Towards a Cost Optimal Design for a 5G Mobile Core Network Based on SDN and NFV

Arsany Basta, Andreas Blenk, Klaus Hoffmann, Hans Jochen Morper, Marco Hoffmann, and Wolfgang Kellerer, *Senior Member, IEEE*

***Abstract*—With the rapid growth of user traffic, service innovation, and the persistent necessity to reduce costs, today's mobile operators are faced with several challenges. In networking, two concepts have emerged aiming at cost reduction, increase of network scalability and deployment flexibility, namely Network Functions Virtualization (NFV) and Software Defined Networking (SDN). NFV mitigates the dependency on hardware, where mobile network functions are deployed as software virtual network functions on commodity servers at cloud infrastructure, i.e., data centers. SDN provides a programmable and flexible network control by decoupling the mobile network functions into control plane and data plane functions. The design of the next generation mobile network (5G) requires new planning and dimensioning models to achieve a cost optimal design that supports a wide range of traffic demands. We propose three optimization models that aim at minimizing the network load cost as well as data center resources cost by finding the optimal placement of the data centers as well the SDN and NFV mobile network functions. The optimization solutions demonstrate the trade-offs between the different data center deployments, i.e., centralized or distributed, and the different cost factors, i.e., optimal network load cost or data center resources cost. We propose a Pareto optimal multi-objective model that achieves a balance between network and data center cost. Additionally, we use prior inference, based on the solutions of the single objectives, to pre-select data center locations for the multi-objective model that results in reducing the optimization complexity and achieves savings in run time while keeping a minimal optimality gap.**

***Index Terms*—Software Defined Networking, Network Functions Virtualization, 5G, mobile core network, optimization.**

I. Introduction

THE next generation 5G requires new concepts and architectures for the mobile network in order to improve the offered performance, to increase its deployment flexibility and to reduce its cost. An essential part that imposes several challenges to mobile operators is the mobile core network. The mobile core network is currently populated with several integrated hardware-based network functions. This limits the mobile core network's scalability to cope with the drastic increase in users' traffic. This also results in long deployment cycles and limits the service innovation and performance improvement. Another limitation in the current core network architecture is the distributed control plane design which contributes to the offered performance to users and induces inflexibility to the network configuration. Therefore, according to these challenges, the current deployment induces a high Total Cost of Ownership (TCO) on operators to build and operate the mobile core network and hinders the innovation in the offered services by the mobile network operators [1].

In networking, two main concepts are being considered for the core network architecture towards the next generation 5G [2], [3], namely Network Functions Virtualization (NFV) and Software Defined Networking (SDN). NFV [4] leverages the concepts of IT virtualization to network functions, where functions can be implemented in software and deployed as Virtual Network Functions (VNF) on commodity hardware at cloud, i.e., data center (DC) infrastructure. NFV offers more flexibility by removing the dependency on the hardware and it enables more possibilities for shorter deployment cycles and service upgrade. Hence, NFV is expected to reduce the cost of mobile networks. SDN [5], decouples the data and control planes of network functions and introduces an open API, e.g., OpenFlow protocol [6] as a current defacto standard, between the decoupled planes. The control plane is realized by SDN controllers that configure the SDN data plane for a mobile core network, what we refer to as SDN+ switches. SDN+ switches implement special purpose data plane functions, e.g., GPRS Tunneling Protocol (GTP) tunneling that encapsulates users' traffic or charging and accounting functions. In this way, SDN offers a programmable network, which simplifies the network operation and control. Furthermore, SDN enables a centralized control view that provides the operators with the possibility to achieve more efficient network control.

Considering the mobile core network architecture based on SDN and NFV, novel optimization models need to be developed for the planning and dimensioning of the SDN and NFV mobile core network architecture. The optimization models are required to consider the new realization of the mobile core functions as well as the new mobile core network infrastructure. Such infrastructure comprises of a mix of networking forwarding elements, i.e., switches, as well as

Manuscript received March 15, 2017; revised June 6, 2017 and July 6, 2017; accepted July 24, 2017. Date of publication July 27, 2017; date of current version December 8, 2017. This work has been performed in part in the framework of the CELTIC EUREKA project SENDATE-PLANETS (Project ID C2015/3-1) funded by the German BMBF (Project ID 16KIS0473), and in part in the framework of the EU project FlexNets funded by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No 647158 - FlexNets). The associate editor coordinating the review of this paper and approving it for publication was F. De Turck. *(Corresponding author: Arsany Basta.)*

A. Basta, A. Blenk, and W. Kellerer are with the Chair of Communication Networks, Technical University of Munich, Munich, Germany (e-mail: arsany.basta@tum.de).

K. Hoffmann, H. J. Morper, and M. Hoffmann are with Nokia Bell Labs, Munich, Germany.

Digital Object Identifier 10.1109/TNSM.2017.2732505

cloud infrastructure, i.e., data centers. The models should also incorporate new traffic models for the data as well as control planes, e.g., additional SDN control plane traffic.

In this work, we propose three optimization models that aim at finding the optimal design for a mobile core network based on SDN and NFV. These models provide optimal cost solutions with respect to the following aspects: a) the optimal placement of the data centers, which host the mobile VNFs and mobile SDN controllers, b) the optimal mapping of VNFs and controllers to each data center, and c) the number and placement of the mobile special purpose SDN+ switches. The proposed optimization models consider latency requirements for both data and control planes. An extensive evaluation is carried out, that generates various possible function chains in order to find the optimal network design that supports the expected wide range of varying traffic in 5G.

There are different cost factors that can be optimized in the new core network design based on SDN and NFV. The first cost factor is the network load cost which represents the cost of the network resources needed to support the data and control plane traffic of the mobile core network. In our previous work [7], we have introduced the optimization model that incorporates both SDN and NFV core network functions. However, we have only considered the optimization of the network load cost. We also focused only on data plane function chains and data plane latency requirements. Hence, in this work, we extend the network load cost optimization model to include control plane functions chains and control latency requirements to provide a more comprehensive overall model for a mobile core network.

The other cost factor, which is introduced by the concepts of SDN and NFV, is the cost of the data centers infrastructure that hosts the VNFs and SDN controllers. In this work, we propose a new optimization model for the data center resources cost to analyze the trade-offs between the network load and data center resources cost factors. Additionally, a multi-objective model is proposed in order to find Pareto optimal cost solutions considering both the network as well as data center resources cost. We also use prior inference, based on the single objective solutions, to pre-select candidate data center locations for the Pareto optimal multi-objective model in order to improve its run time. All three proposed models take into account the data and control plane latency as key performance metrics, as well as the number of data centers that are used for deployment.

The remainder of this paper is structured as follows. Section II presents an overview of the background and related state of the art. In Section III, the architecture of the mobile core network based on SDN and NFV is introduced with an analysis for the data and control planes. Section IV introduces the mathematical formulations and approaches for the proposed models. An extensive evaluation of the models is presented in Sections V and VI. Finally, conclusions and steps for future work are presented in Section VII.

II. Background and Related Work

The state-of-the-art literature can be classified into two areas. The first area is concerned with the architecture designs and implementation designs for SDN or NFV mobile core networks. The second area considers the modeling and optimization of SDN or NFV networks, for mobile networks and for traditional IP networks. In both areas, we could observe a clear split of the work into either SDN or NFV related.

A. SDN and NFV Mobile Network Architectures

In our review, we focus on related work that considers deployment architectures or implementation oriented solutions for SDN or NFV mobile core network. Considering SDN, Softcell [8], MobileFlow [9], SAMA [10] and SoftMoW [11] apply the concept of SDN on the mobile core network by replacing the network functions with SDN controllers and switches that are used to interconnect between the RAN and external packet networks. Reference [12] presents a qualitative discussion to the advantages and drawbacks of using SDN for mobile networks. Kempf *et al.* [13] present an SDN core network architecture with extensions to the OpenFlow protocol to implement GTP to encapsulate users' traffic in the core network. SDMA [14] and TrafficJam [15] are proposals for a core network architecture based on SDN with a focus on user mobility management using OpenFlow. Both argue that an SDN mobility management can improve the core network support for mobile users. Another direction is presented in [16] where Lindholm *et al.* focus on the state, e.g., user data tunnels and charging profiles, that needs to be collected and exchanged between SDN controllers that implement control functions of the mobile core network.

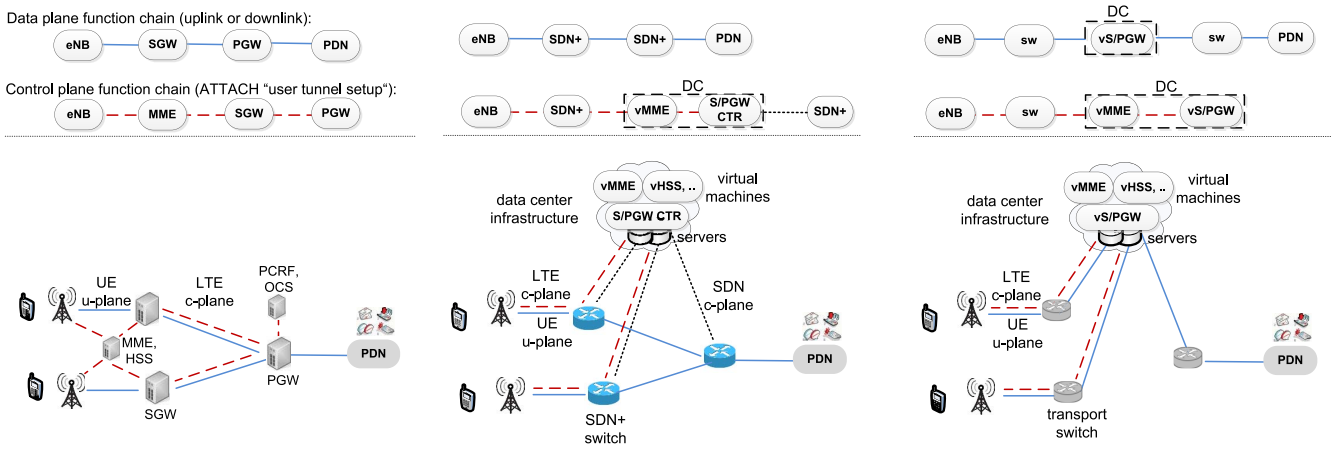
A second group of proposals has investigated an NFV architecture for the mobile core network. Nguyen *et al.* [17] and Baba *et al.* [18] discuss a core network architecture that is fully comprised of virtual network functions and deployed on a cloud infrastructure. The work in [19] present the concept of Software as a Service for a virtualized core network. Wang *et al.* [20] exploit the concepts of NFV and cloud computing to present a virtualized core network that follows mobile users as they move. Furthermore, the work in [21] and [22] present an NFV core architecture that runs alongside a standard legacy core network. The NFV core network in these proposals is used for offloading purposes in case the legacy core network is overloaded.

All proposed mobile core network architectures in the state-of-the-art literature consider either a deployment solely based on SDN or NFV. However, as we have presented in our previous work in [23], an SDN architecture can induce a higher cost due to the additional SDN control plane, while an NFV architecture can violate the network latency requirements due to the consolidation of VNFs in data centers. An architecture that includes both SDN and NFV, where part of the network is selectively operated with SDN and the other part is comprised of VNFs, can exploit the advantages from both concepts and address their limitations.

B. Dimensioning and Resource Allocation Problems

There are two main areas of modeling and optimization related to the use of SDN and NFV in the mobile core network: (a) placement of SDN controllers and switches and (b) resource allocation and placement of VNFs.

The dimensioning and placement of SDN controllers and switches is known as the controller placement problem.



(a) Legacy Core Network Architecture (LTE) (b) SDN Core Network Architecture (c) NFV Core Network Architecture

Fig. 1. Architecture comparison between (a) legacy LTE architecture, (b) SDN mobile core network and (c) NFV mobile core network. The figures additionally illustrate the logical user data plane, LTE control plane as well as SDN control plane function chains of each architecture.

This problem has been introduced in [24] which uses a brute force approach to find the placement of K number of controllers and the assignment of switches to each controller targeting a minimum control plane latency. A controller placement based on a simulated annealing heuristic has been proposed in [25] with a focus on control plane latency and resilience aspects. Sallahi and St-Hilaire [26] provide a mathematical formulation for an optimal controller placement that considers both control latency and controllers load. A controller placement that minimizes the control overhead of sharing network information among distributed controllers is proposed in [27].

Considering the resource allocation and placement of VNFs, Gebert *et al.* [28] demonstrate an optimal placement for virtual core gateways that handle sudden traffic increase in case of large crowd events. Reference [29] presents a mathematical formulation for an optimal placement of virtual function chains. They consider constraints on the network capacity as well as requested latency for a function chain. Reference [30] proposes two algorithms to embed network service chains with a target of minimizing the overall embedding cost. Shi *et al.* [31] use machine learning techniques to find an optimal placement for VNFs given data center resources. An optimal location-aware VNF mapping is proposed in [32], that minimizes the function processing and traffic transmission cost. For mobile networks, an optimization for the network resources, i.e., link and node capacity, has been proposed in [33] for the embedding of virtual mobile core network functions.

Reviewing the existing related literature on modeling and optimization, we can observe that models that jointly consider SDN and NFV are missing. Additionally, only a few proposals incorporate the detailed functions, operations and requirements of the mobile core network as we aim at in our work. Furthermore, there are only a few proposals that investigate the impact of the data plane as well as the control plane latency requirements. There is no existing work, to our knowledge, that is tailored for mobile core functions and considers the joint optimization of VNF function chains as well as SDN controllers and switches, which is the focus of our work.

III. SDN and NFV Core Network Architecture and Analysis

In this section, we discuss more in detail the next generation mobile core network design based on SDN and NFV. Additionally, we analyze the impact of SDN and NFV on both the data as well the control plane of the mobile core network, which is all incorporated in the proposed optimization models.

A. Mobile Core Network Architecture

1) Legacy LTE Mobile Core Network Architecture: The mobile core network, in the latest LTE standard [34], shown in Fig. 1a, comprises of several network functions that implement special operations that are needed for a mobile network. The core network functions can be classified into two categories based on their purpose: (a) functions that handle the control plane only, such as the Mobility Management Entity (MME) or the Home Subsriber Server (HSS) (b) functions that handle both data as well as control planes, such as the Serving Gateway (SGW) and the Packet Data Network Gateway (PGW). The data plane functions implement special purpose processing for mobile networks, i.e., GTP tunneling for the user data in order to differentiate between the users and to be able to provide service quality classes for each user. Other data plane functions include charging and accounting for the user data usage. The control plane functions handle the setup of the user tunnels and mobility management, i.e., tracking area updates and redirection of user tunnels. Additionally, control functions handle user authentication, subscription management and as access control. For more details, we refer to our previous work [23], where we performed a detailed analysis of the LTE mobile core network functions.

In the current LTE mobile core network, what we refer to as legacy, the data and control plane functions are realized by dedicated hardware that implements each specialized function. Moving towards the next generation 5G mobile core network, functions that only handle the control plane, e.g., MME, could be deployed as virtual network functions, i.e., software, on a cloud infrastructure, i.e., data centers. However, regarding the

functions that handle both the data as well as control planes, i.e., SGW and PGW, we consider the two realization options, either SDN based or NFV based.

2) SDN Mobile Core Network Architecture: Considering an SDN based deployment, shown in Fig. 1b, the control plane mobile core functions run as VNFs while the gateway functions, i.e., SGW and PGW, are decoupled into SDN controllers (S/PGW CTR) and special purpose SDN+ switches, as shown in Fig. 1b. The SDN controllers, deployed at the data center infrastructure, configure the SDN+ switches which handle the data plane traffic. The controllers implement the control plane of the core network gateway functions. Thus, the SDN controllers are required to handle the LTE control plane signaling procedures which are defined by the 3GPP standard, i.e., exchange of signaling messages with the radio access network in order to support the user's attachment to the mobile network or user's mobility. According to the signaling procedures, the controllers are responsible to configure the data plane, i.e., SDN+ switches, via the SDN API used by the operator. Additionally, the controllers need to collect the data usage of each user from the data plane switches for the purpose of charging and accounting. On the other hand, the SDN+ switches implement the gateway data plane functions. One important data plane function needed at the SDN+ switches is GTP tunneling which is used to identify data plane traffic of users. The SDN+ switches monitor the data plane statistics for charging and accounting. Additionally, the SDN+ switches need to support the configuration of quality of service classes that can be assigned to users.

3) NFV Mobile Core Network Architecture: In case of an NFV based deployment, as illustrated in Fig. 1c, the control plane mobile core functions as well as the gateway functions, i.e., SGW and PGW, run as VNFs (vS/PGW) on commodity hardware at data centers. This means that the gateway's control plane as well as the data plane processing is running on commodity servers in the cloud. The data plane processing on commodity servers can be accelerated by solutions such as Intel DPDK [35]. Hence, the legacy core network hardware would be replaced by simple forwarding switches, i.e., transport switches, that forward both the data plane and control plane traffic between the radio access network, the data centers and the external network, as illustrated in Fig. 1c. Note that in this architecture, all mobile core network functions are migrated to software running on commodity servers and are fully independent from hardware, i.e., functions which handle control plane only, e.g., MME, and functions that handle both data as well control plane, e.g., SGW and PGW. This implies that there is no processing, i.e., function, implemented on the forwarding switches of the mobile core network.

B. Data Plane Function Chains Analysis

The data plane path within the mobile core depends on the operator's decision for the realization of both the SGW and PGW functions. In case of using SDN, as shown in Fig. 1b, the legacy hardware functions would be replaced with the SDN+ switches which are controlled by the controllers residing in the cloud. This means that the data plane itself would follow the same function chains as the legacy network, i.e., between the SDN+ switches. It also means that the data plane latency depends only on the locations of the SDN+ switches and is decoupled from the location of the data center infrastructure. The data plane traffic in mobile networks can be modeled as uni directional function chains, i.e., uplink or downlink.

On the other hand, following the concept of NFV, the SGW and PGW functions are moved to the cloud. The legacy functions are replaced by simple forwarding transport switches, as shown in Fig. 1c, which transport the data plane traffic towards the data center infrastructure where the data plane processing is carried out by the software gateway functions. This means that the NFV architecture has an impact on the data plane latency as it changes the data plane function chains. The data plane function chains are extended by the links carrying the traffic back and forth between the transport switches and the data centers. Hence, the data plane latency becomes dependent on the data center locations.

C. Control Plane Function Chains Analysis

The LTE control plane procedures in the mobile core network consist of multiple sequential iterations between the network functions. For instance, the *ATTACH* procedure, refer to the 3GPP standard [34], involves mainly the MME, SGW and PGW for the setup of a user GTP tunnel. The *ATTACH* procedure defines the control messages exchanged in order to attach a user to the mobile network and setup its data plane GTP tunnel. It includes 3 control iterations between the RAN and the MME, 2 control iterations between the MME and the SGW and 2 control iterations between the SGW and PGW, respectively. Hence, the control plane is required to be modeled differently from how the control plane is modeled in traditional IP networks. Existing work, as discussed in Section II-B, e.g., [29], models the control plane function chains as uni directional demands. This does not match the control at the mobile core network, where sequential control iterations are required.

Considering an SDN deployment for the mobile core gateway functions, the control plane function chains would be mapped on the path between the RAN, i.e., eNBs, and the data centers which run the virtual control functions, i.e., vMME and the SDN controllers. This makes the control plane latency dependent on the location of the data centers. The control function chains are also extended by the control path between the SDN controllers and their respective SDN+ switches. Whereas, an NFV deployment means that the mobile core VNFs are all consolidated in data centers. Hence, the control plane function chains are mapped on the path between the RAN and the data centers infrastructure. Therefore, the latency of the control plane function chains becomes dependent only on the locations of the data centers.

D. Problem Statement

From the analysis in Sections III-B and III-C, we could observe that SDN and NFV deployments for mobile core networks show trade-offs in terms of data plane or control plane latency, network traffic and data center resources. Hence, novel optimization models are required to find an optimal planning and dimensioning for a mobile core network, that

jointly includes both SDN and NFV deployments, in terms of the network load cost and the data center resources cost. The optimal core network design entails the optimal locations for data centers and the optimal network split between SDN and NFV that supports the expected wide range of traffic demands in 5G. Additionally, the optimal network design has to ensure the mobile core network performance requirements, in terms of data plane and control plane latency.

IV. SDN and NFV Based Mobile Core Dimensioning and Resource Allocation Models

In this section, three optimization models are proposed for the optimal cost dimensioning of the mobile core network based on both SDN and NFV concepts. We introduce the mathematical formulation for the models and the used notations for each of the proposed models. The optimization models are formulated as Mixed Integer Linear Programs (MILP). In general, the aim of the proposed models is to find the optimal dimensioning and resource allocation of the core network that satisfies data plane and control plane latency requirements given a core network topology and number of data centers. The models are used to solve a) the optimal placement of the data centers, which host the mobile VNFs and SDN controllers, b) the optimal mapping of VNFs and controllers to each data center and c) the number and placement of the special purpose SDN+ switches that implement the data plane functions of the core network. The first model targets the optimal network load cost, the second model optimizes the data center resources cost, while the third model is a Pareto optimal multi-objective model that results in Pareto optimal cost for the network load and data center resources. The used notation for sets, parameters and variables is presented in Tables I–III, respectively.

TABLE I
SETS

Notation	Description
$G(V, E)$	core network graph
C	set of nodes (locations) for data centers $C \subseteq V$
V^s	set of SGW nodes (locations) $V^s \subset V$
V^p	set of PGW nodes (locations) $V^p \subset V$
E	Set of physical network edges
D	Set of traffic demands $d = (v^s, v^p) \in D$
$F^d(c, d)$	set of data function chains for demand $d \in D$, DC $c \in C$
$F^c(c, d)$	set of control function chains for demand $d \in D$, DC $c \in C$

TABLE II
PARAMETERS

Notation	Description
K	number of data centers
L^d	data plane latency requirement
L^c	control plane latency requirement
$r(d)$	requested data bandwidth by a demand $d \in D$
$\alpha(d)$	control percentage of $r(d)$ for demand $d \in D$
$l^d(c, f^d, d)$	data plane latency of demand $d \in D$ as a data function chain $f^d \in F^d$ using DC $c \in C$, 0 otherwise
$l^c(c, f^c, d)$	control plane latency of demand $d \in D$ as a control function chain $f^c \in F^c$ using DC $c \in C$, 0 otherwise
$n^d(c, f^d, d)$	data plane load of demand $d \in D$ as a data function chain $f^d \in F^d$ using DC $c \in C$, 0 otherwise
$n^c(c, f^c, d)$	control plane load of demand $d \in D$ as a control function chain $f^c \in F^c$ using DC $c \in C$, 0 otherwise
$r^d(c, f^d, d)$	DC CPU resources for demand $d \in D$ as a data function chain $f^d \in F^d$ using DC $c \in C$, 0 otherwise
$r^c(c, f^c, d)$	DC CPU resources for demand $d \in D$ as a control function chain $f^c \in F^c$ using DC $c \in C$, 0 otherwise
s_{cores}	number of cores in a data center server
p^d_{vnf}	number of cores used by a VNF for data plane
p^c_{vnf}	number of cores needed by a VNF for control plane
p^c_{ctr}	number of cores needed by an SDN controller

A. Graph Model and Notation

A core network graph $G(V, E)$ is considered with a set of nodes V and edges E. The core nodes are classified as SGW nodes $v^s \in V^s \subset V$ and PGW nodes $v^p \in V^p \subset V$. We assume a brownfield scenario where an operator would select a location to deploy a data center (DC) where it already has a deployed node, thus, data center nodes, i.e., locations, $C \subseteq V$. The set D contains flow demands in the core network, where a flow demand $d = (v^s, v^p) \in D$ represents the requested bidirectional and non-splittable data plane traffic flow, i.e., uplink and downlink, between an SGW node v^s and PGW node v^p. The data and control planes of each demand can be realized as SDN or NFV function chains, respectively. For each demand, the set $F^d(c, d)$ contains the SDN and NFV data plane function chains of a demand $d \in D$ using a data center $c \in C$. Similarly, the set $F^c(c, d)$ contains the SDN and NFV control function chains of a demand d using a data center $c \in C$.

Regarding the NFV realization of a demand $d = (v^s, v^p)$, the data plane chain is defined as the path traversing SGW node v^s, the VNFs deployed at the data center nodes c and the PGW node v^p , while the control plane is defined as three times the path between the SGW node v^s and VNF deployed at the data center $c \in C$, as explained for the ATTACH procedure in Section III-C. As for the SDN realization, the data plane function chain represents the path between the SDN+ switches, instead of the SGW node v^s and PGW node v^p. The control plane function chain is defined as three times the path between the SGW node v^s and the SDN controller deployed at data center node c in addition to maximum of the two paths between the controller and switches at v^s and v^p, respectively. All combinations of data and control functions chains with data center locations in the sets $F^d(c, d)$ and $F^c(c, d)$ are calculated for each demand, i.e., calculated and provided as input to the optimization problem in order to simplify the problem and improve the solving time. The end-to-end latency of each function chain is additionally calculated. Assuming an underlying optical transport layer in the mobile core network, the latency $\ell(e)$ of an edge e is calculated as the geographic distance in *kilometers* between any two connected nodes divided by the speed of light 2×10^8 m/s in optical fiber. The latency of a function chain is the sum of latencies $\sum_e \ell(e)$ on the edges e that belong to a data function chain $f^d(c, d) \in F^d(c, d)$ or a control function chain $f^c(c, d) \in F^c(c, d)$. According to our previous measurements and observations in [7], the processing latency of NFV gateways and carrier-grade SDN+ switches are assumed to be insignificant, in the order of *mircoseconds*, compared to the network propagation latency of a wide spread core network topology, which is in the order of *milliseconds*.

B. Network Load Cost Optimization Model

This model aims at optimizing the network cost, i.e., it finds the dimensioning and resource allocation that provides an optimal network cost. The model's cost function, what we call the network traffic load or shortly network load, is defined as the bandwidth-latency product. In this way, we could optimize the network resource allocation, i.e., bandwidth, in addition to the performance, i.e., latency, which would provide performance gains to the users' experience. For each function chain f using a data center c for each demand, the network load is computed as the requested bandwidth by the demand $r(d)$ multiplied by the latency on the function chain. Hence, the load for the data function chain is defined as $n^d(c, f^d, d) = r(d) \cdot l^d(c, f^d, d)$, while the load of a control function chain $n^c(c, f^c, d) = \alpha(d) \cdot r(d) \cdot l^c(c, f^c, d)$ where $\alpha(d)$ denotes the control bandwidth percentage of requested data plane bandwidth for this demand. For SDN function chains, we consider that the percentage of the mobile control traffic, i.e., signaling, can be assumed to be comparably similar to the traffic resulting from SDN control. The constraints used in this model are defined as follows:

1) Function Chain and DC Selection: These constraints ensure that for every demand $d \in D$ there is one function chain selected, i.e., either NFV or SDN, denoted by the binary variables $\delta^d(c, f^d, d)$ and $\delta^c(c, f^c, d)$ for data and control plane, respectively. This function chain must use at most one data center c, i.e., place the VNF at this data center location for an NFV function chain or use this data center to host the controller for the SDN chain of this demand.

$$\sum_{c \in C} \sum_{f \in F^d} \delta^d(c, f^d, d) = 1 \quad \forall d \in D \tag{1}$$

$$\sum_{c \in C} \sum_{f \in F^c} \delta^c(c, f^c, d) = 1 \quad \forall d \in D \tag{2}$$

2) Function Chain Match: This constraint makes sure that the control function chain $f^c \in F^c(c, d)$ matches the selected data plane function chain $f^d \in F^d(c, d)$ for each demand $d \in D$ using a data center location $c \in C$, e.g., if an SDN data plane function chain is selected for a demand, then the control function chain of this demand must be SDN. A function $\pi(f^d, f^c)$ returns the function chain type, i.e., SDN or NFV.

$$\delta^d(c, f^d, d) \leq \delta^c(c, f^c, d) \quad \forall d \in D, c \in C, f^d \in F^d, f^c \in F^c \tag{3}$$

3) DC Selected Flag: A binary variable $\delta(c)$ is utilized in this constraint to flag that this data center location has been selected in case at least one function chain of one demand has selected the data center c to place the VNF or controller.

$$\sum_{f^d \in F^d} \delta^d(c, f^d, d) \leq \delta(c) \quad \forall d \in D, c \in C \tag{4}$$

$$\sum_{f^c \in F^c} \delta^c\big(c, f^c, d\big) \leq \delta(c) \quad \forall d \in D, c \in C \tag{5}$$

4) Number of DCs: This constraint defines the number of data center locations to be used. It ensures that the sum of the binary variable $\delta(c)$, which indicates the overall locations, is equal to a given input parameter K.

$$\sum_{c \in C} \delta(c) = K \tag{6}$$

5) Data and Control Latency: For mobile networks, it is very important to meet the latency performance requirements for both data and control planes, the next two constraints ensure that a selected function chain using a data center $c \in C$ for a demand $d \in D$ satisfies the upper bound for allowed data and control latency.

$$\sum_{f^d \in F^d} \delta^d(c, f^d, d) l^d(c, f^d, d) \leq L^d \quad \forall d \in D, c \in C \tag{7}$$

$$\sum_{f^c \in F^c} \delta^c(c, f^c, d) l^c(c, f^c, d) \leq L^c \quad \forall d \in D, c \in C \tag{8}$$

Network Load Cost Objective: The model's objective is to minimize the network load cost which is defined by the product of carried traffic and the function chain latency. The network load is the sum of the load of both data and control function chains for all demands $d \in D$.

$$C_{net} = \min \sum_{c \in C} \sum_{f^d \in F^d} \sum_{d \in D} \delta^d(c, f^d, d) n^d(c, f^d, d) + \sum_{c \in C} \sum_{f^c \in F^c} \sum_{d \in D} \delta^c(c, f^c, d) n^c(c, f^c, d) \tag{9}$$

Solving this objective results in finding the optimal locations of K data centers. It also finds the optimal functions chains for each demand, i.e., either SDN or NFV, based on the selected data center locations in addition to optimally assign the function chains to the data centers such that the resulting total network load, i.e., data and control traffic, is minimized.

TABLE III
VARIABLES

Notation	Description
$\delta(c)$	binary variable =1 if DC is located at $c \in C$, 0 otherwise
$\delta^d(c, f^d, d)$	binary variable = 1 if data plane of demand $d \in D$ is selected as a function chain $f^d \in F^d$, either SDN or NFV, using DC $c \in C$, 0 otherwise
$\delta^c(c, f^c, d)$	binary variable = 1 if control plane of demand $d \in D$ is selected as a function chain $f^c \in F^c$, either SDN or NFV, using DC $c \in C$, 0 otherwise
$\sigma^d(c)$	integer variable denoting number of servers required for data plane function chains at DC $c \in C$
$\sigma^c(c)$	integer variable denoting number of servers required for control plane function chains at DC $c \in C$
$\mu(c)$	integer variable denoting the total number of servers required for both data and control planes at DC $c \in C$

C. Data Center Resources Cost Optimization Model

This model aims at optimizing the data center infrastructure cost needed to operate a core network given a set of demands and latency requirements. This model reflects the dimensioning of the data centers independently from the network cost, e.g., in case an operator does not control or does not have access to the inter-data center network. As an initial assumption, we only consider the infrastructure cost as the servers cost. The number of servers is proportional to the number of

computational resources, i.e., CPU cores, that are needed for the NFV functions chains, i.e., virtual gateways, or SDN function chains, i.e., controllers. For NFV function chains, the CPU resources needed are computed as the requested bandwidth by a demand multiplied by the number of cores required by a virtual gateway per unit demand $r^d(c, f^d, d) = r(d) \cdot p^d_{vnf}$ while the control plane resources $r^c(c, f^c, d) = \alpha(d) \cdot r(d) \cdot p^c_{vnf}$. As for the SDN function chains, the number of cores needed for the SDN controllers are $r^c(c, f^c, d) = \alpha(d) \cdot r(d) \cdot p^c_{ctr}$, while there are no resources needed for the data plane, i.e., $r^d(c, f^d, d) = 0$. This model additionally aims at balancing the resources among the data centers, in case the number of data centers $K > 1$, by minimizing the largest data center, i.e., the maximum number of servers allocated at a single data center location. This model uses all previous defined constraints, i.e., eqs (1)-(8). It requires additional constraints for the data centers as follows:

1) DC Number of Servers: The number of servers, for data and control planes, at each data center $c \in C$ is calculated by adding the resources $r^d(c, f^d, d)$ or $r^c(c, f^c, d)$ used by function chains of all demands that use this data center. This gives the total number of CPU cores required at this data center, which is divided by the number of cores per server, what we call the server consolidation factor $\frac{1}{s_{cores}}$.

$$\sum_{d \in D} \sum_{f^d \in F^d} \frac{1}{s_{cores}} \Big(\delta^d(c, f^d, d) r^d(c, f^d, d) \Big) \leq \sigma^d(c) \quad \forall c \in C \tag{10}$$

$$\sum_{d \in D} \sum_{f^c \in F^c} \frac{1}{s_{cores}} \Big(\delta^c(c, f^c, d) r^c(c, f^c, d) \Big) \leq \sigma^c(c) \quad \forall c \in C \tag{11}$$

2) Largest DC: The integer variable $\mu(c)$ represents the largest data center, in terms of number of servers, which is lower bounded according to this constraint by the data center $c \in C$ that has the maximum number of allocated servers.

$$\sigma^d(c) + \sigma^c(c) \leq \mu(c) \quad \forall c \in C \tag{12}$$

DC Resources Cost Objective: This model's objective is to minimize the data center resources cost in terms of the total number of servers required at the deployed data centers. Additionally, it aims at minimizing the maximum number of servers allocated at a single data center location in order to achieve a balanced resource distribution.

$$C_{dc} = \min \sum_{c \in C} \Big(\sigma^d(c) + \sigma^c(c) \Big) + \mu(c) \tag{13}$$

Solving this objective results in finding the optimal locations of K data centers. It also finds the optimal functions chains for each demand, i.e., either SDN or NFV, based on the selected data center locations in addition to optimally assign the function chains to the data centers such that the resulting total data center resources, i.e., number of required servers infrastructure, is minimized.

D. Multi-Objective Pareto Optimal Model

This model results in Pareto optimal solutions between the network load cost and data center resource cost objectives to enable operators to choose the right balance between the two objectives. The multi-objective optimization model includes all constraints from the previous two models, i.e., (1)-(8) and (10)-(12). The multi-objective function incorporates both cost functions of the previous two models, where ω_{net} denotes the weight factor for the network load cost objective, while ω_{dc} defines the weight for the data center cost objective. The multi-objective cost function is formally defined as follows:

$$C_{multi} = \min \quad \omega_{net} C_{net} + \omega_{dc} C_{dc} \tag{14}$$

In order to get Pareto solutions for the multi-objective problem, i.e., trade-offs between the optimality of the two objectives, the weights ω can be defined as λ divided by a normalization factor. The parameter λ is a variable that goes from 0 to 1, in order to iterate from the optimality of one objective to the other. Since the two objectives, namely network load cost and data center cost, represent different network metrics and have different units, the normalization factor is used to normalize the two objectives such that they both have the same units and thus contribute similarly to the multi-objective function. In optimization literature, this method is called the weighted sum method for Pareto optimal multi-objective optimization [36]. The details of the proposed model are represented in Algorithm 1.

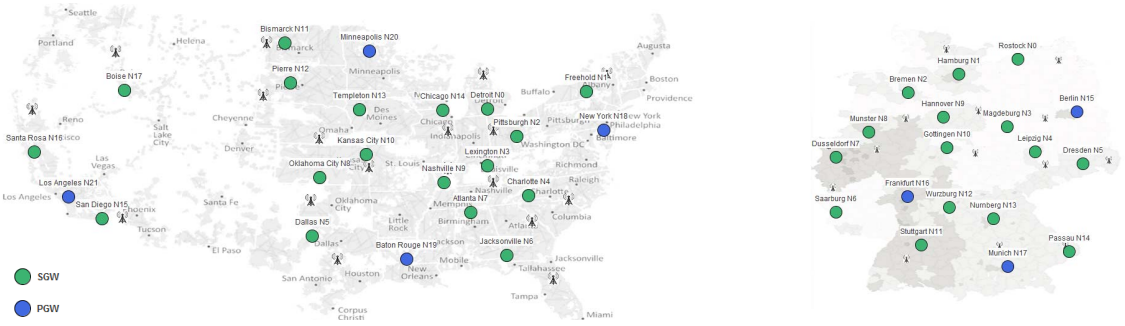
Algorithm 1 Multi-Objective Pareto Optimal Optimization With Pre-Selection Feature for Data Center Locations

Input: no. of DCs K, DC locations $C \subseteq V$,
data and control latency requirements L^d, L^c

1: $min\, C_{net}$, $out\, C_{dc}$, $loc_{net} \leftarrow$ min. network cost C_{net}
2: $out\, C_{net}$, $min\, C_{dc}$, $loc_{dc} \leftarrow$ min. data center cost C_{dc}
3: **(locations pre-selection feature Section IV-E)**
$\{C \leftarrow (loc_{net}, loc_{dc}) || C| = K\}$
4: **for** $\lambda_i = 0 : 0.1 : 1$ **do**
5: $\omega_{net,i} \leftarrow \lambda_i / (out\, C_{net} - min\, C_{net})$
6: $\omega_{dc,i} \leftarrow (1 - \lambda_i) / (out\, C_{dc} - min\, C_{dc})$
7: minimize $C_{multi,i} = \omega_{net,i} C_{net} + \omega_{dc,i} C_{dc}$
8: $C_{net,i} \leftarrow$ post calculation from $C_{multi,i}$
9: $C_{dc,i} \leftarrow$ post calculation from $C_{multi,i}$
10: **end for**

Output: Pareto optimal solutions $(C_{net,i}, C_{dc,i}) \quad \forall \lambda_i = [0, 1]$

In order to get the normalization factors, the single objectives are solved first given a number of data centers K and data as well as control plane latency requirements. Each objective results in an optimal solution for its target and results in an out-turn value for the other target. For instance solving for the network load objective, it results in the optimal network load cost $min\, C_{net}$ and we could calculate the resulting out-turn data center cost $out\, C_{dc}$. Similarly, we solve the data center cost objective and obtain the optimal data centers cost $min\, C_{dc}$ as well the resulting out-turn network cost $out\, C_{net}$. The normalization factor for each objective is defined as the difference between the maximum value for the objective and its optimal solution. The multi-objective function is solved while iterating over λ that ranges from 0 to 1, with a step parameter of 0.1. Each solution from each iteration is unnormalized in order to get the set of Pareto solutions for the network load and data center resources cost, respectively.



(a) Mobile Core topology (USA)

(b) Mobile Core topology (Germany)

Fig. 2. Mobile core topologies considered in the evaluation for both USA and Germany based on the LTE coverage and user population. The figure shows the locations of the SGWs (green) and PGWs (blue). The coverage correlated with the population is depicted by the intensity (grey) on the map background.

E. Data Center Location Pre-Selection Feature for the Multi-Objective Resource Allocation Model

According to previous work and preliminary results, we could observe that each objective function can influence the data center and function chain placement, i.e., locations. Hence, in order to improve the run-time of the multi-objective optimization, we propose a pre-selection for candidate data center locations on the given core network graph from solving the individual objectives, done in steps (1) and (2) of the multi-objective model. The number of pre-selected data center locations is equal to the maximum number of available data centers to be deployed, i.e., size of locations set $|C| = K$.

TABLE IV
EVALUATION PARAMETERS

Parameter	Description
Topology	USA (18 SGWs, 4 PGWs) Germany (15 SGWs, 3 PGWs)
Data Traffic demands	uniform distribution [10 - 50] Gbps
Control and SDN traffic percentage α	uniform distribution [10-30]%
Data plane latency requirement	data plane uni directional 5 ms
Control plane latency requirement	control plane procedure 50 ms
Number of DC locations K	1 - 8 data centers
CPU cores per unit demand (1 Gbps)	p^d_{vnf}=18 cores, p^c_{vnf}=2 cores p^c_{ctr} = 6 cores
Number of cores per server s_{cores}	48 cores per server

V. EVALUATION FOR THE OBJECTIVE TRADE-OFFS

A. Evaluation Setup and Parameters

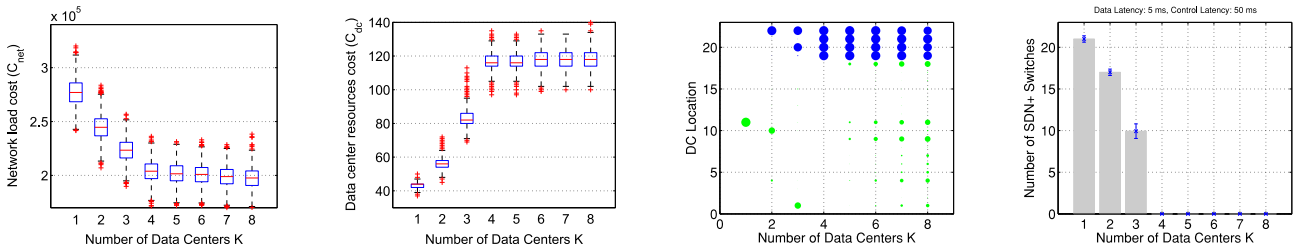
1) Framework: For evaluation, a Java framework has been developed that implements the three proposed optimization models in Section IV. The framework is initialized by reading the graph topology and creating the data plane traffic demands. It also creates the different SDN and NFV function chains, discussed in Sections III-B and III-C, where it computes their associated parameters, i.e., network load, data center resources, data as well as control plane latency. The framework uses Gurobi as the linear optimization solver for the implemented models. Finally, the framework is used to calculate the different parameters and attributes of the solution and forms the resulting SDN and NFV mobile core network.

2) Mobile Core Topologies: For evaluation, we use a mobile core network topology for the USA based on the LTE coverage map in [37], which correlates with the population distribution and considers the locations of Internet Exchange Points (IxP) [38], as illustrated in Fig. 2a. The U.S. topology consists of 18 SGWs and 4 PGWs with a total of 22 nodes, i.e., potential data center locations. For comparison, we use another mobile core network topology for Germany that has 15 SGWs and 3 PGWs with a total of 18 nodes, shown in Fig. 2b. In both topologies, each SGW node is associated to its geographically nearest PGW node, respectively.

3) Data and Control Plane Traffic Demands: In order to evaluate the mobile core network dimensioning cost with respect to the expected traffic increase and the traffic dynamics introduced by SDN and NFV, we consider random traffic requests for each data plane demand. The demands between each SGW and its nearest PGW are uniformly distributed between 10 and 50 Gbps. As for the control as well SDN traffic, we have considered a random control traffic ratio between 10 and 30% of the data traffic demand, which represents conventional LTE control loads and futuristic control loads, e.g., with machine type communication. The traffic assumptions are projections from the predicted data plane and control plane traffic increase in the next generation 5G network [1], [39]. For statistical evidence, the optimization models are solved for multiple runs until a 95% confidence is reached for the optimization solution or at least for 30 runs.

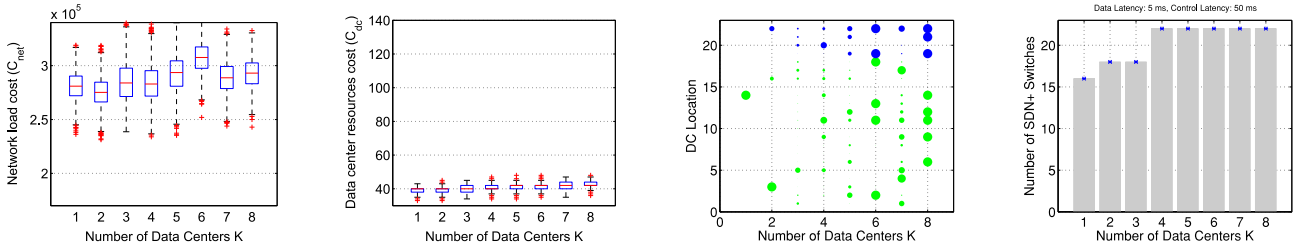
4) Data and Control Plane Latency Requirements: Moving towards the next generation 5G, data and control latency requirements are critical performance metrics that need to be ensured. Hence, we consider the lowest latency that can be achieved by both considered mobile core networks, U.S. and Germany. According to our previous observations and evaluation in [7], we consider a budget of 5 ms for the mobile core network data plane, as a uni directional latency either for uplink or downlink. As for the control latency budget, a 50 ms budget is considered, including SDN control for SDN function chains. The control latency requirement is derived from 3GPP LTE standards [40], [41]. This control latency covers the end-to-end latency to complete the control iterations of the ATTACH procedure as explained in Section III-C.

5) Data Center Resources: It is intuitive to assume that a VNF that handles both data and control planes would need



(a) Network load (optimal) (b) Data center resources (c) Data center location (d) Number of SDN+ switches

Fig. 3. Trade-offs solving for **network load cost** objective for **U.S.** topology, data latency = 5 ms and control latency = 50 ms.



(a) Network load (b) Data center resources (optimal) (c) Data center location (d) Number of SDN+ switches

Fig. 4. Trade-offs solving for **data center resources cost** objective for **U.S.** topology, data latency = 5 ms and control latency = 50 ms.

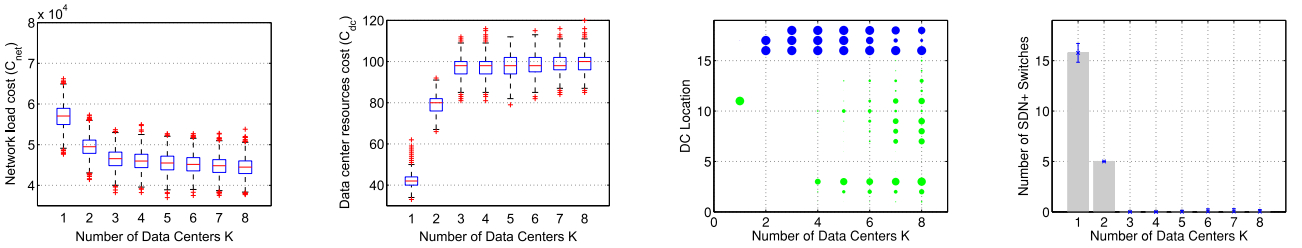
more computational and processing power than an SDN controller that handles the control plane only. Therefore, according to our measurements in [7], we assign 20 cores for the VNF for the processing of 1 unit data traffic demand, i.e., 1 Gbps, with a distribution of 18 cores for data plane p^d_{vnf} and 2 cores for control plane p^c_{vnf}. As for the SDN controller, 6 cores are allocated for the processing of the control plane p^c_{ctr} that corresponds to a unit data plane traffic demand. As for the consolidation factor s_{cores} that defines the number of cores per server, we assumed server sizes of 48 cores that can be typical in current data center deployments [42]. Additional parameters used for the models as well as a summary of the evaluation parameters are presented in Table IV.

B. Trade Offs Between the Network Load and Data Center Resources Cost Objectives

First, we present an evaluation for the trade offs between the two proposed optimization models, i.e., the network load cost objective compared to the data center resources cost objective. We also investigate the impact of the data center deployment by going from a single centralized data center, i.e., $K = 1$, up to a distributed data center deployment with $K = 8$. We start by presenting the results for the U.S. topology considering a data latency requirement of 5 ms and a control latency requirement of 50 ms. The results of optimizing for the network load cost objective are illustrated in Fig. 3, while the results for the data center resources cost objective are illustrated in Fig. 4. For each objective, the results focus on four evaluated criteria which are the network load cost, data center resources cost, data center locations and the number of required SDN+ switches.

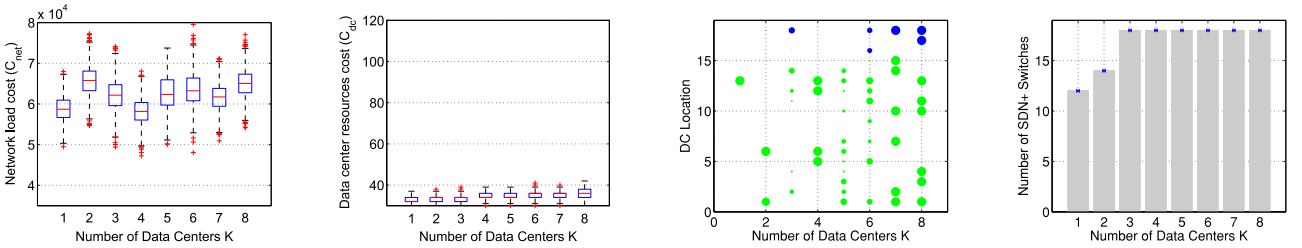
1) Network Load Cost Objective, U.S. Topology: Considering the network load cost objective, Fig. 3a shows that the optimal network load cost is impacted by the data center deployment choice, i.e., the number of data centers. We could observe that the optimal network load cost could be significantly improved by distributing the data center infrastructure, up to 75% at 8 data centers. The reason for this improvement is that, with more available data centers, more VNFs could be deployed under the given latency requirements, refer to Fig. 3d, in order to decrease the additional SDN control traffic and thus decreasing the total network load cost. Additionally, since the network load cost metric considers both the traffic bandwidth and the length of the function chains, deploying distributed data centers can decrease the length of the function chains across the network. Moreover, we can observe that adding more data centers at $K > 4$ does not bring significant improvements to the optimal network load cost.

Considering the resulting data center resources cost, i.e., the number of servers required, as shown in Fig. 3b, a trade off between the optimal network load and the resulting data center resources cost while increasing the number of data centers K can be observed. The resulting number of servers required with 8 distributed data centers is 275% higher than with a single centralized data center. This is again due to the deployment of more VNFs while increasing the number of available data centers, refer to Fig. 3d, which requires more computational CPU cores at the data centers and hence more servers. We can conclude that adding more data centers could optimize and decrease the network load cost further on the expense of needing more servers and increasing the cost for the data centers infrastructure. Throughout the repeated runs of solving the optimization model given random demands, we could observe several trends in the placement of the data centers, i.e., their locations, as shown in Fig. 3c. The frequency of selecting a location for the data centers among the repeated runs is represented by the density of the plotted point, i.e., location, on the figure. The green locations represent the locations of SGWs, while blue locations represent those of PGWs. For instance, at a single data center $K = 1$ and optimizing for the



(a) Network load (optimal) (b) Data center resources (c) Data center location (d) Number of SDN+ switches

Fig. 5. Trade-offs solving for **network load cost** objective for **German** topology, data latency = 5 ms and control latency = 50 ms.



(a) Network load (b) Data center resources (optimal) (c) Data center location (d) Number of SDN+ switches

Fig. 6. Trade-offs solving for **data center resources cost** objective for **German** topology, data latency = 5 ms and control latency = 50 ms.

network load cost, we could observe that there is one dominant location (node 11: Kansas City) that is always selected even with varying random demands. This is due to the geographic centrality of this location, which balances the traffic in the network and optimizes the load cost and could satisfy the data as well as control latency constraints. The other trend that we could observe is that by increasing the number of data centers from $K = [2 - 8]$, the locations of PGWs get more dominant, i.e., they are more frequently selected while varying the input traffic demands. This is because the locations of the PGWs could serve aggregated traffic demands from multiple SGWs, which decreases the distance of transporting the traffic to a different location. Hence, with more than one data center, i.e., distributed deployment, data centers are favored to be placed at the location of PGWs for the network load cost optimization.

Finally, the number of needed SDN+ switches with respect to the number of data centers is illustrated in Fig. 3d. As mentioned before, the network load cost optimization attempts at decreasing the additional control traffic induced by SDN and thus aims at deploying more VNFs. However, according to the data center locations, the data and control latency requirements might not be satisfied for all demands with only VNFs, therefore the need for SDN+ switches. The number of SDN+ switches decreases while increasing the number of data centers K, going from a single centralized data center up to 3 distributed data centers. A network that comprises only of virtual functions is possible starting from 4 data centers.

2) Data Center Resources Cost Objective, U.S. Topology: Here, the same four evaluation metrics as before are used, however, while solving for the optimal data center resources cost, in terms of the total number of servers. The results are shown in Fig. 4. First, we start by discussing the target of this optimization model, i.e., data center resources cost, illustrated in Fig. 4b. We observe that the optimal solutions are less impacted by the available number of data centers. This can be explained by observing the number of SDN+ switches shown in Fig. 4d. Since SDN controllers require less computational cores at the data centers, the model's solution results in almost a full SDN deployment given the data and control latency requirements. This results in decoupling the optimal data center resources cost from their deployment design, i.e., centralized or distributed.

Additionally in fact, the optimal data center resources cost increases slightly while increasing the number of data centers from a centralized $K = 1$ to distributed $K = 8$ deployment. This is due to the possibility of consolidating more cores on servers with centralized data centers which decreases the total number of required servers. With a distributed data center infrastructure, servers are needed at each location without the full utilization of their computational cores. However, the optimal data center resources cost in Fig. 4b is much lower than the resulting data center resources cost while optimizing for the network load cost objective in Fig. 3b, e.g., at $K = 8$, 260% savings in terms of number of servers. As for the resulting network load cost with the data center resources objective, shown in Fig. 4a, we could observe fluctuations in the resulting load cost varying with the number of data centers. This shows the trade off between the network load cost and data center resources cost, where optimizing the data center resources only as an objective results in a quite high network load cost in return. In fact, this points out to the necessity of our third model, i.e., multi-objective Pareto optimization, such that the operator can find Pareto solutions that balance between the network load cost and data center resources cost.

Regarding the locations of the data centers selected through out the repeated runs with varying random traffic demands, Fig. 4c shows that the data center locations are more biased towards the locations of SGWs, i.e., towards the network

edge, while using the data center resources cost objective. We could also observe that the selected locations are more sparse and diverse. These trends are different from what has been observed with the network load objective, refer to Fig. 3c. The data center placement in this case is biased with the control plane latency requirement. Since this model attempts to use more SDN controllers to save on the data center resources, the data centers are placed more towards the edge in order to enable more SDN controllers to satisfy the control plane latency requirement. Finally, Fig. 4d, shows the number of SDN+ switches needed for the data center resources cost objective compared to the number of data centers. We could observe that more SDN+ switches are used in this objective compared to the network load objective. Additionally, the network turns to a full SDN deployment starting at $K = 4$ data centers, which is the same K for the network load objective where the network turns to a full NFV deployment.

C. Trends With Different Topologies

In this section we investigate whether the previously observed trends for the two cost optimization models can also be observed with different topologies. Therefore, we have repeated the previous evaluation for the German topology. Results are illustrated in Fig. 5 and Fig. 6. The results show similar trends for the German topology as the U.S. topology for both network load and data center resources cost objectives. Hence, the repetition of the trends for the evaluated topologies can support our proposed pre-selection of data center locations for the multi-objective optimization model based on the resulting locations from the single objective models. Note that the number of data centers K at which a full NFV deployment, with the network load cost objective, or a full deployment of SDN+ switches, with the data resources cost objective, differs between the two topologies. For the German topology, it is possible starting from $K = 3$ compared to $K = 4$ for the U.S. topology. As previously explained, this is influenced by the number of PGWs that the topology contains, where the German topology contains 3 PGWs, compared to 4 PGWs at the U.S. topology. This can be remarked as a trend observation, where a full deployment, either SDN or NFV depending on the cost objective, is possible starting from K data centers equal to the number of PGWs.

VI. Evaluation for the Multi-Objective Model

A. Gain From Pareto Optimal Multi-Objective Model

First, we investigate the results of the Pareto optimal multi-objective model without data center locations pre-selection and we compare it to the results of the single cost objective models. As explained in Section IV-D, the multi objective method iterates over different weights λ for each objective ranging between [0, 1]. In other words, it explores the solution space starting by solving one single objective, then moving to solve both objectives simultaneously, and stops after solving the other single objective, thus producing the Pareto frontier between the two objectives. For each weight λ, the setup is again repeated with random varying

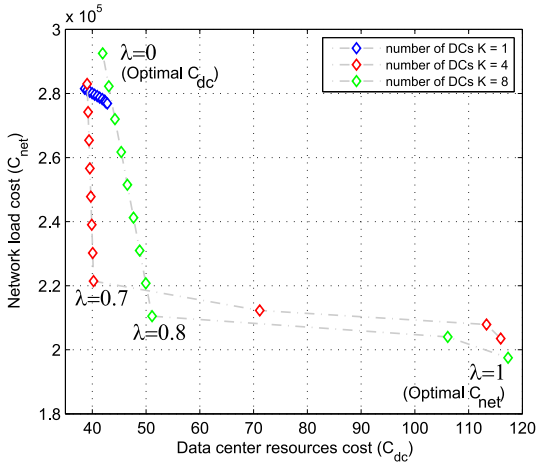


Fig. 7. **Pareto frontier** for the network load cost (C_{net}) and data center resources cost (C_{dc}), solving the Pareto optimal multi-objective model at number of data centers $K = (1, 4, 8)$ for the **U.S.** topology.

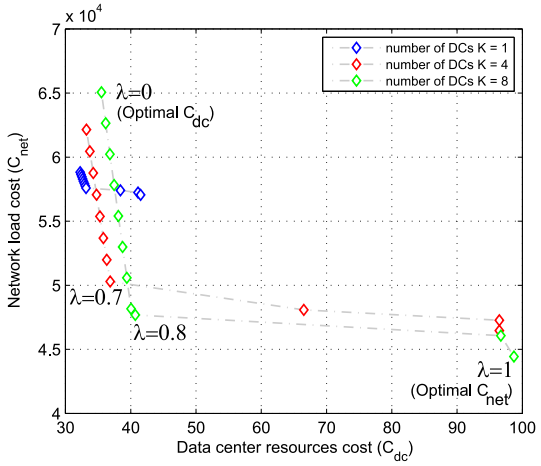
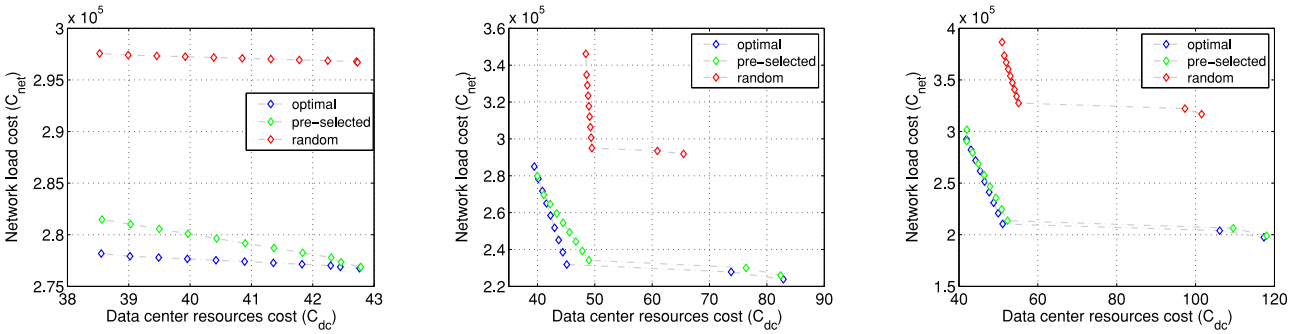


Fig. 8. **Pareto frontier** for the network load cost (C_{net}) and data center resources cost (C_{dc}), solving the Pareto optimal multi-objective model at number of data centers $K = (1, 4, 8)$ for the **German** topology.

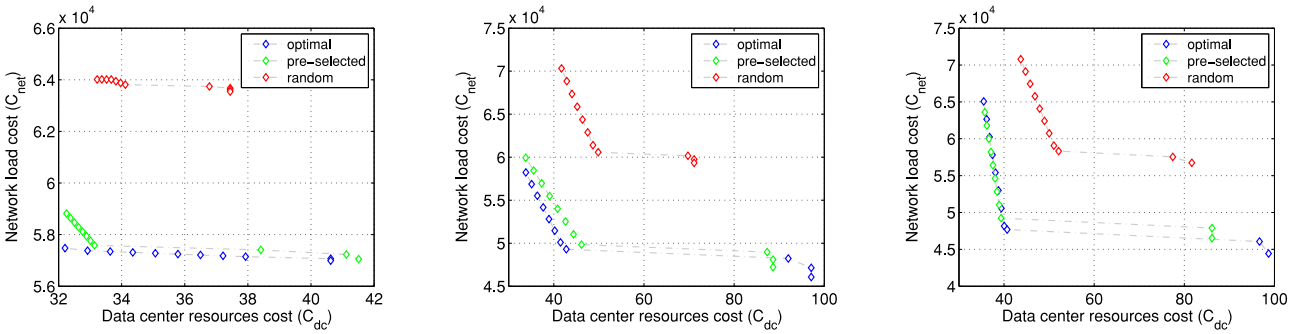
traffic demands till a 95% confidence is reached or at least with 30 runs. Fig. 7 illustrates the Pareto frontier between the network load cost and the data center resources cost, for the U.S. topology and given a data latency requirement of 5 ms and a control latency requirement of 50 ms. The evaluation is assessed for a number of data centers $K = (1, 4, 8)$. We could observe that for a single centralized data center $K = 1$, there is not enough degree of freedom to explore the solution space and provide a balance or trade-off between the network load cost and data center resources cost. This is because with a centralized data center, the locations that satisfy both the data as well control plane latency requirements are quite limited.

Considering a distributed data center infrastructure with $K = 4$, more Pareto solutions offering trade-offs between the two objectives can be observed. For instance the Pareto solution at $\lambda = 0.7$, the network load cost has only an overhead of 3% compared to its optimal solution at $\lambda = 1$, while the data center resources cost results in an overhead of 4% compared to its optimal solution at $\lambda = 0$. Considering more



(a) Number of data centers $K = 1$ (b) Number of data centers $K = 4$ (c) Number of data centers $K = 8$

Fig. 9. **Pareto frontier** for the network load cost (C_{net}) and data center resources cost (C_{dc}) for the **U.S.** topology, comparing the solutions of the optimal multi-objective model with data center locations **pre-selection** and with **random** data center locations.



(a) Number of data centers $K = 1$ (b) Number of data centers $K = 4$ (c) Number of data centers $K = 8$

Fig. 10. **Pareto frontier** for the network load cost (C_{net}) and data center resources cost (C_{dc}) for the **German** topology, comparing the solutions of the optimal multi-objective model with data center locations **pre-selection** and with **random** data center locations.

distributed data centers at $K = 8$, we could observe that there could be more degree of freedom to cover a larger solution space. For instance, considering the Pareto solution at $\lambda = 0.8$, the network load cost has an overhead of 5% compared to its optimal solution, while the data center resources witness an increase of 21% compared to its optimal solution. It is worth mentioning that an operator could go for a different Pareto solution depending on the cost values for each of the network traffic load and the data center resources. In general, the Pareto frontier shows the advantage of finding solutions that could not be easily found through arbitrary weights to each objective in the multi-objective function. This provides operators with the possibility to find the optimal network that balances between the network load cost and data center resources cost.

The evaluation for the Pareto optimal multi-objective model for the German topology is shown in Fig. 8. We demonstrate the Pareto frontier evaluation for the number of data centers $K = (1, 4, 8)$. Similar trends for the Pareto frontiers could be observed as in the U.S. topology. However, more Pareto optimal solutions could be obtained with a centralized single data center at $K = 1$. Since the German topology is geographically smaller than the U.S., this provides more locations to the single data center that could satisfy the data and control latency requirements, thus, find more Pareto solutions for the network load cost and data center resources cost. With a distributed data center at $K = 4$, the Pareto solution with $\lambda = 0.7$ provides an overhead of 6% to the optimal network load cost at $\lambda = 1$ and an overhead of 11% compared to the optimal solution for the data center resources cost at $\lambda = 0$. Considering more distributed data centers with $K = 8$, the Pareto solution at $\lambda = 0.8$ offers a trade off of 11% increase in the optimal network load cost while a 14% increase in terms of the data center resources cost.

B. Gain From Data Center Locations Pre-Selection for the Multi-Objective Optimization Model

We discuss the evaluation of our proposal of data center locations pre-selection for the multi-objective model as explained in Section IV-E. The pre-selected data center locations are a combination of the resulting locations from the solutions of the single objective models, i.e., network load cost model and data center resources cost model. Let us consider an example with a number of data centers $K = 4$. The solution of the network load cost model, with 4 data centers, gives 4 optimal data center locations that minimize the network load cost. Similarly, 4 optimal data center locations are given by solving the data center resources cost model with 4 data centers. Two data center locations are selected arbitrarily from the given solutions of each single objective, respectively. The pre-selected data center locations form the input set to the multi-objective optimization model. Note that in case of a centralized data center $K = 1$, an arbitrary location among the two resulting data center locations from the solution of the two single objectives is pre-selected. The evaluation

focuses on the solution optimality and how much is it impacted by the pre-selection, since the size of the input data center locations set would be $|C| = K$ instead of $|C| = |V|$, i.e., all graph nodes. The evaluation also focuses on how much does the pre-selection improve the run time of the multi-objective model.

1) Optimality Gap With Pre-Selection: Fig. 9 and Fig. 10 illustrate the Pareto frontier evaluation for the optimal multi-objective model compared to the multi-objective model with pre-selection, for the U.S. and German topology, respectively. We also evaluate our proposed pre-selection, based on the solutions of the single objectives, to a random pre-selection. The random pre-selection represents the case where an operator already has fixed locations for the data centers and is solving the multi-objective model for the given locations. The optimality gap is the difference between the three evaluation cases at each Pareto solution. We evaluate the optimality gap at number of data centers $K = (1, 4, 8)$ in order to investigate the impact of centralizing or distributing the data center infrastructure.

For both topologies, we could observe that the proposed pre-selection results in Pareto optimal solutions with a minimal gap compared to the optimal solutions for the evaluated number of data centers $K = (1, 4, 8)$. For instance, at a number of data centers $K = 4$ for the U.S. topology, shown in Fig. 9b, the maximum gap for a Pareto solution with pre-selection is 2% in terms of the network load cost and 6% in terms of data center resources cost. This means that the pre-selection, based on the knowledge from the selected locations of the single objectives, can be used to reduce the problem's complexity while achieving a minimal optimality gap. We could also observe that the optimality gap with pre-selection decreases while adding more data centers, i.e., moving from a centralized to a distributed data center infrastructure. On the other hand, there is a significant optimality gap with the random pre-selection, i.e., given by the operator, compared to the optimal solutions. This observation holds for both topologies as well as for all used number of data centers $K = (1, 4, 8)$. This shows the impact of the data center locations on the resulting optimal cost. Additionally, it supports the importance of the joint placement of the data center infrastructure while solving the placement of the network functions chains.

2) Run Time Improvement With Pre-Selection: Fig. 11 and Fig. 12 illustrate the average run time for the optimal multi-objective model compared to the multi-objective model with pre-selection, for the U.S. and German topology, respectively. The run time is also evaluated for the multi-objective model with random pre-selection. For the U.S. topology, we could observe that the pre-selection could significantly improve the average run time of the multi-objective model, e.g., at $K = 3$, from the order of several seconds to the order of tens of milliseconds. For the German topology, it could improve the run time from the order of hundreds of milliseconds to tens of milliseconds as well. The proposed pre-selection for the data center locations enables operators to use the multi-objective model for online cost optimization, while keeping a minimum gap to the optimal cost. The pre-selection also allows the multi-objective model to scale further for bigger core topology instances or more traffic demand sets.

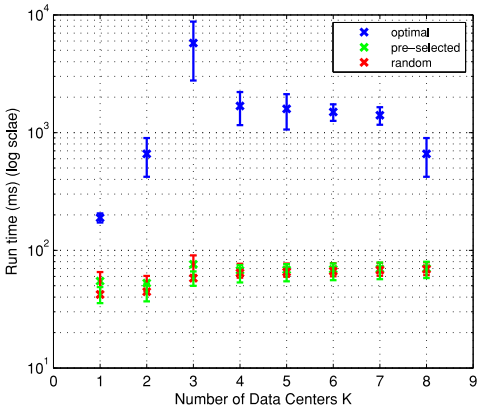


Fig. 11. **Run time** for the multi-objective model for the **U.S.** topology.

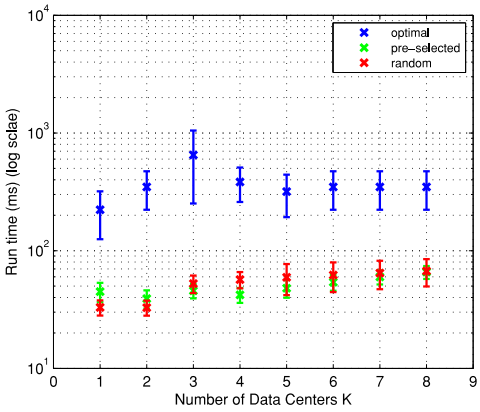


Fig. 12. **Run time** for the multi-objective model for the **German** topology.

VII. Conclusion

In this work, we propose three optimization models that aim at finding the optimal dimensioning and planning for a mobile core network based on SDN and NFV, in terms of network load cost and data center resources cost. The proposed models result in the optimal placement of data centers and the optimal mobile core network split between SDN and NFV. An extensive evaluation has been presented comparing the proposed models in terms of the network load cost and the data center resources cost. Trade-offs between the single objective models could be observed, in terms of the cost factors as well as data center locations. The multi-objective model results in Pareto optimal solutions where a balance between the two cost factors can be achieved. Additionally, solving the multi-objective model with the proposed data center locations pre-selection has shown a significant improvement to the run time while keeping a minimal gap compared to the optimal Pareto solutions. For future work, additional cost factors can be considered for the optimization models such as the cost of the SDN+ switches or the inter-data center links. The set of data centers locations could be extended to arbitrary locations on the core network topology, i.e., not the same locations as the graph nodes. Furthermore, the challenges of the joint

co-existence of SDN and NFV mobile core functions need to be investigated, e.g., orchestration and state distribution. Additionally, a heterogeneous access network can be modeled to represent more realistic use-cases for operators.

REFERENCES

[1] Nokia Bell Labs. (2015). *Nokia Bell Labs Technology Vision 2020*. [Online]. Available: https://networks.nokia.com/innovation/technology-vision

[2] Next Generation Mobile Networks Alliance. (Feb. 2015). *NGMN 5G Initiative White Paper*. [Online]. Available: https://www.ngmn.org/uploads/media/NGMN-5G-White-Paper-V1-0.pdf

[3] *Final Report on Architecturer (Deliverable D6.4)*, Mobile Wireless Commun. Enablers Twenty-Twenty Inf. Soc., Boston, MA, USA, Feb. 2015. [Online]. Available: https://www.metis2020.com/wp-content/uploads/deliverables/METIS-D6.4-v2.pdf

[4] *Network Functions Virtualisation (NFV); Management and Orchestration*, ETSI, Sophia Antipolis, France, 2014. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_NFV-MAN001v010101p.pdf

[5] *SDN Architecture*, Open Netw. Found., Palo Alto, CA, USA, 2014. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR_SDN_ARCH_1.0_06062014.pdf

[6] *OpenFlow Switch Specifications 1.5.1*, Open Netw. Found., Palo Alto, CA, USA, Mar. 2015. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-switch-v1.5.1.pdf

[7] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying NFV and SDN to LTE mobile core gateways, the functions placement problem," in *Proc. 4th Workshop All Things Cellular Oper. Appl. Challenges*, Chicago, IL, USA, 2014, pp. 33–38.

[8] X. Jin, L. E. Li, L. Vanbever, and J. Rexford, "SoftCell: Scalable and flexible cellular core network architecture," in *Proc. 9th ACM Conf. Emerg. Netw. Exp. Technol.*, Santa Barbara, CA, USA, 2013, pp. 163–174.

[9] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 44–53, Jul. 2013.

[10] M. R. Sama *et al.*, "Software-defined control of the virtualized mobile packet core," *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 107–115, Feb. 2015.

[11] M. Moradi, W. Wu, L. E. Li, and Z. M. Mao, "SoftMoW: Recursive and reconfigurable cellular wan architecture," in *Proc. 10th ACM Int. Conf. Emerg. Netw. Exp. Technol.*, Sydney, NSW, Australia, 2014, pp. 377–390.

[12] G. Hampel, M. Steiner, and T. Bu, "Applying software-defined networking to the telecom domain," in *Proc. INFOCOM WKSHPS*, Turin, Italy, 2013, pp. 133–138.

[13] J. Kempf, B. Johansson, S. Pettersson, H. Lüning, and T. Nilsson, "Moving the mobile evolved packet core to the cloud," in *Proc. IEEE 8th Int. Conf. Wireless Mobile Comput. Netw. Commun. (WiMob)*, Barcelona, Spain, 2012, pp. 784–791.

[14] P. Gurusanthosh, A. Rostami, and R. Manivasakan, "SDMA: A semi-distributed mobility anchoring in LTE networks," in *Proc. Int. Conf. Selected Topics Mobile Wireless Netw. (MoWNeT)*, Montreal, QC, Canada, 2013, pp. 133–139.

[15] T. Mahmoodi and S. Seetharaman, "Traffic jam: Handling the increasing volume of mobile data traffic," *IEEE Veh. Technol. Mag.*, vol. 9, no. 3, pp. 56–62, Sep. 2014.

[16] H. Lindholm, L. Osmani, H. Flinck, S. Tarkoma, and A. Rao, "State space analysis to refactor the mobile core," in *Proc. 5th Workshop All Things Cell. Oper. Appl. Challenges*, 2015, pp. 31–36.

[17] V.-G. Nguyen, T.-X. Do, and Y. Kim, "SDN and virtualization-based LTE mobile network architectures: A comprehensive survey," *Wireless Pers. Commun.*, vol. 86, no. 3, pp. 1401–1438, 2016.

[18] H. Baba, M. Matsumoto, and K. Noritake, "Lightweight virtualized evolved packet core architecture for future mobile communication," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, USA, 2015, pp. 1811–1816.

[19] W. Kiess, X. An, and S. Beker, "Software-as-a-service for the virtualization of mobile network gateways," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, 2015, pp. 1–6.

[20] K. Wang *et al.*, "MobiScud: A fast moving personal cloud in the mobile network," in *Proc. 5th Workshop All Things Cell. Oper. Appl. Challenges*, 2015, pp. 19–24.

[21] A. Banerjee *et al.*, "MOCA: A lightweight mobile cloud offloading architecture," in *Proc. 8th ACM Int. Workshop Mob. Evol. Internet Archit.*, Miami, FL, USA, 2013, pp. 11–16.

[22] J. Cho *et al.*, "SMORE: Software-defined networking mobile offloading architecture," in *Proc. 4th Workshop All Things Cell. Oper. Appl. Challenges*, 2014, pp. 21–26.

[23] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, and E.-D. Schmidt, "A virtual SDN-enabled LTE EPC architecture: A case study for S-/P-gateways functions," in *Proc. IEEE SDN Future Netw. Services (SDN4FNS)*, Trento, Italy, 2013, pp. 1–7.

[24] B. Heller, R. Sherwood, and N. McKeown, "The controller placement problem," in *Proc. ACM HotSDN*, 2012, pp. 7–12.

[25] S. Lange *et al.*, "Heuristic approaches to the controller placement problem in large scale SDN networks," *IEEE Trans. Netw. Service Manag.*, vol. 12, no. 1, pp. 4–17, Mar. 2015.

[26] A. Sallahi and M. St-Hilaire, "Optimal model for the controller placement problem in software defined networks," *IEEE Commun. Lett.*, vol. 19, no. 1, pp. 30–33, Jan. 2015.

[27] M. Obadia, M. Bouet, J.-L. Rougier, and L. Iannone, "A greedy approach for minimizing SDN control overhead," in *Proc. 1st IEEE Conf. Netw. Softwarization (NetSoft)*, London, U.K., 2015, pp. 1–5.

[28] S. Gebert *et al.*, "Demonstrating the optimal placement of virtualized cellular network functions in case of large crowd events," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 359–360, 2014.

[29] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manag. (IM)*, Ottawa, ON, Canada, 2015, pp. 98–106.

[30] S. Sahhaf, W. Tavernier, D. Colle, and M. Pickavet, "Network service chaining with efficient network function mapping based on service decompositions," in *Proc. 1st IEEE Conf. Netw. Softwarization (NetSoft)*, London, U.K., 2015, pp. 1–5.

[31] R. Shi *et al.*, "MDP and machine learning-based cost-optimization of dynamic resource allocation for network function virtualization," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, New York, NY, USA, 2015, pp. 65–73.

[32] K. Suksomboon, M. Fukushima, M. Hayashi, R. Chawuthai, and H. Takeda, "LawNFO: A decision framework for optimal location-aware network function outsourcing," in *Proc. 1st IEEE Conf. Netw. Softwarization (NetSoft)*, London, U.K., 2015, pp. 1–9.

[33] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization," in *Proc. 1st IEEE Conf. Netw. Softwarization (NetSoft)*, London, U.K., 2015, pp. 1–9.

[34] "LTE; general packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access)," 3GPP, Sophia Antipolis, France, Tech. Rep. 23.401, 2011. [Online]. Available: http://www.etsi.org/deliver/etsi_ts/124399_123401/ts_123401v130500p.pdf

[35] (Mar. 2017). *Intel Data Plane Development Kit*. Accessed on Mar. 15, 2017. [Online]. Available: http://dpdk.org/

[36] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: New insights," in *Structural and Multidisciplinary Optimization*, vol. 41. Heidelberg, Germany: Springer, 2010, pp. 853–862.

[37] *LTE Coverage Map*, OpenSignal, London, U.K., Sep. 2016, accessed on Mar. 15, 2017. [Online]. Available: http://opensignal.com/

[38] *Data Center IxPs*, IxPs, Princeton, NJ, USA, accessed on Mar. 15, 2017. [Online]. Available: //www.datacentermap.com/ixps.html

[39] (2015). *5G Use Cases and Requirements*. [Online]. Available: https://resources.alcatel-lucent.com/asset/200010

[40] "Requirements for evolved UTRA (E-UTRA) and evolved UTRAN (E-UTRAN)," 3GPP, Sophia Antipolis, France, Tech. Rep. 25.913, 2010. [Online]. Available: http://www.etsi.org/deliver/etsi_tr/125900_125999/tr_125913v090000p.pdf

[41] "Requirements for further advancements for evolved universal," 3GPP, Sophia Antipolis, France, Tech. Rep. 36.913, 2010. [Online]. Available: http://www.etsi.org/deliver/etsi_tr/136900/tr_136913v090000p.pdf

[42] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: Research problems in data center networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 68–73, 2008.

Arsany Basta received the M.Sc. degree in communication engineering and the Dr.-Ing. (Ph.D.) degree (with distinction) from Technische Universität München in 2012 and 2017, respectively, where he joined the Chair of Communication Networks in 2012 as a Research and Teaching Staff Member. His current research focuses on the application of software defined networking, network virtualization, and network functions virtualization to the mobile core toward the next generation network.

Andreas Blenk received the Diploma degree in computer science from the University of Würzburg, Germany, in 2012. He is currently pursuing the Ph.D. degree with Technische Universität München, where he joined the Chair of Communication Networks, in 2012, and is working as a Research and Teaching Associate. and a member of the Software Defined Networking and Network Virtualization Research Group. His research is focused on service-aware network virtualization, virtualizing software defined networks, as well as resource management and embedding algorithms for virtualized networks.

Klaus Hoffmann received the Diploma degree in physics from the University of Kaiserslautern, Germany. He has been engaged as a developer and standardization expert in the telecommunication industry covering circuit switched networks and the 3GPP IMS for over 20 years. He is active in the Research Department, Nokia Bell Labs, in the area of transport networks, EPC, SDN, virtualization, and 5G.

Hans Jochen Morper received the Diploma degree in physics from the University of Würzburg. He has been working for over 25 years in the telecommunication industry mainly in the area of systems engineering and research. He is currently focusing on network virtualization.

Marco Hoffmann received the Dr.Rer.Nat. degree in computer science from Technische Universität München in 2005. In 2004, he joined the Research and Development Department, Siemens. He is currently the Technology Manager for softwarization and cloudification and a Project Manager of international projects with Nokia Bell Labs. He was a Consortium Leader and a Board Member of several national and international research projects.

Wolfgang Kellerer (M'96–SM'11) received the Dipl.-Ing. (Master) and Dr.-Ing. (Ph.D.) degrees from the Technical University of Munich (TUM), in 1995 and 2002, respectively. For over ten years, he was with NTT DOCOMO's European Research Laboratories. He is a Full Professor with TUM, heading the Chair of Communication Networks with the Department of Electrical and Computer Engineering. His research resulted in over 200 publications and 35 granted patents. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and on the Editorial Board of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He is a member of ACM and the VDE ITG.