Rethinking Service Chain Embedding for Cellular Network Slicing

Chrysa Papagianni
Institute for Systems Research
University of Maryland
College Park, Maryland, USA
chrisap@isr.umd.edu

Panagiotis Papadimitriou
Department of Applied Informatics
University of Macedonia, Greece
papadimitriou@uom.edu.gr

John S. Baras
Department of Electrical
and Computer Engineering
University of Maryland
College Park, Maryland, USA
baras@isr.umd.edu

***Abstract*—5G is set out to address the business contexts of 2020 and beyond, by enabling new network and service capabilities. The industry consensus is that 5G will facilitate ubiquitous connectivity, seamlessly integrating wireless technologies and complementary communication networks, while operators will be capable of providing networks on a need-for-service basis. Furthermore, there is a need for operators to exploit new revenue sources and break the traditional business model of a single network infrastructure ownership, by supporting multi-tenancy. Network slicing can provide a solution towards this end; it is considered a key for meeting 5G's diverse requirements, including future-proof scalability and flexibility.**

Provisioning and management of network slices in the transition from Long Term Evolution (LTE) to the emerging 5G systems poses the need for the mapping of service chains that express traffic and processing requirements of LTE slices. In this respect, we follow a different approach to the service chain mapping problem, promoting virtualized network function (NF) sharing among multiple service chains that are associated with a certain network slice. Using mixed-integer linear programming formulations, we show that our approach leads to reduced NF state and management overhead, compared to the common resource allocation practice in virtualized Radio Access Networks.

I. Introduction

Next-generation cellular networks will cater to a wide range of new business opportunities, such as network slice provisioning on a lease basis, in order to support multi-tenancy and meet diverse application requirements. Network Function Virtualization (NFV) and Software-Defined Networking (SDN) have been seen as key enablers towards 5G network slicing, as they allow the creation of customisable network elements which can be subsequently chained together programmatically. These network elements and functions can be easily configured and reused in each network slice to meet certain performance requirements, enabling new business opportunities by facilitating flexible and agile support for multi-service and multi-tenancy.

While the vision is very compelling from an infrastructure, operation and business perspective, the deployment of network slices poses various challenges, inherent to the enabling technologies, specific to the shared physical medium or associated to the application context. Focusing on Software Defined Mobile Networks (SDMN), different tenants issue requests to a mobile network provider for leasing network slices, where each slice as a logical end-to-end construct is self-contained, using network function chains for delivering services to a given group of devices.

Long Term Evolution (LTE) network slicing [1] commonly encompasses the following: (i) virtualizing the mobile core, deploying mobile core elements as virtualized network functions (vNF), and sharing the corresponding physical resources among tenants; (ii) sharing the base station (also termed as eNodeB) resources, where different scenarios can be supported for sharing physical resource blocks in the frequency/time/space domain at Layer 2; and (iii) sharing spectrum resources between different operators.

Considering the deployment of LTE elements as vNFs over virtualized infrastructures, authors in [2] introduce LTE as a Service framework, where both the mobile core services and eNodeB are deployed in a virtualized environment, using Openstack and Linux Containers. Authors in [1] describe the deployment of LTE Components as vNFs with OpenAirInterface (OAI) and the JUJU Framework, including mobile core network elements and 3GPP compliant eNodeBs, decomposed to the baseband unit (BBU) and remote radio head (RRH). Following the principles of the aforementioned approaches, we consider network slicing from the mobile core (termed as Evolved Packet Core - EPC) to the Radio Access Network (RAN), where virtualized eNodeBs are deployed, without, however, looking into aspects of slicing and apportioning the radio spectrum. Baseband processing functions are deployed on the virtualized eNodeBs, which are hosted on general-purpose hardware, supporting the dedicated RRHs implemented using software-defined radio (SDR) technology.

To ensure that network slices can deliver the desired benefits for each tenant, mobile network operators need to employ advanced techniques, which will optimize resource allocation for slice provisioning and also facilitate closed-loop performance maintenance. To this end, new algorithms and solutions need to be devised for allocating network and computing resources among different slices with the objective of meeting the performance and other functional requirements of applications/services, while, at the same time, maximizing the overall utility for the provider. In this respect, we consider a LTE slice composed of a group of service chains (SFC),

ISBN 978-3-903176-08-9 © 2018 IFIP

e.g., each one handling a set of traffic classes, such as voice, media streaming etc. Hence, we tackle this resource allocation problem at the granularity of a service chain, and, thereby, seek to optimize the assignment of service chains onto the virtualized RAN infrastructure. This essentially consists in the placement of virtulized LTE/EPC[1] elements (which are assumed to be implemented as vNFs) and the selection of the corresponding paths between these vNFs.

In most existing approaches (*e.g.*, [3], [4]), separate vNFs are allocated for each service chain, which means that each vNF instance is associated with a single chain. This approach yields: (i) increased overheads associated with vNF provisioning and management, (ii) potentially larger amount of NF state, if the state required by a LTE element is replicated among all the vNF instances in the slice, (iii) inefficient resource utilisation, since certain vNFs may have the required capacity to serve additional service chains, due to the statistical multiplexing of traffic, and (iv) fragmentation of resources, due to the larger number of vNF instances. To alleviate these inefficiencies, we promote the sharing of vNFs among the service chains of a LTE slice, aiming at lower provisioning and management costs as well as NF state reduction. To this end, we present mixed-integer linear programming (MILP) formulations for: (i) LTE service chain mapping with vNF sharing, and (ii) a baseline LTE service mapping that corresponds to the common resource allocation practice in virtualized RANs (*i.e.*, each service chain has its own dedicated vNFs).

The remainder of the paper is organized as follows. Section II provides an overview of the LTE network infrastructure. Section III describes the service chain mapping problem. In Section IV, we present our MILP formulations. In Section V, we compare the efficiency of our proposed method against the baseline using simulations. Section VI provides an overview of related work. Finally, in Section VII, we highlight our conclusions and discuss directions for future work.

II. Background

The term LTE encompasses the evolution of the UMTS radio access to the Evolved-UTRAN (E-UTRAN). This is accompanied by the evolution in the GPRS Core Network, under the name System Architecture Evolution (SAE), which includes the EPC network. LTE and SAE together constitute the Evolved Packet System.

E-UTRAN. The E-UTRAN consists of a network of base stations (termed as eNodeBs – eNBs) that provide radio access to the User Equipment (UE). eNBs provide user and control plane protocol terminations toward the UEs. They communicate with each other by means of the X2 interface. The eNBs are also connected via the S1 interface to the EPC. RANs are usually provisioned for peak loads, leading to inefficient resource utilisation, *i.e.*, up to 80% of CAPEX and 60% of OPEX in mobile networks is spent on RANs [5].

Centralised Radio Access Network (C-RAN) architecture splits the eNB to: (i) the BBU responsible for L1 digital processing of the baseband signal (*i.e.*, radio function) along with performing upper layer functions and interfacing with the backhaul; and (ii) the RRH performing functions such as amplification of RF signals, filtering, and AD/DA conversion. Following the C-RAN approach RRHs, installed close to the antenna, are connected to a centralized BBU pool at macro cell sites or central office locations, using the fronthaul transport network. Different protocols have been standardized for the fronthaul such as the common public radio interface (CPRI) [6]. Virtual RAN extends this flexibility further, through the virtualization of the execution environment [7], where radio functions is a network service running in a virtualized environment (Cloud RAN), potentially delivered as a cloud service (RAN as a service – RANaaS). Advances in the direction of leveraging NFV principles for C-RAN (also known as NFV C-RAN,vRAN) are currently emerging via proof of concept implementations [8].

EPC. The EPC contains user and control plane elements for routing, session establishment, mobility management, and billing. The user plane mainly consists of the Serving Gateway (S-GW) and the Packet Data Network Gateway (P-GW), used for UE traffic forwarding and tunnelling. More specifically, S-GW serves as a mobility anchor, whereas the P-GW routes UE traffic to external Packet Data Networks (PDNs). Mobility Management Entity (MME) is the main control plane element, responsible for UE authentication and authorization, session establishment and mobility management. The QoS level for each transmission path (termed as EPS bearer) between the UE and the P-GW is determined by the P-GW. When a UE is attached to the LTE, a default bearer is established supporting best-effort QoS. The EPS bearer consists of the radio data bearer (*i.e.* between UE and eNB), the S1 data bearer (*i.e.* between eNB and S-GW) and the S5 data bearer (*i.e.* between the SG-W and the P-GW). The GPRS tunneling protocol (GTP) is used for setting up the user plane datapaths between the eNB, S-GW and P-GW. In many cases, application-specific traffic (*e.g.* voice, video) is enforced to traverse a set of NFs, used by operators to differentiate their services [9]. Such NFs are traditionally deployed as specialized network devices, known as middleboxes. The SGi interface signifies the demarcation point between the EPC (P-GW) and the PDN. SGi-LAN refers to the NFs (*e.g.*, NATs, firewalls, caches) deployed by mobile operators on this reference point.

III. Problem Description

In order to create an LTE network slice, we should dynamically allocate, install, program and configure all the LTE network-specific elements. This requires the deployment of virtualized data and control plane functional entities (*i.e.*, BBU in RAN and MME, S/P-GW at the EPC) at the mobile operator's NFV Infrastructure (NFVI), which may span multiple NFVI Points-of-Presence (PoPs), *i.e.*, datacenters (DCs). NFVI-PoPs extend to the operator's WAN infrastructure, such as local or regional PoPs for small or larger-scale NFVI deployments. LTE network slicing further raises the need

[1] We refer to LTE/EPC as LTE in the rest of the paper.

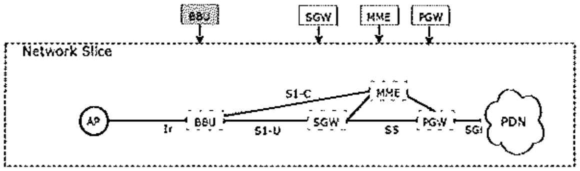

Fig. 1: LTE service chain.

for service chaining (*i.e.,* steering the traffic through a sequence of vNFs that implement the LTE user and control plane elements). Service chaining can be configured using flow tagging or tunneling to overcome the implications of "mangling" middleboxes, as exemplified in recent work [10], [11]. Fig. 1 illustrates such a service chain[2], whose end-points express different levels of abstraction at the mobile fronthaul (*e.g., Aggregation Point* for RRHs using appropriate equipment such as CPRI2Ethernet Gateway and CPRI mux). Based on this description, for a single LTE network slice, we need to efficiently place a set of LTE service chains, defined by the corresponding end-points (*e.g.,* one service chain per RRH aggregation point or RRH/cell).

The optimization problem at hand is the minimization of the resource provisioning cost for the LTE network operator, while allocating CPU and bandwidth for the LTE service chains. The problem is similar to service mapping (*e.g.,* [12]), since LTE service chains can be seen as bi-directional graphs that need to be embedded onto a substrate network [13]. This approach has been followed by recent work on NF placement in a virtualized EPC [4], [14]. However, in this way, the set of vNF forwarding graphs are mapped independently, leading to a potentially large number of vNFs, which in turn yields a substantial management cost for the LTE operator, especially during dynamic re-provisioning. Another downside of this approach is that the state required for each LTE element has to be replicated across a large number of NF instances, which essentially increases the total amount of state that has to be maintained.

In contrast to this common practice and inline with [15], we promote NF sharing across LTE service chains in order to reduce the number of NF instances and, consequently, the provisioning and management cost incurred by network slicing. In particular, we consider that flows from different cells (RRHs or Aggregation Points) can share and reuse NFs. For example, Fig. 2 illustrates two LTE service chains that share common vNFs (P-GW and S-GW). In this respect, we decompose the problem of *resource allocation for LTE network slicing* into the following steps:

- **Slice dimensioning,** which generates the number of NF instances (for each LTE data or control plane element) required to handle the expected traffic load. For example, a typical LTE system at a national level is composed of 10s of PGW, 100s of SGWs, and 1000s of eNBs [16]. The load is defined by the inbound traffic rate and the resource profile for each virtualized EPC functions (*i.e.* CPU cycles / packet).
- **NF placement,** which computes the optimized assignment of the generated NF instances onto the servers of the operator's NFVI.
- **Binding,** which associates the assigned NF instances with the LTE service chains, according to their computational and bandwidth requirements.
- **Path Selection,** which refers to the selection of the data paths through the LTE vNFs placed and bound to the service chains.

Following this approach, we present a MILP formulation for near-optimal LTE service chain mappings, by sharing vNFs among multiple LTE service chains. More specifically, vNF sharing is applicable, *e.g.,* for a set of LTE service chains on the same Tracking Area (TA), which is the logical grouping of neighbour eNBs in LTE networks, or the Tracking Area List (TAL), which is a group of Tracking Areas. TAs manage and locate UEs in a LTE network, when the UE is in *CONNECTED* state. However, in *IDLE* state the UE location is only known at TAL level. Therefore, at any point in time, the corresponding number of UEs in the TAL can provide an estimate of the expected load, which is required for slice dimensioning.

In Section IV, we present a baseline MILP formulation which corresponds to the common practice for LTE service chain mapping, *i.e.,* each LTE service chain is associated with its own individual vNFs. Section V provides a detailed comparison between the two methods and discusses the gains achieved by the MILP that promotes NF sharing.

IV. Problem Formulations for LTE Slicing

In this section, we discuss (i) the MILP formulation that shares NFs among service chains, and (ii) the baseline MILP formulation that allocates separate NFs per chain.

A. Service Mapping with NF Sharing

1) Network and Request Model: In the following, we introduce the network and request model for the MILP formulation with NF sharing.

Network Model. The operator has a number of $|A|$ available NFVI-PoPs interconnected via the provider's network. Each site's infrastructure is represented as a directed weighted graph $G_a = (N_a, E_a)$, where N_a represents the set of all nodes (*i.e.,* routers/switches, and servers) that belong to the operator's NFVI a and E_a the corresponding substrate links. Inter-DC links are denoted as $\{E_{aa'} = (u_a, v_{a'}) | \forall u_a \in N_a, v_{a'} \in N_{a'} \forall a, a' \in A, a \neq a'\}$. We consider a network-wide view of the operator's network; the overall substrate topology is denoted as $G_{S'} = (N_{S'}, E_{S'})$, where $N_{S'} = \cup_{a=1}^{A} N_a$ and $E_{S'} = \left(\bigcup_{\forall a \in A} E_a\right) \bigcup \left(\bigcup_{\substack{\forall a,a' \in A,\\ a \neq a'}} E_{aa'}\right)$.

We consider a set of I RRHs that belong to the same TAL. The length of the fronthaul link between the RRH and the BBU vNF can not exceed a given value; this guarantees the

[2] For the sake of readability, traffic flow direction in the uplink and the downlink is not depicted.

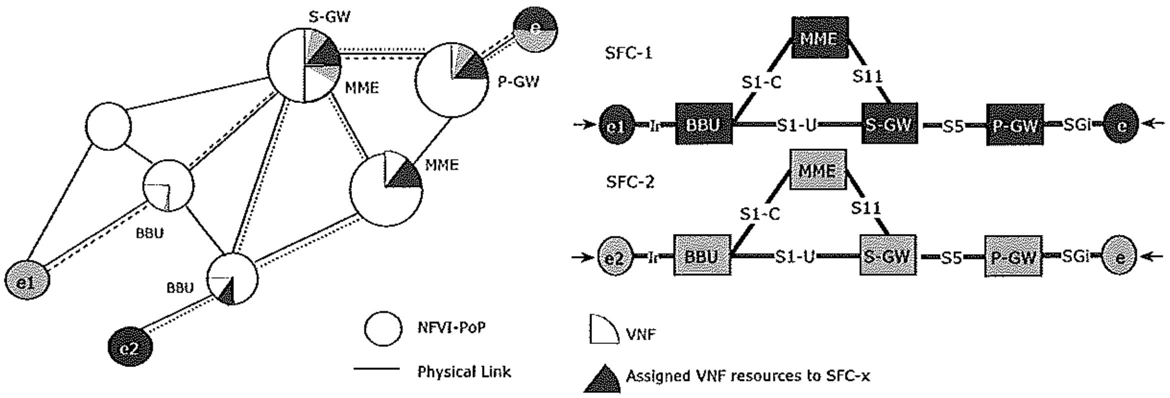

Fig. 2: LTE SFC(s) placed on physical infrastructure.

signal synchronization between RRH and BBU [17]. In this case, this is set to 20km [7]. We consider that there is at least one existing link from an RRH to an NFVI-PoP that meets this requirement. We augment the substrate graph node set with I pseudo nodes $N_S = N_S^{\prime} \cup I$, (zero capacity). Network links between RRHs and PoPs are added to the link set, thus $E_S = E_{S'} \cup \{(u,i) | u \in N_{S'}, i \in I\} \cup \{(i,u) | u \in N_{S'}, i \in I\}$ forming the directed substrate graph $G_S = (N_S, E_S)$. Node sets of specific type are denoted as N_S^x (*i.e.*, routers, switches, RRHs, Internet Exchange Points (IXPs) and servers). Thus, the overall set of physical servers for the substrate is $N_S^{ser} = \cup_{a=1}^{A} N_a^{ser}$. Every node $n \in N_S^x$ and link $(u,v) \in E_S$ is associated with its residual capacity, denoted by r_n and r_{uv}, respectively. The residual capacity for substrate routers, switches and RRHs is set to zero.

Virtual Network Functions. The set N_V represents possible vNFs (*e.g.*, BBU, MME, P-GW, S-GW) that can be deployed at the NFVI-PoPs. Each NF instance is associated with a given amount of computing resources $d^i, i \in N_V$, used by the service chains assigned to that instance. Thus, inline with [15] we have NF instances of the same type (*e.g.*, set N_V^{MME} of MME instances) with different sizes. Each vNF $i \in N_V$ can be instantiated at a substrate node of type N_S^{ser} at most U_i times (*e.g.*, depending on the number of purchased licenses). We extend the set N_V with two additional pseudo vNFs, RRH and IXP, assuming they can be instantiated only at N_S^{RRH} and N_S^{router}, respectively, utilizing minimal computing resources.

Service Chain Model. We use a directed weighted graph $G_f = (N_f, E_f), f \in F$ to express each service chain request, where F represents the set of SFCs. The set of vertices N_f includes two sets: (i) N_f^V: the set of vNFs that belong to either the RAN or the EPC, as well as any NFs (*e.g.*, NAT, firewall) that the traffic has to traverse; and (ii) N_f^S: the set of service chain end-points (RRH and IXP, in this case). Each vertex $k \in N_f$ is associated with a computing demand $g^{f,k}$, which we estimate based on the inbound traffic rate and the resource profile of the LTE element (*i.e.*, CPU cycles / packet), apart from the endpoints (N_f^S) where the computing resources are minimal. The edges are denoted by $(k,m) \in E_f$ while their bandwidth demands are expressed by $g^{f,km}$ for SFC $f \in F$. We further introduce $l_u^{f,k}$ which represents the distance between the preferred location of a function $k \in N_f^S, f \in F$ and the location of the server where this will be hosted, with $u \in N_S$.

2) Problem Formulation: In the MILP formulation, the binary variable $x_u^{i,j}$ expresses the placement of instance j of vNF $i \in N_V$ on the substrate node $u \in N_S$. Furthermore, the binary variable $z_u^{f,k}$ indicates the assignment of vNF $k \in N_f$ required by service chain $f \in F$ to the substrate node $u \in N_S$. The real variable $f_{uv}^{f,km}$ expresses the amount of bandwidth assigned to link $(u,v) \in E_S$ for graph edge (k,m) in the vNF forwarding graph of service chain $f \in F$.

Objective:

$$\text{Min.} \sum_{i \in N_V} \sum_{j \in U_i} \sum_{u \in N_S} d^i x_u^{i,j} + \sum_{f \in F} \sum_{(k,m) \in E_f} \sum_{\substack{(u,v) \in E_S \\ (u \neq v)}} f_{uv}^{f,km} \quad (1)$$

Capacity related Constraints:

$$\sum_{\forall i \in N_V} \sum_{j \in U_i} d^i x_u^{i,j} \leq r_u \quad \forall u \in N_S \quad (2)$$

$$\sum_{f \in F} \sum_{(k,m) \in E_f} f_{uv}^{f,km} \leq r_{uv} \quad \forall (u,v) \in E_S \quad (3)$$

Placement and Assignment related Constraints:

$$\sum_{\forall i \in N_V^{x'}} \sum_{j \in U_i} x_u^{i,j} = 0 \quad \forall u \in N_S^x, x' \neq x \quad (4)$$

$$\sum_{\forall u \in N_S^x} x_u^{i,j} \leq 1 \quad \forall i \in N_V^x, j \in U_i \quad (5)$$

$$\sum_{\substack{\forall f \in F \\ k \in N_f : k = i}} g^{f,k} z_u^{f,k} \leq \sum_{i' \in N_V : i' = i} \sum_{j \in U_i'} d^{i'} x_u^{i',j} \quad \forall u \in N_S, i \in N_V \quad (6)$$

$$z_u^{f,k} \leq \sum_{\substack{j \in U_i \\ i \in N_V : i = k}} x_u^{i,j} \quad \forall k \in N_f, f \in F, u \in N_S \quad (7)$$

$$\sum_{u \in N_S} z_u^{f,k} = 1 \quad \forall k \in N_F, f \in F \quad (8)$$

$$l_u^{f,k} z_u^{f,k} = 0 \quad \forall k \in N_f^S \subset N_f, f \in F, \forall u \in N_S \quad (9)$$

Flow related Constraints:

$$\sum_{\substack{v\in N_S\\(u\neq v)}} (f_{uv}^{f,km} - f_{vu}^{f,km}) = g^{f,km}(z_u^{f,k} - z_u^{f,m})$$
$$m \neq k, \forall(m,k) \in E_f, f \in F, u \in N_S \quad (10)$$

Domain Constraints:

$$x_u^{i,j} \in \{0,1\} \quad \forall i \in N_V, j \in U_i, u \in N_S \quad (11)$$
$$z_u^{f,k} \in \{0,1\} \quad \forall k \in N_f, f \in F, u \in N_S \quad (12)$$
$$f_{uv}^{f,km} \geq 0 \quad \forall(u,v) \in E_S, (k,m) \in E_f, f \in F \quad (13)$$

The optimization objective of the MILP is expressed by the objective function (1). The first term of this function represents the CPU requirements, based on the vNF instances mapped to the infrastructure. The second term of the objective function expresses the accumulated bandwidth assigned to substrate links. Constraint (2) ensures that the sum of CPU required by the vNF instances mapped to substrate node u does not exceed the residucal processing power. Constraint (3) ensures that the allocated bandwidth does not exceed the residual capacity of links. Condition (4) enforces the placement of vNF (and pseudo vNF) instances on substrate nodes that meet the vNF's functional requirements. Constraint (5) ensures that each vNF instance is placed at a single substrate node. Constraint (6) ensures that the sum of processing demands of service chain elements does not exceed the amount of virtual resources provided by vNFs of type i mapped to substrate node u. Constraint (7) ensures that, if a vNF requested by an service chain is assigned to substrate node u, then at least one instance should be placed on u. Constraint (8) ensures that every required service chain (and its respective vNFs) is mapped to the infrastructure. Condition (9) enforces location constraints for the service chain endpoints. Constraint (10) enforces flow conservation, *i.e.*, the sum of all inbound and outbound traffic in switches, routers, and servers that do not host vNFs should be zero. More precisely, this condition ensures that for a given pair of assigned nodes k,m (*i.e.*, vNFs or end-points), there is a path in the network graph where the edge (k,m) has been mapped. Finally, conditions (11), (12) and (13) express the domain constraints for the three variables.

We note that the complexity of the proposed MILP can be reduced by: (i) relaxing the integer domain constraints, and (ii) using a rounding algorithm to extract feasible solutions from non-integer values. Rounding can be performed by employing existing deterministic and randomized techniques used in service mapping [18], [4], [12]. Due to space limitations, we leave this for future work.

B. Baseline Service Mapping

In the following, we discuss the MILP formulation for the baseline service mapping without NF sharing.

1) Request Model: As the *Network Model* is similar to the one described above, we hereby present only the *Request Model.*

Request Model. We use a directed graph $G_F = (N_F, E_F)$ to express a service chain request. The set of vertices N_F includes two sets: (i) N_F^V: the set of vNFs that belong to either the RAN or the EPC, and any other vNFs for additional processing; and (ii) N_F^S: the set of service chain end-points. Each vNF $i \in N_F^V$ can be instantiated at a substrate node of type N_S^{ser}, while RRH and IXP can be instantiated only at the corresponding N_S^{RRH} and N_S^{router}, respectively. Each vertex N_F^V in the graph is associated with a computing demand g^i. The edges are denoted by $(i,j) \in E_F$ while their bandwidth demands are expressed by g^{ij}. We also use l_u^i, as defined in the service chain model in Section IV-A.

2) Problem Formulation: In the following MILP formulation, we use the binary variable x_u^i to express the placement of vNF $i \in N_F$ of the service chain request G_F on the substrate node $u \in N_S$. The real variable f_{uv}^{ij} expresses the amount of bandwidth assigned to link $(u,v) \in E_S$ for the vNF graph edge (i,j).

Objective:

$$\text{Min.} \sum_{i\in N_F}\sum_{u\in N_S} x_u^i + \frac{1}{\sum\limits_{(i,j)\in E_F} g^{ij}} \sum_{(i,j)\in E_F} \sum_{\substack{(u,v)\in E_S\\(u\neq v)}} f_{uv}^{ij} \quad (14)$$

Capacity related Constraints:

$$\sum_{\forall i\in N_F} g^i x_u^i \leq r_u \quad \forall u \in N_S \quad (15)$$
$$\sum_{\forall(i,j)\in E_F} f_{uv}^{ij} \leq r_{uv} \quad \forall(u,v) \in E_S \quad (16)$$

Placement related Constraints:

$$\sum_{\forall i\in N_F^{x'}} x_u^i = 0 \quad \forall u \in N_S^x, x' \neq x \quad (17)$$
$$\sum_{\forall u\in N_S^x} x_u^i = 1 \quad \forall i \in N_F^x \quad (18)$$
$$l_u^i x_u^i = 0 \quad \forall i \in N_F^S \subset N_F, \forall u \in N_S \quad (19)$$

Flow related Constraints:

$$\sum_{\substack{v\in N_S\\(u\neq v)}} (f_{uv}^{ij} - f_{vu}^{ij}) = g^{ij}(x_u^i - x_u^j) \quad i \neq j, \forall(i,j) \in E_F, u \in N_S \quad (20)$$

Domain Constraints:

$$x_u^i \in \{0,1\} \quad \forall i \in N_F, u \in N_S \quad (21)$$
$$f_{uv}^{ij} \geq 0 \quad \forall(u,v) \in E_S, (i,j) \in E_F \quad (22)$$

The optimization objective of the MILP is expressed by the objective function (14). The first term of this function represents the number of assigned servers. The second term of the objective function expresses the accumulated bandwidth assigned to substrate links divided by the total bandwidth demand. Constraint (15) ensures that the sum of processing demands of the vNFs mapped to substrate node u does not exceed the residual computing capacity. Constraint (16) ensures that the allocated bandwidth resources do not exceed the residual link capacity. Condition (17) enforces the placement of vNFs (and pseudo vNFs) on substrate node types that meet the vNF's functional requirements. Constraint (18) ensures that a vNF is placed at most on a substrate node. Condition (19) enforces location constraints for the service chain endpoints. Constraint (20) enforces flow conservation. Finally, conditions (21) and (22) express the domain constraints for the variables. Similar to the previous MILP formulation, relaxation and rounding techniques can be employed to reduce the time complexity.

V. Evaluation

In this section, we evaluate the efficiency and discuss the feasibility of the proposed MILP model, denoted as *NF-Sharing*. The model is compared against the *Baseline* service mapping model without NF sharing. In the following we discuss the evaluation environment (Section V-A), the evaluation metrics (Section V-B) and the evaluation results (Section V-C).

A. Evaluation Environment

We have implemented an evaluation environment in Java including a service chain generator and a cellular network topology generator. We use CPLEX for our MILP models using the branch-and-cut method. Our tests are carried out on a server with one Intel Xeon four-core CPU at 3.5 GHz and 6 GB of allocated main memory.

Given the time complexity of the mixed-integer linear programs, we use a small-scale LTE scenario for the validation/evaluation of the proposed models, based on the real-world scenario presented in [19] that was created using real statistics from a region in Paris, while LTE SFCs are jointly mapped at the Tracking Area List level considering however a single TA per TAL.

NFV Infrastructure. Similar to [4], we have generated a PoP-level network topology with homogeneous NFVI PoPs. Each PoP is essentially a micro-DC with a two-level fat-tree network topology. Table I shows additional NFVI parameters. Regarding the vNF instances for *NF-Sharing*, we consider three distinct levels of LTE vNFs that can support up to 500, 750 and 1000 UEs respectively.

E-UTRAN. We rely on a multi-cell scenario for the RAN. Table II presents the E-UTRAN design parameters. Considering uniform circular cells with an overlapping factor γ of 1.2, the cell radius is $r = \gamma\sqrt{A_t/C\pi}$ (approximately 0.64 km for the above-mentioned settings). In our case, we consider varying user density (up to $rho = 385 UEs/km^2$), so that the number of active UEs per eNB ranges from 200 to 500. We also provide the number of RRHs at Tracking Area level and the Tracking Area size. The maximum length for the RRH-BBU link is limited to 20 km [8].

Traffic Classes. Similar to [4], traffic is classified into three types, *i.e.*, voice, media streaming, and background traffic, with their busy-hour parameters shown in Table III [19] [4]. $Pr\{O\}$ is the probability that a session of a specific application type is originated by a UE.

LTE vNF profiles. The CPU demand for each vNF can be derived based on the inbound traffic rate and the resource profile of the vNF (*i.e.*, CPU cycles per packet). Resource profiles are available for a wide range of NFs (*e.g.*, IPv4 forwarding [20], [21]), while existing profiling techniques (*e.g.*, [22]) can be applied to any flow processing workloads whose computational requirements are not known. We derive the CPU demands for each NF from resource profiles, similar to [12] [4]. We extract the resource profile of the MME using the study on the latency evaluation of a virtualized MME [23]. The BBU processing budget of a GPP platform was based on the study by Nikain on OAI implementation [8] that considers three functions as the main contributors to the BBU processing budget namely; iFFT/FFT, (de)modulation, and (de)coding. The proposed model computes the total BBU uplink and downlink processing time for different physical resource blocks, modulation and coding scheme (MCS) and virtualization environment. We use the particular model considering an Intel SandyBridge architecture with a CPU frequency of 3.2GHz, a channel bandwidth of 20 MHz assuming 64 quadrature amplitude modulation (QAM) in the downlink and 16 QAM in the uplink and Linux Containers platform. This leads to a total processing time of 723.5 us per subframe in the downlink and 1062.4 us per subframe in the uplink.

Service Chains. We generate vNF-forwarding graphs per cell according to Fig. 1 class based on service chain templates. In particular, each service chain contains the main LTE elements (*i.e.*, BBU, S/P-GW, MME) using the aforementioned NF profiles.

Signalling Load and Traffic. We quantify the processing load and the uplink/downlink traffic generated by LTE/EPC data management procedures, using the aforementioned traffic profile based on the analysis provided in [19] and 3GPP LTE/EPC signalling messages and their sizes provided in [24]. In this respect, applications are modelled as ON-OFF state machines, while we assume that each UE is registered in the LTE/EPC network (EMM-registered) and alternates between Connected (*ECM-Connected*) and Idle (*ECM-Idle*) states. In other words, only *Service Request/Release* procedures are taken into account. The RRC inactivity timer defines the inactivity period required for the UE to switch to IDLE state. This timer is adjusted to 40 sec, which is a widely used setting in cellular networks [19].

B. Evaluation Metrics

We use the following metrics for the evaluation of the two service chain mapping methods:

- **vNF Instances** expresses the number of vNF instances that need to be instantiated (which is strongly correlated with the amount of vNF state) in order to support the embedded SFCs.
- **Hop Count** of the vNF forwarding graph edge expresses the length of the physical path where the edge is mapped.
- **Load Balancing Level (LBL)** is defined as the maximum over the average load. We report the (moving average)

TABLE I: NFVI Parameters

NFVI PoPs	2
Servers per DC	20 in 2 racks
Server Capacity	24 · 3.2 GHz
ToR-to-Server link capacity	8 Gbps
Inter-rack link capacity	32 Gbps
Inter-DC link capacity	100 Gbps

TABLE II: User Modeling Parameters

Area Size (A_t)	180 km^2
Total Number of eNBs in the area (C)	200
Active UEs per eNB	200 ... 500
Tracking Area Size	9 km^2
Total Number of eNBs in TA	10

TABLE III: Session Parameters

Application Type	Arrival Rate (1/hour)	Duration (seconds)	Nominal Rate (Kbps)	Pr(0)
Voice	0.67	180	23.85	0.5
Streaming	5	180	2500	1
Background traffic	40	10	1500	0.8

LBL for DCs based on server load. Lower values of LBL represent better load balancing, while $LBL = 1$ designates optimal load balancing.

- **Request Acceptance Rate** is the ratio of successfully embedded requests over the total number of requests.
- **Revenue per Request** is the amount of CPU and bandwidth units specified in the request. In this case we present the aggregated revenue of the successfully embedded SFC requests.

C. Evaluation Results

Fig. 3 illustrates the number of LTE vNF instances used to serve the incoming requests, Fig. 4 depicts the CDF of the hop count of vNF graph edges mapped to physical paths (when all the vNFs of a service chain are collocated we consider the hop count to be 0), whereas Fig. 5 plots the load balancing level across DCs. The NF-Sharing approach reduces significantly (approximately by 47%) the number of vNFs assigned at the NFVI, reducing as a result the corresponding management overhead and provisioning costs associated with vNF instances. According to Fig. 4, the baseline approach employs distinct NF instances per service chain and collocates the vNFs of a service chain in the same host more often than the NF-sharing approach, as means to decrease the cost of the objective function. The behavior of the NF-Sharing approach is consistent with its formulation, attempting at every opportunity to minimize the number of vNF instances used by the incoming batch of service chain requests. However, consolidation leads to a slightly larger number of service chains assigned to vNFs that are placed on different servers; hence, the difference among the hop count CDFs. Therefore, embedding with NF-sharing increases the number of hops onto which vNF graph edges are mapped, although a larger instance of the problem would provide more insight on the particular aspect. At the same time, the NF-sharing approach yields better load balancing, comparing the corresponding load balancing levels for DC1 and DC2 with the baseline.

Fig 6 and 7 illustrate the request acceptance rate of the two approaches, and the corresponding revenue from embedding the service chains, respectively. The baseline leads to an increased acceptance rate and corresponding revenue, due to its intrinsic flexibility, placing independently vNFs per chain. When the DC utilization level increases significantly, the NF-Sharing approach cannot map the corresponding set of service chains per TAL, as opposed to the baseline that embeds chains with finer granularity (approximately 10% higher than NF-Sharing). However, flexibility comes at the cost of a larger number of vNF instances. NF-Sharing results in a trade-off by reducing the number of vNFs assigned at the NFVI, with a proportionally quite smaller reduction at the acceptance rate and revenue. Certainly, a high request acceptance rate is important for the infrastructure provider, since he can increase his revenue by meeting the requirements of multiple Mobile Virtual Network Operators (MVNO) that lease network slices. However, in the process of providing LTE as a Service, operational costs (*e.g.*, slice provisioning/configuration, as the NF state is significantly less with the NF-Sharing approach) need also to be taken into consideration.

Our goal is to decompose the LTE network elements into vNF instances that are easily instantiated based on capacity requirements, but without over-fragmentation that increases the overheads associated with provisioning and NF state management; that is exactly what the NF-sharing approach achieves. Furthermore, optimizing NF placement is particularly important in a dynamic environment where resources become fragmented over time, and it might not be possible for all VNFs in a service chain to be placed in proximity. Based on our results, we believe that the enforced policy on NF placement can potentially change over time in order to reap the benefits of both solutions. More precisely, the NF-Sharing approach is deemed more appropriate for low and medium utilization levels in order to reduce vNF state, while the baseing can be employed under high utilization to exploit its flexible NF placement that eventually leads to higher acceptance rate and revenue.

VI. Related Work

In this section, we discuss related work on EPC and RAN virtualization.

EPC. Research has been conducted on the instantiation of LTE mobile core gateways (S-GW and/or PGW) as vNFs [25], [26], [27]. Alternative approaches in the same direction take into consideration data-plane delay constraints [28], [29]. However, the aforementioned methods optimize the placement only of data-plane functions for various objectives (*e.g.*, minimizing the EPC resource provisioning cost, load balancing). Recently, control-plane EPC NF placement (*e.g.*, MME, PCRF, HSS) along S/P-GWs has been also considered towards a 3GPP-compliant elastic cellular core [3], [14]. In addition, [30], [4] propose MILP formulations for the joint embedding of core

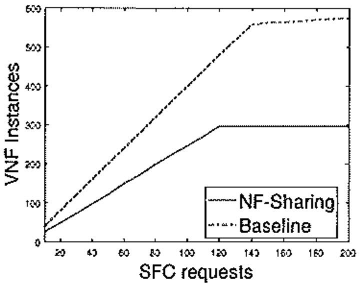

Fig. 3: Number of vNF Instances.

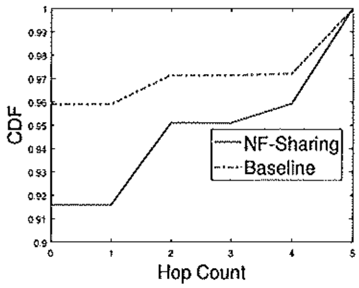

Fig. 4: CDF of hop count

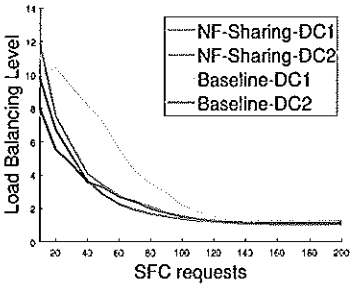

Fig. 5: DC load balancing level.

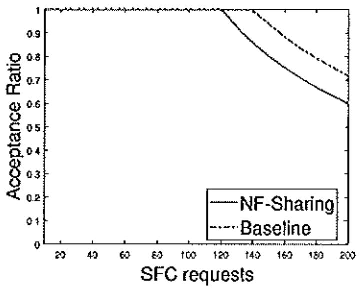

Fig. 6: Request Acceptance Rate.

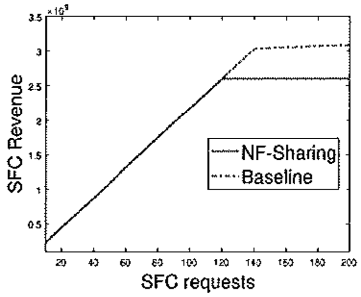

Fig. 7: Aggregated Revenue.

network service chains, taking into account latency budgets between communicating EPC elements, according to 3GPP.

KLEIN [3] presents a orchestration platform for EPC virtualization aiming at load balancing across the operator's datacenters. In terms of NF placement on the virtualized EPC, KLEIN decomposes the placement optimization into three steps (*i.e.,* region, datacenter, and server selection) to cope with the problem complexity at large scale.

RAN. There has been comprehensive research over the past years on minimizing energy consumption in RAN. The problem has been formulated as a joint optimization of RRH selection and power-minimization beamforming [31], [32] or joint optimization of RRH selection and precoding design [33], [34]. Shifting the focus towards the fronthaul and efficient resource usage of the BBU pool, BBU placement has been jointly optimized with the fronthaul transport network [35], [36], [37]. However, these studies are not focused on the placement of virtualized RAN elements.

Following a technology-agnostic approach the problem of BBU placement and RRH assignment in the RAN has been recently investigated. Authors in [38] address the problem in the context of a virtualized RAN, where functions from an eNB (*e.g.,* BBU) are implemented in a shared infrastructure located at either a DC or distributed in network nodes. Specifically this work attempts to minimize (i) the deployment cost of a BBU server, (ii) the cost of setting up the fronthaul links required between the BBUs and RRHs, and (iii) the deviation between the desired and actual latency in the fronthaul links, subject to constraints related to the resource capacities of the physical resources and a corresponding budget for the maximum number of BBU servers. The ILP formulation can be reduced to the maximal covering location problem that is known to be NP-hard. The authors propose a cost-aware greedy algorithm, reaching potentially a suboptimal placement and assignment solution, through a ranking and selection procedure. Authors in [17] strive to minimize the cost of deploying a BBU pool increased by the cost of the corresponding fronthaul links required, while respecting the resource capacities of the physical resources and ensuring that the length of the optical link between the BBU pool and RRHs for signal synchronization can not exceed a predefined maximum value. The problem is formulated as an ILP and solved using a local search heuristic.

In contrast to the aforementioned studies, we provide optimization methods for the joint placement of E-UTRAN and EPC elements onto virtualized infrastructures, as an enabler for 5G network slicing. Our approach is also different, as we enable NF sharing among service chains in order to reduce the number of NF instances and, consequently, the associated provisioning and management cost for cellular network operators.

VII. Conclusions

Towards the delivery of LTE as a service, we tackled the challenging problem of LTE service chain assignment onto the operator's NFV infrastructure, from a different perspective. In this respect, we proposed a MILP formulation for near-optimal LTE service chain mappings, by sharing vNFs among multiple service chains in a network slice, as means to reduce the provisioning and management cost (which is strongly correlated with the number of vNF instances), as well as the fragmentation of resources. To identify potential gains stemming from vNF sharing, we compared our proposed MILP against a baseline MILP which assigns separate vNFs for each service chain.

Our evaluation results corroborate the smaller number of vNF instances allocated with the proposed MILP. This essentially leads to lower overheads with respect to vNF provisioning and management. Additional gains brought by NF sharing include the reduction in the path length and better load balancing in the operator's DCs. Our evaluation further indicates that these gains diminish at high utilization levels, at which the flexibility afforded by a larger number of vNF instances may be preferable by the operator, as it can lead to higher request acceptance rates, and thereby, larger generated revenue. Our evaluation can be used to drive the development of a hybrid LTE service chain mapping approach, at which NF sharing can be enabled depending on the DC load.

In future work, we plan to conduct an experimental evaluation of NF sharing in virtualized RANs in order to quantify the provisioning and management cost savings for the operator. We will further investigate whether NF sharing introduces any implications on resource isolation among the different service chains.

ACKNOWLEDGMENTS

This work is partially supported by the EU-BRA Horizon 2020 NECOS Project (Grant Agreement No. 777067).

REFERENCES

[1] K. Katsalis et al, "Network slices toward 5G communications: Slicing the LTE network," IEEE Communications Magazine, vol. 55, no. 8, 2017, pp.146-154.

[2] N. Nikaein et al., "Network Store: Exploring Slicing in Future 5G Networks," ACM MobiArch, Paris, France, Sep. 2015.

[3] Z. Qazi et al., "KLEIN: A Minimally Disruptive Design for an Elastic Cellular Core," ACM SOSR '16, Santa Clara, CA, USA, March, 2016.

[4] D. Dietrich et al., "Network Function Placement on Virtualized Cellular Cores," IEEE COMSNETS, Bangalore, India, January 2017.

[5] China Mobile Research Institute,"C-RAN The Road Towards Green RAN," White Paper. Version 2.5. [Online]. Available: http://labs.chinamobile.com/cran/wpcontent/uploads/ CRAN_white_paper_v2_5_EN.pdf

[6] FUJITSU, "The Benefits of Cloud-RAN Architecture in Mobile Network Expansion", [Online]. Available: http://www.fujitsu.com/downloads/TEL/fnc/whitepapers/ CloudRANwp.pdf.l [Accessed: Dec. 8, 2017].

[7] N. Nikaein. "Processing Radio Access Network Functions in the Cloud: Critical Issues and Modeling," ACM MCS '15, Paris, France, Sept. 2015.

[8] N. Nikaein, et al., "Demo: Closer to Cloud-RAN: RAN as a Service," ACM MobiCom '15, Paris, France, Sept. 2015.

[9] IETF Service Function Chaining Use Cases in Mobile Networks. [Online]. Available: https:tools.ietf.orghtmldraft-ietf-sfc-use-case-mobility-02 [Accessed: Dec. 8, 2017].

[10] S. Fayazbakhsh et al., Enforcing Network-Wide Policies in the Presence of Dynamic Middlebox Actions using FlowTags, ACM SIGCOMM HotSDN '13, Hong Kong, China, August 2013.

[11] Z. Qazi et al., "SIMPLE-fying middlebox policy enforcement using SDN," ACM SIGCOMM '13, Hong Kong, China, August 2013.

[12] D. Dietrich et al., "Multi-Provider Service Chain Embedding with Nestor," IEEE Transactions on Network and Service Management, vol. 14, no. 1, March 2017, pp. 91-105.

[13] S. Mehraghdam, M. Keller, and H.Karl, "Specifying and placing chains of virtual network functions," CloudNet, London, UK, Oct. 2014.

[14] A. Baumgartner, V.S. Reddy, and T. Bauschert, "Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization," IEEE NetSoft, London, UK, June 2015.

[15] M. C. Luizelli et al., "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," IFIP/IEEE IM, Ottawa, Canada, July 2015.

[16] SONATA D2.1 Use Cases and Requirements. [Online]. Available: http://www.sonata-nfv.eu/content/d21-use-cases-and-requirements [Accessed: Dec. 8, 2017].

[17] S. Xu and S. Wang, "Efficient Algorithm for Baseband Unit Pool Planning in Cloud Radio Access Networks," VTC 2016, Nanjing, China, May 2016.

[18] M. Chowdhury, M. Rahman, and R. Boutaba, "Virtual Network Embedding with Coordinated Node and Link Mapping," IEEE/ACM Transactions on Networking, vol. 20, no. 1, Feb. 2012, pp. 206-219.

[19] W. Diego, I. Hamchaoui, and X. Lagrange, "The Cost of QoS in LTE/EPC Mobile Networks Evaluation of Processing Load," VTC2015, Boston, MA, USA, Sep. 2015.

[20] A. Abujoda, and P. Papadimitriou, "Profiling packet processing workloads on commodity servers", IFIP WWIC 2013, St. Petersburg, Russia, June 2013.

[21] M. Dobrescu, K. Argyarki, and S. Ratnasamy, "Toward Predictable Performance in Software Packet-Processing Platforms," USENIX NSDI, San Jose, CA, USA, March 2016.

[22] Q. Wu and T. Wolf, "Runtime Task Allocation in Multi-Core Packet Processing Systems," IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 10, Oct. 2012, pp. 1934–1943.

[23] J. Prados-Garzon et al., "Latency evaluation of a virtualized MME," IEEE Wireless Days, Toulouse, France, March 2016.

[24] M. R. Sama et al., "Enabling network programmability in LTE/EPC architecture using OpenFlow," IEEE WiOpt, Hammamet, Tunisia, May 2014.

[25] T. Taleb and A. Ksentini, "Gateway relocation avoidance-aware network function placement in carrier cloud," ACM MSWiM '13, Barcelona, Spain, Nov. 2013.

[26] M. Bagaa, T. Taleb, and A. Ksentini, "Service-Aware Network Function Placement for Efficient Traffic Handling in Carrier Cloud," IEEE WCNC, Istanbul, Turkey, Nov. 2014.

[27] F. Yousaf et al., "SoftEPC: Dynamic instantiation of mobile core network entities for efficient resource utilization," IEEE ICC, Budapest, Hungary, June 2013.

[28] A. Basta et al., "Applying NFV and SDN to LTE mobile core gateways, the functions placement problem," ACM AllThingsCellular, Chicago ,IL, USA, August 2014.

[29] T. Taleb, M. Bagaa, and A. Ksentini, "User mobility-aware virtual network function placement for virtual 5G network infrastructure," IEEE ICC, London, UK, June 2015.

[30] A. Baumgartner, V.S. Reddy, and T. Bauschert, "Combined Virtual Mobile Core Network Function Placement and Topology Optimization with Latency Bounds," IEEE EWSDN, Bilbao, Spain, Sept. 2015.

[31] Y. Shi, J. Zhang, and K.B. Letaief, "Group sparse beamforming for green Cloud-RAN," IEEE/ACM Transactions on Wireless Communications, vol. 13, no. 5, May 2014, pp. 2809-2823.

[32] J. Tang, W.P. Tay, and T.Q. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," IEEE/ACM Transactions on Wireless Communications, vol. 14, no. 9, Sept. 2015, pp. 5068-5081.

[33] V.N. Ha, L.B. Le, and Ngo. c-Dung Dao, "Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints," IEEE WCNC, Istanbul, Turkey, Nov. 2014.

[34] V.N. Ha and L.B. Le, "Computation capacity constrained joint transmission design for CRANs," IEEE WCNC, Doha, Qatar, April 2016.

[35] F. Musumeci et al., "Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks," Journal of Lightwave Technology, vol. 34, no. 8, April 2016, pp. 1963-1970.

[36] A. Asensio et al., "Study of the Centralization Level of Optical Network-Supported Cloud RAN," IEEE ONDM, Cartagena, Spain, May 2016.

[37] K. Sundaresan et al., "Fluidnet: A flexible cloud-based radio access network for small cells," IEEE/ACM Transactions on Networking, vol. 24, no. 2, April 2016, pp. 915-928.

[38] R. Mijumbi et al, "Server placement and assignment in virtualized radio access networks," IEEE CNSM 2015, Barcelona, Spain, Nov. 2015.