User Mobility-Aware Virtual Network Function Placement for Virtual 5G Network Infrastructure

Tarik Taleb*, Miloud Bagaa|| and Adlen Ksentini§
* Communications and Networking Department, Aalto University, Finland. Email: talebtarik@ieee.org
|| Department of Theories and Computer Engineering, CERIST, Algiers, Algeria. Email: bagaa@mail.cerist.dz
§ IRISA, University of Rennes 1, Rennes, France. Email: adlen.ksentini@irisa.fr

Abstract—**Cloud offerings represent a promising solution for mobile network operators to cope with the surging mobile traffic. The concept of carrier cloud has therefore emerged as an important topic of inquiry. For a successful carrier cloud, algorithms for optimal placement of Virtual Network Functions (VNFs) on federated cloud are of crucial importance. In this paper, we introduce different VNF placement algorithms for carrier cloud with two main design goals: i) minimizing path between users and their respective data anchor gateways and ii) optimizing their sessions' mobility. The two design goals effectively represent two conflicting objectives, that we deal with considering the mobility features and service usage behavioral patterns of mobile users, in addition to the mobile operators' cost in terms of the total number of instantiated VNFs to build a Virtual Network Infrastructure (VNI). Different solutions are evaluated based on different metrics and encouraging results are obtained.**

I. Introduction

Cloud computing is recently gaining lots of ground as an ever-increasing number of enterprises and individuals are hosting their services and shifting their workloads to cloud service providers. Mobile operators are also considering the usage of cloud computing to extend their services and to cope with the tremendous growth they are experiencing in mobile data traffic [1]. Indeed, virtualizing the core networks represents one of the key visions of the future $5G$ architecture. Thanks to the numerous advantages it offers in terms of network configuration flexibility, scalability, and elasticity, Network Function Virtualization (NFV) has emerged as an important topic of inquiry among different stakeholders in the telecommunications arena.

Several pioneering research work have been conducted to enable the creation and runtime management of mobile networks over the cloud, studying different implementation options [2] and devising an entire framework for the creation of end-to-end mobile services, including mobile transport networks, on the cloud [1]. Software Defined Networking (SDN) has been also considered in virtualization of mobile network functions over OpenFlow-based networks, focusing on the virtualization of the control plane; separately or jointly with the user data plane. For a successful creation of mobile core networks on the cloud, algorithms for optimal placement of Virtual Network Functions (VNFs), forming a Virtual Network Infrastructure (VNI), on federated cloud and within the same datacenter are of crucial importance. This defines the focus of this paper.

The remainder of this paper is organized in the following fashion. Section II presents some related research work. The network model and problem formulation are covered in Section III. The proposed VNF placement strategy is discussed in Section IV. Section V evaluates the performance of the different optimization solutions envisioned in this paper. The paper concludes in Section VI.

II. Related work

In traditional mobile core networks, a wide plethora of mechanisms and algorithms were devised to select, for mobile users, optimal data anchor gateways from within a range of geographically static gateways and that is for the sake of communication efficiency [3]. This gateway selection may be solely based on the geographical proximity of mobile users to gateways and the gateway load [4]. It may also consider the End-to-End connection and/or the application type [5]. In case of cloud-based mobile core networks, created on-demand, operators have more flexibility in deciding where to place VNFs of gateways, rather than just selecting gateways from within a fixed set of static gateways. Such flexibility helps mobile operators to dynamically dimension, plan, and re-plan their mobile networks whenever there is need for that and as per the changing behavior of mobile users, the features of the provisioned services, and according to other metrics relevant to the mobile network performance.

The problem of VNF placement can be studied either within the same datacenter or in case a VNI is to be deployed across federated clouds. With regard to the former, a large library of research work has been conducted for decision on the placement of Virtual Machines, VMs, (not necessarily hosting a VNF) within the same datacenter, having, as objective, cost savings thanks to better utilization of computing resources and less frequent overload situations. In [9], performance isolation (e.g., CPU, memory, storage, and network bandwidth), resource contention properties (amongst VMs on the same physical host), and VMs behavioral usage patterns are taken into account in decisions on VM placement, VM migration, and cloud resource allocations. Generally speaking, VM placement on Physical Machines (PMs) is a well investigated problem. A datacenter may start with an initial configuration and then apply adequate solutions to make a series of live migrations to transit the datacenter from a suboptimal state to an optimal one, similar in fashion to solving an iterative

978-1-4673-6432-4/15/$31.00 ©2015 IEEE

rearrangement problem. Different algorithms can be used, such as N-dimensional set or bin packing [10], the Simulated Annealing algorithms [11], and Ant Colony Optimization [12]. In other research works, optimal placement of VMs, running specific services, on PMs, consider electricity-related costs as well as transient cooling effects [13]. Others do autonomic placement of VMs as per policies specified by the datacenter providers and/or users [14]. Other VM placement strategies consider maximizing the profit under a particular service level agreement and a predetermined power budget [15].

Whilst state of the art solutions for VM placement on PMs are one-dimensional, focusing on the placement of a single VM considering its impact on neighboring VMs sharing the same PM, the VNF placement problem within the same datacenter is more complex. Indeed, in VNF placement strategies, not only the impact of VMs running particular VNFs, forming a particular VNI, on the neighboring VMs is considered but also the interactions among these VNFs and their respective VMs. Indeed, the deployment of VNI is a multi-dimensional process where the VNI may be composed of several VNFs, which in turn may be decomposed into multiple VNF Components (VNFCs), and there is a strict functional relationship between the various VNFCs and performance constraints that may make the deployment process more complex. Unfortunately, there is not much information available that may analyze the impact of deployment strategy during the initial deployment of VNI in a datacenter and how quickly the VNI converges to a functional operational status. In this regard, the work presented in [16] analyzes the impact on the cost of datacenter resources, such as network and compute, by comparing the impact of two constraint-based and heuristically derived deployment strategies namely Vertical Serial Deployment (VSD) and Horizontal Serial Deployment (HSD) strategies.

With regard to the problem of VNF placement across federated clouds, in [6], the authors proposed a VNF placement method, particularly for creating mobile gateway functionalities (Serving Gateway - (S-GW)) and their placement in federated clouds so that the frequency of S-GW relocation occurrences is minimized. In [6], the aim was to conduct an efficient planning of Service Areas (SAs) retrieving a trade-off between minimizing the UE handoff between SAs, and minimizing the number of created instances of the virtual S-GWs. In [7], the focus was on VNF placement and instantiation of another mobile network functionality, namely data anchoring or PDN-GW creation/selection. The work argued the need for adopting application type and service requirements as metrics for (i) creating VNF instances of PDN-GW and (ii) selecting adequate virtual PDN-GWs for UEs receiving specific application types. The placement of PDN-GW VNFs was modeled through a nonlinear optimization problem whose solution is NP-hard. Three heuristics were then proposed to deal with this limitation. In [8], the authors proposed a framework, dubbed softEPC, for flexible and dynamic instantiation of VNFs, where most appropriate, and as per the actual traffic demand.

III. Network model and problem formulation

The envisioned architecture consists of two domains, the cloud domain consisting, in turn, of a number of datacenters, distributed over a geographical area and forming a federated cloud and the Radio Access Network (RAN) domain comprising a number of access points (evolved Node-B – eNBs in the context of the Evolved Packet System – EPS). The datacenters can be macro large-scale ones as well as regional small-scale ones. Whilst the locations of datacenters is outside the scope of this paper, we assume that there are some datacenters collocated with eNBs (i.e., in very busy districts) [17] and some with the potentiality to serve a number of eNBs. On each datacenter, one or multiple VNFs of PDN-GWs and S-GWs can be instantiated on demand and in a dynamic manner to form a VNI.

The objective of this paper is to find efficient solutions that place VNFs of both PDN-GWs and S-GWs on a given topology of distributed datacenters (i.e., federated cloud) for a population of UEs and that is to create on-demand and in a dynamic way an elastic mobile core network compliant with 3GPP standards, i.e., a cloud-based Evolved Packet Core (EPC) as per one of the implementation options described in [2]. As we envision a 3GPP standards-compliant VNI, some 3GPP relevant constraints have to be considered in the VNF placement algorithm.

- A UE cannot have more than one S-GW at the same time.
- A UE has to change its S-GW if it leaves the service area of the current S-GW [19], an operation called S-GW relocation, to be avoided as much as possible [6]. Hereby, a service area consists of a number of Tracking Areas (TAs), whereby each TA consists, in turn, of the coverage areas of a number of eNBs.
- As per the recommendations of SIPTO (Selective IP Traffic Offload [18]), the path between a UE and its corresponding PDN-GW(s) has to be shortened – on both directions – as much as possible, particularly for some service types (e.g., Youtube and Facebook) as determined by the policies of the mobile operator. This can be achieved by having PDN-GWs, alternatively their VNFs, placed nearby RAN nodes, an operation that may push towards the collocation of S-GWs with PDN-GWs in the vicinity of RAN [18].

IV. Solutions description

In this work, our main goal is to devise a VNF placement algorithm for carrier cloud with two main design objectives: $i)$ to minimize path between UEs and PDN-GW VNFs of an underlying VNI to ensure acceptable Quality-of-Experience (QoE) for mobile users and to optimize overall network (i.e., VNI as well as federated cloud) resource utilization; and $ii)$ to minimize S-GW relocation frequency as per 3GPP recommendations. Whilst the first objective can be achieved by placing PDN-GW VNFs nearby RAN, the second objective can be achieved only by having large Service Areas, in other words, placing S-GW VNFs at locations distant enough from RAN. This results in a conflict between the two objectives and a tradeoff is to be retrieved. Different strategies can be also opted based on the mobility features and service usage behavioral patterns of mobile users. Indeed, for UEs not highly mobile, the system may instantiate for them PDN-GW and S-GW VNFs nearby RAN nodes (so their traffic is selectively offloaded as per operator's policies [18]). For UEs with high mobility features but not highly active in terms of IP flow generation and without long-lasting IP flows, the system may also instantiate for them PDN-GW and S-GW VNFs at datacenters nearby their corresponding RAN nodes. As for UEs with high mobility features and many IP flows, particularly long-lasting ones, the system may get them PDN-GW and S-GW VNFs instantiated at distant datacenters while still minimizing the cost of "PDN-GW-to-UE" communication path. As stated earlier, this represents the case of two conflicting objectives and a tradeoff (based on optimization theory) is to be retrieved. This defines the focus of this paper.

We assume that the network and the cloud consist of a set of N eNBs, named $\mathcal{N}$, and a set of M data centers, named $\mathcal{DC}$. Let $h(i,j)$ denote the average frequency of handovers between eNB_i and eNB_j, and w_i denote the amount of traffic generated by all UEs in an eNB_i. Let $c(i,j)$ denote the cost of the path between the data center DC_i and the evolved NodeB eNB_j. The cost can be in terms of packet delivery delay, energy consumption, or a combination of thereof. As we envision a 3GPP standards-compliant VNI, all user data are anchored at their respective PDN-GW VNFs and the mobility of their sessions are managed by their respective S-GW VNFs. It is therefore possible to have a S-GW VNF handling more traffic than a PDN-GW VNF. We denote by VNF_{MAX} the maximum amount of traffic can be handled by a VM running a S-GW VNF. The notations used throughout the paper are summarized in Table I. In the envisioned network architecture, an eNB is deemed to be associated with a datacenter DC if it has a S1-flex interface to a S-GW VNF running on a VM hosted by DC. The relationship of $eNBs$ to datacenters DCs is represented through two matrices, denoted by $S(\mathcal{N},\mathcal{N})$ and $P(\mathcal{N},\mathcal{DC})$, respectively. If eNB_i and eNB_j are associated to the same DC, then $S(i,j)=1$, otherwise $S(i,j)=0$. Additionally, $P(i,j)=1$, if and only if eNB_i is associated with DC_j; otherwise $P(i,j)=0$. Accordingly, the problem of instantiating VNFs of PDN-GW and S-GW on datacenters, such that (i) the path between UEs and PDN-GW is minimized and (ii) the S-GW relocation is minimized, could be formulated according to the following linear program (1):

Notation	Description
$\mathcal{N}$	The set of eNBs in the network.
$\mathcal{DC}$	The set of data centers in the network.
$h(i,j)$	The average frequency of handovers between eNB i and eNB j.
w_i	The total amount of traffic generated by UEs in eNB_i.
$c(i,j)$	The cost of the path between DC_i and eNB_j.
VNF_{MAX}	The maximum capacity of traffic that can be handled by a S-GW VNF.

TABLE I: Notations used.

$$\begin{cases}
\textbf{min} \sum\limits_{i\in\mathcal{N}} \sum\limits_{j\in\mathcal{N}} h(i,j)(1-\mathcal{S}(i,j)) \\
\textbf{min} \sum\limits_{i\in\mathcal{N}} \sum\limits_{j\in\mathcal{DC}} c(i,j)\mathcal{P}(i,j) \\
\textbf{s. t.} \\
\forall i\in\mathcal{N}, \forall j\in\mathcal{N}, \forall t\in\mathcal{DC},\ \mathcal{P}(i,t)+\mathcal{P}(j,t)\le 1+\mathcal{S}(i,j) \\
\forall i\in\mathcal{N}, \forall j\in\mathcal{N}, \forall t\in\mathcal{DC},\ \mathcal{P}(i,t)-\mathcal{P}(j,t)\le 1-\mathcal{S}(i,j) \\
\forall i\in\mathcal{N}, \quad \sum\limits_{t\in\mathcal{DC}} \mathcal{P}(i,t)=1 \\
\forall i\in\mathcal{N}, \forall t\in\mathcal{DC},\ \mathcal{P}(i,t)\in\{0,1\} \\
\forall i\in\mathcal{N}, \forall j\in\mathcal{N},\ \mathcal{S}(i,j)\in\{0,1\} \\
\forall i\in\mathcal{N}, \forall j\in\mathcal{N},\ \mathcal{S}(i,j)=\mathcal{S}(j,i)
\end{cases} \tag{1}$$

The first objective aims at minimizing as much as possible the S-GW relocation. This will increase the likelihood of having eNBs belonging to the same service area, and therefore reduces the frequency of S-GW relocations when UEs perform handoffs between eNBs. The second objective aims at shortening the path between UEs and DCs (particularly those running PDN-GW VNFs) as much as possible. Meanwhile, the constraints in linear programming (1) are used to ensure the following conditions:

1) Constraint 1 ensures that if $\mathcal{S}(i,j)=0$, then eNB_i and eNB_j should not connect to the same datacenter. Formally, $\forall t\in\mathcal{DC}, (\mathcal{S}(i,j)=0)\Rightarrow((\mathcal{P}(i,t)=0)\vee(\mathcal{P}(j,t)=0))$.
2) Constraint 2 ensures that if $\mathcal{S}(i,j)=1$, then eNB_i and eNB_j should connect to the same datacenter. Formally, $\forall t\in\mathcal{DC}, (\mathcal{S}(i,j)=1)\Rightarrow(\mathcal{P}(i,t)=\mathcal{P}(j,t))$.
3) Constraint 3 ensures that each eNB should connect only to one datacenter.
4) Constraints 4 and 5 ensure that the matrices $\mathcal{S}$ and $\mathcal{P}$ are binary.
5) Constraint 6 ensures that the matrix $\mathcal{S}$ is symmetric.

In what follows, we present three solutions to resolve the multi-objectives problem (1). It shall be noted that the results of the three solutions are the same matrices $\mathcal{S}$ and $\mathcal{P}$, however, with different values. The first solution is proposed for networks serving UEs with high mobility features whereby the S-GW relocation avoidance has more priority. The second solution is proposed for networks serving UEs demanding high QoE for their services. The third solution, dubbed FORD (F̲air and O̲ptimal S-GW R̲elocation and data D̲elay transfer), uses Nash bargaining technique to optimize both objectives while ensuring fairness between them.

A. A-SGWR: Avoiding S-GW Relocation

In this solution, we use the min-max approach to minimize the S-GW relocation overhead in the network. We denote by $f(\mathcal{S},\mathcal{P})$ the function that we aim optimizing for the matrices $\mathcal{S}$ and $\mathcal{P}$. Formally, $f(\mathcal{S},\mathcal{P})$ can be defined as the maximum number of S-GW relocation signaling messages that can be tolerated in the network. In this solution, we denote by $DELAY_{MAX}$ the maximum delay tolerated by the network. In case there is no requirement on data delivery delay, $DELAY_{MAX}$ can be set to ∞. In this case, the optimal solution would converge to connecting all eNBs to the same datacenter in order to reduce the S-GW relocation overhead. The optimization model which aims at reducing the S-GW relocation overhead can be formulated according to the following linear program (2):

$$\begin{cases} \textbf{min } f(\mathcal{S},\mathcal{P}) \\ \textbf{s. t.} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) + \mathcal{P}(j,t) \leq 1 + \mathcal{S}(i,j) \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) - \mathcal{P}(j,t) \leq 1 - \mathcal{S}(i,j) \\ \forall i \in \mathcal{N},\ \sum\limits_{t \in \mathcal{DC}} \mathcal{P}(i,t) = 1 \\ \forall i \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) \in \{0,1\} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N},\ \mathcal{S}(i,j) \in \{0,1\} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N},\ \mathcal{S}(i,j) = \mathcal{S}(j,i) \\ \forall i \in \mathcal{N},\ \sum\limits_{j \in \mathcal{DC}} c(i,j)\mathcal{P}(i,j) \leq DELAY_{MAX} \\ \sum\limits_{i \in \mathcal{N}} \sum\limits_{j \in \mathcal{N}} h(i,j)(1 - \mathcal{S}(i,j)) \leq f(\mathcal{S},\mathcal{P}) \end{cases} \tag{2}$$

B. S-PL: Shortening Path Length between eNBs and PDN-GW VNFs

In the linear programming (3), the goal is to optimize the cost of the communication path between UEs and their respective PDN-GW VNFs. Similar to the previous solution, we apply the min-max approach. We define $g(\mathcal{S},\mathcal{P})$ as the function that we aim at optimizing for the matrices $\mathcal{S}$ and $\mathcal{P}$, respectively. We formally define $g(\mathcal{S},\mathcal{P})$ as the maximum cost of the communication path between any datacenter and any eNB. In this solution, we set the maximum amount of S-GW relocation overhead in the network to $SGWR_{MAX}$. Its value could be defined according to the mobility features of UEs in the network. Otherwise, $SGWR_{max}$ can be set to ∞. In this case, the optimal solution would converge to associating each eNB to its nearest datacenter in terms of the cost of the communication path between the eNB and the datacenter. Formally, S-PL is the solution of linear program (3).

$$\begin{cases} \textbf{min } g(\mathcal{S},\mathcal{P}) \\ \textbf{s. t.} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) + \mathcal{P}(j,t) \leq 1 + \mathcal{S}(i,j) \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) - \mathcal{P}(j,t) \leq 1 - \mathcal{S}(i,j) \\ \forall i \in \mathcal{N},\ \sum\limits_{t \in \mathcal{DC}} \mathcal{P}(i,t) = 1 \\ \forall i \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) \in \{0,1\} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N},\ \mathcal{S}(i,j) \in \{0,1\} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N},\ \mathcal{S}(i,j) = \mathcal{S}(j,i) \\ \sum\limits_{i \in \mathcal{N}} \sum\limits_{j \in \mathcal{N}} h(i,j)(1 - \mathcal{S}(i,j)) \leq SGWR_{MAX} \\ \forall i \in \mathcal{N},\ \sum\limits_{j \in \mathcal{DC}} c(i,j)\mathcal{P}(i,j) \leq g(\mathcal{S},\mathcal{P}) \end{cases} \tag{3}$$

C. Trading off S-GW relocation against data path length using Nash bargaining

1) Nash bargaining model and threat value game: Nash bargaining model is a cooperative game with non-transferable utility. We adopt this model to our third solution FORD to find a Pareto efficiency between the S-GW relocation and the data communication path length. For this purpose, S-GW relocation and the length of data communication path are considered as two players in the game. This model is based on two elements, assumed to be given and known to the players. The first element is the set of vector payoffs $\mathcal{V}$ achieved by the players if they agree to cooperate. $\mathcal{V}$ should be a convex and compact set. Formally, $\mathcal{V}$ can be defined as $\mathcal{V} = \{(u(x), v(x)), x = (x_1, x_2) \in X\}$, whereby X is the set of strategies of two players, and $u()$ and $v()$ are the utility functions of the first and second users, respectively. Second, the threat point, $d = (u^*, v^*) = (u((t_1, t_2)), v(t_1, t_2)) \in \mathcal{V}$, that represents the pair of utility whereby the two players fail to achieve an agreement. In Nash bargaining game, we aim at finding a fair and reasonable point, $(\overline{u}, \overline{v}) = f(\mathcal{V}, u^*, v^*) \in \mathcal{V}$ for an arbitrary compact convex set $\mathcal{V}$ and point $(u^*, v^*) \in \mathcal{V}$. Based on Nash theory, a set of axioms are defined that lead to $f(\mathcal{V}, u^*, v^*)$ in order to achieve a unique optimal solution $(\overline{u}, \overline{v})$:

1) **Feasibility:** $(\overline{u}, \overline{v}) \in \mathcal{V}$.
2) **Pareto Optimality:** There is no point $(u(x), v(x)) \in \mathcal{V}$ such that $u(x) \geq \overline{u}$ and $v(x) \geq \overline{v}$ except $(\overline{u}, \overline{v})$.
3) **Pareto Optimality:** If $\mathcal{V}$ is symmetric about the line $u(x) = v(x)$, and $u^* = v^*$, then $\overline{u} = \overline{v}$.
4) **Independence of irrelevant alternatives:** If T is a closed convex subset of $\mathcal{V}$, and if $(u^*, v^*) \in T$ and $(\overline{u}, \overline{v}) \in T$, then $f(\mathcal{V}, u^*, v^*) = (\overline{u}, \overline{v})$.
5) **Invariance under change of location and scale:** If $T = \{(u'(x), v'(x)), u'(x) = \alpha_1 u(x) + \beta_1, v'(x) = \alpha_2 v(x) + \beta_2 \ for \ (u(x), v(x)) \in \mathcal{V}\}$, where $\alpha_1 > 0, \alpha_2 > 0$, and B_1 and B_2 are given numbers, then $f(T, \alpha_1 u^* + \beta_1, \alpha_2 v^* + \beta_2) = (\alpha_1 \overline{u} + \beta_1, \alpha_2 \overline{v} + \beta_2)$.

Moreover, the unique solution $(\overline{u}, \overline{v})$, satisfying the above axioms, is proven to be the solution of the following optimization problem:

$$\begin{cases} \textbf{max } (u(x) - u^*)(v(x) - v^*) \\ \textbf{s. t.} \\ (u(x), v(x)) \in \mathcal{V} \\ (u(x), v(x)) \geq (u^*, v^*) \end{cases}$$

2) FORD: Fair and Optimal SGW Relocation and data delivery Delay: We denote by $d = (SGWR_{worst}, DELAY_{worst})$ the threat point of our bargaining game that resolves FORD. In contrast to traditional bargaining game, in our model the utility function of each player (i.e., S-GW relocation and data delivery delay overhead) is the opposite of its cost. In other words, $(SGWR_{worst}, DELAY_{worst}) \geq (f(\mathcal{S},\mathcal{P}), g(\mathcal{S},\mathcal{P})), \forall \mathcal{V} \in X$. ($SGWR_{worst}$ and $DELAY_{worst}$ values are fixed through the resolution of the linear programs (4) and (5), respectively.

$$\begin{cases} \textbf{min } f(\mathcal{S},\mathcal{P}) \\ \textbf{s. t.} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) + \mathcal{P}(j,t) \leq 1 + \mathcal{S}(i,j) \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) - \mathcal{P}(j,t) \leq 1 - \mathcal{S}(i,j) \\ \forall i \in \mathcal{N},\ \sum\limits_{t \in \mathcal{DC}} \mathcal{P}(i,t) = 1 \\ \forall i \in \mathcal{N}, \forall t \in \mathcal{DC},\ \mathcal{P}(i,t) \in \{0,1\} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N},\ \mathcal{S}(i,j) \in \{0,1\} \\ \forall i \in \mathcal{N}, \forall j \in \mathcal{N},\ \mathcal{S}(i,j) = \mathcal{S}(j,i) \\ \forall i \in \mathcal{N},\ \sum\limits_{j \in \mathcal{DC}} c(i,j)\mathcal{P}(i,j) \leq DELAY_{WORST} \\ DELAY_{WORST} \leq DELAY_{MAX} \\ \sum\limits_{i \in \mathcal{N}} \sum\limits_{j \in \mathcal{N}} h(i,j)(1 - \mathcal{S}(i,j)) \leq SGWR_{BEST} \\ SGWR_{BEST} \leq f(\mathcal{S},\mathcal{P}) \end{cases} \tag{4}$$

$$\left\{\begin{array}{l}
\textbf{min } g(\mathcal{S},\mathcal{P})\\
\textbf{s. t.}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)+\mathcal{P}(j,t)\leq 1+\mathcal{S}(i,j)\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)-\mathcal{P}(j,t)\leq 1-\mathcal{S}(i,j)\\
\forall i\in\mathcal{N},\ \ \sum\limits_{t\in\mathcal{DC}}\mathcal{P}(i,t)=1\\
\forall i\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)\in\{0,1\}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\ \mathcal{S}(i,j)\in\{0,1\}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\ \mathcal{S}(i,j)=\mathcal{S}(j,i)\\
\sum\limits_{i\in\mathcal{N}}\sum\limits_{j\in\mathcal{N}}h(i,j)(1-\mathcal{S}(i,j))\leq SGWR_{WORST}\\
SGWR_{WORST}\leq SGWR_{MAX}\\
\forall i\in\mathcal{N},\ \sum\limits_{j\in\mathcal{DC}}c(i,j)\mathcal{P}(i,j)\leq DELAY_{BEST}\\
DELAY_{BEST}\leq g(\mathcal{S},\mathcal{P})
\end{array}\right. \tag{5}$$

In the FORD solution, the trade-off between S-GW relocation and data delivery delay overheads is achieved through the resolution of a non-convex optimization problem (6).

$$\left\{\begin{array}{l}
\textbf{max } (SGWR_{WORST}-f^{*}(\mathcal{S},\mathcal{P}))\times(DELAY_{WORST}-g^{*}(\mathcal{S},\mathcal{P}))\\
\textbf{s. t.}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)+\mathcal{P}(j,t)\leq 1+\mathcal{S}(i,j)\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)-\mathcal{P}(j,t)\leq 1-\mathcal{S}(i,j)\\
\forall i\in\mathcal{N},\ \ \sum\limits_{t\in\mathcal{DC}}\mathcal{P}(i,t)=1\\
\forall i\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)\in\{0,1\}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\ \mathcal{S}(i,j)\in\{0,1\}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\ \mathcal{S}(i,j)=\mathcal{S}(j,i)\\
\sum\limits_{i\in\mathcal{N}}\sum\limits_{j\in\mathcal{N}}h(i,j)(1-\mathcal{S}(i,j))\leq f^{*}(\mathcal{S},\mathcal{P}))\\
f^{*}(\mathcal{S},\mathcal{P}))\leq SGWR_{WORST}\\
\forall i\in\mathcal{N},\ \sum\limits_{j\in\mathcal{DC}}c(i,j)\mathcal{P}(i,j)\leq g^{*}(\mathcal{S},\mathcal{P})\\
g^{*}(\mathcal{S},\mathcal{P}))\leq SGWR_{WORST}
\end{array}\right. \tag{6}$$

Using the same approach as in [20], the optimization problem (6) can be transformed into a convex-optimization problem without changing the solution. The key idea is to introduce the log function which is an increasing function. Therefore, the optimization problem is reformulated as follows:

$$\left\{\begin{array}{l}
\textbf{max } \log(SGWR_{WORST}-f^{*}(\mathcal{S},\mathcal{P}))+\\
\qquad\qquad\qquad \log(DELAY_{WORST}-g^{*}(\mathcal{S},\mathcal{P}))\\
\textbf{s. t.}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)+\mathcal{P}(j,t)\leq 1+\mathcal{S}(i,j)\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)-\mathcal{P}(j,t)\leq 1-\mathcal{S}(i,j)\\
\forall i\in\mathcal{N},\ \ \sum\limits_{t\in\mathcal{DC}}\mathcal{P}(i,t)=1\\
\forall i\in\mathcal{N},\forall t\in\mathcal{DC},\ \mathcal{P}(i,t)\in\{0,1\}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\ \mathcal{S}(i,j)\in\{0,1\}\\
\forall i\in\mathcal{N},\forall j\in\mathcal{N},\ \mathcal{S}(i,j)=\mathcal{S}(j,i)\\
\sum\limits_{i\in\mathcal{N}}\sum\limits_{j\in\mathcal{N}}h(i,j)(1-\mathcal{S}(i,j))\leq f^{*}(\mathcal{S},\mathcal{P}))\\
f^{*}(\mathcal{S},\mathcal{P}))\leq SGWR_{WORST}\\
\forall i\in\mathcal{N},\ \sum\limits_{j\in\mathcal{DC}}c(i,j)\mathcal{P}(i,j)\leq g^{*}(\mathcal{S},\mathcal{P})\\
g^{*}(\mathcal{S},\mathcal{P}))\leq SGWR_{WORST}
\end{array}\right. \tag{7}$$

V. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed solutions (i.e. A-SGWR, S-PL and FORD). All solutions are evaluated in terms of the following metrics:

1) S-GW relocation cost: The number of serving area update messages generated and forwarded from UEs to VNIs when changing their respective S-GW VNFs.
2) Delay of data delivery: The average delay for delivering data packets from UEs to their respective PDN-GW VNFs.
3) Number of VNFs: The average number of VMs created to run S-GW VNFs and PDN-GW VNFs in each datacenter.

To evaluate the proposed solutions, we have developed a simulator tool based on CPLEX, Matlab and CVX [21]. The linear programs of A-SGWR and S-PL are resolved using CPLEX tools, whereas the optimization problem of FORD is resolved via Matlab and CVX tools. In the simulations, the sites (i.e., eNodeBs) and the datacenters are randomly deployed over the simulated network. The optimization problems are solved varying: (i) the average speed of UEs, which has an impact on the frequency of handovers occurring in the network; (ii) the data delivery delay between different datacenters hosting PDN-GW VNFs and eNBs.

We simulated two scenarios:

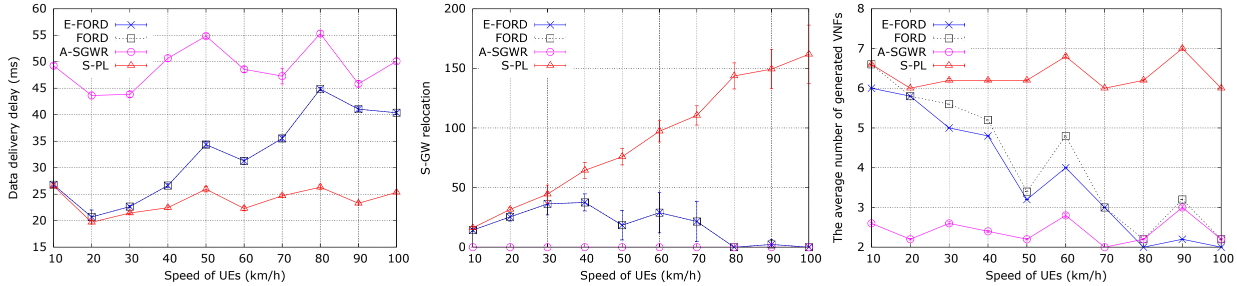
1) Vary the maximum speed of UEs and fix the maximum data delay transfer between the datacenters and the eNBs to 100 ms.
2) Vary the maximum data delivery delay between the datacenters and the eNBs and fix the maximum speed of UEs to 50 km/h.

Figs. 1 and 2 show the performance of the proposed schemes when UEs' speed and the path delay between eNBs and datacenters increase, respectively. From the figures, it becomes apparent that regardless the speed of UEs and the path delay between eNBs and datacenters, S-PL exhibits the best performance in terms of the average data delivery delay; whereas A-SGWR performs best in terms of S-GW relocation avoidance. Moreover, A-SGWR exhibits better performance in terms of the number of VNFs created in each datacenter. This is attributable to the fact that A-SGWR tends to associate most eNBs to the same datacenter in order to reduce the S-GW relocation, and consequently reduces the number of S-GW VNFs.

Figs. $1(a)$, $1(b)$, $2(a)$ and $2(b)$ illustrate the tradeoff achieved by FORD between the two conflicting objectives. FORD has performances near to F-PL when the UEs' speed is slow or when the path delay between eNBs and datacenters is high. However, it has performances near to F-SGWR when the UEs' speed is high or the path delay between eNBs and datacenters is short. FORD always finds an optimal tradeoff between S-GW relocation overhead and the average path delay between eNBs and datacenters running the VNFs of PDN/S-GWs, regardless the UEs' speed or the path delay between eNBs and datacenters. This demonstrates that it successfully achieves its key design goals in placing VNFs at adequate datacenters.

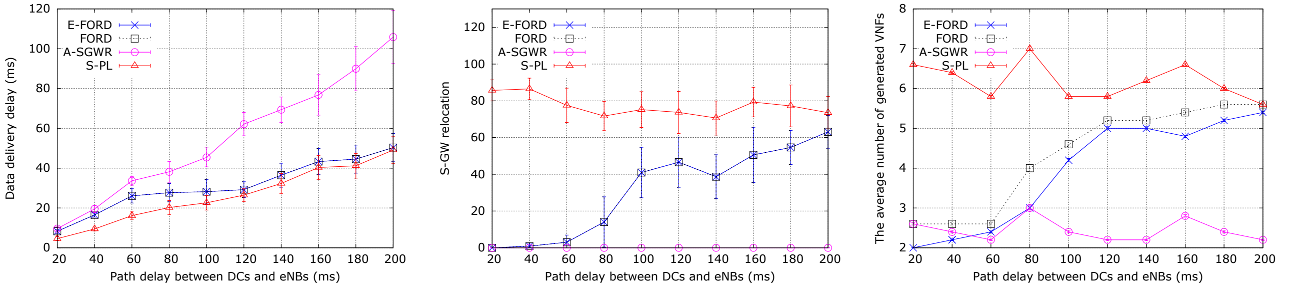
VI. CONCLUSION

In this paper, we have devised a set of solutions to the problem of VNFs placement on federated cloud to create efficient VNIs. The proposed solutions tackle two conflicting objectives, namely the insurance of QoE via the placement of



(a) Data delivery delay vs speed of UEs. (b) S-GW relocation vs speed of UEs. (c) The average number of generated VNFs vs speed of UEs.

Fig. 1: Comparison of the performances of the proposed solutions as a function of the speed of UEs.



(a) Data delivery delay vs path delay between DCs and eNBs. (b) S-GW relocation vs path delay between DCs and eNBs. (c) The average number of generated VNFs vs path delay between DCs and eNBs

Fig. 2: Comparison of the performances of the proposed solutions as a function of path delay between eNBs and datacenters.

VNFs of data anchor gateways (i.e., PDN-GW) closer to UEs and the avoidance of the relocation of mobility anchor gateways (i.e., S-GW) via the placement of their VNFs far enough from UEs. Three solutions were proposed: two solutions favor one objective over the other, whereas the third one aims at finding a fair tradeoff between the two objectives and that is through the use of bargaining Nash theory. Results obtained from the conducted simulations demonstrate the efficiency of each proposed solution in achieving its key design goals with regard to placing VNFs at adequate datacenters as per the strategy of the solution.

REFERENCES

[1] T. Taleb, “Towards Carrier Cloud: Potential, Challenges, & Solutions”, in IEEE Wireless Communications Magazine, Vol. 21, No. 3, Jun. 2014. pp. 80-91.

[2] 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”, in IEEE Network Magazine, Mar/Apr. 2015.

[3] T. Taleb, A. Jamalipour, Y. Nemoto, and N. Kato, “DEMAPS: A Load-Transition Based Mobility Management Scheme for an Efficient Selection of MAP in Mobile IPv6 Networks”, in IEEE TVT J., Vol. 58, No. 2, Feb. 2009. Pp. 954-965.

[4] T. Taleb, K. Samdanis, S. Schmid, “DNS-based solution for operator control of selected IP traffic offload”, in Proc. IEEE ICC, Kyoto, Japan, 2011.

[5] T. Taleb and A. Ksentini, “On Efficient Data Anchor Point Selection in Distributed Mobile Networks”, in Proc. IEEE ICC 2013, Budapest, Hungary, Jun. 2013.

[6] T. Taleb and A. Ksentini, “Gateway Relocation Avoidance-Aware Network Function Placement in Carrier Cloud”, in Proc. ACM MSWIM, Barcelona, Nov. 2013.

[7] M. Bagaa, T. Taleb, and A. Ksentini, “Service-Aware Network Function Placement for Efficient Traffic Handling in Carrier Cloud”, in Proc. IEEE WCNC’14, Istanbul, Turkey, Apr. 2014.

[8] F. Yousef, J. Lessman, P. Loueiro, and S. Schmid, “SoftEPC Dynamic Instantiation of Mobile Core Network Entities for Efficient Resource Utilization”, in Proc. of IEEE ICC 2013, Budapest, Hungary, Jun. 2013.

[9] G. Somani, P. Khandelwal, and K. Phatnani, “VUPIC Virtual Machine Usage Based Placement in IaaS Cloud”, CoRR abs/1212.0085 (2012)

[10] S. Skiena, “The Algorithm Design Manual”, ISBN 0-387-94860-0.

[11] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, “Optimization by Simulated Annealing Science”, 220 (1983) 671-680.

[12] M. Dorigo and T. Sttzle, “Ant Colony Optimization”, ISBN 978-0-262-04219-2.

[13] K. Le, J. Zhang, J. Meng, R. Bianchini, Y. Jaluria, and T.D. Nguyen, “Reducing Electricity Cost Through Virtual Machine Placement in High Performance Computing Clouds”, in Proc. ICHPC, Networking, Storage and Analysis (SC), Seatle, WA, USA, Nov. 2011.

[14] C. Hyser, B. McKee, R. Gardner, and B.J. Watson, “Autonomic Virtual Machine Placement in the Data Center”, in Proc. HP Labs, HPL-2007-189, Feb. 2008.

[15] W. Shi and B. Hong, “Towards Profitable Virtual Machine Placement in the Data Center”, in Proc. 4th IEEE Int’l Conf. on Utility and Cloud Computing, 2011.

[16] F.Z. Yousaf, P. Loreiro, F. Zdarsky, T. Taleb, and M. Leibsch, “Cost Analysis of initial deployment strategies of a Virtual Network Infrastructure in a Datacenter”, submitted to IEEE Communications Mag., Supplement on Communications Standards.

[17] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V.C.M. Leung, “Cache in the air: exploiting content caching and delivery techniques for 5G systems”, in IEEE Communications Magazine, Vol. 52, No. 2, Feb. 2014. pp.131-139.

[18] T. Taleb, Y. Hadjadj-Aoul, and K. Samdanis, “Efficient Solutions for Optimized Data Traffic Management in 3GPP Networks”, to appear in IEEE Systems J.

[19] A. Kunz, T. Taleb, and S. Schmid, “On Minimizing SGW/MME Relocations in LTE”, in Proc. ACM IWCMC’10, Caen, France, Jun. 2010.

[20] Z. Yangming, W. Sheng, X. Sizhong, W. Xiong, G. Xiujiao and Q. Chunming “Load balance vs energy efficiency in traffic engineering: A game Theoretical Perspective”, in Proc. IEEE INFOCOM’13, Turin, Italy, Aril. 2013.

[21] G. Michael and B. Stephen “Graph implementations for nonsmooth convex programs”, in Springer Recent Advances in Learning and Control, 2008.