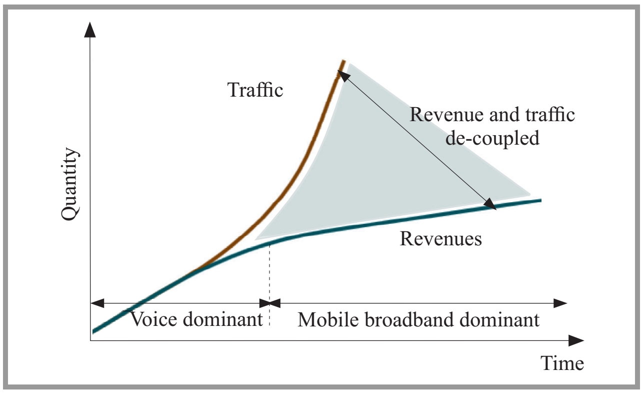


Fig. 3. Coupling of traffic and revenue in mobile networks.

This is to say that operators need techniques to decouple network costs from traffic growth, by monitoring, managing and charging accurately different types of data services. These techniques should also include the control of user's usage, and should have the ability to tackle abuses in consuming network resources. In parallel, these techniques could provide a big support to operators, if they offer the possibility to limit the congestion and enhance the QoS, without any need for capacity expansion. Operators should also investigate ways of increasing subscribers' loyalty, service uptake, and practices that can maximize the revenue even from casual consumers.

The 3rd Generation Partnership Project (3GPP), had the foresight to address these challenges and introduced the network Policy and Charging Rules Function (PCRF) and Policy and Charging Enforcement Function (PCEF), responsible for the management of network QoS, and policy and charging management solutions [9].

Most of the operators start taking control of their network by applying fair usage and traffic management policies, enabled by the PCRF and PCEF functions. However, to achieve the enhanced services' control, charging and quality, operators ended up implementing huge number of complex policies [4]. Proper policies and QoS strategies result in taking back the control over the traffic consumption and charging. Nevertheless, those strategies have the adverse effect of significantly increasing the Diameter signaling of policies, which raises new challenges in managing and executing such policies. Operators look for a centralized policy decision architecture, which enables easy provisioning and high integrity of data. However, many operators start reporting signaling overload after deploying centralized architecture as stated in [10].

Undesirably, even after investing and making a robust Policy Control and Charging (PCC) architecture with a centralized policy decision, operators would not be able to fully utilize this mechanism to efficiently manage traffic growth and services monetization. The current telecom software solutions can run only on a proprietary hardware. Scaling the policy architecture might require multiple racks, full of servers [10]. After exceeding the maximum limit of one PCRF node capacity, the introduction of a new PCRF node will need the addition of a Diameter Routing Agent (DRA), capable of performing binding. To make sure that all messages associated with a user's particular IP Connectivity Access Network (IPCAN) session are processed by the same PCRF node. This increases latency, and adds new nodes to the infrastructure, resulting in an increase of the capital and operational expenditures [11].

Operators will absurdly reach a situation, where the implemented solution generates the same problems as the ones it was designed to solve by impacting the service quality (caused by latency) and increasing the network cost.

3. Introduction of VoLTE Service: Benefits and Impacts

3.1. Advantages of Massive Migration to VoLTE Service

Most of the operators are now aware of the challenging growth of signaling, and start looking for innovative strategies to be implemented in their policy management system to prevent uncontrolled growth of Diameter signaling. But few of them took into consideration the impact of introducing VoLTE service on policy signaling considering the launch of such a service equivalent to the launch of any LTE data service.

Around 80 operators, in 47 countries, are investing in VoLTE deployments [4]. Most of them are still in a small scale deployment, as they are not yet ready to go for a massive introduction of VoLTE, and get rid of their legacy network built over several years. Moreover, currently, only a limited number of handsets are supporting this service. This adds more constraints on the massive move scenario. That is why, the impact of VoLTE service on policy system will not be noticeable at this stage of small migration.

The spectrum bandwidth consumption for a voice call in LTE network is less than the one in circuit-switched network [6]. Operators can use the spectrum resources in a more efficient manner, and will have the ability to nearly double the spectrum capacity with the use of VoLTE service [6]. Furthermore, the migration of a fair percentage of voice load, from legacy networks into LTE, frees up existing bandwidth. The freed bandwidth can be reused to increase the capacity of LTE network and paves the way for technical innovations, focusing in providing very high data speeds to the end users, such as carrier aggregation techniques. The suggestion of "mind commerce" in [6] for a technical case to be made for a massive introduction of VoLTE might attract many operators, especially those having limitations in spectrum resources. To achieve that, operators are making their VoLTE core network, mainly their IP Multimedia Subsystem (IMS), ready to support this massive move. The IMS is an all IP control platform for both fixed and mobile networks, consisting of several logical functions such as Call Session Control Function (CSCF) that controls voice sessions and Application Server (AS) that controls subscribers' services. Detailed VoLTE architecture, including nodes and interfaces, is presented in Fig. 5.

In parallel, handset suppliers are supporting this massive migration scenario, by driving a rapid erosion of average selling prices of smartphones. This is likely to accelerate user migration from basic and feature phones to smartphones [12].

3.2. Impact of VoLTE Introduction on the Policy Management System

Introducing VoLTE in an LTE network will completely change the way policy management works. Unlike the LTE data services, where the policy management handles only one session with PCEF function, the VoLTE needs to correlate two sessions, one with the PCEF functions, and another one with the application functions [11]. For an appropriate network dimensioning, operators planning to launch VoLTE service need to ensure that their infrastructure can properly handle this additional service, by supporting the associated signaling load and generated traffic. Those signaling and traffic loads are impacted by not only the number of subscribers, but also by the subscribers' behavior, their devices, and the services or applications they invoke.

Recall that the Busy Hour Call Attempts (BHCA) value is one of the key values used to design and dimension circuit-switched network [13]. During conducted analysis of the impact of VoLTE introduction in a LTE network, it was assumed that the behavior of subscribers using voice in a circuit-switched network and their BHCA values will remain the same once they move to VoLTE service.

An LTE subscriber provisioned to use VoLTE service needs two default bearers, one for voice and another for data. These default bearers will remain established as long as the subscriber is attached to LTE network. Besides this, a dedicated bearer will be established in the case of mobile originated call and mobile terminated call. This bearer will be released after the end of the call [14].

Handling these additional bearers will be the real challenge that VoLTE service will add to LTE network. Doubling the use of default bearers will increase the number of Diameter signaling messages required to establish or terminate these bearers. Furthermore, the control of dedicated bearers requires a higher number of Diameter signaling messages comparing to the one used to handle default bearers. Consequently, operators should carefully study the impact of VoLTE service introduction, and should not consider it as a normal data service launch.

3.3. A Case Study to Highlight the Impact of VoLTE on Diameter Signaling

To emphasize the severity of Diameter signaling increase, a use case was simulated, with a proposed model for VoLTE subscribers increase (N_v), and another to estimate the forecast of subscribers using LTE data-only (N_d). The proposed subscribers' forecast is close to a massive VoLTE migration scenario.

The average number of Diameter signaling transactions per second during the busy hour at year x, $S_{total}(x)$, generated by LTE subscribers (VoLTE and LTE data-only) can be calculated as:

$$S_{total}(x) = aN_v(x) + bN_d(x)\,, \qquad (1)$$

where $N_v(x)$ is the proposed forecast of VoLTE subscribers' growth at year x, $N_d(x)$ is the forecast of LTE data-only subscribers growth at year x.

The factors a and b are defined as:

$$\begin{aligned} a &= \frac{1}{3600}\big[p_2(t_2+t_1+t_5+t_6)+p_5t_8+ \\ &\quad + p_3(t_4+t_9)+p_1(t_3+p_4t_{10})\big] \\ b &= \frac{1}{3600}\big[p_2(t_1+t_5+t_6+t_7)+p_3t_4\big]\,. \end{aligned}$$

p_1, p_2, p_3, p_4 and p_5 are parameters used to model the users' behavior and LTE coverage. Their descriptions, values and references used to extract those values can be found in Table 1.

$t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9$ and t_{10} are parameters representing the transactions required to handle the default and dedicated bearers' activities (i.e. establishment, modification, termination), between the different LTE core elements (Table 1). The standard 3GPP call flows is used [15]–[19] to calculate these values.

The $N_v(x)$ was simulated using an "S" (sigmoid) curve shape:

$$N_v(x) = \frac{S_{\max}}{1+e^{-\lambda_1(x-x_0)}}\,,$$

where $S_{\max}, x_0$ and λ_1 are respectively the maximum number of subscribers, the inflection year (convexity change in subscriber forecast), and the slope parameter for subscriber forecast. For our case, we use $S_{\max} = 20$ millions, $x_0 = 3$ years and $\lambda_1 = 2.5$, to represent VoLTE subscribers increase following an "S" curve going from 0 to 20 millions over 6 years (Table 2). The $S_{\max}$ value is an estimation of number of subscribers, excluding LTE data-only subscribers, in large operators in Middle East and North Africa (MENA) by first quarter of year 2021 [20].

The $N_d(x)$ was simulated using the sum of two normal distributions, $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, to respectively model the impact of VoLTE introduction on LTE data-only subscribers, and their forecasted growth:

$$\begin{aligned} N_d(x) &= n_1 N(\mu_1, \sigma_1) + n_2 N(\mu_2, \sigma_2)\,, \\ N_d(x) &= \frac{n_1}{\sigma_1\sqrt{2\pi}}\,\mathrm{e}^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}} + \frac{n_2}{\sigma_2\sqrt{2\pi}}\,\mathrm{e}^{\frac{-(x-\mu_2)^2}{2\sigma_2^2}}\,, \end{aligned}$$

where n_1 and n_2 are calculated based on the initial and target values of LTE data-only subscriber's growth. In this case: $n_1 = 25$, $n_2 = 8$, $\mu_1 = 6$, $\mu_2 = -0.1$, $\sigma_1 = 2$ and $\sigma_2 = 1.6$. These values lead to an evolution of the number of LTE data-only subscribers', starting from an initial value of around 2.05 million subscribers [21], where VoLTE was not yet launched in large operators in MENA. This initial value includes subscribers using voice service in circuit-switched network. Such a number will then decrease until it reaches the value of 1.5 million after almost three years. This is justified by the fact that many of those subscribers will start using VoLTE service instead of circuit-switched voice. The authors also believe that during the first three years, the number of LTE data-only migrated to VoLTE will be higher than the forecasted growth of LTE data-only subscribers. In the subsequent years, the number of LTE data-only subscribers will increase as per the forecasted growth until it reaches 4.99 million in year 2021 as the growth in VoLTE subscribers will slow down (Table 2).

The Eq. (1) was calculated based on the sum of signaling Diameter transactions generated by the activities associated to the below functions, where the first element of the function represents the number of transactions related to the Internet default bearer, the second corresponds to the number of transactions pertaining to VoLTE default bearer. The third is association with the number of VoLTE dedicated bearer transactions:

- $F_1(x)$ – number of Diameter transactions to establish default bearers (Internet and VoLTE):

$$\begin{aligned} F_1(x) &= \frac{1}{3600}\big[F_{11} \quad F_{12} \quad F_{13}\big]\,, \\ F_1(x) &= \frac{p_2}{3600}\big[t_1(N_v(x)+N_d(x)) \quad t_2N_v(x) \quad 0\big]\,; \end{aligned}$$

- $F_2(x)$ – number of Diameter transactions to authenticate an LTE subscriber (VoLTE and LTE data-only):

$$\begin{aligned} F_2(x) &= \frac{1}{3600}\big[F_{21} \quad F_{22} \quad F_{23}\big]\,, \\ F_2(x) &= \frac{p_2}{3600}\big[t_5N_d(x) \quad t_5N_v(x) \quad 0\big]\,; \end{aligned}$$

Table 1
Signaling Diameter activities – parameter values and traffic quantification

Network attach parameters	Symbol	Value	References
Number of network attach requests per subscriber during busy hour (BH)	p_2	0.75	[23]
Number of requests to refresh IMS registration per VoLTE subscriber (BH)	p_5	1	[24]
Number of Diameter transactions to establish Internet default bearer (Gx interface) – all listed interfaces are presented in Fig. 5	t_1	1	[15]
Number of Diameter transactions to establish VoLTE default bearer (Gx interface)	t_2	1	[15]
Number of Diameter transactions to authenticate an LTE subscriber (S6a interface)	t_5	2	[16]
Number of Diameter transactions to download spending limit for an LTE subscriber (Sy interface)	t_6	1	[15]
Number of Diameter transactions to charge LTE data service (Gy interface)	t_7	2	[17]
Number of Diameter transactions to register VoLTE subscriber in the IMS (Cx, Sh interfaces)	t_8	6	[17] [19]
VoLTE call parameters description			
Busy hour call attempts	p_1	2	Estimated[1]
Percentage of prepaid VoLTE subscribers	p_4	0.8	[22]
Number of Diameter transactions to establish and release dedicated bearer (Gx and Rx interfaces)	t_3	7	[15]
Number of Diameter transactions to charge a prepaid VoLTE call (Ro interface)	t_{10}	2	[18]
Network detach parameters description			
Number of network detach requests per subscriber (BH)	p_3	0.75	[23]
Number of Diameter transactions to release a default bearer (Gx interface)	t_4	1	[15]
Number of Diameter transactions to deregister a VoLTE subscriber from IMS (Cx interface)	t_9	1	[19]

[1] Estimated similarly to many dimensioning use cases, where the initial value was set to 2 and increased yearly. Authors prefer to keep it constant (conservative forecast) as this value influences a lot the Diameter growth.

- $F_3(x)$ – number of transactions to download the spending limit report from on-line charging system:

$$F_3(x) = \frac{1}{3600}\left[F_{31} \quad F_{32} \quad F_{33}\right],$$

$$F_3(x) = \frac{p_2}{3600}\left[t_6 N_d(x) \quad t_6 N_v(x) \quad 0\right];$$

- $F_4(x)$ – number of Diameter transactions to charge LTE data services:

$$F_4(x) = \frac{1}{3600}\left[F_{41} \quad F_{42} \quad F_{43}\right],$$

$$F_4(x) = \frac{p_2}{3600}\left[t_7 N_d(x) \quad t_7 N_v(x) \quad 0\right];$$

- $F_5(x)$ – number of Diameter transactions to register a VoLTE subscriber in IMS:

$$F_5(x) = \frac{1}{3600}\left[F_{51} \quad F_{52} \quad F_{53}\right]$$

$$F_5(x) = \frac{p_5}{3600}\left[0 \quad t_8 N_v(x) \quad 0\right]$$

- $F_6(x)$ – number of Diameter transactions to release a default bearer (VoLTE and Internet):

$$F_6(x) = \frac{1}{3600}\left[F_{61} \quad F_{62} \quad F_{63}\right],$$

$$F_6(x) = \frac{p_3}{3600}\left[t_4 N_d(x) \quad t_4 N_v(x) \quad 0\right];$$

- $F_7(x)$ – number of Diameter transactions to deregister a VoLTE subscriber from IMS:

$$F_7(x) = \frac{1}{3600}\left[F_{71} \quad F_{72} \quad F_{73}\right],$$

$$F_7(x) = \frac{p_3}{3600}\left[0 \quad t_9 N_v(x) \quad 0\right];$$

- $F_8(x)$ – number of Diameter transactions to establish and release dedicated bearer:

$$F_8(x) = \frac{1}{3600}\left[F_{81} \quad F_{82} \quad F_{83}\right],$$

$$F_8(x) = \frac{p_1}{3600}\left[0 \quad 0 \quad t_3 N_v(x)\right];$$

- $F_9(x)$ – number of Diameter transactions to charge a prepaid VoLTE call:

Table 2
Increase of VoLTE subscribers in years 2015–2021 with used parameters description

Symbol	Description	2015	2016	2017	2018	2019	2020	2021
N_d	Number of subscribers using LTE data-only [million]	2.05	1.79	1.52	1.92	3.10	4.41	4.99
N_v	Number of VoLTE subscribers [million]	0.00	0.30	3.96	16.04	19.70	19.98	20.00
	Average (av.) number (no.) of network attach Diameter TPS (BH)	2558	3166	14261	52542	65451	67960	68732
	Av. no. of VoLTE calls Diameter TPS (BH)	0	1412	18902	76653	94144	95468	95550
	Av. no. of Network detach Diameter TPS (BH)	426	497	1965	7086	8856	9245	9372
	Av. no. of policy Diameter TPS (BH)	1279	2648	21390	84095	103744	105997	106445
S_{total}	Av. no. of LTE subscribers Diameter TPS (BH)	2984	5074	35128	136281	168451	172673	173654
S_{NFV}	Av. no. of Diameter TPS (BH) moved to NFV architecture (to be used in Subsection 4.2)	0	1773	23738	96262	118227	119891	119993

$$F_9(x) = \frac{1}{3600}\begin{bmatrix}F_{91} & F_{92} & F_{93}\end{bmatrix},$$

$$F_9(x) = \frac{p_1}{3600}\begin{bmatrix}0 & 0 & p_4 t_{10} N_v(x)\end{bmatrix}.$$

$S_{total}(x)$ calculated in Eq. (1) can also be written as:

$$S_{total}(x) = \frac{1}{3600}\sum_{j=1}^{j=3}\sum_{i=1}^{i=9} F_{ij}.$$

In Table 1, the values and references of all the parameters used are listed and the $S_{total}(x)$ calculated values from year 2015 to year 2021 are shown.

On the simulated use case, the transactions generated by managing the dedicated bearer's associated to the establishment and the termination of voice call activities are calculated. Also the number of transactions required to handle default bearers are computed, mainly related to attach and detach activities to or from network and influenced by the LTE coverage or subscribers' behaviors. Finally, it is estimated that 80% of the subscribers are prepaid, extracted from the GSMA worldwide prepaid percentage report in [22], and the transactions required to charge them are computed.

The average number of LTE subscribers Diameter transactions per second during busy hour, presented in Fig. 4, will go beyond hundred thousand transactions per second after almost three years from the launch of VoLTE. Around 62% of those transactions are related to Diameter signaling policy. Consequently, despite excluding the impact of the policy techniques, introducing VoLTE in a massive migration scenario will significantly influence the Diameter signaling growth in general and policy signaling management in particular.

By analyzing the signaling growth, for an operator expecting to serve around 25 million subscribers in year 2021, and using basic policy techniques to manage data services, one can conclude that operators having similar number of subscribers, and using complex policies or willing to use them, will not be able to go for a VoLTE massive migration, unless their policy management is scalable enough to accommodate the heavy increase in signaling. Ideally, one single PCRF node should be used to manage the complete signaling policies, in order to avoid problems, listed earlier in Subsection 2.2, of introducing another PCRF node. The current PCRF nodes use proprietary hardware, and cannot be scaled to accommodate both the expected signaling growth that is calculated in presented case, and the flexibility in implementing complex policies. The existing solutions to accommodate the policy signaling growth consist of adding new nodes, either new PCRF node or front-end servers if the PCRF solution is a front-end back-end based solution. However for all those solutions, DRA will be needed to perform binding and to ensure proper communication between network elements (authors do not suggest the addition of DRA as explained earlier in Subsection 2.2).

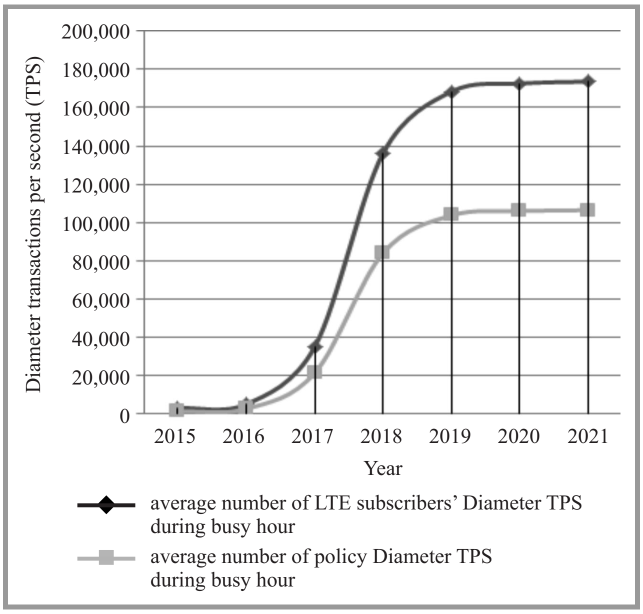


Fig. 4. Prognosed impact of VoLTE on Diameter signaling growth.

4. Challenges of Using NFV Technology to Mitigate the VoLTE Introduction Impact

4.1. Designing an Architecture Model to Overcome VoLTE Challenges

In the current telecom environment, it is not always possible to scale a single point controlling all policies' decision [10]. To have centralized policy decision architecture, able to handle the massive introduction of VoLTE service, it is necessary to find an appropriate environment to build a strong policy system. Ideally, such environment should give the possibility of unlimited scaling, to overcome the unpredictable policy Diameter signaling growth.

Achieving a flexible scaling and giving the possibility to decouple software from hardware in network nodes with a reasonable cost, will help operators to protect their profit margin by preventing expensive network expansions. Adding the possibility of "paying only for what they need, and only when they need it" [25] will be the optimal solution that operators are eager to procure.

Considering the Information Technology (IT) domain experience to solve the issue of inefficient use of servers, by inventing the concept of hardware virtualization and cloud data center; the European Telecommunications Standards Institute (ETSI) published the NFV approach to virtualize telecom network node functions [25]. The cloud environment using the NFV concept seems to be the goal of this, if we could assume that telecom and IT networks' requirements are similar. In telecom networks, it is all about carrier grade systems, where we should meet or exceed "five nines" (99.999%) availability is expected. While in the IT classic cloud, it is not a real necessity [26]. The maturity of the NFV is the biggest constraint blocking the spread of the telecom nodes' virtualization. Many of those nodes require special hardware and software solutions to run properly; and their functioning on a Commercial Off-The-Shelf (COTS) hardware might degrade their performance in contrast to their functioning on a proprietary hardware [25].

The short term virtualization strategy, should focus on network functions allowing minor performance degradation, and where the scaling of control functions overcomes the degradation that might be caused by the use of COTS hardware. Presented NFV architecture is built by analyzing the results of different researches in NFV technology, and taking into consideration the track records' outcomes of the major players in this technology. This architecture aims to mitigate the impact of massive VoLTE introduction, by the efficient use of the facilities acquired while using NFV technology, and by ensuring that the deployment of this architecture will not have any harmful impact on the existing services and infrastructures. Selecting the network functions to be used in this architecture was influenced by the analysis of previous NFV experiments, and by the study of VoLTE introduction impact on the existing network nodes.

During the Mobile World Congress 2015 (MWC 2015), Telefonica company shared many interesting experiences and findings. They confirmed that the classical cloud management can't provide a carrier grade performance. They also confirmed the strong impact of the underlying hardware on network functions' performance. Therefore, they proposed a staged NFV deployment plan [27]. Combining the analysis results of the use case presented in Subsection 3.3, and Telefonica proposals in MWC 2015, a detailed study for the policy management system virtualization is required to achieve authors' goals. Nevertheless the virtualization of any node running business critical services is not yet recommended. This means that the policy management system handling VoLTE service should be different from the one used for LTE data services. Deploying dedicated PCRF and PCEF functions for VoLTE in a virtualized architecture and adding CSCF function to this architecture, appears to be the straightforward solution to avoid all complications accompanying the introduction of VoLTE service, and uncertainty about its way of adoption. Those three functions handle the majority of the VoLTE policy signaling. Important is to highlight that virtualizing those nodes requires a proper assessment of the impact of such virtualization on their performance.

Referring back to Telefonica and other operators' experiences, the signaling functions PCRF and CSCF seem to be a good candidate for virtualization. The reason behind is that even their performance could be impacted by the use of COTS hardware. It can be overcome by scaling up the number of its control plane components. The Mobility Management Entity (MME) is also a pure signaling node, but authors prefer not to include it now, to ensure that proposed architecture's impact on the existing services, nodes and connections is marginal. Dedicating a separated NFV based Home Subscriber Server (HSS) to serve VoLTE subscribers is another interesting initiative; permitting the move of an important part of VoLTE Diameter signaling to virtualized network. But the NFV technology is not yet mature to serve critical systems, such as those handling subscriber databases, as announced by Telefonica [27]. The model will not include HSS, but still use the standard 3GPP Sh and Cx interfaces to communicate with the existing HSS (Fig. 5) will be used.

Dedicating a separate PCEF for VoLTE service requires the appropriation of the complete PGW functions, which are not limited only to signaling control. The PGW is the gateway terminating the connection towards external packet network, in addition to managing policy enforcement, packet filtration, and other control functions. Virtualizing the control part of the PGW can be treated similarly to CSCF and PCRF functions. However, virtualizing its forwarding part is not advisable, especially after Telefonica's experience, where the performance of throughput was significantly degraded using the Hypervisor of the cloud computing environment [28]. This degradation cannot be tolerated in a node requiring high packet-processing rate such as PGW. The Hypervisor is an intermediate software layer between the underlying hardware and the upper vir-

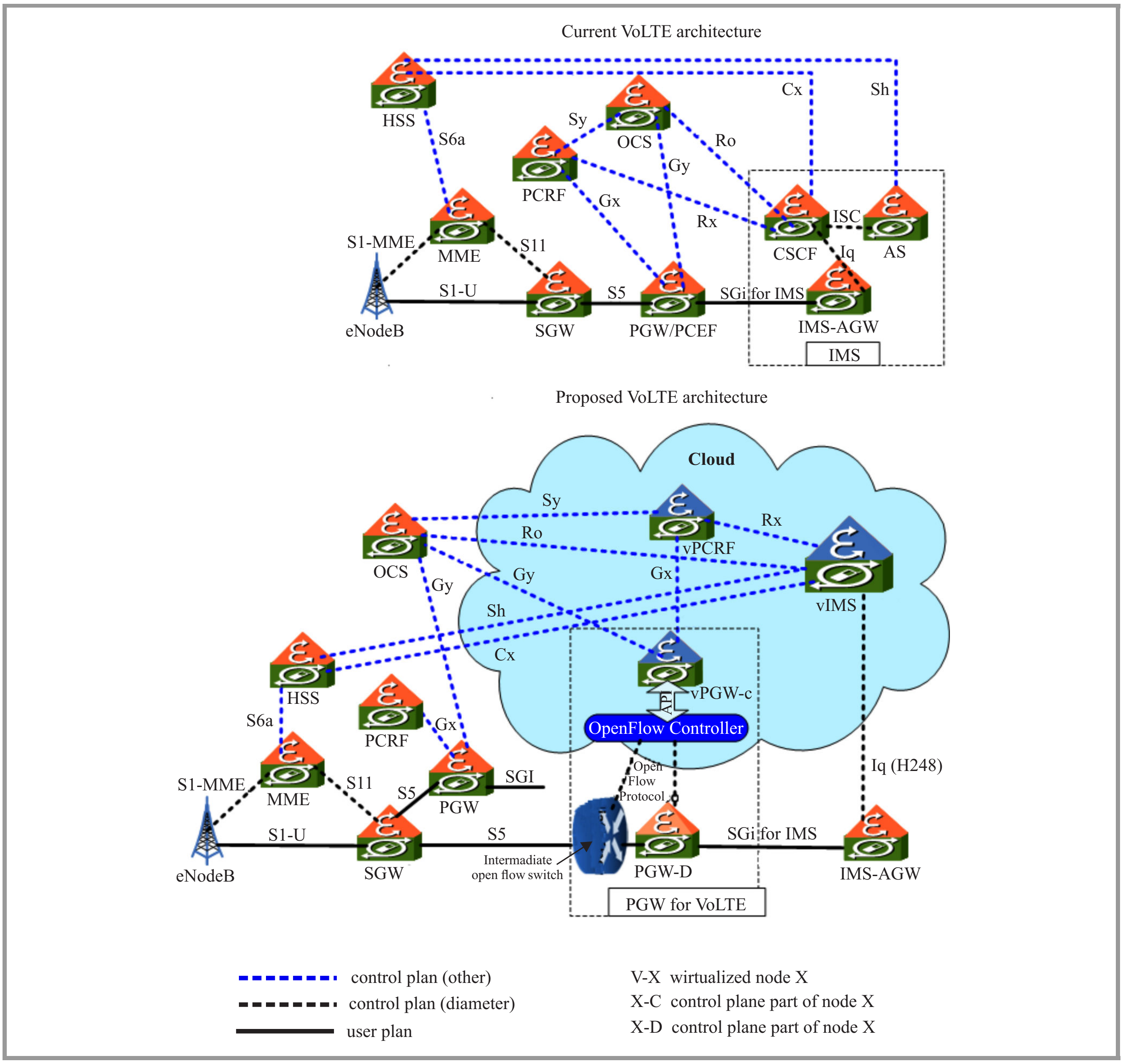


Fig. 5. Proposed network model to mitigate the impact of VoLTE

tual functions, which provides an abstraction of the network hardware, allowing the creation of virtual elements independent from the hardware layer. By bypassing this Hypervisor, Telefonica's NFV Lab has considerably improved the throughput of their virtualized nodes [28]. Such a bypass was performed using the Intel Single Root Input and Output Virtualization (SR-IOV) feature, which allows direct communication between virtual machine and hardware. Nevertheless, this change will have a major impact on the scalability of network functions [29], which is one of the strongest flexibility aspects that NFV technology provides.

Virtualizing the PGW control part allows the move of PCEF functionality to a virtualized domain. But based on Telefonica experience, the forwarding part should be kept out of NFV at this stage of NFV maturity cycle. The authors should look then for a solution to decouple the control and forwarding of the PGW to fulfill proposed requirement, pertaining to virtualizing the PGW control function. The Software Defined Networking (SDN) aims to separate the control and the forwarding in the networking domain. The NFV and SDN are independent approaches [25]. A wide range of research projects consider the SDN as a good enabler for NFV, as it allows the virtualization of the separated control plane software of the network functions [25]. The NFV can achieve the virtualization benefits without requiring SDN implementation, but both technologies complement each other and offer highly flexible network functions [25]. In [30], the authors suggested an SDN architecture to the Evolved Packet Core (EPC), by moving the control plane of the Serving Gateway (SGW) and PGW (SGW-C, PGW-C), and the MME functions into cloud environment. They replaced the data plane of core gateways by switches using OpenFlow protocol, which is an SDN protocol used to centrally control switches and traffic flows in a network where the control plane is centralized and decoupled from the physical layer. This proposal seems to be the best solution for the demanded control plane decoupling, but only the PGW decoupling part will be adopted. The SGW and the MME decoupling will not be considered, and the communication between PGW in SDN architecture and SGW will still use the standard 3GPP S5 interface (Fig. 5). In order to reduce the number of connections between proposed model and the remaining live network elements, such as connections with radio access nodes a move the PGW only to authors' model is preferred.

The IMS has many components and most of them software-based control functions [25], such as CSCF function. In this architecture, all the IMS control functions will be included. As their migration will not have any impact on the existing services and nodes, their performance degradation can be resolved by scaling them up, and they can benefit from NFV advantages. The AS is one of those IMS control functions; where the facilities that NFV environment can offer to mitigate the complexity of its implementation should be explored. Currently, adding different Application Servers to enable the big list of IMS services did not attract operators. The deployment and testing process of those services is time consuming, and require proprietary hardware. Nevertheless, operators don't have any warranty that those services will attract subscribers and generate profits. The IMS pragmatic approach was the ideal solution to revive the IMS concept, where the focus was first on enabling VoLTE, and then other services can be progressively enabled based on the requirements [31]. This pragmatic approach will work perfectly on a virtualized environment, providing the facility of rapid deployment, and consequently a shorter period to confirm the success or the failure of the deployed service [32]. In addition to reducing the investment cost and time, virtualizing the AS will allow the scaling down or the replacement of unsuccessful services. However, the successful ones will benefit from the immediate scaling up. The IMS has also a forwarding part, IMS-Access-Gateway (IMS-AGW), handling the transcoding and the media services. Similarly to the PGW forwarding function case, the IMS-AGW cannot be virtualized. Due to the lack of works proposing an SDN architecture for IMS control and forwarding functions, authors prefer to use the 3GPP architecture using Iq interface to connect the virtualized IMS control functions to IMS-AGW [33]. Figure 5 summarizes the proposed architecture and depicts the key steps mentioned above.

4.2. Applying the Proposed Model to the Massive VoLTE Migration Use Case

In presented model, the network elements are considered where scaling can overcome the degradation of their performance in NFV, and not bearing any business critical services, to move a significant part of Diameter signaling to virtualized network. The amount of LTE Diameter signaling transactions migrated to proposed model is evaluated by adding an NFV weighting factor $\alpha_i = \left[\alpha_{i1}\ \alpha_{i2}\ \alpha_{i3}\right]$ on each element of the F_i functions, described in the case study presented in Subsection 3.3.

α_{ij} is determined by the following rules:

- $\alpha_{ij} = 1$ – means 100% of transactions will be migrated to the virtualized model, if the activities happen between elements in proposed model,
- $\alpha_{ij} = 0.5$ – means 50% of transactions migrated to the virtualized model, if the activities happen between two elements one of them in the model,
- $\alpha_{ij} = 0$ – means 0% of transactions migrated to virtualized model, if the activities happen between elements out of the model.

To calculate $S_{NFV}(x)$, the average number of Diameter transactions per second during busy hour at year x moved to the virtualized architecture, the NFV weighting factor to Eq. (1) will be added:

$$S_{NFV}(x) = \frac{1}{3600}\sum_{j=1}^{j=3}\sum_{i=1}^{i=9}\alpha_{ij}F_{ij}\,,$$

$$S_{NFV}(x) = \frac{1}{3600}\begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \\ F_{41} & F_{42} & F_{43} \\ F_{51} & F_{52} & F_{53} \\ F_{61} & F_{62} & F_{63} \\ F_{71} & F_{72} & F_{73} \\ F_{81} & F_{82} & F_{83} \\ F_{91} & F_{92} & F_{93} \end{bmatrix}[\alpha]\,,$$

$$S_{NFV}(x) = \frac{1}{3600}\left(N_v(x)\begin{bmatrix} p_2t_1 & p_2t_2 & 0 \\ 0 & p_2t_5 & 0 \\ 0 & p_2t_6 & 0 \\ 0 & p_2t_7 & 0 \\ 0 & p_5t_8 & 0 \\ 0 & p_3t_4 & 0 \\ 0 & p_3t_9 & 0 \\ 0 & 0 & p_1t_3 \\ 0 & 0 & p_1p_4t_{10} \end{bmatrix} + \right.$$

$$\left. + N_d(x)\begin{bmatrix} p_2t_1 & 0 & 0 \\ p_2t_5 & 0 & 0 \\ p_2t_6 & 0 & 0 \\ p_2t_7 & 0 & 0 \\ 0 & 0 & 0 \\ p_3t_4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right)[\alpha]\,.$$

The matrix $[\alpha]$ is the matrix grouping the entire weighting factors, defined as:

$$[\alpha] = \begin{bmatrix} \alpha_{11} & \alpha_{21} & \alpha_{31} & \alpha_{41} & \alpha_{51} & \alpha_{61} & \alpha_{71} & \alpha_{81} & \alpha_{91} \\ \alpha_{12} & \alpha_{22} & \alpha_{32} & \alpha_{42} & \alpha_{52} & \alpha_{62} & \alpha_{72} & \alpha_{82} & \alpha_{92} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} & \alpha_{43} & \alpha_{53} & \alpha_{63} & \alpha_{73} & \alpha_{83} & \alpha_{93} \end{bmatrix} .$$

For proposed model, the 3GPP interface corresponding to each activity from Table 1, and the graphical representation of each 3GPP interface in Fig. 5 could be extracted. By combining these inputs, the weighting factors are calculated. The obtained $[\alpha]$ matrix is as follows:

$$[\alpha_{Model}] = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.5 & 0.5 & 0.5 & 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.5 \end{bmatrix} .$$

The values of $S_{NFV}(x)$, number of Diameter transactions per second during busy hour migrated to NFV architecture, from year 2015 till 2021 are calculated and presented in Table 2.

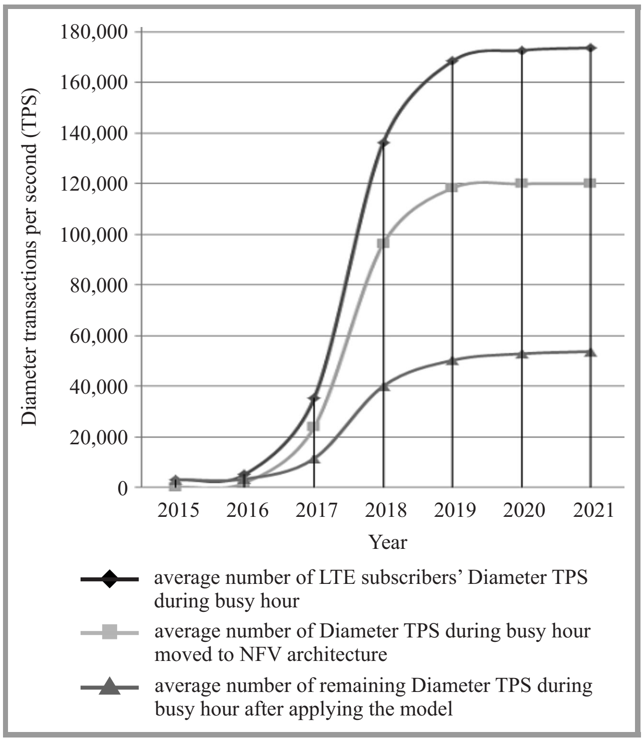


Fig. 6. Mitigating the impact of VoLTE after applying the model.

As shown in Fig. 6, 70% of the generated transactions per second during busy hour will be moved to NFV network. Consequently, operators' existing nodes should deal only with the remaining 30% of the generated transactions. Hence, operators will be relieved from the huge growth of Diameter signaling caused by VoLTE massive introduction.

5. Conclusion

In this paper, the impact of Diameter signaling growth on core network, especially the Diameter signaling traffic generated after massive VoLTE deployment are presented. To anticipate the impact induced by the challenging VoLTE deployment, a model to improve operators' readiness for the adoption of any kind of VoLTE introduction scenarios, either massive or gradual is proposed.

As shown in the model development, up to 70% of the total Diameter signaling transactions will be moved to the proposed architecture. This will relieve operators from the burden of the network cost and scalability, without impacting the running services and without overloading the existing network. Additionally, with this architecture, operators will be able to introduce NFV and SDN technologies and provide valuable contribution in the state-of-the-art of those technologies. NFV and SDN are most likely to redraw the design and implementation guidance for telecom environment in the coming years.

In future work, authors plan to investigate the financial aspects of this NFV based model, and evaluate the possibilities to extend the scope of virtualization to other network elements and services.

References

[1] "Tellabs "End of Profit" study executive summary", Tellabs, Jan. 2011.

[2] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: technical and business perspectives", *IEEE Wirel. Commun.*, vol. 20, no. 2, pp. 104–112, 2013.

[3] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen, "Multiple mobile data offloading through disruption tolerant networks", *IEEE Trans. Mob. Comput.*, vol. 13, no. 7, pp. 1579–1596, 2014.

[4] R. Travis, "Oracle Communications LTE Diameter Signaling Index", 4th ed., Oracle Communications, Oracle, Oct. 2015.

[5] "Cisco Visual Networking Index: Forecast and Methodology, 2014–2019", Cisco, Tech. Rep., May 2015.

[6] J. Rankin, A. Costaiche, J. Zeto, and K. O'Neil, *Validating VoLTE: A Definitive Guide to Succssful Deployments*, 2 ed. IXIA, 2014.

[7] G. Blennerud, "Mobile Broadband: Busting the myth of the scissor effect", Tech. Rep., Ericsson, Dec. 2012.

[8] "Network Policy Control and the Migration to LTE", Sandvine, Tech. Rep., Nov. 2013.

[9] "Quality of Service (QoS) and Policy Management in Mobile Data Networks", IXIA, Tech. Rep. 915-2731-01 Rev. D, Dec. 2013.

[10] "Distributed Decisions in Network Policy Control", Sandvine, Tech. Rep., Nov. 2013.

[11] "New VoLTE Policy Management Architecture: Improve Subscriber Experience and Lower Cost of Growth", Cisco, Tech. Rep. C11-732158-0, Mar. 2015.

[12] J. Gillet and T. Hatt, "Smartphone forecasts and assumptions: 2007-2020", GSMA Intelligence, Tech. Rep., Sept. 2014.

[13] S. Pang, *Successful Service Design for Telecommunications: A Comprehensive Guide to Design and Implementation*. Hoboken, NJ: Wiley, 2009.

[14] "Voice over LTE: Challenges and Opportunities", Sandvine, Tech. Rep. Version 2.0, 2015.

[15] "Policy and charging control signalling flows and Quality of Service (QoS) parameter mapping", TS 29.213 V13.1.0, 3rd Generation Partnership Project (3GPP), Mar. 2015.

[16] "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access", TS 23.401 V13.0.0, 3rd Generation Partnership Project (3GPP), Sept. 2014.

[17] "Policy and charging control architecture", TS 23.203 V13.3.0, 3rd Generation Partnership Project (3GPP), Mar. 2015.

[18] "Telecommunication management; Charging management; Diameter charging applications", TS 32.299 V13.0.0, 3rd Generation Partnership Project (3GPP), Mar. 2015.

[19] "IP Multimedia (IM) Subsystem Cx and Dx Interfaces; Signalling flows and message contents", TS 29.228 V12.0.0, 3rd Generation Partnership Project (3GPP), Sept. 2013.

[20] "Data for Operators Ranking connections, excluding cellular M2M, Total", GSMA Intelligence, Tech. Rep., Q4 2020.

[21] "Data for Operators Ranking connections, excluding cellular M2M, 4G", GSMA Intelligence, Tech. Rep., Q2 2015.

[22] "% connections, excluding cellular M2M, Prepaid, World", GSMA Intelligence, Tech. Rep., Q4 2015.

[23] Y. Ouyang and M. H. Fallah, "An analysis of traffic and throughput for UMTS packet core networks", in *Research, Practice, and Educational Advancements in Telecommunications and Networking*, M. R. Bartolacci and S. R. Powell, Eds. IGI Global, 2012, pp. 82–107.

[24] "IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3", TS 24.229 V13.0.0, 3rd Generation Partnership Project (3GPP), Sept. 2015.

[25] "Bringing Network Function Virtualization to LTE", Tech. Rep., 4G Americas, Nov. 2014.

[26] "MWC 2015 End-to-End NFV Architecture demof", Telefonica, Brocade, Cyad, Intel, and Redhat, Tech. Rep., GSMA Mobile World Congress, Spain, Mar. 2015.

[27] D. Mavrakis, "MWC 2015: NFV gets real", Tech. Rep. TE0006-001035, Ovum, Mar. 2015.

[28] "End to End Network Function Virtualization Architecture Instantiation", Intel, Brocade, Cyan, Red Hat, and Telefonica, Tech. Rep., Feb. 2015.

[29] Z. Huang, R. Ma, J. Li, Z. Chang, and H. Guan, "Adaptive and scalable optimizations for high performance SR-IOV", in *Proc. IEEE Int. Conf. Cluster Comput. CLUSTER 2012*, Beijing, China, 2012, pp. 459–467.

[30] S. B. H. Said, M. R. Sama, K. Guillouard, L. Suciu, G. Simon, X. Lagrange, and J.-M. Bonnin, "New control plane in 3GPP LTE/EPC architecture for on-demand connectivity service", in *Proc. 2nd IEEE Int. Conf. Cloud Netw. CloudNet 2013*, San Francisco, CA, USA, 2013, pp. 205–209.

[31] Y. Jouihri and Z. Guennoun, "Best selection for operators starting LTE deployment towards voice services" in *Proc. Int. Conf. Multimed. Comput. & Syst. ICMCS 2012*, Tangier, Morocco, 2012, pp. 568–573.

[32] G. Brown, "Virtualization of IMS & VoLTE in Mobile Operator Core Networks", Mavenir, Tech. Rep., Mar. 2013.

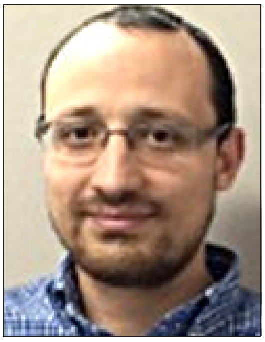
[33] "IMS Application Level Gateway (IMS-ALG) IMS Access Gateway (IMS-AGW); Iq Interface", TS 29. 334 V10.2.0, 3rd Generation Partnership Project (3GPP), Sept. 2011.

Youness Jouihri received his engineering degree in Electronics and Telecommunications from Mohammadia School of Engineering (EMI), Rabat, Morocco in 2005. His research interests include the LTE limitations, voice service evolution, network function virtualization, and cloud convergence between IT and telecom domains. He has a broad international experience in telecom consultancy services.

E-mail: Jouihri.youness@gmail.com
Electronics and Communications Laboratory (LEC)
Mohammadia School of Engineering (EMI)
Rabat Mohammed V University (UM5R)
Rabat, Morocco

Zouhair Guennoun (IEEE senior member) received his engineering degree in Electronics and Telecommunications from Montefiore Institute, Belgium in 1987. He received his M.Sc. and Ph.D. degrees in 1993 and 1996, in Communication Systems from Mohammadia School of Engineering (EMI). He visited, during 1990-1994 the Centre for Communication Research in Bristol University, UK. From 1996, he is working at EMI as a Professor Lecturer and currently in charge of the research laboratory in LEC. His fields of interest are digital signal, speech and image processing, error control coding, communication systems.

E-mail: zouhair@emi.ac.ma
Electronics and Communications Laboratory (LEC) Mohammadia School of Engineering (EMI)
Rabat Mohammed V University (UM5R)
Rabat, Morocco



Youssef Chagh received his engineering degree in Telecommunications from INPT, Rabat, Morocco, in 2005. His fields of interest are IP multimedia subsystem, network function virtualization, and 5G networks.

E-mail: chagh.youssef@gmail.com
Electronics and Communications Laboratory (LEC)
Mohammadia School of Engineering (EMI)
Rabat Mohammed V University (UM5R)
Rabat, Morocco

Driss Zahi received his M.Sc. degree in Electrical Engineering and Telecommunications from Paris ISEP. He also received a M.Sc. degree in Information and Knowledge Systems and an other in Microeconomics from the Business School of the Sorbonne University. He received the Ph.D. degree from the Sorbonne University in Economy and Management Sciences. he has spent more than 25 years in the telecommunications industry as a consultant for large operators and vendors. He also delivered various training curriculum in Universities.

E-mail: driss.zahi2265@gmail.com
GREGOR Research Lab
Pantheon-Sorbonne University
Paris, France