A Queuing based Dynamic Auto Scaling Algorithm for the LTE EPC Control Plane

Jonathan Prados-Garzon*, Abdelquoddouss Laghrissi§, Miloud Bagaa§, and Tarik Taleb§
jpg@ugr.es, abdelquoddouss.laghrissi@aalto.fi, bagmoul@gmail.com, tarik.taleb@aalto.fi
*Research Centre for Information and Communications Technologies of the University of Granada (CITIC-UGR) and Department of Signal Theory, Telematics and Communications, University of Granada, Granada, Spain.
§Aalto University, Espoo, Finland.

***Abstract*—Network softwarization paradigm, whose main enabler is Network Functions Virtualization (NFV), facilitates the automation of the management operations and orchestration of the future networks, thus reducing the operational expenditures of the network. The envisioned management practices include the introduction of automation in the scaling of network services. This may enable operators to handle workload fluctuations to keep the desired performance with great agility and reduced costs. This procedure introduces a non-negligible delay in allocating or releasing virtual resources. Therefore, waiting until the system is overloaded or underutilized so as to scale resources up or down could negatively impact user Quality of Experience, or lead to inefficient resource utilization. In this vein, this paper proposes a novel and agile Dynamic Auto Scaling Algorithm for the Long Term Evolution (LTE) virtualized Evolved Packet Core (vEPC) Control Plane (CP). The resources dimensioning stage of the algorithm is based on an original queuing model for the LTE CP. To model the LTE CP, we use an open network of G/G/m queues. We also provide expressions to derive the steady state transition probabilities of the queuing network. Finally, we validate the proper operation of our solution using accurate simulation tools.**

***Index Terms*—Queuing model, Dimensioning, LTE EPC, Control Plain, NFV, Network Softwarization, Dynamic Resource Provisioning, Dynamic Auto Scaling Algorithm.**

I. Introduction

Fith Generation (5G) mobile networks are expected to play a paramount role in the global industrial digitalization by covering all the vertical market needs in a cost effective and efficient way. Compared to its predecessor (i.e., the Long Term Evolution (LTE) technology), the requirements for 5G systems include, among many others, higher network flexibility and scalability, as well as x100 increase in cost effectiveness and energy efficiency [1]. To meet these goals, network softwarization (NS) tendency is envisaged as the cornerstone to build the 5G technology. The key enabler of the NS concept is Network Functions Virtualization (NFV) paradigm.

NFV paradigm decouples network functions from proprietary hardware enabling them to run as software components, which are called Virtualized Network Functions (VNFs), on commodity servers. NFV facilitates the automation of the management operations and orchestration of the future networks [2]. The envisioned management practices include the automation of the scaling of network services. This may enable operators to handle workload fluctuations to keep the desired performance with great agility and reduced costs. This procedure introduces a non-negligible delay in allocating or releasing virtual resources [3]. Therefore, waiting until the system is overloaded or underutilized so as to scale resources up or down could negatively impact user Quality of Experience (QoE), or lead to inefficient resources utilization. In this regard, analytical models for predicting the performance of softwarized networks are an appropriate and agile solution to this problem.

In this vein, this work proposes an analytical model for the LTE Control Plane (CP) based on queuing theory and its application to the the dynamic resource provisioning (DRP) of the LTE Evolved Packet Core (EPC) CP entities. A DRP algorithm enables a system to adapt its resources autonomously depending on the current workload so that some performance requirements are met.

There are several works that have tackled the modeling of the control plane (CP) of a virtualized LTE network by applying queuing theory [4]–[8]. However, these models do not include all the elements of the LTE CP nor capture the flow of signaling across them. This work proposes an open queue network to model the whole LTE CP, which includes all its elements and interfaces. It thus enables to estimate the end-to-end performance metrics of the system from the aggregated signaling external arrival process. Additionally, [5] and [6] address the DRP of a virtualized EPC (vEPC). Nevertheless, these works address the dimensioning of each component in an isolated way. Our solution considers a processing delay budget for the whole EPC and it automatically distributes this budget among the CP entities. This leads to an optimal dimensioning of the resources, i.e., resources saving. Moreover, our solution allows for the performance requirement defined by the 3GPP for the LTE CP, i.e., the elapsed time to move an User Equipment (UE) from IDLE state to ACTIVE state.

The rest of the paper is organized as follows. Section II describes the system model and formulates the resource dimensioning problem of a vEPC. In Section IV, we present the analytical model for the LTE CP. Section V introduces our solution. Section VI includes the description of our experimental setup and provides some results that show that our solution works properly. Finally, Section VII draws the main conclusions.

II. System Model and Problem Formulation

A. System Architecture

Let us assume an Evolved Terrestrial Radio Access Network (E-UTRAN) already deployed and consisting of eNodeBs (eNBs), which provides connectivity to a set of N_{UEs} UEs to LTE Evolved Packet Core (EPC). The EPC is virtualized and running on a data center.

The LTE network architecture assumed in this work is depicted in Fig. 1. We assume a CUPS architecture for the vEPC, i.e., Control and User Plane (UP) Separation of EPC nodes. This allows the independent scaling between CP and

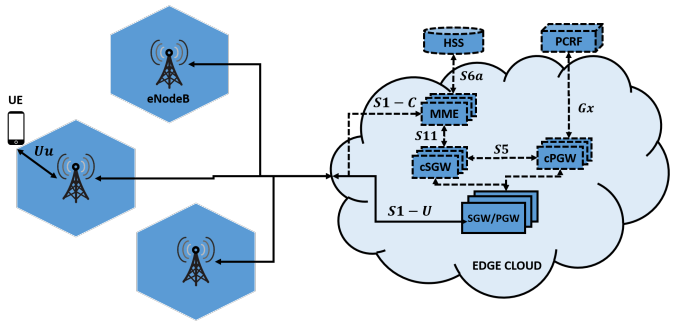

Fig. 1. Assumed LTE network architecture.

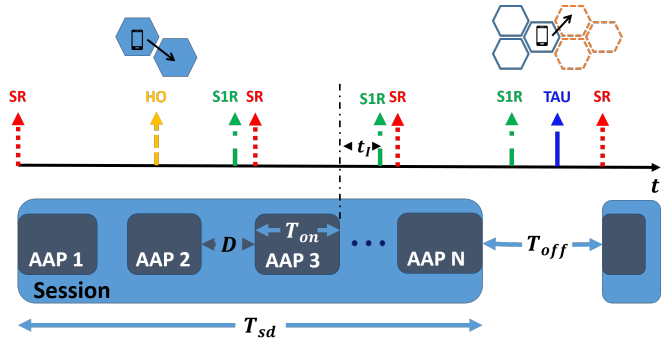

Fig. 2. Workload generation model.

UP functions. Each CP entity (e.g., Mobility Managment Entity -MME-, and the control plane functionalities of the Serving Gateway and Packet Data Network Gateway -cSGW and cPGW-) is implemented separately as a single VNF. Each VNF might have multiple instances, and each instance runs on an isolated virtualization container like a Virtual Machine (VM). Let $m_i^{(e)}$ denote the number of dedicated physical CPU cores allocated to the instance i of the entity $e \in E$, where $E = \{MME, cSGW, cPGW\}$. Since the number of CPU cores of a physical server is finite and they are shared among several VMs, we consider that $m_i^{(e)}$ is limited to m_{max}, i.e., $m_i^{(e)} \leq m_{max}$.

B. Workload generation model

In this paper, we only consider the enhanced mobile broadband (MBB) use case. In this context, the UEs run applications that generate and consume data plane (DP) traffic. We consider the abstraction presented in [7] for such a process (see Fig. 2).

A session with duration T_{sd} is defined as the user activity from it launches an application to it closes it. A session consists of N Application Activity Periods (AAPs) of length T_{on} separated by $N-1$ reading times of duration D. An AAP is a time period in which the application generates or consumes all necessary network traffic to perform a given task (e.g., download the profile of a friend, to send a message or stream a video). A reading time is the temporal interval during which the user performs any action that does not require to generate network traffic such as reading a message or deciding what the next friend profile is to take a look at. The sessions are separated by user inactivity periods of length T_{off}.

Regarding the signaling workload, the user activity and mobility trigger the LTE CP procedures. In this work, we only consider UE-triggered Service Request (SR), S1-Release (S1R), X2-based Handover (HO), and Tracking Area Update (TAU) procedures. Although other procedures such as Attach and S1-based Handover are heavier in terms of computational resources consumption, they do not occur frequently in LTE networks [9].

Once the UE is registered in the network, an SR procedure is triggered during its idle-to-connected transitions. Then, whenever a AAP starts and the UE is in idle mode, an SR procedure takes place (see Fig. 2). Conversely, an S1R procedure occurs during UE's connected-to-idle transitions during which the network releases the UE's resources. Here, we take into account the effects of an inactivity timer whose value is denoted as t_I. That is, the network waits t_I units of time after an AAP finishes before triggering an S1R (see Fig. 2). An HO procedure is triggered when a UE is in connected mode and performs a cell change, but the target cell is attached at the same MME as the source cell. Finally, we assume that a TAU procedure is triggered whenever a UE carries out a tracking area change. These tracking areas are predefined and the same for any UE.

C. Performance Requirements

The considered performance requirement is a bound on the mean CP latency $\overline{T}_{budget}^{(CP)}$ defined by the 3GPP, i.e., the average elapsed time to move an UE from IDLE state to ACTIVE state [10]. In this work, we translate this specification as the required average time to carry out a service request procedure. Moreover, we consider the worst-case scenario for the service request procedure. That is the UE authentication, NAS security setup, and the EPS session modification steps are carried out during the SR.

Let $\overline{T}_c$ and $\overline{T}_{if}$ denote respectively the mean response times of the CP entity $c \in C = \{UE, eNB, MME, cSGW, cPGW, HSS, PCRF\}$ and the LTE interface $if \in IF = \{Uu, S1-C, S11, S6a, S5, Gx\}$. The mean time required to carry out an SR, $\overline{T}^{(SR)}$, in the worst-case scenario can be computed as:

$$\begin{aligned}\overline{T}^{(SR)} &= 5 \cdot \overline{T}_{UE} + 8 \cdot \overline{T}_{eNB} + 5 \cdot \overline{T}_{MME} + 2 \cdot \overline{T}_{cSGW} \\ &+ 2 \cdot \overline{T}_{cPGW} + \overline{T}_{HSS} + \overline{T}_{PCRF} + 8 \cdot \overline{T}_{Uu} + 7 \cdot \overline{T}_{S1-C} \\ &+ 2 \cdot \overline{T}_{S11} + 2 \cdot \overline{T}_{S6a} + 2 \cdot \overline{T}_{S5} + 2 \cdot \overline{T}_{Gx}\end{aligned} \tag{1}$$

The above equation means that during an SR call flow in the worst case scenario the UE, eNB, MME, cSGW, cPGW, HSS, and PCRF entities have to process respectively 5, 8, 5, 2, 2, 1, and 1 control messages. And 8, 7, 2, 2, 2, and 2 control messages have to traverse respectively the LTE Uu, S1-C, S11, S6a, S5, and Gx interfaces [11]. Then, the CP delay requirement can be expressed as $\overline{T}^{(SR)} \leq \overline{T}_{budget}^{(CP)}$.

III. Problem Formulation

In this section, we formulate the resource dimensioning problem for the vEPC CP. The objective is to minimize the required computational resources.

$$minimize \left(\sum_{e \in E} \sum_{i} m_i^{(e)} \right) \tag{2}$$

Subjectto :

$$C1: \quad \overline{T}^{(SR)} \leq \overline{T}_{budget}^{(CP)}, \tag{3}$$

$$C2: \quad m_i^{(e)} \leq m_{max} \quad \forall \quad e \in E, i \in \mathbb{N} \tag{4}$$

Constraint 1 guarantees that the actual mean delay to carry out a service request for the vEPC k (i.e., vEPC instance running on EC k) is lower or equal than the mean CP latency $\overline{T}_{budget}^{(CP)}$. Constraint 2 limits the maximum number of physical cores requested for a single VNFC instance. To have a single VNF instance would be optimal for minimizing the amount of required resources. However, each physical server has a maximum number of physical cores and they are shared among several VMs. Consequently, the higher the number of physical cores requested for a VNFC instance the lower its availability.

IV. Analysis and Modeling

A. LTE CP modeling

We model the CP of the LTE as an open network of G/G/m queues (see Fig. IV-A), where each queue represents an instance of a vEPC entity to be dimensioned (e.g., MME, cSGW, and cPGW) [8]. In Kendall's notation, a G/G/m queue is a queuing node with m servers, arbitrary arrival and service processes, FCFS (First-Come, First-Served) discipline, and infinite capacity and calling population. Each queue has $m_i^{(E)}$ servers which represent different CPU cores processing messages from the same queue. The rest of the LTE CP entities are modeled as infinite servers, i.e., its mean response time is constant and independent of its workload.

The traffic sources are located at the eNB and the UE, since the LTE signaling procedures considered in this work (e.g., SR, S1R, HO, and TAU) are triggered by these entities. Specifically, the TAU and SR procedures are triggered by the UE and the S1R and HO procedures are triggered by the eNB. In the same way, the traffic sinks are placed at the MME instances.

B. EPC CP entities response times estimation

To estimate the mean response times of the vEPC CP entities to be dimensioned (e.g., $\overline{T}_{MME}$, $\overline{T}_{cSGW}$, and $\overline{T}_{cPGW}$), we employ the approximated technique proposed in [12] for the Queuing Network Analyzer, hereinafter referred as QNA method. This methodology was applied and validated to estimate the mean response time of a VNF with several components (VNFCs) [8].

Next, we describe the main steps followed by the QNA method to estimate the mean response time of each individual queue in a network of K G/G/m queues. To that end, QNA method uses a reduced set of input parameters: i) the steady state transition probabilities matrix $P = [p_{ki}]$, where p_{ki} denotes the probability of a packet leaving the node k is next moved to the node i or leaves the network with probability $p_{0k} = 1 - \sum_i p_{ki}$; ii) the mean and squared coefficient of variation (SCV) of the external arrival processes at queue k, λ_{0k} and c_{0k}^2; and iii) the mean and SCV of the service processes at queue k, μ_k and c_{sk}^2.

Please note that, to solve the resulting network of queues modeling the LTE CP, we only need to map each entity to an integer index $k \in [1, K]$. Please note that To simplify the notation in this analysis, we maps each entity instance to an integer index $k \in [1, K]$.

1) Internal flows parameters estimation: The mean arrival rate to each queue k, λ_k, can be computed by solving the flow balance equations:

$$\lambda_k = \lambda_{0k} + \sum_{i=1}^{K} \lambda_i \cdot p_{ik} \tag{5}$$

The most interesting aspect of the QNA method is that it estimates the SCV of the aggregated arrival process to each queue c_{ak}^2 from the following set of linear equations:

$$c_{ak}^2 = a_k + \sum_{i=1}^{K} c_{ai}^2 b_{ik}, \qquad 1 \leq k \leq K \tag{6}$$

$$a_k = 1 + \omega_k \Big\{ (q_{0k} c_{0k}^2 - 1) + \sum_{i=1}^{K} q_{ik}[(1 - p_{ik}) + p_{ik} \rho_i^2 x_i] \Big\} \tag{7}$$

$$b_{ik} = \omega_k q_{ik} p_{ik} (1 - \rho_i^2) \tag{8}$$

$$x_i = 1 + m_i^{-0.5} (max\{c_{si}^2, 0.2\} - 1) \tag{9}$$

$$\omega_k = \big(1 + 4(1 - \rho_k)^2 (\gamma_k - 1)\big)^{-1} \tag{10}$$

$$\gamma_k = \left(\sum_{i=0}^{K} q_{ik}^2 \right)^{-1} \tag{11}$$

where $q_{0k} = \lambda_{0k}/\lambda_k$ and $q_{ik} = (\lambda_i \cdot p_{ik})/\lambda_k$ are respectively the proportion of arrivals to the node k that came from its external arrival process and node i, and $\rho_k = \lambda_k/(\mu_k \cdot m_k)$ is the utilization of the node k.

2) Mean response time computation per node: Once the λ_k and c_{ak}^2 for the aggregated arrival process to each node k are estimated, we can compute the mean response time for each node k.

If the node k has only one server ($m_k = 1$), $\overline{T}_k$ can be estimated as:

$$\overline{T}_k = \frac{\rho_k \cdot (c_{ak}^2 + c_{sk}^2) \cdot \beta}{2 \cdot \mu_k (1 - \rho_k)} + \frac{1}{\mu_k} \tag{12}$$

with

$$\beta = \begin{cases} exp(-\frac{2 \cdot (1 - \rho_k) \cdot (1 - c_{ak}^2)^2}{3 \cdot \rho_k \cdot (c_{ak}^2 + c_{sk}^2)}) & c_{ak}^2 < 1 \\ \beta = 1 & c_{ak}^2 \geq 1 \end{cases} \tag{13}$$

If, by contrast, the node k is a GI/G/m queue ($m_k = m$), $\overline{T}_k$ can be estimated as:

$$\overline{T}_k = 0.5 \cdot \left(c_{ai}^2 + c_{si}^2\right) \cdot W_k^{M/M/m} + \frac{1}{\mu_k} \tag{14}$$

where $W_k^{M/M/m}$ is the mean waiting time for a M/M/m queue, which can be computed as:

$$W_k^{M/M/m} = \frac{C(m_k, \frac{\lambda_k}{\mu_k})}{m_k \mu_k - \lambda_k} \tag{15}$$

and $C(m, \rho)$ represents the Erlang's C formula.

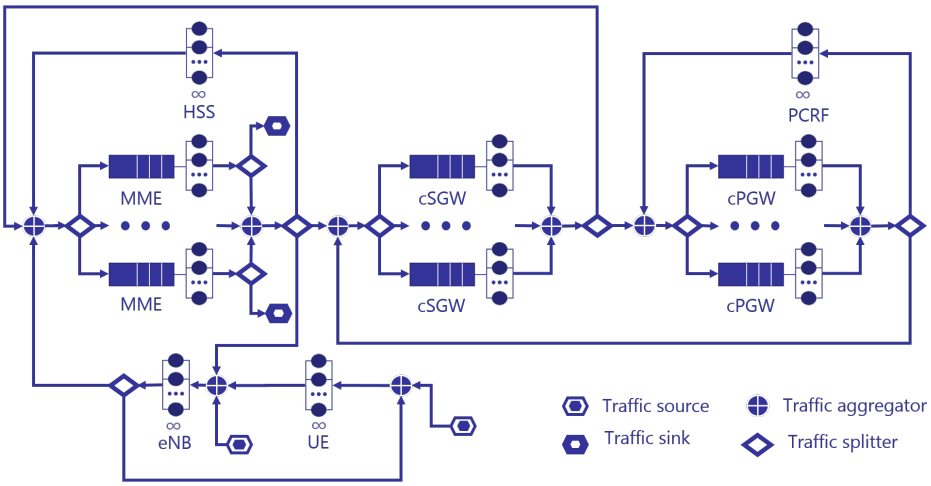

Fig. 3. LTE control plane queuing model.

C. Transtion probabilities for the LTE CP queuing model

In this section, we derive the expressions to compute the transition probabilities for the proposed LTE CP queuing model. These probabilities are used as input parameter to estimate the response times of the network of queues.

Let V_c denote the visit ratio of the LTE CP entity $c \in C$ which is defined as the average number of visits to entity c by a signaling procedure during its lifetime in the network. That is, $V_c = \lambda_c / \sum_c \lambda_{0c} = \lambda_c/(\lambda_{0UE} + \lambda_{0eNB})$. Please note that V_c is equal to the average number of packets to be processed by the LTE CP entity c per control procedure $p \in P = \{SR, S1R, HO, TAU\}$. Then,

$$V_c = \frac{\sum_{p \in P} \lambda_p \cdot n_p^{(c)}}{\sum_{p \in P} \lambda_p} \tag{16}$$

where $n_p^{(c)}$ is the number of packets to be processed by the LTE CP entity c for the control procedure p.

The visit ratios and the transition probabilities are related through (5) (flow balance equations):

$$V_c = \frac{\lambda_{0c}}{\sum_{c \in C} \lambda_{0c}} + \sum_{cs \in C} V_{cs} \cdot p_c^{cs} \tag{17}$$

where p_c^{cs} denotes the transition probability from entity cs to entity c.

The transition probabilities also satisfy

$$p_{0c} + \sum_{cd \in C} p_{cd}^{c} = 1 \tag{18}$$

Assuming that the workload is distributed among the instances of the different entities (e.g., MME, cSGW, and cPGW) according to their capacities, i.e., $V_{e_l} = m_l^{(e)}/(\sum_i m_i^{(e)}) \cdot V_e$ and using (17) and (18), we can compute the transition probabilities for our LTE CP queuing model. They are given by the following expressions:

$$p_{UE}^{eNB} = \frac{V_{UE} - \frac{\lambda_0^{(UE)}}{\sum_E \lambda_0^{(E)}}}{V_{eNB}} \tag{19}$$

$$p_{MME_l}^{eNB} = \frac{m_l^{(MME)}}{\sum_l m_l^{(MME)}} \cdot (1 - p_{UE}^{eNB}) \tag{20}$$

$$p_{eNB}^{MME_l} = \frac{V_{eNB} - \frac{\lambda_0^{(eNB)}}{\sum_{CE} \lambda_0^{(CE)}} - V_{UE}}{V_{MME}} \tag{21}$$

$$p_{cSGW_l}^{MME_l} = \frac{m_l^{(cSGW)}}{\sum_l m_n^{(cSGW)}} \cdot \left(1 - p_{eNB}^{MME_l} - p_{eNB}^{MME_l} - \frac{1}{V_{MME}}\right) \tag{22}$$

$$p_{HSS}^{MME_l} = \frac{V_{HSS}}{V_{MME}} \tag{23}$$

$$p_{MME_l}^{cSGW_l} = \frac{m_l^{(MME)}}{\sum_m m_m^{(MME)}} \cdot \left(1 - \sum_l p_{cPGW_l}^{cSGW_l}\right) \tag{24}$$

$$p_{cPGW_l}^{cSGW_l} = \frac{m_l^{(cPGW)}}{\sum_m m_m^{(cPGW)}} \cdot \frac{(V_{PGW} - V_{PCRF})}{V_{SGW}} \tag{25}$$

$$p_{cSGW_l}^{cPGW_l} = \frac{m_l^{(cSGW)}}{\sum_m m_m^{(cSGW)}} \cdot \left(1 - \frac{V_{PCRF}}{V_{PGW}}\right) \tag{26}$$

$$p_{PCRF}^{PGW_l} = \frac{V_{PCRF}}{V_{PGW}} \tag{27}$$

$$p_{MME_l}^{HSS} = \frac{m_l^{(MME)}}{\sum_m m_m} \tag{28}$$

$$p_{cPGW_l}^{PCRF} = \frac{m_l^{(cPGW)}}{\sum_n m_n^{(cPGW)}} \tag{29}$$

Please note that the transition probabilities depend on the average number of packets to be processed for each LTE CP entity per control procedure, which is equal to the visit ratio of the entity; the external arrival processes λ_{0UE} and λ_{0eNB}; and the number of processing instances assigned to each EPC entity instance $m_i^{(e)}$.

V. Dynamic Resource Provisioning Algorithm

The goal of our DRP algorithm for a vEPC is to allocate sufficient resources to the virtualized CP entities (e.g., MME, cSGW, and cPGW) so that the 3GPP performance requirement for the LTE CP can be met. The main stages of the algorithm are shown in Fig. 4.

The signaling workload predictor is in charge to predict the peak control traffic demand until the next decision to

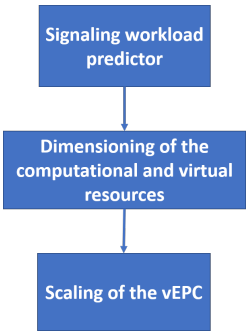

Fig. 4. Overall Dynamic Resource Provisioning algorithm for the vEPC.

provision is taken. The decisions of when to provision will depend on the dynamics of mobile networks workloads. For instance, this predictor could be implemented by using Artificial Intelligence (AI) techniques. As output it has to provide the mean arrival rates and SCVs of the external arrival processes for the predicted peak signaling demand. That is, λ_{0UE}, λ_{0eNB}, c^2_{0UE}, and c^2_{0eNB}. Then, we can estimate the parameters of the aggregated signaling workload to each entity to be dimensioned (e.g., λ_{MME}, λ_{cSGW}, λ_{cPGW}, c^2_{aMME}, c^2_{acSGW}, and c^2_{acPGW}) using (5)-(11).

The next stage is the dimensioning of the computational and virtual resources. To that end, we propose a novel algorithm for the dimensioning vEPC CP computational resources, which is based on the analytical model for the LTE CP described in Section IV (see Algorithm 1). As input, it requires the processing delay budget for the vEPC CP $T^{(CP)}_{proc-budget}$, the mean and SCV of the external arrival processes provided by the predictor (i.e., λ_{0UE}, λ_{0eNB}, c^2_{0UE}, and c^2_{0eNB}), and the mean and SCV of the service processes for each entity to be dimensioned (i.e., μ_{MME}, μ_{cSGW}, μ_{cPGW}, c^2_{sMME}, c^2_{scSGW}, and c^2_{scPGW}).

To estimate $T^{(CP)}_{proc-budget}$, assuming that the response times of the LTE interfaces and the other entities (e.g., UE, eNB, HSS, and PCRF) are known, we can evaluate $\overline{T}^{(SR)}$ in (1) for $\overline{T}_{MME}$, $\overline{T}_{cSGW}$, and $\overline{T}_{cPGW}$ equal to zero. That is, $\overline{T}^{(SR)}_0 = \overline{T}^{(SR)}(\overline{T}_{MME}=0, \overline{T}_{cSGW}=0, \overline{T}_{cPGW}=0)$. Then,

$$T^{(CP)}_{proc-budget} = \overline{T}^{(CP)}_{budget} - \overline{T}^{(SR)}_0 \quad (30)$$

The dimensioning algorithm searches for the minimum number of processing instances to be allocated to the vEPC CP for a given EC so that the processing delay budget $T^{(CP)}_{proc-budget}$ be met. The algorithm iterates until the processing delay budget is fulfilled. At each iteration it increments by one the number of processing instances M_{CP} allocated to the vEPC CP. For a given M_{CP}, the algorithm explores different combinations to distribute these instances among the different entities to be dimensioned (e.g., MME, cSGW, cPGW), and choose that one providing the lowest processing delay. To achieve linear complexity, the search space is limited at each iteration (see line 12 of Algorithm 1). In the algorithm, $T_{mme}(m)$, $T_{cSGW}(n)$, and $T_{cPGW}(l)$ respectively denote the mean response times of the MME, cSGW, and cPGW for a given number of allocated processing instances m, n, and l. These mean response times are estimated by using the QNA method (refer to Section IV).

Please note that, although it is not explicitly included in Algorithm 1, for each 'processing instances allocation (m, n, l) the mean and SCV of the aggregated signaling workload at each entity to be dimensioned (i.e., (e.g., λ_{MME}, λ_{cSGW}, λ_{cPGW}, c^2_{aMME}, c^2_{acSGW}, and c^2_{acPGW}) are re-estimated by using (5)-(11). The same applies to the transition probability matrix, which is re-computed by using (19)-(29).

Observe also that the number of instances or, equivalently, the number of virtualization containers for each vEPC entity might change for a given processing instances allocation. The number of instances for a given allocation can be simply computed as: $\lceil m_{MME}/m_{max} \rceil$, $\lceil m_{cSGW}/m_{max} \rceil$, and $\lceil m_{cPGW}/m_{max} \rceil$.

Algorithm 1 Dimensioning Algorithm

Input: $T^{(CP)}_{proc-budget}$, λ_{0UE}, λ_{0eNB}, c^2_{0UE}, c^2_{0eNB}, μ_{MME}, c^2_{sMME}, μ_{cSGW}, c^2_{scSGW}, μ_{cPGW}, and c^2_{scPGW}.
Output: number of physical cores assigned per network entity m_{MME}, m_{cSGW}, and m_{cPGW}.

1: **Initialization** $m_{MME} = \lceil \lambda_{MME}/\mu_{MME} \rceil$, $m_{cSGW} = \lceil \lambda_{cSGW}/\mu_{cSGW} \rceil$, $m_{cPGW} = \lceil \lambda_{cPGW}/\mu_{cPGW} \rceil$, $M_{CP} = m_{MME} + m_{cSGW} + m_{cPGW}$, $T_{RT-CP} = 8 \cdot T_{MME}(m_{MME}) + 3 \cdot T_{cSGW}(m_{cSGW}) + 2 \cdot T_{cPGW}(m_{cPGW})$;
2: **while** $T_{RT-CP} > T^{(CP)}_{proc-budget}$ **do**
3: $M_{CP} \Leftarrow M_{CP} + 1$
4: **for** each $m \in \{m_{MME}, ..., M_{CP} - m_{cSGW} - m_{cPGW}\} \cap \mathbb{N}$ **do**
5: **for** each $n \in \{m_{cSGW}, ..., M_{CP} - m_{MME} - m_{cPGW}\} \cap \mathbb{N}$ **do**
6: $l = M_{CP} - m - n$
7: $T_{aux} = 8 \cdot T_{MME}(m) + 3 \cdot T_{cSGW}(n) + 2 \cdot T_{cPGW}(l)$
8: **if** $T_{RT-CP} > T_{aux}$ **then**
9: $T_{RT-CP} \Leftarrow T_{aux}$, $m_{MME} \Leftarrow m$, $m_{cSGW} \Leftarrow n$, $m_{cPGW} \Leftarrow l$
10: **end if**
11: **end for**
12: **end for**
13: **end while**

Finally, based on the dimensioning algorithm output the scaling of the vEPC is carried out by allocating or releasing resources.

VI. Results

A. Experimental setup

To validate the proper operation of our solution, we employed two software tools: i) the "Network Slice Planner" NSP [13], and ii) a system-level simulator of an LTE network.

NSP is a simulation tool that implements accurate models for the users' behavior and mobility, and a compound traffic model for cellular networks. We used the NSP [13] to generate synthetic signaling workload in an LTE network. We extended the compound traffic model of this tool by including the traffic models employed in [7].

The system-level LTE network simulator was developed within the ns3 environment. It implements the messages exchange between the main LTE network entities. The traces generated from the NSP are used as input of the simulator to emulate the workload generation in the LTE network. Each instance of the vEPC CP entity to be dimensioned is simulated as a First Come First Served (FCFS) queue with multiples generic servers. The rest of the LTE entities (e.g.,

TABLE I
PARAMETERS CONFIGURATION

Service processes and mean response times for CP	
CPU instance service rate for the MME, cSGW, and cPGW (μ_{MME}, μ_{cSGW}, and μ_{cPGW})	6700 packets per second
Squared coefficient of variation of the service time for the MME, cSGW, and cPGW ($c^2_{s_{MME}}$, $c^2_{s_{cSGW}}$, and $c^2_{s_{cPGW}}$)	0.65
Mean response times for the UE, eNB, HSS, and PCRF ($\overline{T}_{UE}$, $\overline{T}_{eNB}$, $\overline{T}_{HSS}$, and $\overline{T}_{PCRF}$)	1 ms
Mean delays for the interfaces S6a and Gx ($\overline{T}_{S6a}$ and $\overline{T}_{Gx}$)	1.5 ms
Mean delays for the interfaces S11 and S5 ($\overline{T}_{S11}$ and $\overline{T}_{S5}$)	30 μs
QoS requirements	
$\overline{T}^{(CP)}_{budget}$	25 ms

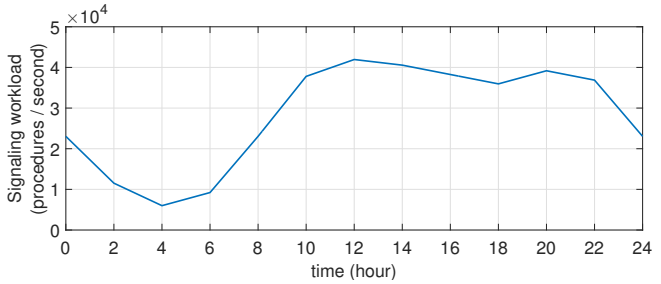

Fig. 5. Signaling Workload.

UE, eNB, HSS, and PCRF) and the network delays (e.g., transmission, propagation, and switches processing) between any couple of EPC entities are simulated as infinite servers, i.e., constant processing delay without queuing waiting time. Table I includes the configuration of the main parameters for the simulator.

B. Dynamic Resource Provisioning Algorithm Evaluation

To verify that our solution works properly, we carried out a simulation with the simulation time set to one day. The simulation scenario had 2000000 UEs and a population density of 1000 inhabitants per km^2. Please note that the HO generation rate depends on the E-UTRAN density, which depends on the population density in our simulation tools. The aggregated signaling workload was modulated according to the temporal distribution measured in [14] (see Fig. 5). We considered an ideal signaling workload predictor.

Figure 6 depicts the required computational resources predicted by our DRP algorithm. The MME has to process a higher number of messages per control procedure, thus it presents the greatest demand of resources. As it is shown in Fig. 7, the LTE CP delay budget is always met, thus validating the operation of our DRP algorithm.

VII. CONCLUSION

In this work we have proposed an open queuing network of G/G/m queues to model the whole LTE CP and the interactions between its entities. To solve the resulting network of queues, we have used the QNA method [12]. Moreover, we have derived expressions for the steady state transition probabilities of the queuing network. We have showed that this probabilities depend on the average number of packets to be processed for

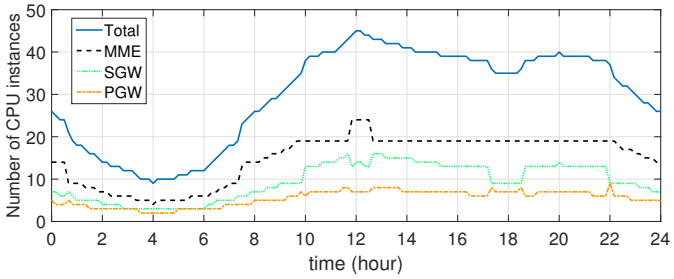

Fig. 6. Number of dedicated CPU instances per entity of the EPC CP.

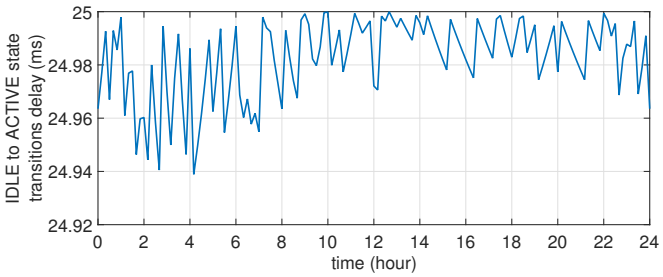

Fig. 7. Mean response times per CP VNFCs.

each LTE CP entity per control procedure, the external arrival processes, and the number of processing instances assigned to each vEPC entity instance.

We have also proposed a DRP solution for the vEPC CP. This solution includes a novel resources dimensioning algorithm, which is based on the aforementioned model. Finally, we have validated the correct operation of DRP algorithm by simulation.

ACKNOWLEDGMENT

This work is partially supported by the European Unions Horizon 2020 research and innovation programme under the 5G!Pagoda project with grant agreement No. 723172, the Spanish Ministry of Education, Culture and Sport (FPU Grant 13/04833), and the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (TEC2016-76795-C6-4-R).

REFERENCES

[1] *5G Vision and Requirements*, White paper, IMT-2020 (5G) Promotion Group, May 2014.

[2] *Network Functions Virtualisation (NFV); Management and Orchestration*, ETSI GS NFV-MAN 001 V1.1.1, December 2014.

[3] R. Mijumbi, S. Hasija, S. Davy, A. Davy, B. Jennings, and R. Boutaba, "A connectionist approach to dynamic resource management for virtualised network functions," in *Network and Service Management (CNSM), 2016 12th International Conference on*. IEEE, 2016, pp. 1–9.

[4] A. Rajan *et al.*, "Understanding the bottlenecks in virtualizing cellular core network functions," in *Local and Metropolitan Area Networks (LANMAN), 2015 IEEE Int. Workshop on*, April 2015, pp. 1–6.

[5] Y. Ren, T. Phung-Duc, J. C. Chen, and Z. W. Yu, "Dynamic auto scaling algorithm (dasa) for 5g mobile networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[6] K. Tanabe, H. Nakayama, T. Hayashi, and K. Yamaoka, "An optimal resource assignment for c/d-plane virtualized mobile core networks," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[7] J. Prados-Garzon, J. J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado, and J. M. Lopez-Soler, "Modeling and dimensioning of a virtualized MME for 5G mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4383–4395, 2017.

[8] J. Prados-Garzon, P. Ameigeiras, J. J. Ramos-Munoz, P. Andres-Maldonado, and J. M. Lopez-Soler, "Analytical modeling for virtualized network functions," in *Communications Workshops (ICC Workshops), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–7.

[9] B. Hirschman *et al.*, "High-performance evolved packet core signaling and bearer processing on general-purpose processors," *IEEE Network*, vol. 29, no. 3, pp. 6–14, May 2015.

[10] *5G; Study on Scenarios and Requirements for Next Generation Access Technologies*, 3GPP TR 38.913 V14.2.0, 2017.

[11] *General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access*, 3GPP TS 23.401 Rel 12, 2014.

[12] W. Whitt, "The queueing network analyzer," *Bell System Tech. J.*, vol. 62, no. 9, pp. 2779–2815, Nov. 1983.

[13] A. Laghrissi, T. Taleb, M. Bagaa, and H. Flinck, "Towards edge slicing: Vnf placement algorithms for a dynamic realistic edge cloud environment," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[14] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin, "Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks," in *Proceedings of the 7th International Workshop on Hot Topics in Planet-scale mObile Computing and Online Social neTworking*, ser. HOTPOST '15. New York, NY, USA: ACM, 2015, pp. 19–24.