Optimal Planning of Virtual Mobile Networks

Jocelyne Elias*, Fabio Martignon†‡, Michele Mangili§ and Antonio Capone§

* LIPADE Laboratory, Paris Descartes University, France
jocelyne.elias@parisdescartes.fr

† University of Bergamo, Italy
fabio.martignon@unibg.it

‡ IUF, Institut Universitaire de France

§ DEIB, Politecnico di Milano, Italy
{michele.mangili, antonio.capone}@polimi.it

***Abstract*—The explosive growth of smartphones and other portable devices, along with new traffic types generated by M2M applications, are creating huge volumes of mobile data traffic and signaling overhead, therefore requiring a radical change to the current mobile network architecture. This has promoted new *virtualization paradigms*, which combine diverse packet core services, and provide network functions implemented in software, rather than in dedicated hardware appliances, in order to scale capacity and introduce new services in a fast and cost-effective way.**

In this paper we study the optimization and resource allocation problems taking into account the deployment of virtualization structures. Our aim is to develop a theoretical framework of resource orchestration for mobile access networks, deriving the fundamental performance limits as well as the tradeoffs among the key system parameters. We therefore study optimal, time-varying placement and chaining of network functions. With respect to existing works, our optimization framework provides a much more precise system modeling, with, among others, a separation between control and data plane functions.

We perform an extensive numerical analysis using both real traffic traces provided by a mobile operator (Vodafone UK) and real positions for radio access points for the UK area, and discuss the impact of network parameters on the system performance. Numerical results show that our proposed optimization framework permits to carefully model key aspects of network virtualization and service deployment/chaining in such scenarios, thus representing a very promising framework for the design of efficient and cost-effective mobile networks.

I. Introduction

Mobile traffic from smartphones and portable devices, along with Machine to Machine (M2M) applications, are creating huge volumes of mobile data traffic. The signaling overhead necessary for handling these diverse applications, which is even more challenging than the capacity needs, requires a radical change in the actual mobile network architecture (i.e., the Evolved Packet Core of LTE network). This indeed has encouraged mobile operators to leverage virtualization techniques (i.e., Network Function Virtualization (NFV) and Software Defined Networking (SDN)) in their network infrastructure, where diverse packet core functions are provided as virtualized services in order to scale capacity and introduce new services in a fast and cost-effective way.

A key feature of *mobile core network function virtualization* is the ability to provide intelligent resource management and network orchestration by dynamically scaling packet core functions to adapt the system to actual needs, in a flexible way: instead of building out a packet core infrastructure dimensioned for peak capacity, virtualization permits mobile operators to elastically create or take down resources on-the-fly. It also lowers both CAPEX and OPEX, since it lets replacing purpose-built hardware with standardized computing and storage platforms while, at the same time, helping the packet core infrastructure run more efficiently, reducing the network footprint, and simplifying network configuration and maintenance.

Using virtualization in core mobile networks in order to increase network flexibility and performance while reducing services deployment cost has been investigated by several works [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], which will be reviewed in the next section.

In this work, we focus on the optimal planning (following dynamic changes in traffic demand) of the data-center structure which must be deployed in order to implement a virtualized network infrastructure (i.e., the virtual Evolved Packet Core, vEPC [1], [2], [3]). Two possible configurations exist to deploy such Data-Centers (DC)s: i) Few DCs are distributed on a geographic scale, in order to aggregate traffic coming from Base Stations or ii) several micro-DCs are considered at the edge of the network. In both cases, simply evaluating the aggregate traffic profiles (that can be obtained from mobile operators) is likely to be insufficient. It is, instead, very interesting to study (and enhance) the capability that such network has to cope with sudden traffic demand changes. Furthermore, another key aspect that we tackle regards the functions placement, as well as their interconnections. More specifically, we will answer to questions like: Where is it better to instantiate network functions? How to interconnect them?

For all these reasons, in this paper, we address the resource management and network orchestration problem using a vEPC architecture and considering time-varying traffic patterns. The main goal is to minimize the total cost expressed in terms of i) the cost of opening data centers on which vEPC functions are executed, and ii) the resource computational cost necessary to satisfy the different applications requirements.

Differently from existing works, we provide clear quantitative insights regarding the structure of the underlying computational infrastructure, as well as the interrelations between the different elements composing the EPC in a more detailed, quantitative manner, while the analysis provided in earlier works was only qualitative. In particular, we propose novel optimization models that consider time-varying traffic patterns based on real traces provided by Vodafone, taking into account the correlation between consecutive time slots, while bounding the maximum delay between specific pairs of functions.

We provide a thorough performance analysis of the proposed model using both real traffic traces provided by Vodafone UK and real radio access points' positions in the UK area, and we show the tradeoffs between key mobile network parameters on the overall system performance. Numerical results confirm that our proposed model captures several important aspects of network virtualization and service deployment in mobile networks, thus representing a very promising framework for the design of efficient and cost-effective mobile networks.

The remainder of this paper is organized as follows. Section II discusses related work. Section III introduces the system model as well as the notation and assumptions made in the paper. Section IV describes the mixed integer linear programming model for the resource allocation problem. Numerical results are provided in Section V. Finally, concluding remarks and future research directions are discussed in Section VI.

II. Related Work

Virtualization in core mobile networks permits to increase network flexibility and performance while reducing services deployment cost [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. The work in [1] describes the key elements to realize the architectural vision of EPC as a Service, an implementation option of the Evolved Packet Core, as specified by 3GPP, which can be deployed in cloud environments. In [2] the authors propose to integrate NFV with SDN and software-defined radio (SDR), for 4G and 5G networks. In [3] the authors introduce an NFV framework, and discuss the challenges and requirements of its use in mobile networks. In particular, in order to reduce signaling traffic and achieve better performance, this work proposes a criterion to bundle multiple functions of virtualized evolved packet-core in a single physical device or a group of adjacent devices.

The work in [11] presents novel architectural design patterns towards open, cloud-based 5G communications. The white paper [12] identifies key findings and challenges that need to be overcome so as to meet the 5G requirements.

With the precise aim to ease the design and management of cellular data networks, while supporting new services, Li et al. suggested in [4] to use the services provided by SDN. Similarly, Jin et al. in [5] presented SoftCell: a design for a scalable architecture for a mobile core-network that supports fine-grained policies for mobile devices using SDN.

The problem of virtual network function placement for service chains is studied in [10] for the purpose of energy and traffic-aware cost minimization. This problem is formulated as an optimization problem, and it is solved by proposing a sampling-based Markov approximation (MA) approach, and combining MA with matching theory.

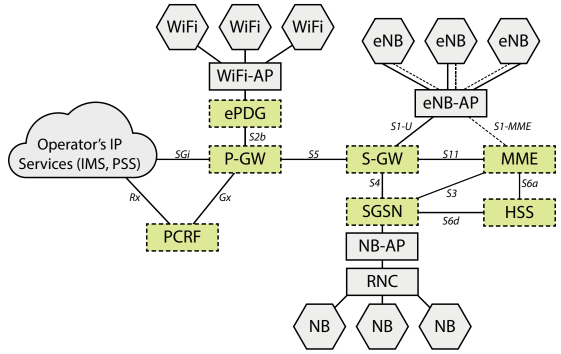


Figure 1. *System Model considered in this work: an LTE-EPC network with radio access and core nodes, which run EPC network functionalities on virtualized platforms.*

In order to reduce the Total Cost of Ownership (TCO), Basta et al. evaluated in [6] the possibility to move some functions of the EPC to a data-center using NFV and SDN. They also proposed to extend OpenFlow to support GTP tunneling. Furthermore, they discussed different deployment architectures for the different functions composing the EPC, and described the tradeoff between the cost reduction and the increase in the end-to-end delay as well as the exchanged data volume between the different elements composing the EPC. Our contribution differs from [6] because we provide clear, quantitative insights regarding the structure of the underlying computational infrastructure, as well as the inter-relations between the different elements composing the core network, while the analysis provided in [6] is only qualitative.

In [7] the same authors tackled the data center placement problem to run virtualized instances of the EPC gateways, to minimize the transport network load under a data-plane delay budget. They proposed to separate the control-plane and data-plane of the SGW and PGW, to either virtualize the gateway function (and run them in the data-center), or run only the control-plane functions in the data-center. The numerical analysis showed that virtualizing the gateway functions can reduce the traffic overhead, but it also can increase the service latency. An extended formulation with time-varying traffic patterns was considered in [8].

We emphasize that our approach considers the inter-relation between the different elements composing the core network in a (more) fine-grained manner. In particular, we bound the maximum delay on each network link. Furthermore, we take into account the fact that there exists correlation between one time slot and the next, since it is unrealistic to assume that all the functions can be migrated at any time in the infrastructure. Furthermore, from a different, and more practical perspective, works [13], [14], [15], [16] conduct experimentation and measurements with the aim to investigate latency, signaling overhead of control-data plane and load balancing issues in 3G/LTE network.

III. System Model

In this section, we describe the system model and the notation used throughout the paper.

We consider, in particular, the LTE-EPC architecture illustrated in Figure 1; we observe, however, that our model can be applied straightforwardly also to currently proposed 5G architectures, which share with LTE-EPC several key essential features.

In such architecture, a set of radio access nodes (viz., eNodeBs, NodeBs and WiFi access points) generate (time varying) traffic that must be served by, and routed through, the core network in an efficient manner, passing through different network functionalities, guaranteeing a bounded end-to-end delay. The traffic outgoing from the radio access nodes is aggregated into *aggregation points* represented by eNB-APs, NB-APs, and WiFi-APs, respectively, in the figure. The core network of the EPC architecture is assumed to be virtual in the sense that all EPC network functionalities, including the Mobility Management Entity (MME), Home Subscriber Service (HSS), Serving Gateway (SGW), PDN Gateways (PGW), Service GPRS Support Node (SGSN), Policy and Charging Rules Function (PCRF), run as virtualized instances on data centers (DCs). The number, locations, and capacities of these data centers (as well as their interconnections/chaining) will be optimized by our model described in the next section. We refer to [9] for a detailed description of these functionalities.

Let us represent the architecture in Figure 1 by a graph $G = (N, A)$, where N is the set of nodes, and A is the set of arcs. For node $i \in N$ we denote with $FS(i)$ the set of forward arcs $(i, j) \in A$, whereas $BS(i)$ is the set of backward arcs $(j, i) \in A$.

We denote by $N^{DC} (\subseteq N)$ the set of test points (candidate sites), where DCs can be deployed. $N^{AGP} (\subseteq N)$ is the set of aggregation points and $N^{IMS-PSS} (\subseteq N)$ is the set of IMS-PSS nodes (IP Multimedia Subsystem-Packet Switch Streaming). The radio access and core data network functionalities are represented by the set of functions $F = F_{FIX} \cup \{\text{S-GW,MME,SGSN,HSS,P-GW,PCRF,IMS-PSS}\}$, where F_{FIX} is the set of fixed functions dedicated to the access ($F_{FIX} = \{\text{eNB,WiFi,NB}\}$).

We assume that the traffic is *time-varying*, and we thus define the set of time slots as $\mathcal{T}$, and let $d^{n,f,\tau}$ be the traffic (in Mbps) entering the core network through aggregation point $n \in N^{AGP}$ for a fixed function $f \in F_{FIX}$ at time $\tau \in \mathcal{T}$. To handle the traffic arriving at the core network, the functions in F should interact with each other in an established and *ordered* manner. Therefore, we define the set of pairs of legitimate adjacent functions as $\hat{F} = \{\langle f_1, f_2\rangle | f_1 \in F, f_2 \in F\}$. Note that not necessarily all the traffic arriving at an ingress arc of a node (with some core network functionalities) should be forwarded at its egress arcs. This can be modeled by introducing $K^{\langle f_1, f_2\rangle} \in [0, 1]$, which is a splitting parameter that specifies the fraction of the total traffic for flows sent from f_1 to f_2.

The one-way propagation delay on arc a (l^a) is bounded by the maximum tolerated delay ($l_{max}^{\langle f_1, f_2\rangle}$) between any two functions f_1 and f_2, $\forall \langle f_1, f_2\rangle \in \hat{F}$. The values of $l_{max}^{\langle f_1, f_2\rangle}$ can be set according to the 3GPP standards [17], [18] and delay measurements results summarized in Section II.

The cost for opening a data center and the per-unit cost for using its computational resources are denoted by C^O and C^F, respectively. Finally, let R^f be the computational resources needed to serve one unit of flow for $f \in F$ and M be a large constant.

IV. Optimization Model

We now present the optimization model we propose to minimize the overall network cost (composed of CAPEX and OPEX costs), while serving time-varying traffic demands and taking into account capacity and latency constraints.

Decision variables include:

- Data center opening variables

$$x_n = \begin{cases} 1 & \text{if a data center is opened in } n \in N^{DC} \\ 0 & \text{otherwise.} \end{cases}$$

- Connection functions variables

$$z_{s,t}^{\langle f_1, f_2\rangle,\tau} = \begin{cases} 1 & \text{if functions } \langle f_1, f_2\rangle \in \hat{F} \text{ exchange} \\ & \text{traffic on arc } (s,t) \in A \text{ during slot } \tau \\ 0 & \text{otherwise.} \end{cases}$$

- Flow variables between connections
  $r_{s,t}^{\langle f_1, f_2\rangle,\tau}$: flow between functions $\langle f_1, f_2\rangle \in \hat{F}$ on arc $(s, t) \in A$ during time slot τ

Given the above definitions and notations, the optimal vEPC planning problem can be stated as follows:

$$\min \sum_{n \in N^{DC}} \left(C^O x_n + C^F \sum_{a \in BS(n)} \sum_{\tau \in \mathcal{T}} \sum_{\langle f_1, f_2\rangle \in \hat{F}} R^{f_2} r_a^{\langle f_1, f_2\rangle,\tau} \right) \tag{1}$$

subject to:

$$K^{\langle f_1, f_2\rangle} d^{n,f_1,\tau} = \sum_{a \in FS(n)} r_a^{\langle f_1, f_2\rangle,\tau} \qquad \forall n \in N^{AGP}, \tau \in \mathcal{T}, \langle f_1, f_2\rangle \in \hat{F} | f_1 \in F_{FIX} \tag{2}$$

$$z_{n_1,n_2}^{\langle f_1, f_2\rangle,\tau} = 0 \qquad \forall (n_1, n_2) \in A, \tau \in \mathcal{T}, \langle f_1, f_2\rangle \in \hat{F} | n_1 \in N^{AGP}, f_1 \notin F_{FIX} \tag{3}$$

$$K^{\langle f_1, f_2\rangle} \sum_{\substack{a \in BS(n) \\ \langle f_s, f_1\rangle \in \hat{F}}} r_a^{\langle f_s, f_1\rangle,\tau} = \sum_{a \in FS(n)} r_a^{\langle f_1, f_2\rangle,\tau}, \forall n \in N \setminus N^{AGP}, \tau \in \mathcal{T}, \langle f_1, f_2\rangle \in \hat{F} \tag{4}$$

$$r_a^{\langle f_1, f_2\rangle,\tau} \leq M z_a^{\langle f_1, f_2\rangle,\tau}, \qquad \forall a \in A, \tau \in \mathcal{T}, \langle f_1, f_2\rangle \in \hat{F} \tag{5}$$

$$\sum_{a \in FS(n)} z_a^{\langle f_1, f_2\rangle,\tau} \leq 1, \qquad \forall n \in N, \tau \in \mathcal{T}, \langle f_1, f_2\rangle \in \hat{F} \tag{6}$$

$$z_a^{\langle f_1, f_2\rangle,\tau} \leq x_n \qquad \forall n \in N^{DC}, a \in \{BS(n), FS(n)\}, \tau \in \mathcal{T}, \langle f_1, f_2\rangle \in \hat{F} \tag{7}$$

$$l^a z_a^{\langle f_1, f_2\rangle,\tau} \leq l_{max}^{\langle f_1, f_2\rangle} \qquad \forall a \in A, \tau \in \mathcal{T}, \langle f_1, f_2\rangle \in \hat{F} \tag{8}$$

$$z_{n_1,n_2}^{\langle f_1,f_2\rangle,\tau}=0 \qquad \forall(n_1,n_2)\in A,\tau\in\mathcal{T},\langle f_1,f_2\rangle\in\hat{F} \; |n_1\in N^{DC},f_1\in F_{FIX} \tag{9}$$

$$z_{n_1,n_2}^{\langle f_1,f_2\rangle,\tau}=0 \qquad \forall(n_1,n_2)\in A,\tau\in\mathcal{T},\langle f_1,f_2\rangle\in\hat{F} \; |n_2\in N^{DC},f_2=\text{IMS-PSS} \tag{10}$$

$$z_{n_1,n_2}^{\langle f_1,f_2\rangle,\tau}=0 \qquad \forall(n_1,n_2)\in A,\tau\in\mathcal{T},\langle f_1,f_2\rangle\in\hat{F} \; |n_2\in N^{IMS-PSS},f_2\neq\text{IMS-PSS} \tag{11}$$

$$x_n\in\{0,1\} \qquad \forall n\in N^{DC} \tag{12}$$

$$z_a^{f,\tau}\in\{0,1\}\ ,\ r_a^{f,\tau}\in\mathbb{R}^+ \qquad \forall a\in A,\tau\in\mathcal{T},f\in\hat{F}. \tag{13}$$

The objective function (1) minimizes the overall cost, including the CAPEX for the datacenter opening costs, as well as the OPEX related to the operation of the virtualized EPC.

The set of constraints (2) forces traffic demands to be served, while in (3) we make sure that only fixed functions (i.e., eNB, WiFi and NB) can be implemented in the aggregation points.

In (4) we impose the flow conservation conditions: for each legitimate functions pair, we route a fraction $K^{\langle f_1,f_2\rangle}$ of the ingress flow on egress arcs. Constraints (5) make sure that a flow is routed on an arc only if the corresponding functions are available in the end-nodes.

In (6) we make sure that the flow for the function pair $\langle f_1,f_2\rangle$ is forwarded only on a single arc. In (7) we force data-centers to open whenever they host at least one function, whereas in (8) latency constraints are imposed.

The set of constraints in (9) guarantees that fixed functions are not deployed in DCs. Similarly, in (10) we prevent IMS-PSS to be deployed in DCs. Therefore, in (11), we force IMS-PSS functions to be only deployed in the IMS-PSS nodes.

Lastly, binary conditions on the DC activation and connection variables are imposed in (12)-(13), as well as non-negativity constraints for the flow variables.

Note that we can define a variation of the above model by further introducing the following capacity constraints on the total traffic entering in a DC node (where c_n denotes the overall amount of computational resources available at the candidate site $n\in N^{DC}$):

$$\sum_{a\in BS(n)}\sum_{\langle f_1,f_2\rangle\in\hat{F}} R^{f_2}r_a^{\langle f_1,f_2\rangle,\tau}\leq c_n,\forall n\in N^{DC},\tau\in\mathcal{T} \tag{14}$$

We will refer to this variant as *capacitated* problem in the numerical results section.

V. Numerical Results

In this section we provide a numerical evaluation of the proposed optimization model, which we implemented in OPL, and solved using the CPLEX commercial solver on a server equipped with an Intel CPU at 2.60GHz and 64 GByte of RAM.

We collected the base station deployment data for the Vodafone UK network from the public "OpenCellID" dataset [19]. According to such dataset, 67508 cells belong to the Vodafone UK network, in particular 64.3% of them are GSM radios, 35.4% UMTS, and 0.3% LTE. Furthermore, we collected a 1-week traffic trace (1 hour sampling interval) from a subset of Vodafone UK network (1431 GSM cells and 1487 UMTS cells, deployed in London), and this allows us to set up the mobile data traffic entering in NB, eNB and WiFi nodes. Finally we collected 2011 census data for UK population density (dataset KS101UK).

By leveraging the official census data, we projected the collected traffic traces with respect to the entire country, assuming, as done by Basta et al. in [8], that there exists a proportional dependence between the population density and the amount of traffic generated in the cellular network.

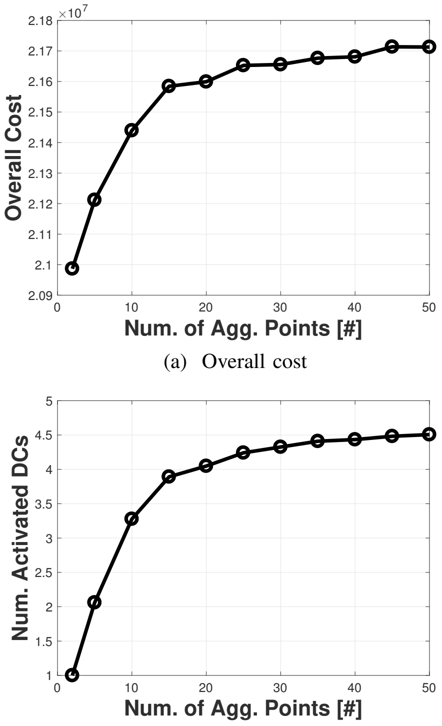
We could not obtain more detailed information on the topological structure of the aggregation points and, for this reason, we assumed that they are deployed in "central" locations within the networking infrastructure. In order to do so we leveraged the K-Means model, since it is frequently used in the literature to optimally deploy network nodes in central locations (especially when dealing with problems related to content distribution). In other words, we deploy "k" aggregation centers in central locations with respect to the geographical placement of the cells, by leveraging the K-Means model. Note that WiFi access points are scattered randomly in the network, taking into account both population density and cell towers (or aggregation points) locations. The results discussed in the following sub-sections are averaged over 200 network instances.

A. Effect of the Number of Aggregation Points (APs)

We start by evaluating the impact of the number of APs on the model's solution. To do so, we fix the number of candidate sites where DCs and IMS-PSS nodes can be installed to 10 and 5, respectively. The DC opening cost C^O, the DC computation cost per traffic unit C^F, and the DC computational resource needed by one unit of flow of network function f, R^f, are set to 2e+5, 0.1 and 1, respectively. Finally, the maximum latency $l_{max}^{\langle f_1,f_2\rangle}$ is fixed to 10 ms. This parameters setting is used throughout this section if not stated otherwise. Figures 2a and 2b show the effect of the number of aggregation points on the objective function value (the overall cost) and the number of activated data centers, respectively. As expected, it can be seen that the overall cost increases with the number of aggregation data points, since there is more mobile data traffic injected in the network. Furthermore, the trend of the number of activated DCs is very similar to the one corresponding to the overall cost, and this is logic since more and more DCs are opened when increasing the number of aggregation points in order to guarantee low latency and limited congestion (load balancing).

B. Effect of the Number of Candidate Data Centers

To highlight the effect of the number of candidate DCs on the model, we vary this parameter value in the range [10,20], for the original (uncapacitated) and capacitated version of the optimal planning problem. As for the number of APs, we set



(a) Overall cost

(b) Num. of Activated DCs

Figure 2. *Effect of the Number of Agg. Points*. Plots 2a-2b show the effect of the number of aggregation points on the overall cost (Fig. 2a) and the number of data centers activated (Fig. 2b), for the proposed optimal vEPC model.

it to 30. The obtained results are illustrated in Figures 3a, 3b, 3c, and 3d. Since the solution space of the capacitated problem is smaller than the one of the uncapacitated version (due to introduction of capacity constraints (14)), the overall network cost is higher than the one obtained in this latter case. Moreover, since DCs have some capacity limits, the capacitated model needs to open more DCs with respect to the uncapacitated version to ensure low latency and traffic balancing among nodes; this behavior is reflected by the curve in Figure 3d. Finally, we observe that the computation time necessary to find an optimal solution ranges between 20 s and 380 s, and becomes larger for the capacitated version of the problem, ranging between 8 s and 1400 s.

C. Effect of the DC Opening Cost and the Maximum Tolerated Delay ($l_{max}^{\langle f_1,f_2\rangle}$)

Finally, we evaluate the impact of both the DC opening cost and the maximum tolerated delay on the proposed model. To this aim, we consider 4 different values for the maximum tolerated delay ($l_{max}^{\langle f_1,f_2\rangle} \in \{1, 2, 5, 10\}$ ms), and vary the installation cost of DCs in the range [200,1e+6]. All other parameters are the same as in the previous sub-section. Figures 4a, 4b, and 4c show respectively the overall cost (installation and computation cost), the computation cost only, and the number of activated DCs. Note that, in all these figures, the curves corresponding to maximum delay equal to 1 and 2 ms are practically overlapping. For quite large delay values the model tends to open more DCs, and as a consequence the overall cost increases when the delay takes larger values. For the computation cost, when the tolerated delay increases, and hence more DCs are opened, network functions can reside on different and potentially distant DCs. This generates more traffic among DCs and requires more computational resources. In fact, on one extreme case, all network functions could be activated on the same DC, while on the other extreme, distinct functions could be activated on distinct DCs, requiring intense traffic exchange and more computational resources. Finally, we observe in Figure 4c that the number of activated DCs is much more sensitive to latency than to DC opening cost, especially for high latency values.

VI. Conclusion

In this paper, we developed an optimization framework tailored for mobile operators that leverage virtualization techniques to enhance their access network infrastructure. We studied the optimal, time-varying placement and chaining of network functions, providing a precise system modeling with a separation between control and data plane functions.

We performed a thorough performance analysis of our optimization model, considering both real traffic traces and base station positions (for the UK area), and we measured the performance tradeoffs between key system parameters, like number and position of candidate sites where Data Centers can be activated, of traffic aggregation points, the DCs opening cost, the maximum tolerated delay, among others.

Numerical results show that our optimization framework permits to carefully model service deployment and chaining, thus representing a promising framework for the design of efficient and cost-effective mobile networks featuring network virtualization.

References

[1] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, T. Magedanz, EASE: EPC as a service to ease mobile core network deployment over cloud, IEEE Network 29 (2) (2015) 78–88.

[2] S. Sun, M. Kadoch, L. Gong, B. Rong, Integrating network function virtualization with SDR and SDN for 4G/5G networks, in: IEEE Network, 2015, pp. 54–59.

[3] H. Hawilo, A. Shami, M. Mirahmadi, R. Asal, NFV: State of the Art, Challenges and Implementation in Next Generation Mobile Networks (vEPC), in: IEEE Network, 2014, pp. 18–26.

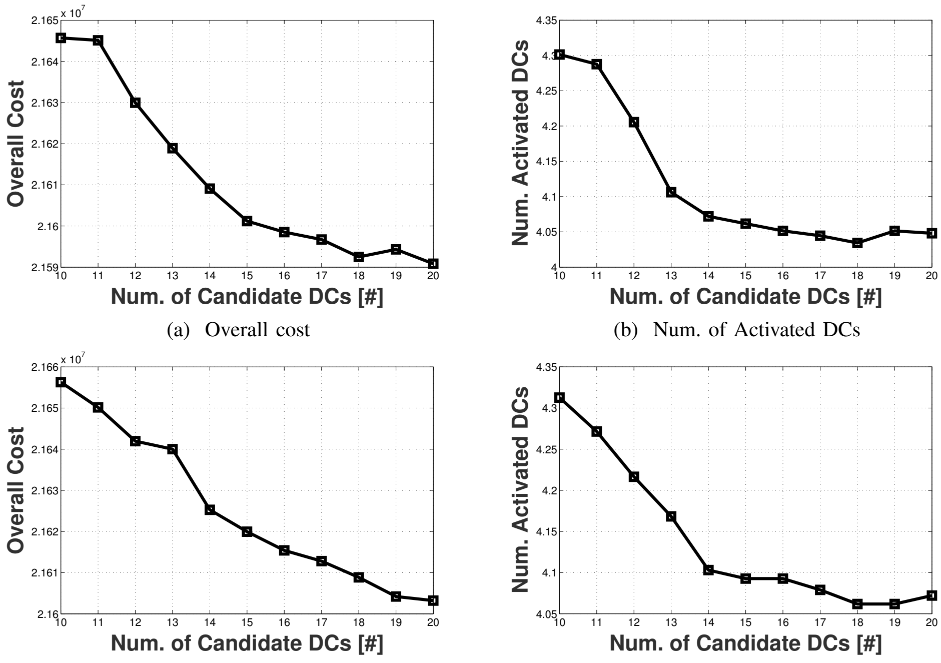
[4] L. Li, Z. Mao, J. Rexford, Toward Software-Defined Cellular Networks, in: European Workshop on Software Defined Networking (EWSDN), 2012, pp. 7–12.

[5] X. Jin, L. Li, L. Vanbever, J. Rexford, SoftCell: Scalable and Flexible Cellular Core Network Architecture, in: Proc. of the 9th ACM conference on Emerging networking experiments and technologies (CoNEXT), 2013, pp. 163–174.

[6] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, E.-D. Schmidt, A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions, in: Proc. of the IEEE Conf. on Software Defined Networks for Future Networks and Services (SDN4FNS), 2013, pp. 1–7.

[7] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, 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, Applications, and Challenges, AllThingsCellular'14, 2014, pp. 33–38.

[8] A. Basta, A. Blenk, M. Hoffmann, H. J. Morper, K. Hoffmann, W. Kellerer, SDN and NFV Dynamic Operation of LTE EPC Gateways for Time-varying Traffic Patterns, in: Proc. of the 6th Int.l Conf. on Mobile Networks and Management, 2014, pp. 63–76.



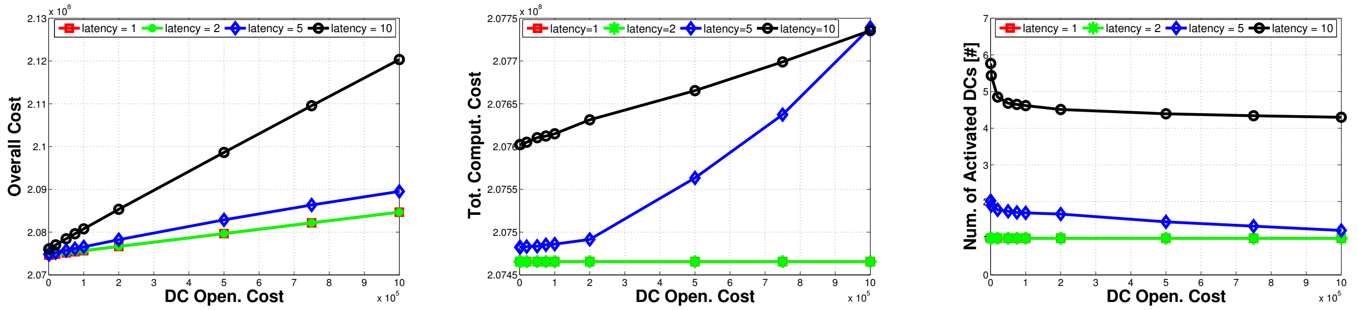
(a) Overall cost

(b) Num. of Activated DCs

(c) Overall cost (Capacitated pb.)

(d) Num. of Activated DCs (Capacitated pb.)

Figure 3. *Effect of the Number of Candidate Data Centers (DCs).* Plots 3a-3d show the effect of the number of candidate DCs on the overall cost (3a-3c) and the number of data centers activated (3b-3d) for both the uncapacitated and capacitated versions of the vEPC planning optimization problem.



(a) Overall cost

(b) Tot. Computation Cost

(c) Num. of Activated DCs

Figure 4. *Effect of the DC opening cost and the maximum tolerated delay* ($l_{max}^{\langle f_1, f_2 \rangle}$). Plots 4a-4c show the effect of the DC opening cost on the overall cost (Fig. 4a), the total computation cost (Fig. 4b), and the number of DCs activated (Fig. 4c), for different values of the maximum tolerated delay $l_{max}^{\langle f_1, f_2 \rangle} \in \{1, 2, 5, 10\}$ ms, for the proposed optimal vEPC planning model.

[9] S. Sesia, I. Toufik, M. Baker, LTE, The UMTS Long Term Evolution: From Theory to Practice, 2009.

[10] C. Pham, N. H. Tran, S. Ren, W. Saad, C. Hong, Traffic-aware and Energy-efficient vNF Placement for Service Chaining: Joint Sampling and Matching Approach, IEEE Transactions on Services Computing PP (99) (20 Feb. 2017) 1–1.

[11] K. Katsalis, N. Nikaein, E. S. R. Favraud, T. Braun, 5G Architectural Design Patterns, in: Proc. of IEEE International Conference on Communications, 3rd International Workshop on 5G Architecture, 2016.

[12] 5G PPP Architecture Working Group, View on 5G Architecture (Version 2.0), in: Architecture White Paper, 2017.

[13] M. Laner, P. Svoboda, P. Romirer-Maierhofer, N. Nikaein, F. Ricciato, M. Rupp, A Comparison Between One-Way Delays in Operating HSPA and LTE Networks, in: Proc. of the 10th Int.l Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2012, pp. 286–292.

[14] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, O. Spatscheck, An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance, in: Proc. of the ACM SIGCOMM Conference, 2013, pp. 363–374.

[15] Q. Xu, J. Huang, Z. Wang, F. Qian, A. Gerber, Z. M. Mao, Cellular Data Network Infrastructure Characterization and Implication on Mobile Content Placement, in: Proc. of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems, 2011, pp. 317–328.

[16] M. Sama, S. Ben Hadj Said, K. Guillouard, L. Suciu, Enabling Network Programmability in LTE/EPC Architecture Using OpenFlow, in: Proc. of the 12th Int.l Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014, pp. 389–396.

[17] 3GPP, 3GPP TR 25.913 (2010).

[18] 3GPP, 3GPP TR 25.912 (2014).

[19] OpenCellID Website, `http://opencellid.org/`, Last accessed: Sep. 2017.