A Complete LTE Mathematical Framework for the Network Slice Planning of the EPC

Jonathan Prados-Garzon, Abdelquoddouss Laghrissi, Miloud Bagaa, Tarik Taleb, and Juan M. Lopez-Soler

***Abstract*—5G is the next telecommunications standards that will enable the sharing of physical infrastructures to provision ultra short-latency applications, mobile broadband services, Internet of Things, etc. Network slicing is the virtualization technique that is expected to achieve that, as it can allow logical networks to run on top of a common physical infrastructure and ensure service level agreement requirements for different services and applications. In this vein, our paper proposes a novel and complete solution for planning network slices of the LTE EPC, tailored for the enhanced Mobile BroadBand use case. The solution defines a framework which consists of: i) an abstraction of the LTE workload generation process, ii) a compound traffic model, iii) performance models of the whole LTE network, and iv) an algorithm to jointly perform the resource dimensioning and network embedding. Our results show that the aggregated signaling generation is a Poisson process and the data traffic exhibits self-similarity and long-range-dependence features. The proposed performance models for the LTE network rely on these results. We formulate the joint optimization problem of resources dimensioning and embedding of a virtualized EPC and propose a heuristic to solve it. By using simulation tools, we validate the proper operation of our solution.**

***Index Terms*—LTE, EPC, Network Slicing, NFV, Softwarized Networks, Mobile Networks, Traffic characterization, Resources dimensioning, and Network embedding.**

I. Introduction

FIFTH Generation (5G) mobile networks play a paramount role in the forthcoming global industrial digitalization. 5G will cover all the vertical market needs in a cost effective manner. 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 [1]–[4]. To meet these challenging goals, network softwarization (NS) is envisaged as the cornerstone to build the 5G technology [5], [6]. The concept of NS is mainly based on i) Network Function Virtualization (NFV), which decouples network functions from proprietary hardware enabling them to run as software on virtualization containers such as virtual machines (VMs) [7], and ii) Software Defined Networking (SDN), which fully separates control and data planes in network nodes allowing network programmability.

Under the NS approach, isolated, fully automated, programmable, flexible, and service-customized networks known as *network slices* can be deployed on top of a common physical infrastructure [8]–[10]. This approach is referred to as *network slicing*. It will allow the mobile operators to cover the different market scenarios and use cases that demand heterogeneous, diverse and possibly mutually incompatible requirements [5].

The adoption of network slicing in 5G mobile networks requires optimal solutions for planning the slices according to the different use cases requirements. This mainly involves the dimensioning of the resources and its embedding in a given infrastructure. Furthermore, these processes have to be done in a manner that ensures the Quality of Service (QoS) requirements for each use case. Likewise, faced with a decreasing Average Revenue Per User (ARPU), operators are challenged to reduce, or even optimize, i) the acquirement and maintenance of the physical infrastructure (*i.e.*, capital expenditures -CAPEX-), and ii) the ongoing expenses to properly operate the network equipment (*i.e.*, operating expenditures -OPEX-). Many techno-economic models have been proposed to reduce the CAPEX and OPEX such as in [11], [12].

Our work aims to design a complete solution for network slices planning of the LTE Evolved Packet Core (EPC), which is tailored for the enhanced Mobile BroadBand (eMBB) use case [7], [13]. To that end, we propose a framework consisting of the following components:

- An abstraction of the LTE workload generation process, for both Control Plane (CP) and Data Plane (DP), along with a compound traffic model that includes the most representative services consumed in current cellular networks. This is required to estimate the service consumption when there is no previous knowledge of the workload demand. This component is also useful to generate synthetic workloads for experimentation (*e.g.*, to stress a virtualized LTE network).
- Holistic analytical models to predict the performance (e.g., packet loss probability and response time) of a virtualized EPC (vEPC). We apply queuing theory and stochastic network calculus to develop the CP and DP models, respectively. For a given workload and a set of QoS requirements, our models facilitate resources dimensioning.
- The corresponding formulation and heuristic to solve the joint optimization problem of resources dimensioning and embedding of the vEPC. We have suggested a multi-objective optimization problem that minimizes the work-

Jonathan Prados-Garzon and Juan M. Lopez-Soler are with the Research Centre for Information and Communications Technologies of the University of Granada (CITIC-UGR); and the Department of Signal Theory, Telematics and Communications of the University of Granada, Granada, 18071 Spain (email: jpg@ugr.es, juanma@ugr.es).

Abdelquoddouss Laghrissi, Miloud Bagaa, and Tarik Taleb are with the Department of Communications and Networking, School of Electrical Engineering, Aalto University, Espoo, Finland. Tarik Taleb is also with the Centre for Wireless Communications (CWC), University of Oulu, 90014 Oulu, Finland, and also with the Computer and Information Security Department, Sejong University, 143-747 Seoul, AQ3 South Korea. (emails: abdelquoddouss.laghrissi@aalto.fi, bagmoul@gmail.com, tarik.taleb@aalto.fi).

load imbalances among a set of candidate Edge Clouds (ECs) (*i.e.*, Data Centers (DCs) deployed close to end users) and maximizes the resources utilization, on the network side, and Quality of Experience (QoE), on the end user's side. These objectives are subject to meet a set of QoS requirements. For the CP, the QoS requirements are defined as an upper bound on the average elapsed time to move a User Equipment (UE) from *IDLE* to *ACTIVE* states. For the DP, the QoS requirements considered are the limit on the maximum one-way network delay and a maximum packet loss probability at the vEPC. Additionally, we impose a condition to limit the maximum number of Central Processing Unit (CPU) cores to be assigned to a single Virtual Network Function Component (VNFC) instance. That is to take into account the actual limitation on the number of CPU cores of the Physical Machines.

Sharing network resources between different users has proven to reduce CAPEX and OPEX [14]–[16]. The abovementioned features, namely the performance-predictive models, the load balancing among ECs, and the maximization of resource utilization will certainly induce considerable cost savings. Also, the maximum number of CPU cores constraint will have an impact on reducing the costs due to OPEX [17]. Last, although the NS paradigm enables operators to dynamically adapt the resources allocated to each network slice and services [18], the on-demand plans offered by infrastructure providers are more expensive than the reservation plans. Specifically, resources can be purchased as a reservation for up to 70% off the on-demand price [19]. Thus, the network slices planning is crucial for operators to save money.

As a starting point, this work is meant to enhance the "Network Slice Planner" (NSP) [20]. NSP is a simulation tool that implements accurate models for the users' behavior, mobility, and data consumption in cellular networks. Specifically, we extend its data consumption model to include the most representative services consumed in current mobile networks. Then, by using NSP we characterize stochastically the aggregated workload generation processes for the CP and DP. Under our workload generation model, the results show that the aggregated signaling generation process follows a Poisson distribution and the aggregated DP workload exhibits Self-Similarity (SS) and Long-Range Dependence (LRD) features.

Based on the aforesaid results, we develop holistic performance models of a virtualized LTE network. The CP is modeled following the same technique as in [21] for chains of Virtual Network Functions (VNFs). The model includes the main LTE entities and their messages exchange. The DP is modeled as a queue fed by a fractional Brownian Motion (fBm) process [22]. These comprehensive models allow us to define efficient resources dimensioning algorithms.

Finally, the heuristic proposed in this work to solve the planning for vEPC relies on the aforementioned performance models. The algorithm is dubbed "Planner for the EPC as a Service" (PES). By using a system-level LTE simulator, we validate the correct operation of PES. We also show that PES embedding algorithm reduces the workload imbalances among candidate ECs in contrast to other baseline techniques.

The remainder of the paper is organized as follows. Section II briefly reviews the related literature. Section III describes the system model. Section IV includes the formulation of the joint optimization problem of resource dimensioning and embedding for the vEPC. In Section V, the modeling and analysis to estimate the performance of the CP and DP are presented. Next, in Section VI, we introduce the proposed heuristic to perform the planning of the vEPC. Section VII explains the experimental setup. Section VIII provides numerical results that show the proper operation of our solution. Finally, Section IX summarizes the main conclusions.

II. Related Works

This section briefly reviews the related literature. In particular, we focus on performance models and embedding algorithms (*i.e.*, on how to map VNFC instances to physical infrastructures) for the vEPC.

A. Modeling of the vEPC

Analytical models constitute an agile way to predict the performance of a system in advance. There are several proposals in the literature tackling the analytical modeling of parts or the entire vEPC [21], [23]–[27]. Invariably, these works employ queuing theory.

In [24], Rajan *et al.* model the EPC as a D/D/m node. They conclude that when simply replacing existing EPC elements with virtualized equivalents, severe performance bottlenecks occur. In [26], [27], Prados *et al.* analyze the performance of a virtualized Mobility Management Entity (vMME) with a three-tier design, inspired by web services, and using a Jackson's network (*i.e.*, a network of M/M/m queues). Each queue represents a tier or VNFC of the vMME. The authors show that the proposed model provides fairly good results for computational resources dimensioning. In [21], the same authors enhance the previous model by extending its applicability domain to any chain of VNFs, increasing its flexibility, and using a more accurate technique of analysis. Specifically, each VNFC instance is modeled as a G/G/m queue. The resulting network of queues is solved by using the approximated technique proposed by Whitt *et al.* in [28] for the Queuing Network Analyzer referred to, hereinafter, as the QNA method. For the abovementioned use case (a three-tiered vMME), the authors show the QNA method outperforms Jackson's networks and Mean Value Analysis techniques in terms of the response time estimation error. Tanabe *et al.* propose in [23] a bi-class (*i.e.*, Machine-to-Machine and Mobile Broadband -MBB- communications) queuing model for the vEPC. The CP and DP of the vEPC are modeled as M/M/m/m and M/D/1 nodes, respectively. This model constitutes the core of the vEPC-ORA method which aims to optimize the resource assignment for the CP and DP of the vEPC. Finally, in [25], Ren *et al.* propose a dynamic resource provisioning algorithm for the vEPC considering the capacity of legacy network equipment already deployed. To evaluate the performance of their solution, they model each vEPC element as a M/M/m/K queue and assume that the VNF instantiation time is exponentially distributed.

The aforementioned works only model parts of the EPC and/or do not capture the interactions among its elements.

In this paper, this gap is covered. We consider the main elements of the LTE network CP (*i.e.*, UE, evolved Node B -eNB-, MME, Serving Gateway -SGW-, Packet Data Network Gateway -PGW-, Home Subscriber Server -HSS-, and Policy and Charging Rules Function -PCRF-) as well as their interactions. In this way, it is possible to predict the performance of the whole LTE CP from the aggregated signaling generation process. In [23], [25], the resources dimensioning of the vEPC is also visited. Nevertheless, these works address the dimensioning of each component in an isolated way. Only then, it is necessary to define a processing delay budget for each entity to be dimensioned in advance. Our holistic model for an LTE network overcomes this limitation by enabling the resources dimensioning algorithm to consider an overall processing delay budget for the whole EPC. This leads to resources savings.

For the DP, we leverage the results obtained for the analysis of the LTE data traffic traces to derive its performance metrics. Specifically, the vEPC DP is modeled as a single queue fed by a fBm process. To the best knowledge of the authors, this is the first work that uses stochastic network calculus results for analyzing the performance of a vEPC.

B. Algorithms for the vEPC embedding

There is a rich literature proposing algorithms to embed the whole vEPC or some of its entities in a physical infrastructure [29]–[38]. In [29], Taleb *et al.* propose a heuristic algorithm for virtualized SGWs (vSGWs) embedding. The algorithm tries to minimize the frequency of mobility gateway relocations while ensuring that a maximum capacity for each vSGW, which handles the traffic load of a serving area, is not exceeded. This work is extended in [32] where some additional objectives and restrictions are considered. Regarding the objectives, the path between UEs and PGWs is minimized, and the overall network resource utilization is optimized. Concerning the restrictions, this work was a pioneer in considering some relevant third generation partnership project (3GPP) constraints.

In [30], Bagaa *et al.* address the embedding of the virtualized PGW (vPGW). The embedding problem is formulated as a multi-objective non-linear optimization problem which minimizes the costs for the network operators, maximizes the network performance, and balances the load equally among the vPGW instances. To solve the problem, three heuristic algorithms are proposed to achieve near-optimal solutions. In [31], Basta *et al.* investigate different approaches to deploy the core gateways (*i.e.*, SGW and PGW) in the DCs. Specifically, they consider a fully and partially virtualization approaches for the gateways. The former consists in moving the CP and DP functionalities of each gateway to a DC. The latter decouples CP and DP functionalities by using the SDN paradigm and only the CP part is hosted within a DC. In the same context; relying on an SDN framework that decouples the CP from the DP, Datsika *et al.* propose in [39] a Matching Theoretic Flow Prioritization algorithm that aims to improve the grade of service level and delay induced by the core network congestion. This approach allows over the top service providers to intervene in the virtual slices allocation process. It has permitted to achieve efficient flows prioritization with respect to the service providers' policies and QoS demands. In [33], Martini *et al.* formulate the problem of choosing the VNF instances provided by a distributed set of DCs to serve a given service chain request. The objective is to minimize the overall latency of the chain. This optimization problem can be formulated as a resource constrained shortest path problem. In [34] [35], Baumgartner *et al.* formulate the joint optimization problem of the virtual mobile core network topology composition and embedding. The formulation guarantees a maximum end-to-end latency and takes into account the processing, queuing, and propagation delays. In [36], Bagaa *et al.* address the placement of virtual instances of 4G (MME, SGW, PGW) and 5G (AMF, SMF and AUSF) core network elements over a federated cloud based on Mixed Integer Linear Programming and coalitional formation game. Finally, Dietrich *et al.* [37] formulate a mixed-integer linear program for the vSGW and vMME embedding. To reduce its time complexity, they transform it into a linear program by employing relaxation and rounding techniques. Their proposal mitigates the load imbalance in today's mobile networks, which improves request acceptance and resource utilization.

The resources dimensioning and embedding are treated throughout the literature as separate problems. These two stages of resources allocation are closely related and performing them in a coordinated way brings benefits. For instance, there is a trade-off between the workload balance among a set of candidate DCs (*i.e.*, propagation delays) and the resources utilization (*i.e.*, processing delays) when an overall delay budget to be met is partitioned among these two stages. Herein, we formulate the joint optimization problem for planning the vEPC to address this trade-off.

TABLE I: Notation.

Notation	Description
$m_l^{(c)}$	Number of dedicated physical CPU cores allocated to instance l of the VNFC $c \in C$.
m_{max}	Maximum number of dedicated physical CPU cores to be allocated to a single virtualization container.
$\overline{T}_e$	Actual mean response time for the CP entity $e \in E$.
$\overline{T}_{if}$	Actual mean response time for the LTE interface $if \in IF$.
$\overline{T}^{(SR)}$	Actual mean delay for the CP to carry out an SR procedure.
$\overline{T}_{budget}^{(CP)}$	Mean delay budget for the CP.
$T_u^{(max)}$	Actual maximum response time for the DP entity $u \in U = \{UE, eNB, DPGW\}$.
$T_{if}^{(max)}$	Actual maximum response time for the LTE DP interface $if \in IFU = \{Uu, S1-U\}$.
$T_{max}^{(DP)}$	Actual maximum delay for the DP.
$T_{budget}^{(DP)}$	Maximum delay budget for the DP.
$P^{(EPC)}$	Actual EPC packet loss probability.
$P_{budget}^{(EPC)}$	EPC packet loss probability budget.

III. System Model

A. System Architecture

Let us assume an evolved universal terrestrial radio access network (E-UTRAN), already deployed with I eNBs, which provides connectivity to a set of J UEs to the LTE EPC (see Fig. 1). Each UE j is attached to an eNB i.

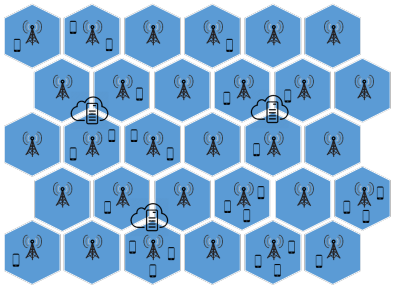


Fig. 1: E-UTRAN deployment and ECs sites.

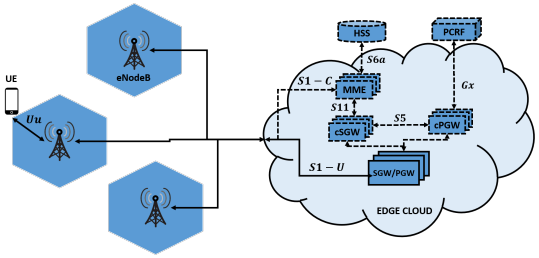


Fig. 2: Assumed LTE network architecture.

Let u_{ji} be a binary variable indicating whether the UE j is attached to the eNB i ($u_{ji} = 1$) or not ($u_{ji} = 0$). We consider the coverage map of this E-UTRAN as a rectangular area A with height h and width w.

Within A, there are already deployed K ECs (see Fig. 1). Let $\mathbf{r_i^{(eNB)}} = (x_i^{(eNB)}, y_i^{(eNB)}) \ \forall \quad i \in \mathbb{N} \cap \{1, .., I\}$, $\mathbf{r_j^{(UE)}} = (x_j^{(UE)}, y_j^{(UE)}) \ \forall \quad j \in \mathbb{N} \cap \{1, .., J\}$, and $\mathbf{r_k^{(EC)}} = (x_k^{(EC)}, y_k^{(EC)}) \ \forall \quad k \in \mathbb{N} \cap \{1, .., K\}$ denote two dimensional vectors representing the positions of eNBs, UEs, and ECs within A, respectively.

The MME, SGW, and PGW of the EPC will be implemented as a set of VNFs that makes up a network service [18], hereafter referred to as vEPC, and deployed on the candidate ECs. We discard the option of deploying the vEPC as a single VNF with several components (VNFCs), since, in this work, we will assume that the LTE EPC internal interfaces such as S11 and S5 will remain unchanged. Other EPC entities, such as the HSS and the PCRF, might be located outside of the ECs and implemented either as VNFs or physical network functions (PNFs).

The aggregated workload generated by the J UEs attached to the E-UTRAN is distributed among the K candidate ECs. This workload distribution is performed at the granularity of eNBs (i.e., each eNB i is assigned to a candidate EC k). Let v_{ik} be a binary variable indicating whether the eNB i is assigned to the EC k (i.e., $v_{ik} = 1$) or not (i.e., $v_{ik} = 0$). To serve its corresponding workload, a vEPC is instantiated on each EC.

The LTE network architecture deemed in this work is depicted in Fig. 2. We consider that the CP and DP of the vEPC are fully decoupled. Also, we assume the interfaces, between the CP functional entities, as the ones defined in the 3GPP LTE standards. Consequently, each CP entity (e.g., the MME, and the control functionalities of the SGW and PGW -cSGW and cPGW-) are implemented separately as a single VNF with a single component (VNFC). The DP functionalities of the SGW and PGW are integrated on a single VNF, with only one VNFC, that exposes the LTE S1-U and SGi interfaces. We assume that all VNFCs of the vEPC execute CPU-intensive tasks. Each VNFC might have multiple instances. Considering the ETSI NFV architectural framework and terminology [40][18] and without loss of generality, each VNFC instance is supposedly running on an isolated virtualization container such as a VM. Let $m_l^{(c)}$ denote the number of dedicated physical CPU cores allocated to the instance l of the VNFC $c \in C = \{MME, cSGW, cPGW, DPGW\}$. Since the number of CPU cores of a physical server is finite and the latter are shared among several VMs, we consider that $m_l^{(c)}$ is limited to m_{max} (i.e., $m_l^{(c)} \leq m_{max}$).

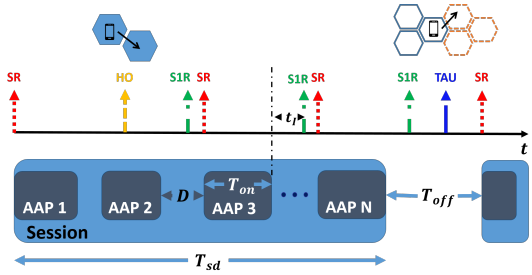


Fig. 3: Workload generation model.

B. Workload generation model

In this paper, we address the eMBB use case. In this context, the UEs run applications that generate and consume DP traffic. We consider the abstraction presented in [27] for such a process (see Fig. 3).

A session with duration T_{sd} is defined as the user's activity beginning from the time an application is launched to the time it closes. 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, or send an instant message). A reading time is the temporal interval during which the user performs any action that does not require to generate network traffic such as deciding which friend's profile to visit next or reading a message.

Regarding the signaling workload, the users' activity and mobility trigger the LTE CP procedures. In this work, we only consider the 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 [41].

Once the UE is registered in the network, an SR procedure is triggered during its idle-to-connected (i.e., *IDLE* to *ACTIVE*) transitions. Then, whenever an AAP starts while the UE is in idle mode, an SR procedure takes place (see Fig. 3).

Conversely, an S1R procedure occurs during UE's connected-to-idle transitions during which the network releases the UE's resources. We also take into account the effects of an inactivity timer. Its value is denoted as t_I. The network waits t_I units of time after that an AAP finishes before triggering an S1R (see Fig. 3). A HO procedure is triggered when a UE is in connected mode and performs a cell change, but the target cell is attached to the same MME as the source cell's. Finally, we assume that a TAU procedure is triggered whenever a UE carries out a Tracking Area (TA) change. These TAs are predefined and are the same for any UE.

C. Performance Requirements

The LTE network has to meet a set of performance requirements in terms of latency and packet loss probability [42]. For the CP, the considered performance requirement is an upper bound on the mean CP latency $\overline{T}^{(CP)}_{budget}$ defined by the 3GPP, i.e., the average elapsed time to move an UE from IDLE state to ACTIVE state [42]. 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, where the UE authentication, NAS (Non-Access Stratum) security setup, and the EPS (Evolved Packet System) session modification steps occur during the SR.

Let $\overline{T}_e$ and $\overline{T}_{if}$ denote, respectively, the mean response times of the CP entity $e \in E = \{UE, eNB, MME, cSGW, cPGW, HSS, PCRF\}$ 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. Also, 8, 7, 2, 2, 2, and 2 control messages have to traverse, respectively, the LTE Uu, S1-C, S11, S6a, S5, and Gx interfaces [43]. Then, the CP delay requirement can be expressed as $\overline{T}^{(SR)} \leq \overline{T}^{(CP)}_{budget}$.

For the DP, the performance requirements considered are the maximum DP delay budget $T^{(DP)}_{budget}$ and the packet loss probability at the vEPC $P^{(EPC)}_{budget}$. We consider $T^{(DP)}_{budget}$ as the maximum time it takes for a packet to travel from the SGi interface at the SGW/PGW VNFC to the UE application. The $P^{(EPC)}_{budget}$ is the maximum allowable packet loss at the DPGW VNFC receive buffer.

Let $T^{(DP)}_{max}$ and $P^{(EPC)}$ denote the actual maximum delay of the DP and the packet loss probability of the EPC, respectively. We can compute $T^{(DP)}_{max}$ as:

$$T^{(DP)}_{max} = T^{(max)}_{UE} + T^{(max)}_{eNB} + T^{(max)}_{DPGW} + T^{(max)}_{Uu} + T^{(max)}_{S1-U} \tag{2}$$

where: $T^{(max)}_{UE}$, $T^{(max)}_{eNB}$, and $T^{(max)}_{DPGW}$ are respectively the actual maximum DP packet processing delay at the UE, eNB, and DPGW. And $T^{(max)}_{Uu}$ and $T^{(max)}_{S1-U}$ are the actual maximum delays for the DP radio and backhaul interfaces, respectively. Then, the DP requirements can be expressed as $T^{(DP)}_{max} \leq T^{(DP)}_{budget}$ and $P^{(EPC)} \leq P^{(EPC)}_{budget}$.

IV. Problem formulation

In this section, we formulate the joint optimization problem to distribute the aggregated workload generated by the E-UTRAN among the candidate ECs and to perform the dimensioning of the required resources for each vEPC instance. Taking into account the defined system model, it can be formulated as follows:

Objectives :

$$minimize \left(\sum_{k=1}^{|K|} \left| \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} v_{ik} u_{ij} - \frac{|J|}{|K|} \right| \right) \tag{3a}$$

$$minimize \left(\sum_{k=1}^{|K|} \sum_{i=1}^{|I|} v_{ik} \cdot d_{ik} \right) \tag{3b}$$

$$minimize \left(\sum_{k=1}^{|K|} \sum_{c \in C} \sum_{l} m_l^{(c)} \right) \quad m_l^{(c)} \in \mathbb{N} \tag{3c}$$

where $d_{ik} = ||r_i^{(eNB)} - r_k^{(EC)}||$ is the Euclidean distance between eNB i and EC k.

Constraints :

CP :

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

DP :

$$C2: \quad max\left(T^{(DP)}\right) \leq \overline{T}^{(DP)}_{budget}, \tag{3e}$$

$$C3: \quad P^{(EPC)} \leq P^{(EPC)}_{budget}, \tag{3f}$$

Others

$$C4: \quad m_l^{(c)} \leq m_{max} \quad \forall \quad k \in [1, |K|] \cap \mathbb{N} \tag{3g}$$

$$C5: \quad \sum_{k=1}^{|K|} \sum_{i=1}^{|I|} v_{ik} = |I|, \quad v_{ik} \in \{0, 1\} \tag{3h}$$

The decision variables of the optimization problem are v_{ik} and $m_l^{(c)}$. Objective (3a) aims to distribute the workload as equally as possible or to minimize the workload imbalances across the candidate ECs. The goal is optimally achieved when the same number of users ($|J|/|K|$) is assigned to every EC $k \in K$. Objective (3b) aims to minimize the propagation delays. The corresponding objective function is minimized when every eNB $i \in I$ is assigned to the nearest EC $k^* \in K$, where $k^* = argmin_{k \in K}(d_{ik})$. Last, objective (3c) intends to minimize the total number of CPU instances allocated to the vEPC or, equivalently, to maximize the utilization of the computational resources.

Constraints (3d), (3e), and (3f) guarantee that the QoS requirements are fulfilled. Specifically, Constraint (3d) ensures

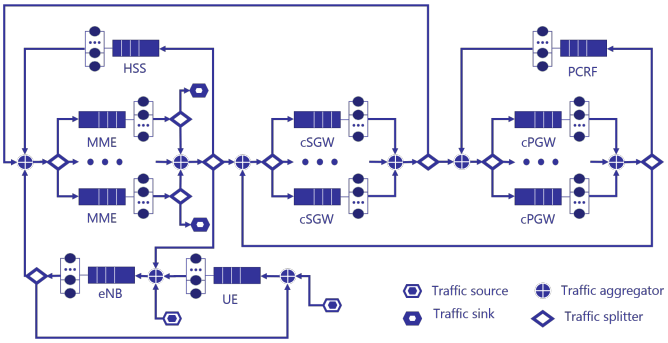


Fig. 4: LTE control plane model.

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 (3e) and (3f) ensure that the maximum DP delay budget and the packet loss probability at the EPC are met, respectively. Constraint (3g) limits the maximum number of physical cores requested for a single VNFC instance. Having a single VNFC instance would be optimal for minimizing the amount of required resources (statistical multiplexing). However, each physical server has a maximum number of physical cores, i.e., the number of physical cores we can request per VNFC instance is limited. Moreover, in general, the higher is the number of physical cores requested for a VNFC instance, the lower is its availability. Finally, Constraint (3h) guarantees that all eNBs are assigned to a candidate EC k (or vEPC instance k).

V. Analysis and Modeling

A. LTE CP modeling

We model the CP of the LTE as an open network of G/G/m[1] queues (see Fig. 4), where each queuing node represents an instance of a given entity of the LTE network. The MME, cSGW, and cPGW might have several instances, each of which is modeled as a G/G/m queuing node with $m_l^{(c)}$ servers. The servers of a queuing node represent the CPU instances, allocated to the entity instance, processing control messages in parallel. As stated in Section III-A, $m_l^{(C)} \leq m_{max}$. For the sake of simplicity, only one instance is considered for the rest of LTE CP entities (e.g., UE, eNB, HSS, and PCRF). The corresponding G/G/m queuing node that models the instance of these entities might have an arbitrary number of servers as they might be deployed as PNFs.

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.

To solve the network of queues, we employ the QNA method [28] which is described in Appendix A. This technique was applied and validated in [21] to estimate the mean response time of a VNF with several VNFCs. In this work, we use the QNA method to estimate the mean response times of the LTE CP entities $\overline{T}_e \, \forall \, e \in E$. To that end, the QNA method uses a reduced set of the following input parameters:

- The steady state transition probabilities matrix $P = [p_{ki}]$, where p_{ki} denotes the probability of a packet to leave node k to node i and $p_{0k} = 1 - \sum_i p_{ki}$ denotes the probability of a packet at node k to leave the network. In this work, we provide the expressions to compute P for the LTE CP (refer to Appendix B).
- The mean and squared coefficient of variation (SCV) of the external arrival processes at node k, λ_{0k}, and c_{0k}^2. Please note that only the UE and the eNB have external arrival processes in our model (see Fig. 4). Considering the abstraction described in section III-B for the signaling generation process, we found that these arrival processes are Poissonian (see Section VIII-A). Then, $c_{0k}^2 = 1 \, \forall \, k$.
- The mean and the SCV of the service processes at each queue k, μ_k and c_{sk}^2.

B. LTE DP modeling

For the considered architecture, the LTE DP consists of three network entities namely, UE, eNB, and DPGW, which are connected in tandem. Since the focus is on the vEPC dimensioning, we assume that the UE and eNB entities have constant maximum delays.

The same methodology as applied to model the LTE CP cannot be used to model the vEPC DP as it can only provide overall mean performance metrics of a queuing network, but not the performance bounds such as those defined in Section III-C (e.g., $T_{max}^{(DP)}$ and $P^{(EPC)}$) for the DP. Moreover, the stochastic characterization of the aggregated DP traffic carried out in this work (see Section VIII-A) shows that the vEPC DP workload arrival process exhibits SS and LRD features. Conventional queuing theory does not comprise such kind of arrival process [44]. Then, we model the DPGW as a single queue fed by a fBm process. More precisely, we use the model that was first reported in [22] and also derived in [44] from stochastic network calculus results. This model can provide the performance bounds of a tandem of queues with SS and LRD input in an effective and simple way.

To characterize the arrival process, we adopt the model proposed in [22]. Let A_t denote the cumulating arrival process to the DPGW queue, i.e., the cumulative amount of traffic (i.e., in number of packets) arriving at the DPGW in the time interval $[0, t]$. The following model is considered for A_t [22]:

$$A_t = \lambda \cdot t + \sqrt{\lambda \cdot \alpha} \cdot Z_t \tag{4}$$

where Z_t is a normalized fBm parameter with Hurst parameter $H \in (1/2, 1]$, $\lambda > 0$ is the mean input rate, and $\alpha > 0$ is a variance coefficient.

Under the above packet arrival model and considering a constant rate server with capacity C, the violation probability $\epsilon = P[B > b]$ of a backlog bound b can be approximated as [22], [44]:

$$\epsilon \approx exp\left(-\frac{(C-\lambda)^{2H}}{2 \cdot \kappa(H)^2 \cdot \lambda \cdot \alpha} b^{2-2H} \right) \tag{5}$$

[1] 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.

where $\kappa(H) = H^H(1-H)^{1-H}$. The above equation gives us an approximation for the probability of saturation of a buffer of size b packets or equivalently the packet loss probability at a queue fed with a fBm arrival process.

Finally, the maximum response time of a queuing node with buffer size b and constant rate server with capacity C can be computed as:

$$T^{(max)} = \frac{b+1}{C} \tag{6}$$

By using (5) and (6), we can perform the dimensioning of the required capacity of the DPGW.

VI. PES: Planner for the EPC as a Service

Algorithm 1 PES Algorithm

Input: eNBs positions $r_i^{(eNB)}$ along with the number of UEs they serve $\mathbf{N}_{eNB}^{UE}(i) = \sum_j u_{ji}$, and the QoS specs $T_{budget}^{(DP)}$, $P_{budget}^{(EPC)}$, and $\overline{T}_{budget}^{(CP)}$.

Output: eNBs assigment (i.e., v_{ik}), and total number of processing instances allocated to each vEPC entity per EC (e.g., $\mathbf{m}_{MME}$, $\mathbf{m}_{cSGW}$, $\mathbf{m}_{cPGW}$, and $\mathbf{m}_{DPGW}$).

1: $[\mathbf{N}_{EC}^{UE}, v_{ik}] \Leftarrow Partitioning(\mathbf{r}^{(eNB)}, \mathbf{N}_{eNB}^{UE})$
2: **for** each $k \in K$ **do**
3: Compute the processing delay budgets for the vEPC CP and DP, $T_{proc-budget}^{(CP)}$ and $T_{proc-budget}^{(DP)}$, using (7) and (8).
4: For $N_U = \mathbf{N}_{EC}^{UE}(k)$, estimate the external arrival processes ($\lambda^{(CP)}$, $\lambda^{(DP)}$, $\alpha^{(DP)}$, and $H^{(DP)}$) using (9)-(15)
5: $[\mathbf{m}_{MME}(k), \mathbf{m}_{cSGW}(k), \mathbf{m}_{cPGW}(k), \mathbf{m}_{DPGW}(k)] \Leftarrow Dimensioning(\lambda^{(CP)}, \lambda^{(DP)}, \alpha^{(DP)}, H^{(DP)}, T_{proc-budget}^{(CP)}, T_{proc-budget}^{(DP)}, P_{budget}^{(EPC)})$
6: **end for**

In this section, we propose a heuristic method to find a sub-optimal solution of the problem formulated in Section IV. To achieve a method with low-complexity, we decouple the process of workload distribution among the candidate ECs and the resources dimensioning of the vEPC at each EC. The heuristic method, depicted in Algorithm 1, proceeds as follows. Initially, the partitioning algorithm assigns each eNB to a candidate EC (see Algorithm 2). The idea in this algorithm is to distribute the workload as equally as possible among the candidate ECs, while guaranteeing a maximum propagation delay for the backhaul network $t_{prop-backhaul}^{(max)}$. The algorithm initializes the workload assigned to each EC k $\mathbf{N}_{EC}^{UE}(k)$, which is measured as the number of assigned UEs, to zero. Then, it iteratively finds the candidate EC k^* with the lowest workload allocated and its nearest eNB i^* being not assigned yet. If the propagation delay limit between the EC k^* and the eNB i^* is not violated, then, the eNB i^* is attached to the EC k^* ($v_{i^*k^*} = 1$). Otherwise, the EC k^* is excluded from the set of candidate ECs K. The algorithm ends when all eNBs are allocated. Observe that, in the worst case scenario, the algorithm requires $N_{eNB} + N_{EC}$ iterations to assign all eNBs.

Please note that the number of UEs attached to each eNB is assumed to be known. On the one hand, if the E-UTRAN is in the operation phase, the operator can know accurately the average number of UEs attached to each eNB. On the other hand, if the E-UTRAN is not in the operation phase, the operator can estimate the average number of UEs attached to each eNB from the population density map of the coverage geographical area and the expected market shares.

Algorithm 2 E-UTRAN Partitioning Algorithm

Require: All eNBs of the set I have to be assigned to an EC of the set K.

Input: eNBs positions $r_i^{(eNB)}$ along with the number of UEs they serve $\mathbf{N}_{eNB}^{UE}(i) = \sum_j u_{ji}$, the ECs positions $r_k^{(EC)}$, and the maximum propagation time for the backhaul network $t_{prop-backhaul}^{(max)}$.

Output: eNBs assigment, i.e., v_{ik}

Initialization $\mathbf{N}_{EC}^{UE} = \overrightarrow{0}$, $v_{ik} = 0$
while $I \neq \emptyset$ **do**
  $k^* = \arg\min_{k \in K}(\mathbf{N}_{EC}^{UE}(k))$
  $i^* = \arg\min_{i \in I} ||\mathbf{r}_{k^*}^{(EC)} - \mathbf{r}_i^{(eNB)}||$
  if $||\mathbf{r}_{k^*}^{(EC)} - \mathbf{r}_{i^*}^{(eNB)}|| \leq t_{prop-backhaul}^{(max)} \cdot c$ **then**
    $I \Leftarrow I \backslash i^*$, $v_{i^*k^*} = 1$
    $\mathbf{N}_{EC}^{UE}(k^*) \Leftarrow \mathbf{N}_{EC}^{UE}(k^*) + \mathbf{N}_{eNB}^{UE}(i^*)$
  else
    $K \Leftarrow K \backslash k^*$
  end if
end while

Once the eNBs assignment is carried out, the processing time budgets for the vEPC CP $T_{proc-budget}^{(CP)}$ and DP $T_{proc-budget}^{(DP)}$ can be computed. To that end, we can evaluate $\overline{T}^{(SR)}$ and $T_{max}^{(DP)}$, in (1) and (2), for $\overline{T}_{MME}$, $\overline{T}_{cSGW}$, $\overline{T}_{cPGW}$, and $T_{DPGW}^{(max)}$, equal to zero, respectively. Formally, $\overline{T}_0^{(SR)} = \overline{T}^{(SR)}(\overline{T}_{MME} = 0, \overline{T}_{cSGW} = 0, \overline{T}_{cPGW} = 0)$ and $T_{max0}^{(DP)} = T_{max}^{(DP)}(T_{DPGW}^{(max)} = 0)$. Then,

$$T_{proc-budget}^{(CP)} = \overline{T}_{budget}^{(CP)} - \overline{T}_0^{(SR)} \tag{7}$$

$$T_{proc-budget}^{(DP)} = T_{bugdet}^{(DP)} - T_{max0}^{(DP)} \tag{8}$$

Then, once there is an estimation of the number of UEs to be served by each EC, we can also estimate the aggregated external arrival processes, for both the LTE CP and DP, which are inputs to the resources dimensioning algorithm. We use an abstraction of the LTE workload generation process, along with a compound traffic model, to perform such an estimation. We characterize stochastically these arrival processes in Section VIII-A, where the curve fittings are provided to estimate the main parameters to model them as a function of the users' number.

Finally, the resources dimensioning is carried out (see Algorithm 3). The dimensioning of the vEPC CP and DP is performed separately. Since we are considering only one VNFC for the vEPC DP, its dimensioning simply requires solving numerically (5). For the CP, we propose a novel algorithm which searches for the minimum number of processing instances to be allocated to the vEPC CP for a given EC so that a processing delay budget $T_{proc-budget}^{(CP)}$ is met. The algorithm

Algorithm 3 Dimensioning Algorithm

Input: Processing delay budgets for the vEPC CP $T^{(CP)}_{proc-budget}$ and DP $T^{(DP)}_{proc-budget}$; $P^{(EPC)}_{budget}$; External arrival processes characterization for CP and DP ($\lambda^{(CP)}$, $\lambda^{(DP)}$, $\alpha^{(DP)}$, and $H^{(DP)}$).

Output: number of physical cores allocated to each vEPC entity m_{MME}, m_{cSGW}, m_{cPGW}, and m_{DPGW}

1: {DATA PLANE:}
2: Solve (5) numerically for $b \leq T^{(DP)}_{proc-budget} \cdot C - 1$ and $\epsilon \approx P^{(EPC)}_{budget}$ to obtain the required DP processing capacity C. Then, $m_{DPGW} = \lceil C/\mu_{DPGW} \rceil$.
3: {CONTROL PLANE:}
4: **Initialization** $m_{MME} = \lceil \lambda_{MME}/\mu_{MME} \rceil$, $m_{cSGW} = \lceil \lambda_{cSGW}/\mu_{cSGW} \rceil$, $m_{cPGW} = \lceil \lambda_{cPGW}/\mu_{cPGW} \rceil$, $M_{CP} = m_{MME} + m_{cSGW} + m_{cPGW}$, $T^{(CP)}_{proc} = 8 \cdot T_{MME}(m_{MME}) + 3 \cdot T_{cSGW}(m_{cSGW}) + 2 \cdot T_{cPGW}(m_{cPGW})$;
5: **while** $T^{(CP)}_{proc} > T^{(CP)}_{proc-budget}$ **do**
6: $M_{CP} \Leftarrow M_{CP} + 1$
7: **for** each $m \in \{m_{MME}, ..., M_{CP} - m_{cSGW} - m_{cPGW}\} \cap \mathbb{N}$ **do**
8: **for** each $n \in \{m_{cSGW}, ..., M_{CP} - m_{MME} - m_{cPGW}\} \cap \mathbb{N}$ **do**
9: $l = M_{CP} - m - n$
10: $T_{aux} = 8 \cdot T_{MME}(m) + 3 \cdot T_{cSGW}(n) + 2 \cdot T_{cPGW}(l)$
11: **if** $T^{(CP)}_{proc} > T_{aux}$ **then**
12: $T^{(CP)}_{proc} \Leftarrow T_{aux}$, $m_{MME} \Leftarrow m$, $m_{cSGW} \Leftarrow n$, $m_{cPGW} \Leftarrow l$
13: **end if**
14: **end for**
15: **end for**
16: **end while**

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 VNFCs to be dimensioned (e.g., MME, cSGW, cPGW), and choose the one providing the lowest processing delay. To achieve the linear complexity, the search space is limited at each iteration (see line 12 of Algorithm 3). In the algorithm, $T_{mme}(m)$, $T_{cSGW}(n)$, and $T_{cPGW}(l)$ denote, respectively, 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 Appendix A). Please note that, although it is not explicitly included in Algorithm 3, for each 'processing instances allocation (m, n, l), it is necessary to re-estimate both the internal flow parameters at each queue, using (16)-(22), and the transition probability matrix, using (31)-(41).

The number of instances or, equivalently, the number of virtualization containers for each vEPC entity at a given EC can be simply computed as follows: $\lceil m_{MME}/m_{max} \rceil$, $\lceil m_{cSGW}/m_{max} \rceil$, and $\lceil m_{cPGW}/m_{max} \rceil$, and $\lceil m_{DPGW}/m_{max} \rceil$.

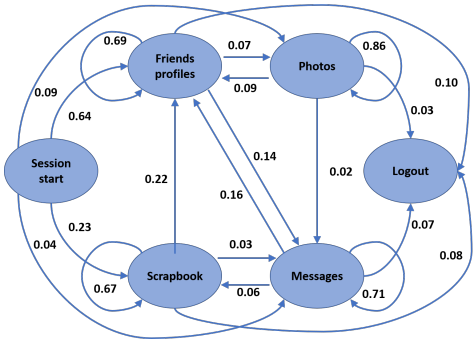


Fig. 5: Markov chain based model for social networking.

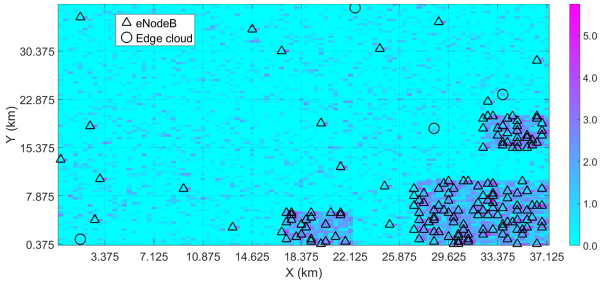


Fig. 6: Scenario realization with a population density of 1000 users per km^2.

VII. Experimental Setup

To validate the models developed in this work and to assess our solution for EPC slices planning, we employed two software tools: i) the NSP [20], and ii) a system-level simulator of an LTE network.

A. Network Slice Planner

We used the NSP [20] to generate the synthetic signaling and data traffic in an LTE network. We extended the compound traffic model of this tool by including the traffic models employed in [27]. The setup for each service type (see Table II and Fig. 5) relies on models taken from the literature, which are derived from real traces. Specifically, the main references used for the different services setup are [46] for social networking; [47] for Mobile Instant Messaging; [48] for web browsing; [49] and [45] for video streaming; and [50] and [51] for video calls. According to [52], the services considered account for more than 70% of the peak aggregate traffic in the American mobile access networks.

The output traces of the NSP were used to characterize the aggregate packet arrival processes at the LTE CP and DP. These traces are also used as inputs for our system-level LTE network simulator.

B. LTE network simulator

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 inputs of the simulator

TABLE II: Compound traffic model for Mobile Broadband users

Traffic Type	Parameters	Statistical Characterization
Social Networking $P_{app} = 0.20$	Inter-arrival session times (T_{sst})	Log-normal distribution: μ=2.245 σ=1.333 (samples in seconds)
	Number of APPs per session (N)	From Markov chain depicted in Fig. 5.
	Reading time (D)	Log-normal distribution: $\mu = 1.789$, $\sigma = 2.366$ (samples in seconds).
	AAPs length (T_{on})	Request data consumption (Markov Chain): i) friend page = 1300 kB, ii) message page = 1 MB, iii) scrapbook page = 2 MB, and iv) photo page = 750 kB.
Video streaming $P_{app} = 0.20$	Inter-arrival session times (T_{sst})	Log-normal distribution: $\mu = 2.1$, $\sigma = 1.3$ (samples in seconds).
	Number of APPs per session (N)	1 (Constant)
	Reading Time (D)	Since $N = 1$, *no reading times*
	AAPs length (T_{on})	• Video length: power-law ($x_{min} = 32.8285$, $\alpha = 2.2619$) (samples in seconds) • Video resolutions: i) 360p: 3 Mb/min, ii) 480p: 5 Mb/min, iii) 720p: 10 Mb/min, and iv) 1080p: 15 Mb/min. • Download model according to [45]
Mobile Instant Messaging $P_{app} = 0.20$	Inter-arrival session times (T_{sst})	Log-normal distribution: $\mu = 2.411$, $\sigma = 2.276$ (samples in seconds)
	Number of APPs per session (N)	1 (constant)
	Reading Time (D)	Since $N = 1$, *no reading times*
	AAPs length (T_{on})	Message length (in KB): Power-law distribution ($x_{min} = 0.4823$ KB, $\alpha = 2.2566$).
Web browsing $P_{app} = 0.20$	Inter-arrival session times (T_{sst})	Exponential distribution: $\lambda^{-1} = 1200$ seconds
	Number of AAPs per session (N)	Geometric distribution: $p = 0.893$
	Reading times (D)	Exponential distribution: $\lambda^{-1} = 30$ seconds
	AAPs length (T_{on})	• Main object size: Truncated log-normal distribution: $\mu = 15.098$, $\sigma = 4.39 \cdot 10^{-5}$, $min = 100$ B, $max = 6$ MB (samples in bytes). • Embedded object size: Truncated log-normal distribution: $\mu = 6.17$, $\sigma = 2.36 \cdot 10^{-5}$, $min = 50$ B, $max = 2$ MB (samples in bytes). • Number of embedded objects per webpage: Truncated Pareto distribution: $mean = 22$, $shape = 1.1$. • Parsing time: Exponential distribution: $\lambda^{-1} = 0.13$ seconds.
Video calling $P_{app} = 0.20$	Inter-arrival session times (T_{sst})	Exponential distribution: $\lambda^{-1} = 1200$ seconds
	Number of APPs per session (N)	1 (constant)
	Reading Time (D)	Since $N = 1$, *no reading times*
	AAPs length (T_{on})	Pareto distribution: $k = -0.39$, $s = 69.33$, and $m = 0$ (samples in seconds)

to emulate the workload generation in the LTE network. To distribute the users through the coverage area of the E-UTRAN, we employed the model presented in [53]. To generate Radio Access Network (RAN) deployment (i.e., the distribution of the eNBs), we adapted the heuristic proposed in [54]. Fig. 6 shows the synthetic E-UTRAN scenario considered in this work for a population density of 1000 UEs/km^2. The scenario consists of three urban zones where most of the population is concentrated. Additionally, four candidate ECs are considered, and their positions are randomly generated.

Each LTE functionality deployed as a VNFC of the vEPC is simulated as a First Come First Served (FCFS) queue with multiple generic servers. The rest of the LTE entities (e.g., UE, eNB, HSS, and PCRF) and the network delays (e.g., transmission, propagation, and switches processing), among any couple of EPC entities, are simulated as infinite servers (i.e., constant processing delay without a queuing waiting time). Table III includes the configuration of the main parameters for the simulator. The distribution of the service time for each entity to be deployed was obtained experimentally.

VIII. Numerical Results

In this section, some numerical results are reported to assess the proposed solution for the planning of LTE EPC slices.

A. Workload Characterization

By using NSP, we generated signaling and data traffic traces for 100000 UEs and different population densities. The simulated measurement period was set to 10000 seconds. Following the analysis of the traces in [55], we depicted the rate process on 6 different time scales for both CP and DP traffics. The chosen time scales were 1 ms, 10 ms, 100 ms, 1 s, 10 s, and 100 s. From these representations, we concluded that the DP traffic showed self-similarity (i.e., statistically indistinguishable on different time scales). The same phenomenon was not observed for CP traces. Based on the aforementioned observations, we measured $\lambda^{(DP)}$, $\alpha^{(DP)}$, and $H^{(DP)}$ parameters of the traffic model introduced in Section V-B for the DP traces (see Fig. 7).

To estimate the mean rate $\lambda^{(DP)}$, we simply counted the number of packets collected in the trace and divided it by the simulated measurement period. We obtained that in average, each UE generates around 5.1121 packets per second considering the compound traffic model defined in Table II. We got a similar result regardless of the number of users N_U and the population density. Consequently, $\lambda^{(DP)} = 5.1121 \cdot N_U$.

To measure $\alpha^{(DP)}$ and $H^{(DP)}$, we followed the same procedure as in [22]; we performed a linear regression from the logarithms of the sample variances of the increments of A_t for the 6 different time scales considered. Note that it is

TABLE III: Parameters Configuration

eNB configuration	
Maximum Tx power	20 W
Noise power	$4 \cdot 10^{-21}$ W/Hz
Number of antennas	1
Antenna gain	10 dB
Carrier frequency	2.3 GHz
Bandwidth	20 MHz
Noise figure	8 dB
Std of log-normal shadowing	8 dB
Spectral efficiency	10 $bits/Hz$
Minimum SNR Requirement	3.5 dB
Inactivity timer value	10 s
Service processes and mean response times for CP	
μ_{MME}, μ_{cSGW}, and μ_{cPGW}	6700 packets per second
$c^2_{s_{MME}}$, $c^2_{s_{cSGW}}$, and $c^2_{s_{cPGW}}$	0.65
$\overline{T}_{UE}$, $\overline{T}_{eNB}$, $\overline{T}_{HSS}$, and $\overline{T}_{PCRF}$	1 ms
$\overline{T}_{S6a}$ and $\overline{T}_{Gx}$	1.5 ms
$\overline{T}_{S11}$ and $\overline{T}_{S5}$	30 μs
Service processes for DP	
μ_{DPGW} (per CPU instance)	1813236 packets per second
$T^{(max)}_{UE}$	100 μs
$T^{(max)}_{eNB}$	200 μs
Propagation delays	
Speed of light in air	$3 \cdot 10^8$ m/s
Speed of light in fiber	$2 \cdot 10^8$ m/s
QoS requirements	
$\overline{T}^{(CP)}_{budget}$	25 ms
$T^{(DP)}_{budget}$	1 ms
$P^{(EPC)}_{budget}$	10^{-6}

supposed that A_t defined in (4) has stationary increments [22], and $VAR[A_t] = \lambda \cdot \alpha \cdot t^{2H}$. Figure 7 shows the measured $\alpha^{(DP)}$ and $H^{(DP)}$ for different numbers of users.

The measured values for $H^{(DP)}$ versus N_U range from 0.7 to 0.93 confirming the LRD of the DP traffic and the fBm process is suitable to model the aggregated DP traffic generation process. We fitted the following model:

$$H^{(DP)} = \left(H_0^{(DP)} - H_{max}^{(DP)}\right) \cdot e^{-\delta \cdot N_U} + H_{max}^{(DP)}$$

to the experimental $H^{(DP)}$ curve (i.e., $H^{(DP)}$ versus N_U). Where $H_{max}^{(DP)} = \lim_{N_U \to \infty} H^{(DP)}$, $H_0^{(DP)}$ is the value of $H^{(DP)}$ when $N_U = 0$, and δ is a rate constant. This model is appropriate to fit $H^{(DP)}$ because it is bounded (we know beforehand that $0 \leq H^{(DP)} \leq 1$) and its shape fits the experimental data (we obtained an R-squared of 96.82%).

Regarding $\alpha^{(DP)}$, the slope of $\alpha^{(DP)}$ versus N_U decreases when N_U increases, though the rate of change of this slope tends to zero in the range of N_U observed. Then, it seems reasonable to estimate $\alpha^{(DP)}$ at any N_U by using the linear function defined from the last two points of measurements (e.g., $N_U = 75000$ and $N_U = 100000$).

In summary, for a given number of users N_U, we can estimate the main parameters of the fBm process that models the aggregated data traffic for our compound traffic model as the following:

$$\lambda^{(DP)} = 5.1121 \cdot N_U \qquad packets/second \tag{9}$$

$$\alpha^{(DP)} = 0.00216 \cdot N_U + 1637.7 \qquad packets \cdot second \tag{10}$$

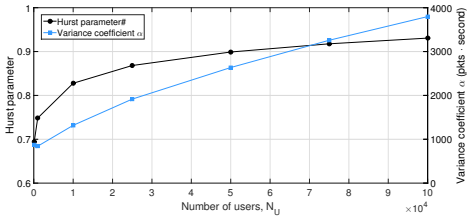


Fig. 7: Variance coefficient α and Hurst parameter H measurements versus the number of users N_U for the aggregated DP arrival process.

$$H^{(DP)} = 0.9172 - 0.2025 \cdot e^{-6.9430 \cdot 10^{-5} \cdot N_U} \tag{11}$$

For the CP, we measured the mean $\overline{S}_0^{(CP)} = 1/\lambda_0^{(CP)}$ and the standard deviation $\sigma_{0s}^{(CP)}$ of the control procedures inter-generation times $S_0^{(CP)}$ and observed that $\overline{S}_0^{(CP)} \approx \sigma_{0s}^{(CP)}$. This result suggests that the generation process of LTE signaling procedures is Poissonian. Then, we computed the empirical cumulative distribution of the $S_0^{(CP)}$ and fitted $S_0^{(CP)}$ into an exponential distribution. As it is shown in Fig. 9, both curves are overlapped. Additionally, we performed a Kolmogorov-Smirnov test to check whether the $S_0^{(CP)}$ samples come from an exponential distribution. The test failed to reject the null hypothesis at the 1% significance level. The same experiment was conducted for different values of N_U and the same result was obtained. Specifically, we swept N_U from 100 to 100000. Consequently, the LTE CP workload generation process, under the assumption that it follows the abstraction described in Section III-B, follows a Poisson distribution. Then, the aggregated arrival process at CP is fully characterized by the signaling generation rate.

Finally, following the same procedure to measure $\lambda^{(DP)}$, we estimated the signaling rates per control procedure for different population densities (see Fig. 8). It is observed that, unlike the SRs and S1Rs rates per user, the HOs and TAUs rates per user depend on the population densities. This fact is due to the increase in the E-UTRAN density (i.e., number of eNBs per km^2) when the population density increases.

Let λ_{SR}, λ_{S1R}, λ_{HO}, and λ_{TAU} be respectively the aggregated generation rate of the SR, S1R, HO, TAU procedures. For our setup and a given number of users N_U, these rates can be estimated as:

$$\lambda_{SR} = \lambda_{S1R} = 0.0044 \cdot N_U \tag{12}$$

$$\lambda_{HO} = 4.2466 \cdot 10^{-6} \cdot N_U^2/(w \cdot h) + 0.003272 \cdot N_U \tag{13}$$

$$\lambda_{TAU} = 2.6281 \cdot 10^{-6} \cdot N_U^2/(w \cdot h) + 0.002025 \cdot N_U \tag{14}$$

And the aggregated signaling procedure generation rate $\lambda^{(CP)}$ as

$$\lambda^{(CP)} = \lambda_{SR} + \lambda_{S1R} + \lambda_{HO} + \lambda_{TAU} \tag{15}$$

B. EPC Network Slices Planning

To gauge the performance of our solution, we considered the scenario which consists of the rural and urban zones layout and the ECs positions depicted in Fig. 6. The assessment

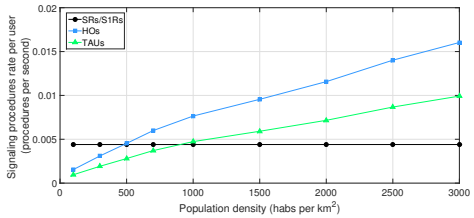


Fig. 8: Generation rate per UE for the different LTE signaling procedures versus the population density for the considered scenario.

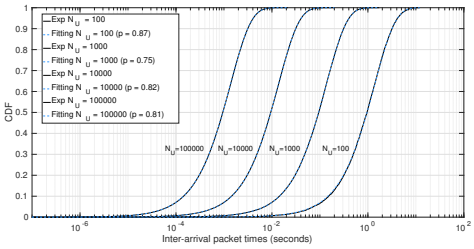


Fig. 9: CDF of the inter-arrival signaling procedures for a population density of 500 $habs/km^2$.

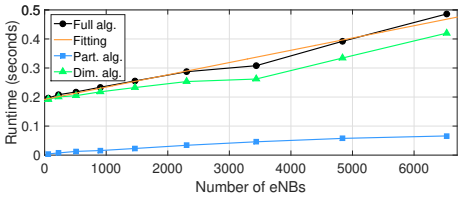


Fig. 10: Algorithm execution time of the distributed approach.

metrics are the algorithm runtime, the dimensioning of the network and computational resources, and the network QoS requirements defined in Section III-C. These metrics were measured for different population densities. Additionally, we compared our solution with two baseline approaches for the workload partitioning:

- Workload partitioning based on proximity: each eNB is assigned to the closest EC using Voronoi diagram. This approach is labeled as "Voronoi" in the figures.
- A fully centralized approach: eNBs are assigned to the same EC. The chosen EC is the nearest to the largest concentration of users. This approach is labeled as "Centralized" in the figures.

First, we assessed the computational complexity of our algorithm. Fig. 10 shows the runtime of our solution versus the number of eNBs deployed at each considered population density. We repeated each measurement five times and calculated the average. The dimensioning algorithm (labeled as "Dim. alg.") and, more specifically, the CP dimensioning was the heaviest part of the full algorithm in terms of complexity. Thanks to the limitation of the search space at each iteration, the CP dimensioning algorithm was able to achieve a linear complexity. The partitioning algorithm (labeled as "Part. alg.") took a linear time, as it was expected, since that in the worst case, it requires $I + K$ iterations to assign all eNBs.

Second, we analyzed the computational resources estimated by PES as depicted in Figs. 11a and 12a. The CP has a higher demand of computational resources than DP in the considered scenario, though the throughput demand is three orders of magnitude higher for DP (see Figs. 11c and 12c), owing to the fact that the processing of the control messages is heavier than that one required for a data packet (see Table III). It is worth mentioning that the computational resources allocated to the CP depend quadratically on the population density (see Fig. 11a). This is attributable to the fact that the HO and TAU procedures per user increase proportionally with the density of the RAN, as shown in (13) and (14). Moreover, since the MME has the highest visit ratio ($V_{MME} = 2.4196$, while $V_{cSGW} = 1.3585$ and $V_{cPGW} = 0.7170$), it was observed that PES allocated most of the CP computational resources to MME entity, followed by cSGW, and so forth. Third, we validated the proper operation of PES. Figs. 11b and 12b show the values of the QoS metrics obtained via simulation for the different population densities studied. As it can be observed, the target performance metrics ($\overline{T}^{(CP)}_{budget} = 25\,ms$, $T^{(DP)}_{budget} = 1\,ms$, and $P^{(EPC)}_{budget} = 10^{-6}$) are always met, thus validating the proper operation of PES.

As mentioned, we compared three different approaches to distribute the workload among the candidate ECs. In general, the Voronoi approach offers the best performance in terms of delay (see Figs. 11b and 12b) owing to its minimization of the propagation delays. Regarding the centralized approach, it requires the lowest amount of computational resources (see Figs. 11a and 12a). That is because it consolidates the workload in a single EC, which leads to a better resources utilization and facilitates the statistical multiplexing of the computational resources (i.e., a lower number of virtualization containers are required as shown in Fig. 11a). Finally, the distributed approach (see Algorithm 2) minimizes the workload imbalances among the candidate ECs as shown in Figs. 11c and 12c. In scenarios where the candidate ECs have a limited capacity, the distributed approach, included in PES, would improve the request acceptance and infrastructure utilization [37]. It is also appropriate for the planning of large geographical areas where the centralized approach could not meet the delay constraints, due to the high propagation delays.

IX. Conclusions

In this article, we proposed an integral solution for planning the network slices of the LTE EPC. We characterized stochastically the LTE signaling and data traffic workload, designed accurate and detailed models to predict the performance of the LTE networks, and formulated the joint optimization problem of resources dimensioning and vEPC embedding for a set of candidate ECs. Using this framework, we proposed a heuristic for planning the vEPC, dubbed "Planner for the EPC as a Service" (PES).

Regarding the LTE workload characterization, we described an abstraction for the workload generation process and designed a compound traffic model which includes the most

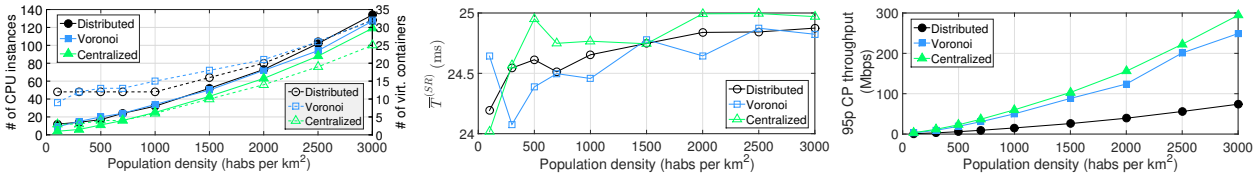


Fig. 11: PES validation for the vEPC control plane: a) Total number of CP dedicated CPU instances. b) Mean time to move a UE from IDLE state to ACTIVE state. c) 95 percentile of the CP workload for the most loaded EC.

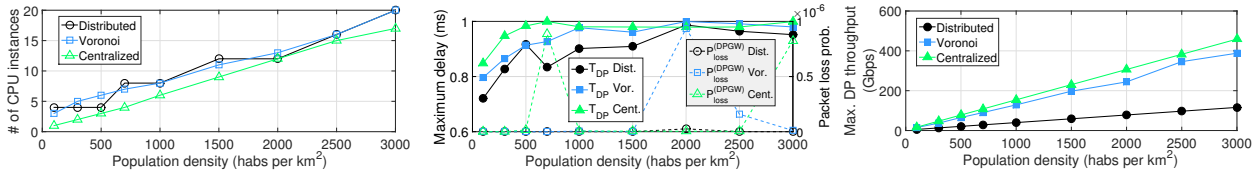


Fig. 12: PES validation for the vEPC data plane: a) Total number of DP dedicated CPU instances. b) QoS of the vEPC DP. c) Maximum DP workload for the most loaded EC.

representative services in current LTE networks. Specifically, the services considered account for more than 70% of the peak aggregate traffic in the American mobile access networks [52]. We enhanced the NSP simulation tool by including the proposed compound traffic model and generated synthetic signaling and data LTE traces. From these traces, a stochastic characterization of the aggregated arrival processes at LTE CP and DP has been carried out. The results show that the aggregated signaling generation process is roughly Poissonian and the DP workload exhibits self-similarity and long-range dependence features. Based on the workload characterization results, we modeled the vEPC DP as a queue fed by a fBm process and the LTE CP as a network of queues following a similar technique to the one proposed in [21].

The planning of the EPC network slices has been formulated as a multi-objective optimization problem that minimizes the workload imbalances among a set of candidate ECs, and maximizes the resources utilization and the end user QoE, while ensuring the set of QoS requirements defined in the 3GPP LTE specs. Finally, a system-level LTE network simulator was developed to validate the proper operation of PES.

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), the Spanish Ministry of Economy and Competitiveness, the European Regional Development Fund (TEC2016-76795-C6-4-R), the Academy of Finland's Flagship programme 6Genesis under grant agreement no. 318927, and the Academy of Finland Project CSN under grant agreement no. 311654.

References

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

[2] T. Taleb, A. Ksentini, and R. Jantti, "Anything as a service for 5g mobile systems," *IEEE Network*, vol. 30, no. 6, pp. 84–91, November 2016.

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

[4] 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," *IEEE Network*, vol. 29, no. 2, pp. 78–88, Mar. 2015.

[5] 3GPP TR23.799 V14.0.0. (2016) Study on Architecture for Next Generation System.

[6] P. Ameigeiras, J. J. Ramos-munoz, L. Schumacher, J. Prados-Garzon, J. Navarro-Ortiz, and J. M. Lopez-soler, "Link-level access cloud architecture design based on sdn for 5g networks," *IEEE Network*, vol. 29, no. 2, pp. 24–31, March 2015.

[7] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "Nfv: state of the art, challenges, and implementation in next generation mobile networks (vepc)," *IEEE Network*, vol. 28, no. 6, pp. 18–26, Nov 2014.

[8] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2429–2453, thirdquarter 2018.

[9] T. Taleb, B. Mada, M. Corici, A. Nakao, and H. Flinck, "Permit: Network slicing for personalized 5g mobile telecommunications," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 88–93, May 2017.

[10] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," *Comm. Mag.*, vol. 55, no. 5, pp. 80–87, May 2017.

[11] T. M. Knoll, "A combined CAPEX and OPEX cost model for LTE networks," in *2014 16th Int. Telecommun. Network Strategy and Planning Symp. (Networks)*, Sept. 2014, pp. 1–6.

[12] M. El-Sayed, A. Mukhopadhyay, C. Urrutia-Valds, and Z. J. Zhao, "Mobile data explosion: Monetizing the opportunity through dynamic policies and QoS pipes," *Bell Labs Tech. J.*, vol. 16, no. 2, pp. 79–99, Sept. 2011.

[13] H. Hawilo, L. Liao, A. Shami, and V. C. M. Leung, "Nfv/sdn-based vepc solution in hybrid clouds," in *2018 IEEE Middle East and North Africa Commun. Conf. (MENACOMM)*, April 2018, pp. 1–6.

[14] R. Sherwood, G. Gibb, K.-K. Yap, G. Appenzeller, M. Casado, N. McKeown, and G. Parulkar, "FlowVisor : A Network Virtualization Layer," 2009.

[15] A. Al-Shabibi, M. D. Leenheer, M. Gerola, A. Koshibe, W. Snow, and G. Parulkar, "OpenVirteX: A Network Hypervisor," in *Open Networking Summit 2014 (ONS 2014)*, Santa Clara, CA, 2014.

[16] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso, and C. Verikoukis, "Multi-Tenant Slicing for Spectrum Management on the Road to 5G," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 118–125, Oct. 2017.

[17] ——, "Cooperation incentives for multi-operator C-RAN energy efficient sharing," in *2017 IEEE Int. Conf. on Commun. (ICC)*, May 2017, pp. 1–6.

[18] O. Adamuz-Hinojosa, J. Ordonez-Lucena, P. Ameigeiras, J. J. Ramos-Munoz, D. Lopez, and J. Folgueira, "Automated Network Service Scaling in NFV: Concepts, Mechanisms and Scaling Workflow," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 162–169, JULY 2018.

[19] AWS. Amazon EC2 Pricing. [Online]. Available: https://aws.amazon.com/ec2/pricing/

[20] 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 Commun. Conf.*, Dec. 2017, pp. 1–6.

[21] J. Prados-Garzon, P. Ameigeiras, J. J. Ramos-Munoz, P. Andres-Maldonado, and J. M. Lopez-Soler, "Analytical modeling for Virtualized Network Functions," in *2017 IEEE Int. Conf. on Commun. Workshops (ICC Workshops)*, May 2017, pp. 979–985.

[22] I. Norros, "On the use of fractional brownian motion in the theory of connectionless networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 953–962, Aug 1995.

[23] K. Tanabe, H. Nakayama, T. Hayashi, and K. Yamaoka, "An optimal resource assignment for C/D-plane virtualized mobile core networks," in *2017 IEEE Int. Conf. on Commun. (ICC)*, May 2017, pp. 1–6.

[24] A. S. Rajan, S. Gobriel, C. Maciocco, K. B. Ramia, S. Kapury, A. Singhy, J. Ermanz, V. Gopalakrishnanz, and R. Janaz, "Understanding the bottlenecks in virtualizing cellular core network functions," in *The 21st IEEE Int. Workshop on Local and Metropolitan Area Networks*, April 2015, pp. 1–6.

[25] 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 Commun. Conf. (GLOBECOM)*, Dec 2016, pp. 1–6.

[26] J. Prados-Garzon, J. J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado, and J. M. Lopez-Soler, "Latency evaluation of a virtualized MME," in *2016 Wireless Days (WD)*. IEEE, March 2016, pp. 1–3.

[27] ——, "Modeling and Dimensioning of a Virtualized MME for 5G Mobile Networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4383–4395, 2017.

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

[29] T. Taleb and A. Ksentini, "Gateway Relocation Avoidance-aware Network Function Placement in Carrier Cloud," in *Proc. of the 16th ACM Int. Conf. on Modeling, Anal. & Simulation of Wireless and Mobile Syst.*, ser. MSWiM '13, 2013, pp. 341–346.

[30] M. Bagaa, T. Taleb, and A. Ksentini, "Service-aware network function placement for efficient traffic handling in carrier cloud," in *2014 IEEE Wireless Commun. and Networking Conf. (WCNC)*, April 2014, pp. 2402–2407.

[31] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying NFV and SDN to LTE Mobile Core Gateways, the Functions Placement Problem," in *Proc. of the 4th Workshop on All Things Cellular: Operations, Appl., & Challenges*, ser. AllThingsCellular '14, 2014, pp. 33–38.

[32] T. Taleb, M. Bagaa, and A. Ksentini, "User mobility-aware Virtual Network Function placement for Virtual 5G Network Infrastructure," in *2015 IEEE Int. Conf. on Commun. (ICC)*, June 2015, pp. 3879–3884.

[33] B. Martini, F. Paganelli, P. Cappanera, S. Turchi, and P. Castoldi, "Latency-aware composition of Virtual Functions in 5G," in *Proc. of the 2015 1st IEEE Conf. on Network Softwarization (NetSoft)*, April 2015, pp. 1–6.

[34] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization," in *Proc. of the 2015 1st IEEE Conf. on Network Softwarization (NetSoft)*, April 2015, pp. 1–9.

[35] ——, "Combined Virtual Mobile Core Network Function Placement and Topology Optimization with Latency Bounds," in *2015 Fourth Eur. Workshop on Software Defined Networks*, Sept 2015, pp. 97–102.

[36] M. Bagaa, T. Taleb, A. Laghrissi, A. Ksentini, and H. Flinck, "Coalitional game for the creation of efficient virtual core network slices in 5g mobile systems," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 469–484, March 2018.

[37] D. Dietrich, C. Papagianni, P. Papadimitriou, and J. S. Baras, "Near-Optimal Placement of Virtualized EPC Functions with Latency Bounds," in *Communication Syst. and Networks*, 2017, pp. 200–222.

[38] A. Laghrissi, T. Taleb, and M. Bagaa, "Conformal mapping for optimal network slice planning based on canonical domains," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 519–528, March 2018.

[39] E. Datsika, A. Antonopoulos, D. Yuan, and C. Verikoukis, "Matching theory for over-the-top service provision in 5g networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5452–5464, Aug. 2018.

[40] ETSI GS NFV 003 V1.3.1. (2018, January) Network Functions Virtualisation (NFV): Terminology for Main concepts in NFV.

[41] B. Hirschman, P. Mehta, K. B. Ramia, A. S. Rajan, E. Dylag, A. Singh, and M. Mcdonald, "High-performance evolved packet core signaling and bearer processing on general-purpose processors," *IEEE Netw.*, vol. 29, no. 3, pp. 6–14, May 2015.

[42] 3GPP TR 38.913 V14.2.0. (2017) 5G; Study on Scenarios and Requirements for Next Generation Access Technologies.

[43] 3GPP TS 23.401 Rel 12. (2014) General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access.

[44] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tut.*, vol. 17, no. 1, pp. 92–105, Firstquarter 2015.

[45] J. J. Ramos-Muñoz, J. Prados-Garzon, P. Ameigeiras, J. Navarro-Ortiz, and J. M. López-Soler, "Characteristics of mobile youtube traffic," *IEEE Wireless Communications*, vol. 21, no. 1, pp. 18–25, 2014.

[46] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proc. of the 9th ACM SIGCOMM conf. on Internet measurement conf.*, 2009, pp. 49–62.

[47] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Understanding the nature of social mobile instant messaging in cellular networks," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 389–392, 2014.

[48] G. feng Zhao, Q. Shan, S. Xiao, and C. Xu, "Modeling Web Browsing on Mobile Internet," *IEEE Commun. Lett.*, vol. 15, no. 10, pp. 1081–1083, October 2011.

[49] A. Rao, A. Legout, Y.-s. Lim, D. Towsley, C. Barakat, and W. Dabbous, "Network characteristics of video streaming traffic," in *Proc. of the Seventh COnf. on emerging Networking EXperiments and Technologies*. ACM, 2011, p. 25.

[50] T. D. Dang, B. Sonkoly, and S. Molnar, "Fractal analysis and modeling of VoIP traffic," in *11th Int. Telecommun. Network Strategy and Planning Symp (NETWORKS 2004)*. IEEE, June 2004, pp. 123–130.

[51] D. Bonfiglio, M. Mellia, M. Meo, and D. Rossi, "Detailed analysis of skype traffic," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 117–127, 2009.

[52] Sandvine. (2016) 2016 Global Internet Phenomena: Latin America & North America.

[53] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Communications*, vol. 21, no. 1, pp. 80–88, 2014.

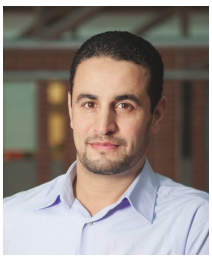
[54] D. Lee, S. Zhou, and Z. Niu, "Spatial modeling of scalable spatially-correlated log-normal distributed traffic inhomogeneity and energy-efficient network planning," in *IEEE 2013 Wireless Commun. and Networking Conf. (WCNC)*, 2013, pp. 1285–1290.

[55] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic," *SIGCOMM Comput. Commun. Rev.*, vol. 23, no. 4, pp. 183–193, Oct. 1993.

Jonathan Prados-Garzon is a Ph.D. student at the Department of Signal Theory, Telematics and Communications of the University of Granada (Spain). He received his B. Sc. in Telecommunications Engineering from the University of Granada (Spain), and M. Sc. in Multimedia Systems from the University of Granada in 2011 and 2012, respectively. He was granted a Ph.D. fellowship by the Education Spanish Ministry on September 2014 and started his Ph.D. studies. In 2017 he stayed at the MOSA!C Lab headed by Professor Tarik Taleb investigating the adoption of network softwarization paradigm in future mobile networks and its applications. His research interests are focused on 5G mobile network architectures, network functions virtualization, software defined networking, and Internet of Things.

Abdelquoddouss Laghrissi received his bachelor's degree in mathematics and computer science and the master's degree in applied computer science with a dissertation on empathy in vehicular ad hoc networks from the School of Sciences, Mohamed V. University, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering, Aalto University, Finland. From 2014 to 2015, he was a Volunteer Member of the cloud computing working group within a Euro-Mediterranean project MOSAIC on Cooperation with Mediterranean Partners to build Opportunities around ICT and Societal and Industrial Challenges of H2020. He is currently a MOSA!C Lab member, involved in the European project 5G!Pagoda on network slicing and 5G in the context of H2020. His research interests include network softwarization and slicing mechanisms, mobile cloud computing, network function virtualization, software defined networking, and vehicular ad hoc networks.

Dr. Miloud Bagaa received his Engineers, Masters, and Ph.D. degrees from the University of Science and Technology Houari Boumediene (USTHB), Algiers, Algeria, in 2005, 2008, and 2014, respectively. From 2009 to 2015, he was a researcher with the Research Center on Scientific and Technical Information (CERIST), Algiers. From 2015 to 2016, he was granted a postdoctoral fellowship from the European Research Consortium for Informatics and Mathematics, and worked at the Norwegian University of Science and Technology, Trondheim, Norway. Currently, he is a senior researcher in Aalto University. His research interests include wireless sensor networks, Internet of Things, 5G wireless communication, security, and networking modeling.



Prof. Tarik Taleb is currently Professor at the School of Electrical Engineering, Aalto University, Finland. He is the founder and director of the MOSA!C Lab (www.mosaic-lab.org). He is also working as part time professor at the Center of Wireless Communications, University of Oulu. Prior to his current academic position, he was working as Senior Researcher and 3GPP Standards Expert at NEC Europe Ltd, Heidelberg, Germany. He was then leading the NEC Europe Labs Team working on R&D projects on carrier cloud platforms, an important vision of 5G systems. Before joining NEC and till Mar. 2009, he worked as assistant professor at the Graduate School of Information Sciences, Tohoku University, Japan, in a lab fully funded by KDDI. From Oct. 2005 till Mar. 2006, he worked as research fellow at the Intelligent Cosmos Research Institute, Sendai, Japan. He received his B. E degree in Information Engineering with distinction, M.Sc. and Ph.D. degrees in Information Sciences from Tohoku Univ., in 2001, 2003, and 2005, respectively.

Prof. Talebs research interests lie in the field of architectural enhancements to mobile core networks (particularly 3GPPs), network softwarization & slicing, mobile cloud networking, network function virtualization, software defined networking, mobile multimedia streaming, inter-vehicular communications, and social media networking. Prof. Taleb has been also directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPPs System Architecture working group. Prof. Taleb is a member of the IEEE Communications Society Standardization Program Development Board.

Prof. Taleb is/was on the editorial board of the IEEE Transactions on Wireless Communications, IEEE Wireless Communications Magazine, IEEE Journal on Internet of Things, IEEE Transactions on Vehicular Technology, IEEE Communications Surveys & Tutorials, and a number of Wiley journals.

Prof. Taleb is the recipient of the 2017 IEEE ComSoc Communications Software Technical Achievement Award (Dec. 2017) for his outstanding contributions to network softwarization. He is also the (co-) recipient of the 2017 IEEE Communications Society Fred W. Ellersick Prize (May 2017), the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher award (Jun. 2009), the 2008 TELECOM System Technology Award from the Telecommunications Advancement Foundation (Mar. 2008), the 2007 Funai Foundation Science Promotion Award (Apr. 2007), the 2006 IEEE Computer Society Japan Chapter Young Author Award (Dec. 2006), the Niwa Yasujirou Memorial Award (Feb. 2005), and the Young Researcher's Encouragement Award from the Japan chapter of the IEEE Vehicular Technology Society (VTS) (Oct. 2003). Some of Prof. Talebs research work have been also awarded best paper awards at prestigious IEEE-flagged conferences.

Prof. Juan M. Lopez-Soler received the B.Sc. degree in physics (electronics) and the Ph.D. degree in signal processing and communications, both from the University of Granada, Granada, Spain, in 1995. He is a Full Professor with the Department of Signals, Telematics and Communications, University of Granada. During 1991-1992, he joined the Institute for Systems Research (formerly SRC), University of Maryland, College Park, MD, USA, as a Visiting Faculty Research Assistant. Since its creation in 2012, he has been the Head of the Wireless and Multimedia Networking Laboratory, University of Granada. He has participated in 11 public and 13 private funded research projects and is the coordinator in 14 of them. He has advised five Ph.D. students and has published 24 papers in indexed journals and contributed to more than 40 workshops/conferences. His research interests include real-time middleware, multimedia communications, and networking.

APPENDIX A
QNA METHOD

This appendix describes the main steps followed by the QNA method to estimate the mean response time of each individual queue in a network of G/G/m queues.

A. Internal flows parameters estimation

As in the case of Jackson's networks, the mean arrival rate to each queue λ_k can be computed by solving the flow balance equations:

$$\lambda_k = \lambda_{0k} + \sum_{i=1}^{K} \lambda_i \cdot p_{ik} \tag{16}$$

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

$$c_{ak}^2 = a_k + \sum_{i=1}^{K} c_{ai}^2 b_{ik}, \qquad 1 \leq k \leq K \tag{17}$$

$$a_k = 1 + \omega_k \Bigg\{ (q_{0k} c_{0k}^2 - 1) + \sum_{i=1}^{K} q_{ik}[(1 - p_{ik}) + p_{ik}\rho_i^2 x_i] \Bigg\} \tag{18}$$

$$b_{ik} = \omega_k q_{ik} p_{ik} (1 - \rho_i^2) \tag{19}$$

$$x_i = 1 + m_i^{-0.5}(max\{c_{si}^2, 0.2\} - 1) \tag{20}$$

$$\omega_k = \left(1 + 4(1 - \rho_k)^2(\gamma_k - 1)\right)^{-1} \tag{21}$$

$$\gamma_k = \left(\sum_{i=0}^{K} q_{ik}^2 \right)^{-1} \tag{22}$$

where $q_{0k} = \lambda_{0k}/\lambda_k$ and $q_{ik} = (\lambda_i \cdot \cdot p_{ik})/\lambda_k$ are respectively the proportion of arrivals to the node k coming 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.

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

$$\overline{T}_k = \frac{\rho_k \cdot (c_{ak}^2 + c_{sk}^2) \cdot \beta}{2 \cdot \mu_k (1 - \rho_k)} + \frac{1}{\mu_k} \tag{23}$$

with

$$\beta = \begin{cases} exp(-\frac{2 \cdot (1-\rho_k) \cdot (1-c_{ak}^2)^2}{3 \cdot \rho_k \cdot (c_{ak}^2 + c_{sk}^2)}) & c_{ak}^2 < 1 \\ \beta = 1 & c_{ak}^2 \geq 1 \end{cases} \tag{24}$$

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

$$\overline{T}_k = 0.5 \cdot \left(c_{ai}^2 + c_{si}^2\right) \cdot W_k^{M/M/m} + \frac{1}{\mu_k} \tag{25}$$

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

$$W_k^{M/M/m} = \frac{C(m_k, \frac{\lambda_k}{\mu_k})}{m_k \mu_k - \lambda_k} \tag{26}$$

and $C(m, \rho)$ represents the Erlang's C formula which has the following formulation:

$$C(m, \rho) = \frac{\left(\frac{(m \cdot \rho)^m}{m!}\right) \cdot \left(\frac{1}{1-\rho}\right)}{\sum_{k=0}^{m-1} \frac{(m \cdot \rho)^k}{k!} + \left(\frac{(m \cdot \rho)^m}{m!}\right) \cdot \left(\frac{1}{1-\rho}\right)} \tag{27}$$

APPENDIX B
TRANSITION PROBABILITIES FOR THE LTE CP QUEUING MODEL

This appendix includes expressions to compute the transition probabilities for the proposed LTE CP queuing model.

Let V_E denote the visit ratio of the CP entity $E \in \{UE, eNB, MME, cSGW, cPGW, HSS, PCRF\}$ which is defined as the average number of visits to entity E by a signaling procedure during its lifetime in the network. Formally, $V_E = \lambda_E / \sum_E \lambda_{0E} = \lambda_E/(\lambda_{0UE} + \lambda_{0eNB})$. Please note that V_E is equal to the average number of packets to be processed by the LTE CP entity E per control procedure. Then,

$$V_E = \frac{\sum_{CP} \lambda_{CP} \cdot n_{CP}^{(E)}}{\sum_{CP} \lambda_{CP}} \tag{28}$$

where $n_{CP}^{(E)}$ is the number of packets to be processed by the LTE CP entity E for the control procedure $CP \in \{SR, S1R, HO, TAU\}$.

The visit ratios and the transition probabilities are related through (16) (flow balance equations):

$$V_E = \frac{\lambda_{0E}}{\sum_E \lambda_{0E}} + \sum_E V_E \cdot p_{E_1 \to E_2} \tag{29}$$

The transition probabilities also satisfy

$$p_{0E_1} + \sum_{E2} p_{E_1 \to E_2} = 1 \tag{30}$$

Assuming that the workload is distributed among the instances of the VNFCs (e.g., MME, cSGW, and cPGW), according to their capacities, i.e., $V_{E_l} = m_l^{(E)}/(\sum_l m_l^{(E)}) \cdot V_E$, and using (29) and (30), we can compute the transition probabilities for our LTE CP queuing model by the following equations:

$$p_{UE}^{eNB} = \frac{V_{UE} - \frac{\lambda_0^{(UE)}}{\sum_E \lambda_0^{(E)}}}{V_{eNB}} \tag{31}$$

$$p_{MME_l}^{eNB} = \frac{m_l^{(MME)}}{\sum_l m_l^{(MME)}} \cdot (1 - p_{UE}^{eNB}) \tag{32}$$

$$p_{eNB}^{MME_l} = \frac{V_{eNB} - \frac{\lambda_0^{(eNB)}}{\sum_E \lambda_0^{(E)}} - V_{UE}}{V_{MME}} \tag{33}$$

$$p_{cSGW_l}^{MME_l} = \frac{m_l^{(cSGW)}}{\sum_l m_n^{(cSGW)}} \cdot \left(1 - p_{eNB}^{MME_l} - p_{HSS}^{MME_l} - \frac{1}{V_{MME}}\right) \tag{34}$$

$$p_{HSS}^{MME_l} = \frac{V_{HSS}}{V_{MME}} \tag{35}$$

$$p_{MME_l}^{cSGW_l} = \frac{m_l^{(MME)}}{\sum_m m_m^{(MME)}} \cdot \left(1 - \sum_l p_{cPGW_l}^{cSGW_l}\right) \tag{36}$$

$$p_{cPGW_l}^{cSGW_l} = \frac{m_l^{(cPGW)}}{\sum_m m_m^{(cPGW)}} \cdot \frac{(V_{PGW} - V_{PCRF})}{V_{SGW}} \tag{37}$$

$$p_{cSGW_l}^{cPGW_l} = \frac{m_l^{(cSGW)}}{\sum_m m_m^{(cSGW)}} \cdot \left(1 - \frac{V_{PCRF}}{V_{PGW}}\right) \tag{38}$$

$$p_{PCRF}^{PGW_l} = \frac{V_{PCRF}}{V_{PGW}} \tag{39}$$

$$p_{MME_l}^{HSS} = \frac{m_l^{(MME)}}{\sum_m m_m} \tag{40}$$

$$p_{cPGW_l}^{PCRF} = \frac{m_l^{(cPGW)}}{\sum_n m_n^{(cPGW)}} \tag{41}$$

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 VNFC instance $m_l^{(C)}$.