A Coverage-aware Resource Provisioning Method for Network Slicing

Quang-Trung Luu, *Student Member, IEEE,* Sylvaine Kerboeuf, Alexandre Mouradian, and Michel Kieffer, *Senior Member, IEEE*

***Abstract*—Network slicing appears as a key enabler for the future 5G networks. Mobile Network Operators create various slices for Service Providers (SP) to accommodate customized services. As network slices are operated on a common network infrastructure owned by some Infrastructure Provider (InP), sharing the resources across a set of network slices is important for future deployment. Moreover, in many situations, slices have to be deployed over some geographical area: coverage as well as minimum per-user rate constraints have then to be taken into account.**

Usually, the various Service Function Chains (SFCs) belonging to a slice are deployed on a best-effort basis. Nothing ensures that the InP will be able to allocate enough resources to cope with the increasing demands of some SP. This paper takes the InP perspective and proposes a slice resource *provisioning* approach to cope with multiple slice demands in terms of computing, storage, coverage, and rate constraints.

The resource requirements of the various Service Function Chains to be deployed within a slice are aggregated within a graph of Slice Resource Demands (SRD). Coverage and rate constraints are also taken into account in the SRD. Infrastructure nodes and links have then to be provisioned so as to satisfy all types of resource demands. This problem leads to a Mixed Integer Linear Programming formulation. A two-step deployment approach is considered, with several variants, depending on whether the constraints of each slide to be deployed are taken into account sequentially or jointly. Once provisioning has been performed, any slice deployment strategy may be considered on the reduced-size infrastructure graph representing the nodes and links on which resources have been provisioned.

Simulation results demonstrate the effectiveness of the proposed approach compared to a more classical direct slice embedding approach.

***Index Terms*—Network slicing, resource provisioning, coverage constraints, wireless network virtualization, 5G, linear programming.**

I. Introduction

NETWORK Function Virtualization (NFV) is attracting widespread interest due to the overall equipment and management cost reductions it allows [1] and to the increased network flexibility it provides [2]. Using NFV, network functions are decoupled from their hosting hardware and are offered as virtualized services decomposed in *Virtual Network Functions* (VNFs) on general-purpose servers. With cloud networks, infrastructure is also evolving to integrate edge and central data centers onto which VNFs may be deployed by using IT technologies. With the help of virtualization, many dedicated end-to-end network services can co-exist and share the same physical infrastructure, while relying on different network capabilities, protocols and network architecture optimized towards customized requirements. This paves the way to the network slicing concept which has emerged as a new paradigm for 5G networks [3, 4, 5]. Slicing can be applied for deploying business cases such as multi-tenants sharing the same network infrastructure, where tenants, *i.e.*, vertical actors, can operate and manage their own network slice to address applications in energy, e-health, smart city, connected cars [4].

A network slice can be seen as a collection of *Service Function Chains* (SFCs) and a set of physical network resources, which are dynamically allocated to build a customized logically isolated virtual network. Each SFC consists of several interconnected VNFs describing the processing applied to a data flow related to a given service. With cloudification technology, SFCs and VNFs can be easily and flexibly initialized, launched, chained, and scaled to meet changeable workload requests [6]. Iterative SFC deployment strategies are well-suited to such dynamic slice management. Nevertheless, when several concurrent slices are managed in parallel, nothing ensures that enough infrastructure resources will be available to deploy a new SFC. Such best-effort slice management makes it difficult to satisfy a *Service Level Agreement* (SLA) expressed by tenants in terms, *e.g.*, of guaranteed amount of serviced users. More flexible models and mechanisms for network service provisioning and deployment are needed [7]. Additionally, research challenges remain when network slicing incorporates the wireless part of legacy or 5G networks [8, 9], where multiple network segments including the radio access, transport, and core network, have to be considered.

This paper studies the way to efficiently provision and deploy end-to-end network slices on radio and cloud network infrastructure in a multi-tenancy context. Our work focuses on the problem of slice resource provisioning, *i.e.*, reservation. By provisioning we ensure that enough resource is reserved for allocation to further SFC deployment. Enough resource has to be provisioned to deploy SFCs in the network while satisfying coverage constraints for mobile end-users of the slice service. A two-step method is proposed for efficient slice deployment: A resource provisioning process is followed by an SFC embedding process involving any state-of-the-art deployment approach, but considering a simplified infras-

Parts of this work have been presented at *IEEE GLOBECOM 2018*.

Q.-T. Luu is with Nokia Bell Labs and the Laboratoire des Signaux et Systèmes (L2S), CNRS-CentraleSupélec-Univ Paris-Sud-Univ Paris-Saclay, France, e-mail: quang_trung.luu@nokia.com.

S. Kerboeuf is with Nokia Bell Labs, France, e-mail: sylvaine.kerboeuf@nokia-bell-labs.com.

A. Mouradian and M. Kieffer are with the Laboratoire des Signaux et Systèmes (L2S), CNRS-CentraleSupélec-Univ Paris-Sud-Univ Paris-Saclay, France, e-mail: {michel.kieffer, alexandre.mouradian}@l2s.centralesupelec.fr.

tructure network reduced to the nodes and links which have provisioned resource. The SFC embedding time may then be much smaller.

The rest of the paper is structured as follows. Section II presents the system architecture, analyzes related work, and highlights our main contributions. The model of the infrastructure network and of the slice resource demands are presented in Section III. The slice resource provisioning problem is then formulated in Section IV as a mixed integer linear programming problem accounting for cloud network and radio resource constraints for the deployment of multiple slices. Numerical results are presented in Section V. Finally, Section VI draws some conclusions and perspectives.

II. System Architecture, Related Work, and Main Contributions

A. System Architecture

Several entities are involved in network slicing, as described in Figure 1 [1]. The *Infrastructure Provider* (InP) owns and manages the wireless and wired infrastructure such as the cell sites, the fronthaul and backhaul networks, and cloud data centers. The *Mobile Network Operator* (MNO) owns the

Fig. 1. System architecture

virtual network functions and leases the physical resources provided by the InPs to setup and manage the slices. Finally, the *tenant* (or *Service Provider,* SP) exploits the slices supplied by the MNO, and provides to his customers the required services that are running within the slices. Service needs are forwarded by the tenant to the MNO within an SLA. The SLA describes, at a high level of abstraction, characteristics of the service with the desired QoS, the number of devices (or the device/user density), the geographical region where the service has to be made available for the end-users, *etc*. The MNO translates the tenant high-level demands into SFCs able to fulfill the service requirements. SFCs are then deployed on the network infrastructure so that QoS requirements are satisfied. To perform this deployment, the InP has to identify the infrastructure nodes on which the VNFs are deployed and the links able to transmit data between these nodes. Given a set of SFC demands, this consists in finding *i*) the placement of the VNFs on the data center nodes and *ii*) the routing of data flows between the VNFs, while respecting the structure of SFCs and optimizing a given objective (*e.g.*, minimizing the infrastructure and software fees cost). In particular, it has also to take into account the coverage constrains for the service and to select the base station nodes accordingly.

In this work, we adopt the *Cloud Radio Access Network* (C-RAN) architecture, a cloud architecture for future mobile network, illustrated in Figure 2. The C-RAN nodes (*i.e.*, eNB for 4G and gNB for 5G) mainly consists of two parts: The distributed *Remote Radio Heads* (RRHs) plus antennas deployed at the cellular radio sites and the centralized *Base Band Unit* (BBU) pool hosted in an edge cloud data center [10]. The BBU pool is hosting multiple virtual BBUs, handling higher layer processing functions, whereas all basic radio functions remain at the cellular radio station with the RRH. In 4G, the BBU handles all the L1-L2-L3 functional layers whereas radio frequency functions reside at the RRH. Within 5G, the gNB is split in three parts, namely *Central Unit* (CU), *Distributed Unit* (DU), and *Radio Unit* (RU), and different functional splits are under study where in some options the RU can support some L2 functions thus reducing the capacity required for the fronthaul link [11]. The link (interface) between the BBU and the RRH is known as the fronthaul whereas the backhaul network connects the BBU with the core network functions hosted in the regional or central cloud.

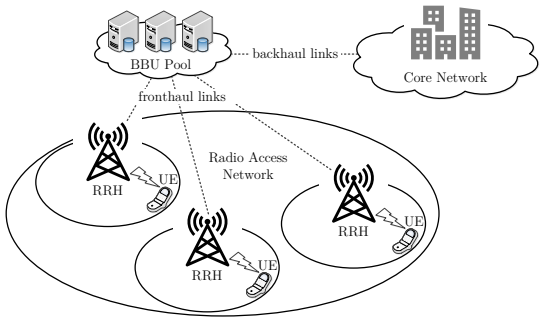


Fig. 2. General architecture of C-RAN

B. Related Work

Early results on assigning infrastructure network resources to virtual network components may be found, *e.g.*, in [12, 13]. Due to its capability of sharing efficiently network resource in 5G networks, the concept of network virtualization has gained renewed attention in the literature [5, 14, 15, 16] via the concept of network slicing.

Network slice resource allocation is a complex problem. When a slice instance is seen as a collection of SFCs, slice embedding needs to deploy the SFCs on a shared infrastructure while satisfying various constraints. Most of prior works related to SFC and VNF deployment do not account for coverage constraints. For example, in [17, 18], computing, storage, and aggregated wireless resource demands of SFCs are considered. The minimization of the SFC embedding cost is formulated either as an *Integer Linear Programming* (ILP) [18, 19, 20] or as a *Mixed Integer Linear Programming* (MILP) problem [13, 21], which are known to be NP-hard [22]. In [23], the VNF placement problem is expressed as an *Integer Quadratic Programming* (IQP) problem with a set of energy consumption constraints, and then is transformed to a solvable linear form.

To address the high computational complexity resulting from the ILPs or MILPs, various heuristics have been proposed, see, *e.g.*, [17, 18, 19]. For example, [17] introduced an heuristic based on the search of shortest paths to sequentially embed the SFCs. In [18], the candidate infrastructure nodes

are sorted to find the best node, in terms of deployment cost, to host a given VNF. Its neighbors are then considered as candidates to deploy the next VNF.

The *Column Generation* (CG) technique has been widely studied to solve large ILP problems [24]. In the CG approach, the original ILP is decomposed into a *Master Problem* (MP) and a *Pricing Problem* (PP). The MP is the original problem where only a subset of variables is considered. The PP is a new problem created to identify a new variable*, i.e.*, a column, to add to the MP to improve the current solution. In [24] or [25], CG has been used to relax ILP-based SFC embedding or reconfiguration problems. Specifically, in [24], the SFC embedding problem is addressed. Only core capacity and bandwidth resources for infrastructure nodes and links are considered. In [25], the embedding of new SFCs and the re-adjustment of in-service SFCs are both considered. Re-adjustment of in-service SFCs may imply the migration of VNFs and virtual links may need to be updated to meet changes of resource demands. This problem is again formulated as an ILP where the objective is to minimize the deployment as well as the migration costs. Only linear SFCs are allowed and any node with radio ressource may serve as access point for the users, which makes difficult the satisfaction of coverage constraints. Moreover, possible paths in the network are assumed to be available, which needs some computational effort before the deployment.

In [26], the join VNF and virtual link placement is formulated as a *Weighted Graph Matching Problem* (WGMP), where the SFC graph and the infrastructure graph are modeled as weighted graphs, on which each node and each link have their own weight corresponding to their required resource (for the SFC graph), or their available resource (for the infrastructure graph). An eigendecomposition-based method is then proposed to solve the WGMP problem, which aim is to find, with a significantly reduced complexity, the optimum matching between the SFC graph and the infrastructure graph. In [24], [25], and [26], a unique type of resource is considered at infrastructure nodes (processing) and at links (bandwidth). Radio resource is not considered.

The resource allocation problem among competing slices in an heterogeneous cloud infrastructure is addressed in [27]. Slice resource demands are aggregated in a vector of VNF resource demands in the slice multiplied by a coefficient linked to the number of services to be processed per time unit. The considered types of resource are CPU, memory, bandwidth, and storage. The resource allocation among multiple slices is performed considering two different approaches. The first approach involves a centralized convex optimization problem, which objective is to maximize the total slice utility. Nevertheless, as pointed out in [27], such centralized lacks of scalability, is not robust to a failure of the central optimizer, and is prone to non-collaborative slice providers which may harm the system. For these reasons, a distributed method based on game theory is considered to improve robustness and scalability. Optimization is performed in a decentralized way among the data centers and slice providers, and the results provided by all entities will provide the final resource allocation for all slices. Nevertheless, the placement of VNFs in data centers is predetermined by the MNO and again, wireless resources are not considered. A resource aggregation scheme similar to that in [27] has been introduced in our previous work [28], where infrastructure resources are provisioned to satisfy slice resource demand constraints. Radio resources are considered, but radio coverage constraints are still ignored.

The design of efficient allocation mechanisms for virtualized radio resources has been recently addressed in [29]. This paper aims at minimizing the leasing cost of BSs so as to meet SP demands, while providing, with a given probability, a minimum data rate for any user located in their coverage area. The rate constraint is expressed as a linear function of the BS load (number of users served by the BS), of the distance from user to the nearest BS, and of the downlink interference. This linear approximation, however, requires some assumptions. For instance, a user of an SP is assumed to be served by its nearest BS among the set of BSs allocated to the SP. This reduces somehow the potentiality of achieving the optimal sharing of the radio resource.

In [30], an heterogeneous spatial user density is considered, and the joint BS selection and adaptive slicing are formulated as a two-stage stochastic optimization problem. The first stage aims at defining the set of BSs to activate. The second stage aims at allocating wireless resources of the BSs to each point of the region to be covered by the SP. Several random realizations of user locations are generated to get a reduced-complexity deterministic optimization problem. A genetic algorithm is then used for the optimization.

In [31], a network slicing framework for *Multi-Tenant Heterogeneous CRAN* (H-CRAN) is introduced. The sharing of radio resources in terms of data rate is considered, with some constraints related to the fronthaul capacity, the transmission power budget of RRHs, or the tolerable interference threshold of an RRH on a sub-channel. Slicing is formulated as a weighted throughput maximization problem, which aims at maximizing the total rate obtained by users connected to given RRHs on given sub-channels. Nevertheless, the proposed framework does not consider computing and memory resource associated to the processing within the BBUs. Such resource is assumed to be properly scaled so as to support the required service rate. Moreover, the proposed framework addresses only downlink data services.

The wireless network slicing problem is also addressed in [32]. A game theory-based distributed algorithm to solve the problem is proposed. The proposed algorithm accounts for the limited availability of wireless resources and considers different aspects such as congestion, deployment costs and the RRH-user distance. This work considered the coverage area of RRH, but ignored the possible coverage constraints required by the slices.

C. Main Contributions

Compared to previous works, we consider slice resource demands in terms of coverage and traffic requirements in the radio access part of the network as well as network, storage, and computing requirements from a cloud infrastructure of interconnected data centers for the rest of the network. This

work borrows the slice resource provisioning approach introduced in [28], and adapts it to the joint radio and network infrastructure resource provisioning. Constraints related to the infrastructure network considered in [17, 18, 27, 28] are combined with coverage and radio resource constraints introduced in [29, 30, 31, 32].

In this work, we assume that the resource requirements for the various SFCs that will have to be deployed within a slice may be aggregated and represented by a Slice Resource Demand (SRD) graph that mimics the graph of SFCs. These SRDs are evaluated by the MNO to satisfy the QoS requirements imposed by the tenant. The InP has then to provision enough infrastructure resources to meet the SLA. Due to the fact that nodes or links of the graph of SRDs represent aggregate requirements, several infrastructure nodes may have to be gathered and parallel physical links have to be considered to satisfy the various SRDs. This is the main difference with respect to the traditional service chain embedding approach considered for example in [17, 18], where each VNF is deployed on a single node. In [17, 18], virtual nodes and links are mapped on the infrastructure network to allocate resources to VNFs and virtual links. In this paper, one provisions a sufficient number of infrastructure nodes and links, so that the aggregated provisioned resources meet the slice demands represented by the graph of SRDs.

When provisioning slices, we consider coverage constraints, in which slices are assumed to cover a specific region in the considered geographical area, that is part of the SLA with the tenant. We devise the special case of the cloud RAN architecture with RRHs which are nodes having radio resources. In our model, radio resource blocks are allocated and the channel between the RRH nodes and users is taken into account. Compared with [29], the selected BS is not necessarily the nearest one. Moreover, both downlink and uplink traffic are considered for the service rate model.

III. System Model

Consider a set of SPs which aim is to provide different services, indexed by $\sigma = 1, \ldots, |\mathcal{S}|$, to mobile users. The geographical area under study is denoted $\mathcal{A}$ and the subarea over which service σ has to be made available is denoted $\mathcal{A}^\sigma$. For that purpose, each SP forwards his service requirements to an MNO, which aim is to design a network slice able to satisfy these requirements. The MNO sends to the InP a Slice Resource Demand (SRD). This SRD consists of (*i*) an SRD graph accounting for the structure and SLA of the slice, and (*ii*) SRD coverage information related to the area $\mathcal{A}^\sigma$ over which the service will have to be made available. The InP is then in charge of provisioning enough infrastructure resources to deploy the SFCs, which resource demands have been described by the SRD graph.

This section details the model of the infrastructure provided by the InP and the way a service with wireless coverage constraints can be mapped to a slice with specific SRD graph.

A. Infrastructure model

Consider an infrastructure network managed by some InPs. This network is represented by a directed graph $\mathcal{G}_\mathrm{I} = (\mathcal{N}_\mathrm{I}, \mathcal{E}_\mathrm{I})$, where $\mathcal{N}_\mathrm{I}$ is the set of infrastructure nodes and $\mathcal{E}_\mathrm{I}$ is the set of infrastructure links, which correspond to the wired connections between nodes and within nodes (loopback links) of the infrastructure network.

Each infrastructure node $i \in \mathcal{N}_\mathrm{I}$ is characterized by a given amount of computing and storage resources, denoted as $a_\mathrm{c}(i)$ and $a_\mathrm{s}(i)$, which may be allocated to network slices. Radio resources are exclusively provided by a subset $\mathcal{N}_\mathrm{Ir} \subset \mathcal{N}_\mathrm{I}$ of RRH nodes, which location in some Cartesian frame attached to $\mathcal{A}$ is denoted x_i^r. The cost associated to the use of an infrastructure node i consists of a fixed part $c_\mathrm{f}(i)$ for node disposal and variable parts $c_\mathrm{c}(i)$, $c_\mathrm{s}(i)$, and $c_\mathrm{r}(i)$, which depend linearly on the amount of computing, storage, and radio resources provided by that node.

Each infrastructure link $ij \in \mathcal{E}_\mathrm{I}$ connecting node i to j has a bandwidth of $a_e(ij)$, and an associated per-unit bandwidth cost $c_e(ij)$. Several distinct VNFs of the same slice may be deployed on a given infrastructure node. When communication between these VNFs is required, an internal (loopback) infrastructure link $ii \in \mathcal{E}_\mathrm{I}$ can be used at each node $i \in \mathcal{N}_\mathrm{I}$, as in [33], in the case of interconnected virtual machines (VMs) deployed on the same host. The associated per-unit bandwidth cost, in that case, is $c_\mathrm{e}(ii)$.

B. SRD Model

An SRD is defined on the basis of an SLA between a tenant and an MNO. In what follows, we consider SLAs involving coverage constraints, related to the geographical distribution of the end-users of the service provided by the slice. This distribution is described by the user density function $\rho^\sigma(x)$, with $x \in \mathcal{A}$. The SLA is also expressed in terms of the supported service type and the targeted QoS such as a minimum average data rate $\underline{R}_\mathrm{u}^\sigma$ and $\underline{R}_\mathrm{d}^\sigma$ for the wireless uplink and downlink traffic for each client.

As an example, Figure 3 illustrates three geographical subareas over which three different services have to be deployed.

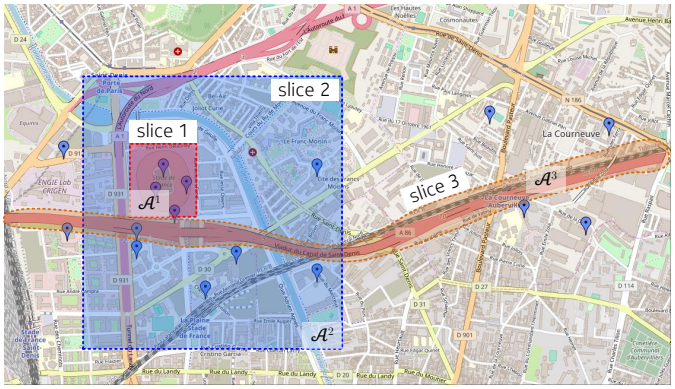


Fig. 3. The considered metropolitan area including the Stade de France (covered by the red rectangle representing $\mathcal{A}^1$), its surrounding (blue rectangle representing $\mathcal{A}^2$), and part of the A86 highway (orange shape representing $\mathcal{A}^3$). Blue markers show the location of RRH nodes of Orange.

A collection of SFCs has to be deployed to support the SLA of the slice. One assumes that the resource requirements for these SFCs can be aggregated and represented by an SRD

graph that mimics the graph of SFCs. The SRD graph for slice σ is an oriented graph denoted $\mathcal{G}_{\mathrm{V}}^{\sigma} = (\mathcal{N}_{\mathrm{V}}^{\sigma}, \mathcal{E}_{\mathrm{V}}^{\sigma})$, where $\mathcal{N}_{\mathrm{V}}^{\sigma}$ and $\mathcal{E}_{\mathrm{V}}^{\sigma}$ are respectively the set of (virtual) SRD nodes and links. The SRD graph has a structure close to the SFC graph, with SRD nodes corresponding to the VNFs of the SFC. Each SRD node $v \in \mathcal{N}_{\mathrm{V}}^{\sigma}$ is characterized by a given amount of *required* computing and storage resources, denoted $r_{\mathrm{c}}(v)$ and $r_{\mathrm{s}}(v)$ to sustain the aggregated demand for all instances of a given VNF in the slice. Each link $vw \in \mathcal{E}_{\mathrm{V}}^{\sigma}$, connecting node v to w in the SRD graph, is characterized by the bandwidth $r_{\mathrm{b}}(vw)$ required to sustain the aggregated traffic demand between the VNFs associated to v and w.

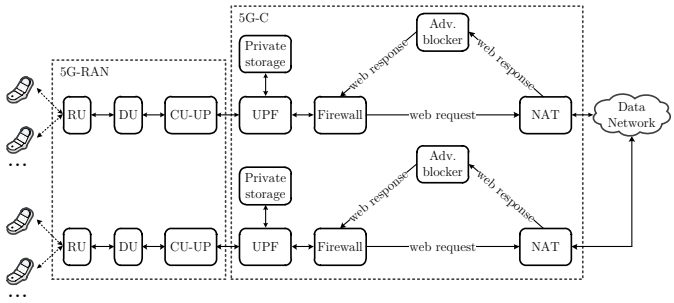
SFCs will be deployed on the infrastructure nodes and links which have provisioned resources. Enough resources should be provisioned by each node to be able to host a minimum number of VNFs. Consequently, when resources are provisioned on a node, a minimum amount $\underline{r}_{\mathrm{c}}(v)$ and $\underline{r}_{\mathrm{s}}(v)$ of such resources should be provisioned.

One also assumes that the uplink and downlink radio resource demands are represented by a single node v_{r} of the SRD graph. The aggregated uplink and downlink data rates $r_{\mathrm{u}}(v_{\mathrm{r}})$ and $r_{\mathrm{d}}(v_{\mathrm{r}})$ are associated to the coverage constraint of slice σ

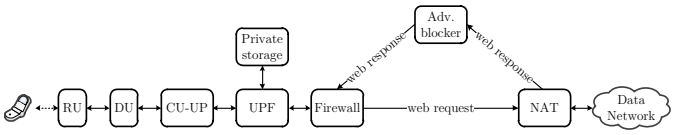
$$r_{\mathrm{u}}(v_{\mathrm{r}}) = \underline{R}_{\mathrm{u}}^{\sigma} \int_{\mathcal{A}^{\sigma}} \rho^{\sigma}(x)\,\mathrm{d}x,$$
$$r_{\mathrm{d}}(v_{\mathrm{r}}) = \underline{R}_{\mathrm{d}}^{\sigma} \int_{\mathcal{A}^{\sigma}} \rho^{\sigma}(x)\,\mathrm{d}x. \quad (1)$$

Figure 4 illustrates the SFCs required for the deployment of a web browsing service with advertisement removal inspired by [34] and its associated SRD graph. Figure 4a describes the eight VNFs to be deployed, including: three RAN VNFs namely a *Radio Unit* (RU) to handle RF operations, a *Data Unit* (DU) and a *Centralized Unit for User-Plane Functions* (CU-UP) to handle computing and processing loads; and five VNFs placed in the core network, namely a *User-Plane Function* (UPF), a private storage management function, a firewall, an advertisement blocker, and a *Network Address Translation* (NAT). Each of these VNFs is characterized by computing and storage requirements. Some links are bidirectional, *e.g.*, between the UPF and the firewall, others are unidirectional, *e.g.*, the uplink traffic from users has not to go through the advertisement blocker. The corresponding SRD graph is represented in Figure 4b. All identical instances of SFCs deployed within the slice are represented by a single graph which structure is identical to the SFC graph. The requirements in terms of storage, computing, and wireless capacity of each component of the SRD graph aggregate the corresponding requirements of the components of the SFC graph. See Section V for more details.

A second example is provided in Figure 5, which represents the SFCs required for the deployment of an adaptive wireless video streaming service and its associated SRD graph taken from [35]. Figure 5a represents the VNFs for the user-plane of the 5G-RAN (RU, DU, CU-UP), the 5G-Core (UPF), and the server and *Video Optimization Controller* (VOC) placed in the data network. The server archives videos with different qualities (bitrate). Using the information received from users such as the bandwidth or end-to-end latency, the VOC dynamically adjusts the video bitrate to provide to the users. Figure 5b describes the associated SRD graph.

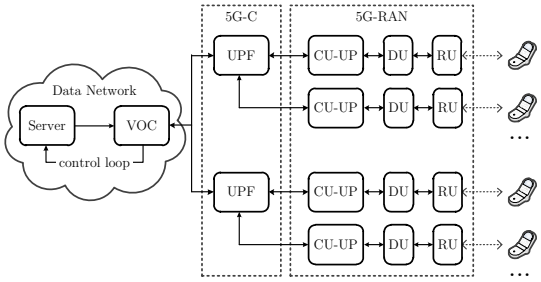


(a) Graph of SFCs.

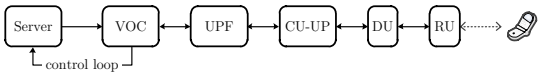


(b) Corresponding SRD graph.

Fig. 4. SFCs required for the deployment of a secured web browsing service with advertisement removal and its associated SRD graph.



(a) Graph of SFCs.



(b) Corresponding SRD graph.

Fig. 5. SFCs required for the deployment of an adaptive wireless video streaming service and its associated SRD graph.

Table I summarizes all parameters involved in the description of the infrastructure network and the graph of SRDs for a slice.

IV. Problem Formulation

Our aim is to provision resources from the infrastructure for the slices managed by an MNO. This provisioning is represented by a mapping between the infrastructure graph $\mathcal{G}_{\mathrm{I}}$ and the SRD graph $\mathcal{G}_{\mathrm{V}}^{\sigma}$, as illustrated in Figure 6. In this example, the slice σ is described by an SRD graph aggregating the demands of several linear SFCs. The constraints which have to be satisfied by this mapping are detailed in the following sections.

A. Accounting for SRD Coverage Constraints

For the slice σ, the InP has to provide a minimum average data rate ($\underline{R}_{\mathrm{u}}^{\sigma}$ for uplink and $\underline{R}_{\mathrm{d}}^{\sigma}$ for downlink) to each

TABLE I
INFRASTRUCTURE NETWORK AND SLICE PARAMETERS.

Resource types: n	
n	Type of node resources: computing (c), storage (s), and radio (r)
Infrastructure network graph: $\mathcal{G}_{\mathbf{I}} = (\mathcal{N}_{\mathbf{I}}, \mathcal{E}_{\mathbf{I}})$	
$\mathcal{N}_{\mathrm{I}}$	Set of infrastructure nodes
$\mathcal{E}_{\mathrm{I}}$	Set of infrastructure links
$a_n(i)$	Available resource of type n at node $i \in \mathcal{N}_{\mathrm{I}}$
$a_{\mathrm{b}}(ij)$	Available bandwidth of link $ij \in \mathcal{E}_{\mathrm{I}}$
$c_n(i)$	Per-unit cost of resource of type n for node $i \in \mathcal{N}_{\mathrm{I}}$
$c_{\mathrm{b}}(ij)$	Per-unit cost for link $ij \in \mathcal{E}_{\mathrm{I}}$
$c_{\mathrm{f}}(i)$	Fixed cost for using node $i \in \mathcal{N}_{\mathrm{I}}$
SRD graph for slice σ: $\mathcal{G}_{\mathrm{V}}^{\sigma} = (\mathcal{N}_{\mathrm{V}}^{\sigma}, \mathcal{E}_{\mathrm{V}}^{\sigma})$	
$\mathcal{N}_{\mathrm{V}}^{\sigma}$	Set of SRD nodes of slice σ
$\mathcal{E}_{\mathrm{v}}^{\sigma}$	Set of SRD links of slice σ
v_{r}	SRD node aggregating uplink and downlink radio resource demand, $v_{\mathrm{r}} \in \mathcal{N}_{\mathrm{V}}^{\sigma}$
$r_n(v)$	Resource demand of type n at node $v \in \mathcal{N}_{\mathrm{V}}^{\sigma}$
$r_{\mathrm{b}}(vw)$	Bandwidth demand at link $vw \in \mathcal{E}_{\mathrm{V}}^{\sigma}$
$\mathcal{A}^{\sigma}$	Coverage area of slice σ
$\mathcal{Q}^{\sigma}$	Set of all divided subareas in $\mathcal{A}^{\sigma}$
q	Subarea index, $q \in \mathcal{Q}^{\sigma}$
$\mathcal{A}_q^{\sigma}$	Subarea q
σ	Slice index
$\mathcal{S}$	Set of all slices σ

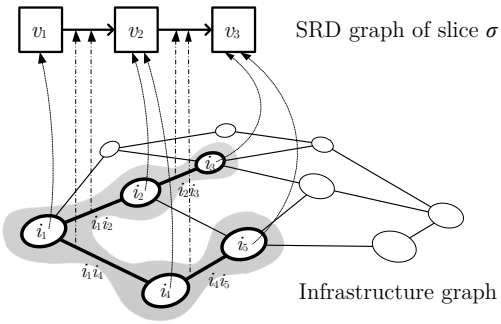


Fig. 6. Provisioning of infrastructure resource to an SRD graph: Resources from the infrastructure node i_1 is provisioned for SRD node v_1; Resources from i_2 and i_4 are provisioned for SRD node v_2; and resources from i_3 and i_5 are provisioned for SRD node v_3. Correspondingly, the infrastructure links $i_1 i_2$ and $i_1 i_4$ are provisioned for SRD link $v_1 v_2$ and resources from links $i_2 i_3$ and $i_4 i_5$ are provisioned for SRD link $v_2 v_3$.

mobile user spread over $\mathcal{A}^{\sigma}$ with a density $\rho^{\sigma}(x)$. For that purpose, the InP will have to provision resources from the RRH nodes in $\mathcal{N}_{\mathrm{Ir}}$. One assumes that every RRH node is able to provide a fixed amount $a_{\mathrm{r}}(i)$ of resource blocks (RB) per time unit to exchange data (up and downlink) with users. The amount of data transmitted using a single RB depends on the characteristics of the RRH, of the User Equipment (UE), and on the transmission channel between the RRH and the user.

During the resource provisioning phase, the locations of users are unknown. To address this problem, [30] considers different realizations of a point process representing the location of users. Here an approach inspired by the subarea partitioning technique introduced in [36] is considered. $\mathcal{A}^{\sigma}$ is partitioned into Q^{σ} convex subareas $\mathcal{A}_q^{\sigma}$, $q \in \mathcal{Q}^{\sigma} = \{1, \ldots, Q^{\sigma}\}$. Instead of allocating RBs to users, RRH nodes allocate RBs to subareas. The way the partitioning is performed is not detailed here. One may consider, *e.g.*, a partitioning into squares of equal surfaces or a partitioning based on ρ^{σ} that provides an equal average number of users per subarea.

For slice σ, the proportion of RBs provisioned by RRH i to the users in $\mathcal{A}_q^{\sigma}$ is denoted $\eta_{\mathrm{u}}^{\sigma}(i,q) \in [0,1]$ and $\eta_{\mathrm{d}}^{\sigma}(i,q) \in [0,1]$ for uplink and downlink traffic, respectively. The summed proportions of RBs provided by a given RRH i must be less than one

$$\sum_{\sigma \in \mathcal{S}} \sum_{q \in \mathcal{Q}^{\sigma}} \left(\eta_{\mathrm{u}}^{\sigma}(i,q) + \eta_{\mathrm{d}}^{\sigma}(i,q) \right) \leq 1, \, \forall i \in \mathcal{N}_{\mathrm{Ir}}. \tag{2}$$

For each slice σ and each subarea $\mathcal{A}_q^{\sigma}$, the total data rate provided by the allocated resource blocks should satisfy the minimum average user demand. Then, $\forall q \in \mathcal{Q}^{\sigma}, \forall \sigma \in \mathcal{S}$, one should have

$$\begin{aligned} \sum_{i \in \mathcal{N}_{\mathrm{Ir}}} \eta_{\mathrm{u}}^{\sigma}(i,q) \, a_{\mathrm{r}}(i) \, b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) &\geq \underline{R}_{\mathrm{u}}^{\sigma} \int_{\mathcal{A}_q^{\sigma}} \rho^{\sigma}(x) \, \mathrm{d}x, \\ \sum_{i \in \mathcal{N}_{\mathrm{Ir}}} \eta_{\mathrm{d}}^{\sigma}(i,q) \, a_{\mathrm{r}}(i) \, b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) &\geq \underline{R}_{\mathrm{d}}^{\sigma} \int_{\mathcal{A}_q^{\sigma}} \rho^{\sigma}(x) \, \mathrm{d}x, \end{aligned} \tag{3}$$

which corresponds to the satisfaction of the geographical coverage constraints for uplink and downlink. Here, $b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$ and $b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$ denote the amount of data (bits) carried by a RB for a user located in $\mathcal{A}_q^{\sigma}$ for up and downlink. Depending on the level of conservatism, $b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$ and $b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$ may represent the minimum or the average amount of data evaluated over the possible locations of users in $\mathcal{A}_q^{\sigma}$. The terms $b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$, $b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$, and $\int_{\mathcal{A}_q^{\sigma}} \rho^{\sigma}(x) \, \mathrm{d}x$ are fixed quantities that only depend on the RRH location x_i^{r}, on the user density ρ^{σ}, and on the way the partitioning of $\mathcal{A}^{\sigma}$ has been performed. These terms may thus be evaluated in advance, see Section V-B3. Summing (3) over all $q \in \mathcal{Q}^{\sigma}$ and using (1), one gets

$$\begin{aligned} \sum_{q \in \mathcal{Q}^{\sigma}} \sum_{i \in \mathcal{N}_{\mathrm{Ir}}} \eta_{\mathrm{u}}^{\sigma}(i,q) \, a_{\mathrm{r}}(i) \, b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) &\geq r_{\mathrm{u}}(v_{\mathrm{r}}), \\ \sum_{q \in \mathcal{Q}^{\sigma}} \sum_{i \in \mathcal{N}_{\mathrm{Ir}}} \eta_{\mathrm{d}}^{\sigma}(i,q) \, a_{\mathrm{r}}(i) \, b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) &\geq r_{\mathrm{d}}(v_{\mathrm{r}}), \end{aligned} \tag{4}$$

which ensures, for slice σ, the satisfaction of the part of the SRD graph related to the uplink and downlink radio resource demands.

For each RRH i, the amount of provisioned uplink and downlink resources should be proportional to the demand expressed in the SRD graph through $r_{\mathrm{u}}(v_{\mathrm{r}})$ and $r_{\mathrm{d}}(v_{\mathrm{r}})$. This avoids provisioning RRH resources taking care only of the uplink or only of the downlink traffic. This has to be ensured for all subareas $q \in \mathcal{Q}^{\sigma}$

$$\frac{\eta_{\mathrm{u}}^{\sigma}(i,q) \, a_{\mathrm{r}}(i) \, b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)}{r_{\mathrm{u}}(v_{\mathrm{r}})} = \frac{\eta_{\mathrm{d}}^{\sigma}(i,q) \, a_{\mathrm{r}}(i) \, b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)}{r_{\mathrm{d}}(v_{\mathrm{r}})}. \tag{5}$$

To identify whether a RRH $i \in \mathcal{N}_{\mathrm{Ir}}$ has provisioned some RBs to any subarea for slice σ, one introduces the variables $\widetilde{\eta}^{\sigma}(i) \in \{0,1\}$, with $\widetilde{\eta}^{\sigma}(i) = 1$ if $\sum_{q \in \mathcal{Q}^{\sigma}} \eta^{\sigma}(i,q) > 0$, and $\widetilde{\eta}^{\sigma}(i) = 0$ otherwise. The variables $\eta^{\sigma}(i,q)$ and $\widetilde{\eta}^{\sigma}(i)$ are gathered in the sets $\boldsymbol{\eta} = \{\eta^{\sigma}(i,q)\}_{i \in \mathcal{N}_{\mathrm{Ir}}, q \in \mathcal{Q}^{\sigma}, \sigma \in \mathcal{S}}$ and $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}^{\sigma}(i)\}_{i \in \mathcal{N}_{\mathrm{Ir}}, \sigma \in \mathcal{S}}$. The relation between $\eta^{\sigma}(i,v,q)$ and

$\widetilde{\eta}^{\sigma}(i)$ is nonlinear. Nevertheless, both quantities can be linked with the following linear constraints

$$0 \leq \widetilde{\eta}^{\sigma}(i) - \sum_{q \in \mathcal{Q}^{\sigma}} \eta^{\sigma}(i,q) < 1, \, \forall i \in \mathcal{N}_{\mathrm{Ir}}, \forall \sigma \in \mathcal{S}, \tag{6}$$

where

$$\eta^{\sigma}(i,q) = \eta_{\mathrm{u}}^{\sigma}(i,q) + \eta_{\mathrm{d}}^{\sigma}(i,q). \tag{7}$$

The cost related to the radio resource provisioning for a given slice σ gathers the fixed costs $c_{\mathrm{f}}(i)\,\widetilde{\eta}^{\sigma}(i)$ related to the use of a RRH by the slice and the variable costs $c_{\mathrm{r}} a_{\mathrm{r}}(i)\,\eta^{\sigma}(i,q)$ related to the amount of RBs provided by each RRH to the slice. Some bias towards allocation of RBs by RRHs to subareas which will provide a high data rate is obtained by the introduction of a rate-related discount $\lambda b\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) a_{\mathrm{r}}(i)\,\eta^{\sigma}(i,q)$. The resulting cost function for the *radio* resources is

$$\begin{aligned} c_{\mathrm{rr}}(\boldsymbol{\eta}^{\sigma}, \widetilde{\boldsymbol{\eta}}^{\sigma}) &= \sum_{\sigma \in \mathcal{S}} \sum_{i \in \mathcal{N}_{\mathrm{Ir}}} c_{\mathrm{f}}(i)\,\widetilde{\eta}^{\sigma}(i) \\ &+ \sum_{\sigma \in \mathcal{S}} \sum_{i \in \mathcal{N}_{\mathrm{Ir}}} \sum_{q \in \mathcal{Q}^{\sigma}} \left[c_{\mathrm{r}} - \lambda b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)\right] a_{\mathrm{r}}(i)\,\eta_{\mathrm{u}}^{\sigma}(i,q) \\ &+ \sum_{\sigma \in \mathcal{S}} \sum_{i \in \mathcal{N}_{\mathrm{Ir}}} \sum_{q \in \mathcal{Q}^{\sigma}} \left[c_{\mathrm{r}} - \lambda b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)\right] a_{\mathrm{r}}(i)\,\eta_{\mathrm{d}}^{\sigma}(i,q). \end{aligned} \tag{8}$$

B. Accounting for Other SRD Constraints

This section introduces a set of constraints which have to be satisfied to address the other resource demands for each $\sigma \in \mathcal{S}$, while being consistent with the coverage constraints.

For that purpose, one introduces first $\boldsymbol{\Phi}_{\mathrm{n}} = \{\phi_n^{\sigma}(i,v)\}_{i \in \mathcal{N}_{\mathrm{I}}, v \in \mathcal{N}_{\mathrm{V}}^{\sigma}, \sigma \in \mathcal{S}, n \in \{\mathrm{c},\mathrm{s}\}}$, where $\phi_n^{\sigma}(i,v)$ represents the proportion of resources of type $n \in \{\mathrm{c},\mathrm{s}\}$ provisioned on the infrastructure node $i \in \mathcal{G}_{\mathrm{I}}$ for the SRD node $v \in \mathcal{N}_{\mathrm{V}}^{\sigma}$ of slice σ. Second, let $\boldsymbol{\Phi}_{\mathrm{b}} = \{\phi_{\mathrm{b}}^{\sigma}(ij,vw)\}_{ij \in \mathcal{E}_{\mathrm{I}}, vw \in \mathcal{E}_{\mathrm{V}}^{\sigma}, \sigma \in \mathcal{S}}$, where $\phi_{\mathrm{b}}^{\sigma}(ij,vw)$ represents the proportion of bandwidth of the infrastructure link $ij \in \mathcal{E}_{\mathrm{I}}$ provisioned for the SRD link $vw \in \mathcal{E}_{\mathrm{V}}^{\sigma}$ of slice σ. Both $\boldsymbol{\Phi}_{\mathrm{n}}$ and $\boldsymbol{\Phi}_{\mathrm{b}}$ are sets of non-negative real variables ranging from 0 to 1. When one of the variables holds zero, there is no mapping between the infrastructure and the SRD node/link.

The sum of resources provided by each infrastructure node $i \in \mathcal{N}_{\mathrm{I}}$ mapped to an SRD node v should satisfy its resource demands. This leads to

$$\sum_{i \in \mathcal{N}_{\mathrm{I}}} a_n(i)\,\phi_n^{\sigma}(i,v) \geq r_n(v), \, \forall n \in \{\mathrm{c},\mathrm{s}\}, \forall v \in \mathcal{N}_{\mathrm{V}}^{\sigma}, \forall \sigma \in \mathcal{S}. \tag{9}$$

Since, the summed proportions of resources provisioned by a given infrastructure node i cannot exceed one, we have

$$\sum_{\sigma \in \mathcal{S}} \sum_{v \in \mathcal{N}_{\mathrm{V}}^{\sigma}} \phi_n^{\sigma}(i,v) \leq 1, \, \forall n \in \{\mathrm{c},\mathrm{s}\}, \forall i \in \mathcal{N}_{\mathrm{s}}. \tag{10}$$

Similarly, the cumulative proportions of resources provisioned by a given infrastructure link ij cannot exceed one

$$\sum_{\sigma \in \mathcal{S}} \sum_{vw \in \mathcal{E}_{\mathrm{V}}^{\sigma}} \phi_{\mathrm{b}}^{\sigma}(ij,vw) \leq 1, \, \forall ij \in \mathcal{E}_{\mathrm{I}}. \tag{11}$$

The amount of resources provided by a given infrastructure node i to an SRD node v has to be equal to an integer multiple of the minimum amount of resources $\underline{r}_n$ for VNFs associated to the SRD node v

$$\begin{aligned} &a_n(i)\phi_n^{\sigma}(i,v) = \underline{r}_n(v)\kappa_n^{\sigma}(i,v), \\ &\forall i \in \mathcal{N}_{\mathrm{I}}, \forall v \in \mathcal{N}_{\mathrm{V}}^{\sigma}, \forall n \in \{\mathrm{c},\mathrm{s}\}, \forall \sigma \in \mathcal{S}, \end{aligned} \tag{12}$$

where $\kappa_n^{\sigma}(i,v)$ is a positive integer. This ensures that enough resources are provisioned by an infrastructure node i to be able to deploy $\kappa_n^{\sigma}(i,v)$ VNFs associated to the SRD node v.

Moreover, resources of each type have to be provisioned in a balanced way by an infrastructure node for an SRD node, consistent with its demands. For example, if an infrastructure node provides 10% of the computing demand of a given SRD node, it should also provide 10% of the storage demand of this SRD node. This translates into the following resource provisioning proportionality constraints $\forall i \in \mathcal{N}_{\mathrm{I}}$ and $\forall v \in \mathcal{N}_{\mathrm{V}}^{\sigma}, \forall \sigma \in \mathcal{S}$,

$$\frac{a_{\mathrm{c}}(i)}{r_{\mathrm{c}}(v)} \phi_{\mathrm{c}}^{\sigma}(i,v) = \frac{a_{\mathrm{s}}(i)}{r_{\mathrm{s}}(v)} \phi_{\mathrm{s}}^{\sigma}(i,v). \tag{13}$$

Additionally, considering the SRD node v_{r}, the computing and storage resources provisioned by an infrastructure node $i \in \mathcal{N}_{\mathrm{Ir}}$ should be commensurate with the provisioned wireless resources, $\forall \sigma \in \mathcal{S}$,

$$\begin{aligned} &\frac{a_{\mathrm{c}}(i)}{r_{\mathrm{c}}(v_{\mathrm{r}})} \phi_{\mathrm{c}}^{\sigma}(i,v_{\mathrm{r}}) = \frac{a_{\mathrm{s}}(i)}{r_{\mathrm{s}}(v_{\mathrm{r}})} \phi_{\mathrm{s}}^{\sigma}(i,v_{\mathrm{r}}) \\ &= \frac{a_{\mathrm{r}}(i)}{r_{\mathrm{r}}(v_{\mathrm{r}})} \sum_{q \in \mathcal{Q}^{\sigma}} \left(\eta_{\mathrm{u}}^{\sigma}(i,q)\, b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) + \eta_{\mathrm{d}}^{\sigma}(i,q)\, b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)\right). \end{aligned} \tag{14}$$

The constraints (13) and (14) ensure a balanced resource provisioning by infrastructure nodes. In (14), $r_{\mathrm{r}}(v_{\mathrm{r}})$ is the total radio resource demand of v_{r} in both up and downlink, *i.e.*, $r_{\mathrm{r}}(v_{\mathrm{r}}) = r_{\mathrm{u}}(v_{\mathrm{r}}) + r_{\mathrm{d}}(v_{\mathrm{r}})$.

Moreover, link resources should be consistently provisioned with the radio resource of the RRH for both uplink and downlink. Thus, for downlink traffic (links with RRH as egress), one should have

$$\begin{aligned} \sum_{i \in \mathcal{N}_{\mathrm{I}} \setminus \mathcal{N}_{\mathrm{Ir}}} \frac{a_{\mathrm{b}}(ij)}{r_{\mathrm{b}}(vv_{\mathrm{r}})} \phi_{\mathrm{b}}^{\sigma}(ij, vv_{\mathrm{r}}) &= \left(\frac{r_{\mathrm{b}}(vv_{\mathrm{r}})}{\sum_{uv_{\mathrm{r}} \in \mathcal{E}_{\mathrm{V}}^{\sigma}} r_{\mathrm{b}}(uv_{\mathrm{r}})}\right) \frac{a_{\mathrm{r}}(j)}{r_{\mathrm{d}}(v_{\mathrm{r}})} \\ &\times \sum_{q \in \mathcal{Q}^{\sigma}} \eta_{\mathrm{d}}^{\sigma}(j,q)\, b_{\mathrm{d}}\left(x_j^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right), \end{aligned} \tag{15}$$

$\forall j \in \mathcal{N}_{\mathrm{Ir}}$, $\forall vv_{\mathrm{r}} \in \mathcal{E}_{\mathrm{V}}^{\sigma}$, and $\forall \sigma \in \mathcal{S}$. In (15), the term $a_{\mathrm{r}}(j) \sum_{q \in \mathcal{Q}^{\sigma}} \eta_{\mathrm{d}}^{\sigma}(j,q)\, b_{\mathrm{d}}\left(x_j^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) / r_{\mathrm{d}}(v_{\mathrm{r}})$ represents the proportion of downlink radio resources provided by RRH j to satisfy the downlink demand of v_{r}. When several SRD links feed v_{r}, the term $r_{\mathrm{b}}(vv_{\mathrm{r}}) / \sum_{uv_{\mathrm{r}} \in \mathcal{E}_{\mathrm{V}}^{\sigma}} r_{\mathrm{b}}(uv_{\mathrm{r}})$ represents the proportion of (downlink) traffic demand associated to the SRD link vv_{r}. The right-hand side of (15) represents thus the proportion of the data traffic that *has to be provisioned* for the SRD link vv_{r} to satisfy the part of the downlink radio resource provided by RRH j to satisfy the part of the downlink demand of v_{r}. The left-hand side of (15), represents the proportion of the data traffic that *is provided* by all infrastructure links ij, $i \in \mathcal{N}_{\mathrm{I}} \setminus \mathcal{N}_{\mathrm{Ir}}$ for the SRD link vv_{r}. Both terms have thus to be equal.

For uplink traffic (links with RRH as ingress), one has

$$\sum_{j\in\mathcal{N}_\mathrm{I}\backslash\mathcal{N}_\mathrm{Ir}} \frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(v_\mathrm{r}v)}\phi_\mathrm{b}^\sigma(ij,v_\mathrm{r}v) = \left(\frac{r_\mathrm{b}(v_\mathrm{r}v)}{\sum_{v_\mathrm{r}u\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(v_\mathrm{r}u)}\right)\frac{a_\mathrm{r}(i)}{r_\mathrm{u}(v_\mathrm{r})} \times \sum_{q\in\mathcal{Q}^\sigma} \eta_\mathrm{u}^\sigma(i,q)\, b_\mathrm{u}\left(x_i^\mathrm{r},\mathcal{A}_q^\sigma\right), \tag{16}$$

$\forall i \in \mathcal{N}_\mathrm{Ir}$, $\forall v_\mathrm{r}v \in \mathcal{E}_\mathrm{V}^\sigma$, and $\forall\sigma\in\mathcal{S}$. In (16), the term $a_\mathrm{r}(i)\sum_{q\in\mathcal{Q}^\sigma}\eta_\mathrm{u}^\sigma(i,q)\, b_\mathrm{u}\left(x_i^\mathrm{r},\mathcal{A}_q^\sigma\right)/r_\mathrm{u}(v_\mathrm{r})$ represents now the proportion of uplink radio resources provided by RRH i to satisfy the uplink demand of v_r. When several SRD links depart from v_r, the term $r_\mathrm{b}(v_\mathrm{r}v)/\sum_{v_\mathrm{r}u\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(v_\mathrm{r}u)$ represents the proportion of (uplink) traffic demand associated to the SRD link $v_\mathrm{r}v$. The right-hand side of (16) represents thus the proportion of the data traffic that *has to be provisioned* for the SRD link $v_\mathrm{r}v$ to convey the part of the uplink radio resource provided by RRH i to satisfy the part of the uplink demand of v_r. The left-hand side of (16), represents the proportion of the data traffic that *is provided* by all infrastructure links ij, $j\in\mathcal{N}_\mathrm{I}\backslash\mathcal{N}_\mathrm{Ir}$ for the SRD link $v_\mathrm{r}v$. Both terms have again to be equal.

Finally, flow conservation constraints have to be satisfied when resources are provisioned on the infrastructure link ij for the SRD link vw. That is, for each SRD link $vw\in\mathcal{E}_\mathrm{v}$, a path of infrastructure links must be provisioned between *each* pair of infrastructure nodes that are mapped to the pair (v,w) of SRD nodes.

Consider first an infrastructure node i which provisions resources for two SRD nodes v and w. The corresponding VNFs will have to exchange information within the considered node i via the internal link ii. For such internal link providing resources to an SRD link, one should have $\forall\sigma\in\mathcal{S}$

$$\frac{a_\mathrm{b}(ii)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ii,vw) = \left(\frac{r_\mathrm{b}(vw)}{\sum_{vu\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(vu)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(v)}\phi_\mathrm{c}^\sigma(i,v) \tag{17}$$

$$= \left(\frac{r_\mathrm{b}(vw)}{\sum_{uw\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(uw)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(w)}\phi_\mathrm{c}^\sigma(i,w). \tag{18}$$

In (17), the term $a_\mathrm{c}(i)\phi_\mathrm{c}^\sigma(i,v)/r_\mathrm{c}(v)$ represents the proportion of computing resources provided by the infrastructure node i to meet the demand of the SRD node v. When several SRD links depart from v, the term $r_\mathrm{b}(vw)/\sum_{vu\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(vu)$ represents the proportion of traffic demand that departs from v associated to the SRD link vw. The right-hand side of (17) represents thus the proportion of the data traffic that *has to be provisioned* for the SRD link vw to satisfy the corresponding proportion of computing resources provided by i to satisfy the part of the demand of v. The left-hand side of (16) represents the proportion of the data traffic that *is provided* by the internal link ii for the SRD link vw. The constraint (18) can be justified similarly.

Consider now an SRD link vw and an infrastructure node i which provisions resources either for only one of the SRD nodes v or w, or for none of them. Focusing again on the computing resource, three cases have to be considered.

Assume first that i provisions resources for v. Then one has the following constraint $\forall\sigma\in\mathcal{S}$

$$\sum_{j\in\mathcal{N}_\mathrm{I}} \frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ij,vw) = \left(\frac{r_\mathrm{b}(vw)}{\sum_{vu\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(vu)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(v)}\phi_\mathrm{c}^\sigma(i,v). \tag{19}$$

The right-hand side of (19) is the same as that of (17). The left-hand side of (19) represents the proportion of the traffic provisioned by all links ij, $j\in\mathcal{N}_\mathrm{I}$ (leaving node i) for the SRD link vw.

Assume second that i provisions resources for w. Then one has the following constraint $\forall\sigma\in\mathcal{S}$

$$\sum_{j\in\mathcal{N}_\mathrm{I}} \frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ji,vw) = \left(\frac{r_\mathrm{b}(vw)}{\sum_{uw\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(uw)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(w)}\phi_\mathrm{c}^\sigma(i,w). \tag{20}$$

The right-hand side of (20) is the same as that of (18). The left-hand side of (20) represents now the proportion of the traffic provisioned by all links ji, $j\in\mathcal{N}_\mathrm{I}$ (feeding node i) for the SRD link vw.

Assume finally that i provisions resources neither for v nor for w. Then the following flow-conservation constraint

$$\sum_{j\in\mathcal{N}_\mathrm{I}}\left[\frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ij,vw) - \frac{a_\mathrm{b}(ji)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ji,vw)\right] = 0 \tag{21}$$

must be satisfied $\forall\sigma\in\mathcal{S}$.

The constraints (19-21) can be gathered in the following single constraint

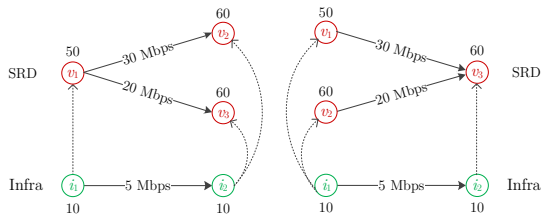
$$\begin{aligned}&\sum_{j\in\mathcal{N}_\mathrm{I}}\left[\frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ij,vw) - \frac{a_\mathrm{b}(ji)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ji,vw)\right]\\ &= \left(\frac{r_\mathrm{b}(vw)}{\sum_{vu\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(vu)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(v)}\phi_\mathrm{c}^\sigma(i,v)\\ &- \left(\frac{r_\mathrm{b}(vw)}{\sum_{uw\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(uw)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(w)}\phi_\mathrm{c}^\sigma(i,w).\\ &\forall i\in\mathcal{N}_\mathrm{I},\forall vw\in\mathcal{E}_\mathrm{V}^\sigma,\forall\sigma\in\mathcal{S}.\end{aligned} \tag{22}$$

In (17), (18), and (22), the consistency with the other provisioned resources is ensured by (13).

Figures 7 illustrate two resource provisioning examples for simplified SRD graphs with branched topologies. In Figure 7a, the node i_1 is mapped onto v_1, i_2 is mapped onto the pair (v_2,v_3); and the link i_1i_2 is mapped onto the pair (v_1v_2,v_1v_3). Considering the infrastructure node i_1 and the SRD link v_1v_2, the constraint (22) leads to $\frac{5}{30}\phi_\mathrm{b}(i_1i_2,v_1v_2) - 0 = \frac{30}{50}\frac{10}{50}\phi_n(i_1,v_1) - 0$, hence $\phi_\mathrm{b}(i_1i_2,v_1v_2) = \frac{18}{25}\phi_n(i_1,v_1)$. Similarly, considering i_1 and v_1v_3, one gets $\phi_\mathrm{b}(i_1i_2,v_1v_3) = \frac{8}{25}\phi_n(i_1,v_1)$. The largest amount of resource of type n node i_1 can provision to v_1 is then $\phi_n(i_1,v_1) = \frac{25}{26}$, which leads to $\phi_\mathrm{b}(i_1i_2,v_1v_2) = \frac{8}{26}$, and $\phi_\mathrm{b}(i_1i_2,v_1v_3) = \frac{18}{26}$. Considering $\phi_n(i_1,v_1) = 1$ would lead to $\phi_\mathrm{b}(i_1i_2,v_1v_2)+\phi_\mathrm{b}(i_1i_2,v_1v_3) = \frac{26}{25} > 1$, which is not consistent with (11).

In Figure 7b, an SRD graph with a merge topology is depicted. Through similar calculations, one gets $\phi_n(i_1,v_1) = \frac{1}{2}$, $\phi_n(i_1,v_2) = \frac{2}{5}$, $\phi_n(i_2,v_3) = 1$, $\phi_\mathrm{b}(i_1i_2,v_1v_3) = \frac{9}{15}$, and $\phi_\mathrm{b}(i_1i_2,v_2v_3) = \frac{4}{15}$. The proportions of provisioned

infrastructure node and link resources are then consistent with the proportions of node and link resource demands. The proportionality of provisioned resources for links entering or leaving the same vertices is also ensured.



(a) Fork topology (b) Merge topology

Fig. 7. Illustration to the constraint (22) related to flow conservation considering an SRD graph with (a) a fork topology and (b) merge topology. The nodes i_1, i_2 and the link $i_1 i_2$ belong to the infrastructure graph; The nodes v_1, v_2, v_3 and the links connecting them belong to the SRD graph.

To indicate whether infrastructure nodes have provisioned resources for some SRD node, one introduces the set of binary variables $\widetilde{\boldsymbol{\Phi}} = \left\{\widetilde{\phi}^{\sigma}(i,v)\right\}_{i\in\mathcal{N}_{\mathrm{I}}, v\in\mathcal{N}_{\mathrm{V}}^{\sigma}, \sigma\in\mathcal{S}}$, which represents node mapping indicators, *i.e.*, $\widetilde{\phi}^{\sigma}(i,v) = 1$ if at least one of the elements of $\{\phi_{\mathrm{c}}^{\sigma}(i,v), \phi_{\mathrm{s}}^{\sigma}(i,v)\}$ is strictly positive, and $\widetilde{\phi}^{\sigma}(i,v) = 0$ otherwise. The relation between $\phi_n^{\sigma}(i,v)$ and $\widetilde{\phi}_n^{\sigma}(i,v)$ is again nonlinear. As in (6), both quantities may be linearly related as follows

$$\sum_{n\in\{\mathrm{c,s}\}} \frac{\phi_n^{\sigma}(i,v)}{2} \leq \widetilde{\phi}(i,v) < \sum_{n\in\{\mathrm{c,s}\}} \frac{\phi_n^{\sigma}(i,v)}{2} + 1, \quad \forall i \in \mathcal{N}_{\mathrm{I}}, \forall v \in \mathcal{N}_{\mathrm{V}}^{\sigma}, \forall \sigma \in \mathcal{S}. \tag{23}$$

The cost related to the provisioning of computing, storage, and bandwidth resources in the *wired* part of the infrastructure network for all slices in $\mathcal{S}$ can be expressed as

$$\begin{aligned} c_{\mathrm{w}}\left(\boldsymbol{\Phi}_{\mathrm{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\mathrm{b}}\right) = & \\ & \sum_{\sigma\in\mathcal{S}} \sum_{i\in\mathcal{N}_{\mathrm{I}}} \sum_{v\in\mathcal{N}_{\mathrm{V}}^{\sigma}} \sum_{n\in\{\mathrm{c,s}\}} a_n(i)\phi_n^{\sigma}(i,v)c_n(i) \\ & + \sum_{\sigma\in\mathcal{S}} \sum_{ij\in\mathcal{E}_{\mathrm{I}}} \sum_{vw\in\mathcal{E}_{\mathrm{V}}^{\sigma}} a_{\mathrm{b}}(ij)\phi_{\mathrm{b}}^{\sigma}(ij,vw)c_{\mathrm{b}}(ij) \\ & + \sum_{\sigma\in\mathcal{S}} \sum_{i\in\mathcal{N}_{\mathrm{I}}} \sum_{v\in\mathcal{N}_{\mathrm{V}}^{\sigma}} \widetilde{\phi}^{\sigma}(i,v)c_{\mathrm{f}}(i), \end{aligned} \tag{24}$$

where the first and second terms indicate the total cost for leasing resources from infrastructure nodes and links, while the third term represents the cost for deploying VNFs in infrastructure nodes.

C. Single-step vs Two-step Provisioning

The global provisioning problem has to account for storage and computing constraints, as well as coverage constraints. It leads to the minimization of the sum of the costs (24) and (8)

$$c_{\mathrm{tot}} = c_{\mathrm{rr}}\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right) + c_{\mathrm{w}}\left(\boldsymbol{\Phi}_{\mathrm{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\mathrm{b}}\right) \tag{25}$$

with the constraints introduced in Sections IV-A and IV-B.

Nevertheless, when the number of variables in $(\boldsymbol{\Phi}_{\mathrm{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\mathrm{b}})$ and $(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ increases, the problem may become intractable. Thus, a two-step provisioning algorithm denoted CARP (Coverage-Aware Resource Provisioning), Algorithm 1, is introduced to minimize both terms of (25) separately. The *Radio resource Provisioning* problem, denoted RP, with the radio coverage constrains introduced in Section IV-A, is solved first. Then, using the solution of the RP problem, the *Network resource Provisioning*, denoted NP, considering the other resources constraints introduced in Section IV-B, is solved by CARP.

Algorithm 1: Coverage-Aware Resource Prov. (CARP)

Input: $\mathcal{A}$, $\mathcal{G}_{\mathrm{I}} = (\mathcal{N}_{\mathrm{I}}, \mathcal{E}_{\mathrm{I}})$, $\mathcal{S} = \{\sigma_1, \sigma_2, \cdots\}$, $\{\mathcal{G}_{\mathrm{V}}^{\sigma}, \sigma \in \mathcal{S}\}$
Output: provisioning result of all slice $\sigma \in \mathcal{S}$

 # Sub-area partitioning
for *each slice* σ *in* $\mathcal{S}$ **do**
  Retrieve coverage information of σ ($\mathcal{A}^{\sigma}$) from SRD message;
  Divide $\mathcal{A}^{\sigma}$ into Q^{σ} convex subareas $\mathcal{A}_q^{\sigma}$, $q \in \mathcal{Q}^{\sigma} = \{1, \ldots, Q^{\sigma}\}$;
end
 # Provision radio resources
Solve $(\widehat{\boldsymbol{\eta}}, \widehat{\widetilde{\boldsymbol{\eta}}}) = \arg\min_{\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}}} c_{\mathrm{r}}(\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}})$, subject to: (2)-(7) in a jointly (JR) or sequential (SR) fashion;
 # Obtain the provisioned radio resources
for *each slice* σ *in* $\mathcal{S}$ **do**
  Report the list of reserved RRHs for σ: $\mathcal{N}_{\mathrm{Ir}}^{\sigma} = \{i \in \mathcal{N}_{\mathrm{Ir}} : \tilde{\eta}_i^{\sigma} = 1\}$;
end
Compute provisioned radio resource of RRH $i \in \mathcal{N}_{\mathrm{Ir}}^{\star}$ for each σ:

$$\sum_{q\in\mathcal{Q}^{\sigma}} \eta_{\mathrm{u}}^{\sigma}(i,q)\, a_{\mathrm{r}}(i)\, b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) \quad \text{(uplink)}$$

$$\sum_{q\in\mathcal{Q}^{\sigma}} \eta_{\mathrm{d}}^{\sigma}(i,q)\, a_{\mathrm{r}}(i)\, b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) \quad \text{(downlink)}$$

 # Provision other network resources
Solve $(\widehat{\boldsymbol{\Phi}}_{\mathrm{n}}, \widehat{\widetilde{\boldsymbol{\Phi}}}, \widehat{\boldsymbol{\Phi}}_{\mathrm{e}}) = \arg\min_{\boldsymbol{\Phi}_{\mathrm{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\mathrm{e}}} c_{\mathrm{w}}(\boldsymbol{\Phi}_{\mathrm{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\mathrm{e}})$, subject to: (9)-(23), in a jointly (JN) or sequential (SN) fashion;
Report $\widehat{\boldsymbol{\Phi}}_{\mathrm{n}}, \widehat{\widetilde{\boldsymbol{\Phi}}}, \widehat{\boldsymbol{\Phi}}_{\mathrm{e}}$;

During initialization of CARP, the slice coverage information ($\mathcal{A}^{\sigma}$) is obtained from the SRD, and $\mathcal{A}^{\sigma}$ is partitioned into Q^{σ} convex subareas $\mathcal{A}_q^{\sigma}$, $q \in \mathcal{Q}^{\sigma} = \{1, \ldots, Q^{\sigma}\}$;

In Step 1, the values of $\boldsymbol{\eta}$ and $\widetilde{\boldsymbol{\eta}}$ minimizing $c_{\mathrm{rr}}\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right)$ while satisfying all constraints related to radio provisioning ((2)-(7)) are evaluated;

In Step 2, the values of $\boldsymbol{\Phi}_{\mathrm{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\mathrm{b}}$ minimizing $c_{\mathrm{w}}\left(\boldsymbol{\Phi}_{\mathrm{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\mathrm{e}}\right)$, subject to the constraints (9)-(23) are evaluated. The constraints (14), (15), (16) are evaluated with the help of $\boldsymbol{\eta}$ and $\widetilde{\boldsymbol{\eta}}$ obtained at Step 1.

When the resource provisioning problem has to be solved for several slices, each of the RP and NP problems can be addressed either sequentially for each slice, or jointly for all slices. Denote SR and JR for sequential and joint RP, and similarly SN and JN for sequential and joint NP. Combining these methods gives four variants of the provisioning algorithm (SR_SN, SR_JN, JR_SN, and JR_JN), as summarized in Table II with the number of RP and NP problems and the corresponding number of variables per problem to be handled by each variant. In Table II, the variables $\kappa_n^{\sigma}(i,v)$ introduced in (12) are not taken into account, since they are directly related to $\phi_n^{\sigma}(i,v)$.

The sequential variants (SR and SN) require to solve $|\mathcal{S}|$ more optimization problems, but with $|\mathcal{S}|$ less variables. Since

each problem is NP-hard, the sequential variants may obviously be solved faster than the joint variants (JR and JN). Section V-B compares these variants on simulations.

TABLE II
VARIANTS OF THE PROVISIONING ALGORITHM.

Variant	*Nb of problem*	*Nb of variables/problem*
SR_SN	$\|\mathcal{S}\|$ RP	$\|\mathcal{N}_{\mathrm{Ir}}\|\,(1+\|\mathcal{Q}^{\sigma}\|)$
	$\|\mathcal{S}\|$ NP	$3\|\mathcal{N}_{\mathrm{I}}\|\|\mathcal{N}_{\mathrm{V}}^{\sigma}\|+\|\mathcal{E}_{\mathrm{I}}\|\|\mathcal{E}_{\mathrm{V}}^{\sigma}\|$
SR_JN	$\|\mathcal{S}\|$ RP	$\|\mathcal{N}_{\mathrm{Ir}}\|\,(1+\|\mathcal{Q}^{\sigma}\|)$
	1 NP	$\|\mathcal{S}\|\left(3\|\mathcal{N}_{\mathrm{I}}\|\|\mathcal{N}_{\mathrm{V}}^{\sigma}\|+\|\mathcal{E}_{\mathrm{I}}\|\|\mathcal{E}_{\mathrm{V}}^{\sigma}\|\right)$
JR_SN	1 RP	$\|\mathcal{S}\|\|\mathcal{N}_{\mathrm{Ir}}\|\,(1+\|\mathcal{Q}^{\sigma}\|)$
	$\|\mathcal{S}\|$ NP	$3\|\mathcal{N}_{\mathrm{I}}\|\|\mathcal{N}_{\mathrm{V}}^{\sigma}\|+\|\mathcal{E}_{\mathrm{I}}\|\|\mathcal{E}_{\mathrm{V}}^{\sigma}\|$
JR_JN	1 RP	$\|\mathcal{S}\|\|\mathcal{N}_{\mathrm{Ir}}\|\,(1+\|\mathcal{Q}^{\sigma}\|)$
	1 NP	$\|\mathcal{S}\|\left(3\|\mathcal{N}_{\mathrm{I}}\|\|\mathcal{N}_{\mathrm{V}}^{\sigma}\|+\|\mathcal{E}_{\mathrm{I}}\|\|\mathcal{E}_{\mathrm{V}}^{\sigma}\|\right)$

V. EVALUATION

In this section, one evaluates via simulations the performance of the proposed provisioning algorithms. First, a simplified scenario addressing only the NP problem is considered in Section V-A to illustrate the benefits of provisioning prior to SFC embedding. Then, the variants of the provisioning algorithm introduced in Section IV-C considering also coverage constraints are compared in Section V-B. All simulations are performed with the CPLEX MILP solver interfaced with MATLAB.

A. Resource Provisioning vs Direct Embedding

In this section, no coverage constraints are considered. Radio resources are provisioned in the same way as computing and storage resources, *i.e.,* independently of any coverage constraint. Infrastructure resources are first provisioned for different slices. Then, the SFCs of these slices are embedded considering the infrastructure subgraph provided by the provisioning algorithm. The performance of this prov_emb approach is compared to a classical approach, direct_emb, where the SFCs are directly embedded in the infrastructure without prior provisioning. Two different resource provisioning schemes are considered. In the sequential approach (SN), resources are provisioned slice-by-slice, while in the joint approach (JN), provisioning is performed taking the constraints of all slices simultaneously and solving the resulting MILP, see (9)-(23). The ILP-based SFC embedding algorithm is adapted from [17] and [18].

1) Infrastructure Topology: As in [17, 37], a k-ary fat-tree infrastructure topology is considered, see Figure 8. The leaf nodes represent the RRHs. The other nodes represent the edge, regional, and central data centers. Infrastructure nodes and links provide a given amount of computing, storage, and possibly radio resources $(a_{\mathrm{c}}, a_{\mathrm{s}}, a_{\mathrm{r}})$ expressed in available number of CPUs, Gbytes of available storage, and Gbps of transmission capacity, depending on the level they belong to. The cost of leasing each unit of infrastructure resource is set to 1.

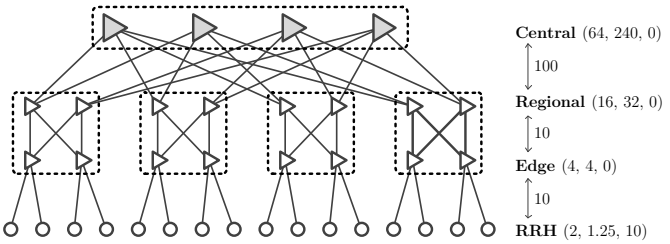


Fig. 8. Description of the k-ary fat-tree infrastructure network in case $k = 4$. Nodes provide a given amount of computing a_{c}, storage a_{s}, and radio resource a_{r} measured in number of used CPUs, Gbytes, and Gbps respectively. Links are assigned with a given amount of bandwidth a_{b} measured in Gbps.

2) Slice Resource Demand (SRD): One considers a slice devoted to the delivery of HD video streaming services at 4 Mbps to 4000 users. Within the slice, each SFC to be deployed consists of three chained VNFs: a virtual Video Optimizer Controller (vVOC), a virtual Gateway (vGW) and a virtual BBU (vBBU). The minimum resource requirements for running each VNF instance correspond to the aggregated resource required by an SFC managing 200 users connected to a single virtual BS.

As detailed in Section III, the resource requirements for the various SFCs that will have to be deployed within a slice are aggregated within an SRD graph that mimics the graph of SFCs. SRD nodes and links are characterized by the aggregated resource needed to support the targeted number of users. Details of each resource type as well as associated SRD graph are given in Table III. Numerical values in Table III have been adapted from [35]. In the following, we consider multiple demands from different SPs for this type of slice.

TABLE III
SRD FOR HD VIDEO STREAMING SERVICE.

Node	r_{c}	$\underline{r}_{\mathrm{c}}$	r_{s}	$\underline{r}_{\mathrm{s}}$	*Link*	r_{b}
VOC	21.6	1.08	60	3.0	VOC→GW	2.0
GW	3.60	0.18	2.0	0.1	GW→BBU	2.0
BBU	3.2	0.16	2.0	0.1		

3) Results: Figures 9a, 9b, and 9c illustrate respectively the infrastructure node utilization, the infrastructure link utilization, and the computing time for different number $|\mathcal{S}|$ of HD video streaming slices. Infrastructure node or link utilization reflects the percentage of infrastructure nodes or links provisioned for the considered slices. The SN and JN approaches are compared.

Figure 9b shows that the joint provisioning scheme provides a slightly reduced node and link utilization compared to the sequential approach. The difference increases with the number of slices. Specifically, when $|\mathcal{S}| = 2$, both methods give the same performance. When $|\mathcal{S}| = 3$ and $|\mathcal{S}| = 4$, the joint provisioning leads to a node and link usage reduced by 0.27% and 0.53%, compared to the sequential method.

Nevertheless, as can be seen in Figure 9c, the sequential provisioning method outperforms the joint approach in terms of computing time. As explained in Section IV-C, in the case of JN, the number of possible configurations to consider increases

with the number of slices to deploy and therefore increases the computation duration.

Figures 10a and 10b show respectively the cost and the consumed computing time for different number of SFCs to be embedded (ranging from 2 to 10). The embedding cost reflects the amount of infrastructure node and link resources used for embedding these SFCs. The proposed method, `prov_emb`, where provisioning is done before VNF embedding, has similar performance as a direct embedding (`direct_emb`) without prior provisioning. Nevertheless, as depicted in Figure 10b, the `prov_emb` approach is faster than `direct_emb`. The difference increases with the number of SFCs to embed. Note that in the `prov_emb` approach, the computing time for the provisioning step has been taken into account.

B. Accounting for coverage constraints

1) Infrastructure Topology: In this section, one accounts for coverage constraints. For that purpose, consider the $1.43\,\text{km}\times 4.95\,\text{km}$ area around the Stade de France in Seine-Saint-Denis (suburban area of the city of Paris) shown in Figure 3. The map includes real coordinates of RRH nodes (indicated by blue markers) taken from the open database provided by the French National Agency of Frequencies[1].

The k-ary fat-tree infrastructure topology considered in Section V-A is used here again. The amount of network infrastructure resource available at each node and link of the infrastructure remains the same. Only the RRH nodes are represented in Figure 3. The locations of the remaining parts of the infrastructure network (central, regional, and edge nodes) are not displayed.

2) Slice Resource Demand (SRD): We consider three types of slices:

- Slice 1 covers the *Stade de France* and aims to provide HD video streaming services at 4 Mbps for at most 500 VIP users within the stadium (downlink traffic);
- Slice 2 is dedicated to provide an SD video streaming services at 0.5 Mbps, and covers the blue-highlighted area in Figure 3 (downlink traffic);
- Slice 3 aims to provide video surveillance and traffic monitoring service at 1 Mbps for 50 cameras installed on the A86 highway (uplink traffic).

Similar to the slice used in Section V-A, the first two slices address a video streaming service, and thus have the same function architecture with 3 virtual functions: a vVOC, a vGW, and a vBBU. The third slice consists of five virtual functions: a vBBU, a vGW, a virtual Traffic Monitor (vTM), a vVOC, and a virtual Intrusion Detection Prevention System (vIDPS).

The minimum resource requirements for running each VNF instance correspond to the aggregated resource required by an SFC managing 50, 600, and 5 users connected to a same virtual RRH, for Slice 1, 2, and 3, respectively.

The coverage area $\mathcal{A}^\sigma$ associated to each slice is partitioned into rectangular subareas $\mathcal{A}^\sigma_q$ of $89.58\,\text{m}\times 103.32\,\text{m}$.

Functional structure and resource requirements of three slices are described in Table IV.

[1]L'Agence Nationale des Fréquences (ANFR): https://data.anfr.fr/, retrieved July 19, 2019.

TABLE IV
SLICE RESOURCE DEMAND.

Slice 1: HD video streaming

Node	r_c	$\underline{r}_\text{c}$	r_s	$\underline{r}_\text{s}$	*Link*	r_b
VOC	2.7	0.27	7.5	0.75	VOC→GW	2.0
GW	0.45	0.045	0.25	0.025	GW→BBU	2.0
BBU	2.0	0.2	0.25	0.025		

Slice 2: SD video streaming

Node	r_c	$\underline{r}_\text{c}$	r_s	$\underline{r}_\text{s}$	*Link*	r_b
VOC	6.51	0.651	11.3	1.130	VOC→GW	2.8
GW	1.09	0.109	0.38	0.038	GW→BBU	2.8
BBU	2.00	0.200	0.38	0.038		

Slice 3: Video surveillance and traffic monitoring

Node	r_c	$\underline{r}_\text{c}$	r_s	$\underline{r}_\text{s}$	*Link*	r_b
BBU	0.20	0.020	0.01	0.001	BBU→GW	0.05
GW	0.05	0.005	0.01	0.001	GW→TM	0.05
TM	0.67	0.067	0.01	0.001	TM→VOC	0.05
VOC	0.27	0.027	0.19	0.019		

3) Rate Function: Models $b_\text{d}\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right)$ and $b_\text{u}\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right)$, introduced in Section IV-A, for the amount of data carried by a RB for a user located in $\mathcal{A}^\sigma_q$ and served by a RRH located in x^r_i are now considered. Let $d\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right)$ be the distance between x^r_i and the center of each rectangle $\mathcal{A}^\sigma_q$. Focusing on downlink traffic, according to [38], one assumes that

$$b_\text{d}\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right) = W_i \log_2\left(1 + \frac{P_{\text{rx,d}}\left(d\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right)\right)}{P_\text{n}}\right), \quad (26)$$

where W_i is the bandwidth (in Hz) of a RB provided by RRH i, P_n is the noise power given by $P_\text{n} = W_i N_0$, where N_0 is the noise power spectral density. $P_\text{rx}(d)$ is the obtained signal power at the receiver evaluated as

$$P_{\text{rx,d}}(d) = P_{\text{tx,d}} + G_{\text{tx,d}} + G_{\text{rx,d}} - PL(d), \quad (27)$$

where P_tx is the transmission power of the transmitter, G_tx and G_rx are the antenna gains of the transmitter and the receiver, and $PL(d)$ is the Path Loss given by the adapted $\alpha\beta\gamma$-model introduced in [39] for 5G mobile network

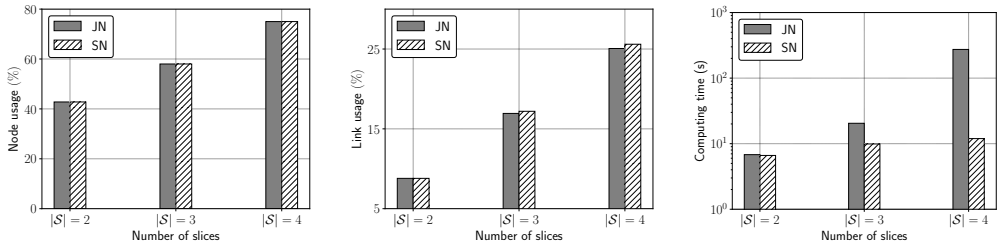
$$PL(d) = 10\alpha \log_{10}(d) + \beta + 10\gamma \log_{10}(f_i), \quad (28)$$

where α and γ are respectively coefficients accounting for the dependency of the path loss with distance and frequency f_i, β is an optimized offset value for path loss (dB). PL, d, and f_i are expressed in dB, meters, and GHz, respectively. An expression similar to (26) may be derived for $b_\text{u}\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right)$.

All RRH $i \in \mathcal{N}_\text{Ir}$ and all UEs are assumed to be identical. The parameters for the models $b_\text{d}\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right)$ and $b_\text{u}\left(x^\text{r}_i, \mathcal{A}^\sigma_q\right)$ are summarized in Table V and have been partly taken from [40].

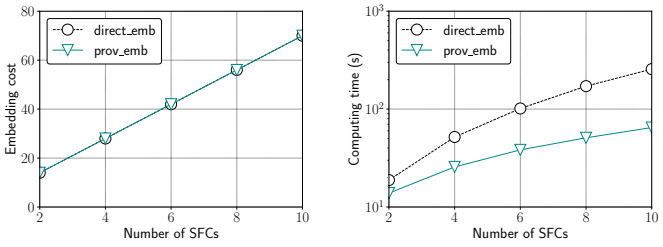
4) Results: This section illustrates the performance of the the four variants of the provisioning algorithm described in Table II, when only Slices 1 and 2 have to be deployed and when all three slices have to be deployed.

Figures 11a, 11b illustrate respectively the wireless and wired cost related to the `RP` (radio resource provisioning) and the `NP` (other network resource provisioning) problems. First, we observe that the joint `RP` scheme yields the minimum radio and wired cost whatever the `NP` allocation method (`JR_SN` and



(a) (b) (c)

Fig. 9. Performance comparison of JN and SN in terms of (a) node utilization, (b) link utilization, and (c) computing time



(a) (b)

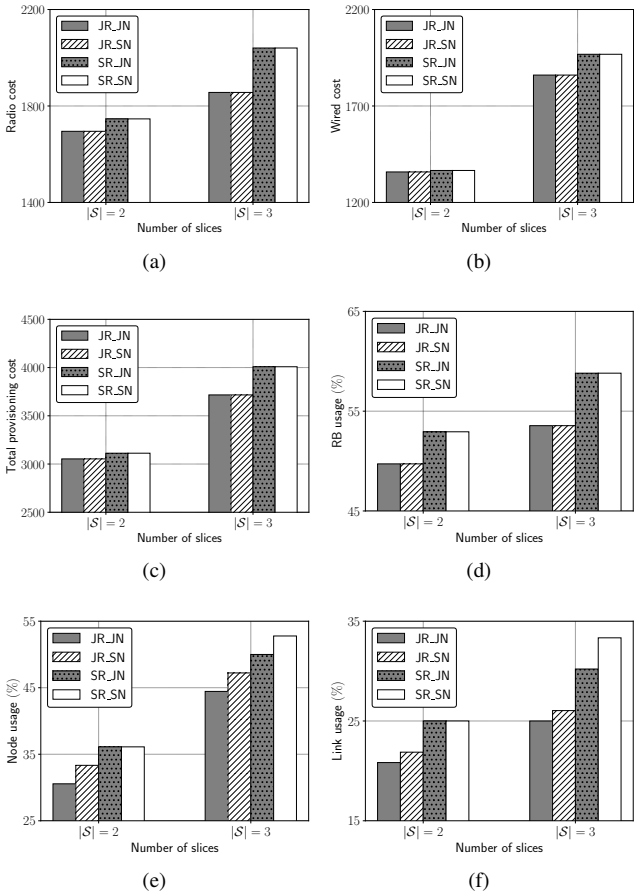
Fig. 10. (a) Embedding costs, and (b) computing time of prov_emb and of the direct_emb approaches as a function of the number of SFCs to embed

TABLE V
PARAMETERS OF RRH, UE, AND $\alpha\beta\gamma$-MODEL.

Parameter	*Definition*	*Value*
$a_{\mathrm{r}}(i)$	Number of RBs available at RRH i	100
f_i	Carrier frequency of RRH i	2.6 GHz
W_i	Bandwidth of a RB of RRH i	0.2 MHz
$P_{\mathrm{tx,d}}$	Antenna transmit power of each RRH	43 dBm
$G_{\mathrm{tx,d}}$	Antenna gain of each RRH	15 dBi
$P_{\mathrm{tx,u}}$	Antenna transmit power of each UE	23 dBm
$G_{\mathrm{tx,u}}$	Antenna gain of each UE	3 dBi
N_0	Noise power spectral density	-174 dBm/Hz
(α, β, γ)	$\alpha\beta\gamma$-model parameters	$(3.6, 7.6, 2)$

JR_JN). It also provides the minimum total provisioning cost, as shown in Figure 11c.

The use of radio resource blocks is depicted in Figure 11d. Again, one observes that the joint RP (JR_SN and JR_JN) outperforms the sequential approach (SR_JN and SR_SN), since the joint RP aims at finding the optimal solution for the whole problem, *i.e.*, provisioning for all the slices, while the sequential method only accounts for the constraints of each slice sequentially. The joint RP approach also leads to an efficient utilization of infrastructure nodes and links when solving the NP problem, as shown in Figures 11e and 11f. The difference in performance of these two sets of methods (JR_SN and JR_JN versus SR_SN and SR_JN) becomes more significant when the three slices are simultaneously considered. For instance, with two slices, a difference of 4.17% in terms of link utilization is observed in favor of the JR_JN approach, see Figure 11f, whereas with three slices, the difference is 8.33%. Overall, the JR_JN approach provides the best performance in terms of deployment costs and resource usage among the four proposed methods.



(a) (b) (c) (d) (e) (f)

Fig. 11. Performance comparison of 4 variants in terms of (a) radio cost, (b) wired cost, (c) total provisioning cost, utilization of (d) RBs, (e) infrastructure nodes, and (f) infrastructure links.

Nevertheless, as expected, the methods involving sequential provisioning (SR and SN) perform better than the joint approaches (JR and/or JN) in terms of computing time. Increasing of the number of slices leads to an increase of the cardinality of the variable sets $\left(\boldsymbol{\eta}_{iq}^{\sigma}, \tilde{\boldsymbol{\eta}}_{i}^{\sigma}\right)$ and $\left(\boldsymbol{\Phi}_{\mathrm{n}}^{\sigma}, \widetilde{\boldsymbol{\Phi}}^{\sigma}, \boldsymbol{\Phi}_{\mathrm{e}}^{\sigma}\right)$, and therefore increases the computing time. In sequential provisioning, slices are considered successively. Therefore, the SR_SN approach and the JR_JN approach are respectively the least and most time-consuming among the four methods, as shown in Figure 12.

Figure 13 focuses on the RP problem and shows the max-

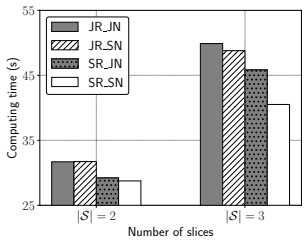


Fig. 12. Computing time of the 4 proposed provisioning variants

imum supported data rate in the case of sequential and joint radio resource provisioning (*i.e.,* SR and JR) as a function of the aggregated data rate demand of users in the three slices, *i.e.*, $\sum_{\sigma \in \mathcal{S}} u^\sigma \underline{R}^\sigma$. $\underline{R}^\sigma$ remains constant for each slice σ. The total number of users u^σ associated to each slice varies, but their relative proportions among slices remain constant. With the joint approach, a larger aggregated data rate is supported: provisioning of slices with more users is then possible.

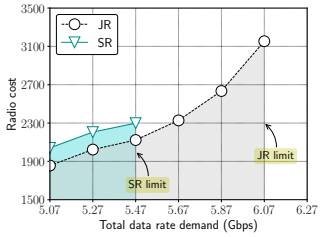


Fig. 13. Maximum supported data rate associated to the SR and JR provisioning approaches when 3 slices have to be deployed

VI. Conclusions

This paper considers the problem of infrastructure resource provisioning for network slicing in future mobile networks. Contrary to previous best-effort approaches where SFCs are deployed sequentially in the infrastructure network, here infrastructure resources are provisioned to accommodate slice resource demands. For that purpose, a graph of Slice Resource Demands is defined. This graph describes the aggregated resource requirements of the SFCs that will be deployed by the MNO for a given slice. Radio coverage constraints, to ensure a minimum data rate for users in the geographical areas where services have to be deployed, as well as other constraints related to the infrastructure network are combined in an MILP problem formulation.

A two-step approach is proposed to address this problem. Resources on RRH are provisioned first to ensure the satisfaction of the coverage constraints. Constraints as defined by the SRD graph are then considered. When resources have to be provisioned for several concurrent slices, two variants have again been considered. At each step, constraints related to each slice may be considered either sequentially, or jointly. Due to the exponential worst-case complexity in the number of variables of the MILP, as expected, sequential methods are shown, through simulation, to better scale to network topologies of realistic size. The price to be paid is a somewhat degraded link utilization and a higher provisioning cost compared to the joint approach.

Once resources have been provisioned, the approach introduced in [17, 37] may be used to deploy SFCs, but considering only a simplified infrastructure network reduced to the nodes and links which have provisioned resources. Simulations show that provisioning and then deploying is more efficient in terms of computing time than direct SFC embedding.

Several issues will be addressed in future work. First, due to the two-step approach, radio coverage constraints are taken into account first, which leads to the provisioning of RRH resources. This provisioning may not be compliant with other infrastructure network constraints. To address this issue, both coverage and infrastructure network constraints have to be really taken into account simultaneously. Alternatively, an iterative approach has to be introduced to eliminate in the first provisioning step, the RRH provisioning schemes which have been shown, in the second step, to lead to infrastructure provisioning issues. Second, only static provisioning is considered in this paper. Adaptive provisioning techniques should be proposed to cope with time-varying aggregated demands.

References

[1] C. Liang and F. R. Yu, "Wireless Network Virtualization: A Survey, Some Research Issues and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 358–380, 2015.
[2] 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 on All Things Cellular: Operations, Applications, & Challenges*, 2014, pp. 33–38.
[3] 5G America, "Network Slicing for 5G and Beyond," *White Paper*, 2016.
[4] IETF, "Network Slicing Architecture," *Internet-Draft*, pp. 1–27, 2017.
[5] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, 2017.
[6] D. H. Luong, H. T. Thieu, A. Outtagarts, and Y. Ghamri-Doudane, "Cloudification and Autoscaling Orchestration for Container-Based Mobile Networks toward 5G: Experimentation, Challenges and Perspectives," in *Proc. IEEE Vehicular Technology Conference*, vol. 2018-June, 2018, pp. 1–7.
[7] N. F. S. De Sousa, D. A. L. Perez, R. V. Rosa, M. A. S. Santos, and C. E. Rothenberg, "Network Service Orchestration: A Survey," *Computer Communications*, 2019.
[8] X. Li, M. Samaka, A. H. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain, "Network Slicing for 5G: Challenges and Opportunities," *IEEE Internet Computing*, vol. 21, no. 5, pp. 20–27, 2017.
[9] A. Kaloxylos, "A Survey and an Analysis of Network Slicing in 5G Networks," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 60–65, 2018.
[10] T. X. Tran, A. Younis, and D. Pompili, "Understanding the Computational Requirements of Virtualized Baseband Units Using a Programmable Cloud Radio Access Network Testbed," in *Proc. 2017 IEEE International Conference on Autonomic Computing (ICAC)*, 2017, pp. 221–226.
[11] ITU-T, "GSTR-TN5G: Transport Network Support of IMT-2020/5G," *ITU Technical Report*, 2018.
[12] Y. Zhu and M. Ammar, "Algorithms for Assigning Substrate Network Resources to Virtual Network Components," in *Proc. 2006 EEE Conference on Computer Communications (INFOCOM)*, 2006.
[13] M. Chowdhury, M. R. Rahman, and R. Boutaba, "ViNEYard: Virtual Network Embedding Algorithms," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 206–219, 2012.
[14] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.
[15] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa, "End-to-end Network Slicing for 5G Mobile Networks," *Journal of Information Processing*, vol. 25, no. 0, pp. 153–163, 2017.
[16] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.
[17] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling Wireless Virtual Networks Functions," *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, pp. 240–252, 2016.
[18] P. Vizarreta, M. Condoluci, C. M. Machuca, T. Mahmoodi, and W. Kellerer, "QoS-driven Function Placement Reducing Expenditures in NFV Deployments," in *Proc. IEEE International Conference on Communications*, 2017.

[19] R. Cohen, L. Lewin-Eytan, J. S. Naor, and D. Raz, "Near Optimal Placement of Virtual Network Functions," in *Proc. 2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 1346–1354.

[20] J. F. Riera, J. Batall, F. Liberati, A. Giuseppi, A. Pietrabissa, A. Ceselli, A. Petrini, M. Trubian, P. Papadimitrou, D. Dietrich, A. Ramos, and J. Meli, "TeNOR: Steps Towards an Orchestration Platform for Multi-PoP NFV Deployment," in *Proc. IEEE NetSoft Conference and Workshops*, 2016, pp. 243–250.

[21] J. Kang, J. Kang, and O. Simeone, "On the Trade-Off between Computational Load and Reliability for Network Function Virtualization," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1767–1770, 2017.

[22] A. Fischer, J. F. Botero, M. Till Beck, H. De Meer, and X. Hesselbach, "Virtual Network Embedding: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1888–1906, 2013.

[23] M. M. Tajiki, S. Salsano, L. Chiaraviglio, M. Shojafar, and B. Akbari, "Joint Energy Efficient and QoS-aware Path Allocation and VNF Placement for Service Function Chaining," *IEEE Transactions on Network and Service Management*, no. July, pp. 1–20, 2018.

[24] N. Huin, B. Jaumard, and F. Giroire, "Optimization of Network Service Chain Provisioning," in *Proc. IEEE International Conference on Communications*, 2017.

[25] J. Liu, W. Lu, F. Zhou, P. Lu, Z. Zhu, and S. Member, "On Dynamic Service Function Chain Deployment and Readjustment," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 543–553, 2017.

[26] M. Mechtri, C. Ghribi, and D. Zeghlache, "A Scalable Algorithm for the Placement of Service Function Chains," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 533–546, 2016.

[27] H. Halabian, "Distributed Resource Allocation Optimization in 5G Virtualized Networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 3, pp. 627–642, 2019.

[28] Q.-T. Luu, M. Kieffer, A. Mouradian, and S. Kerboeuf, "Aggregated Resource Provisioning for Network Slices," in *Proc. 2018 IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, 2018.

[29] S. Chatterjee, M. J. Abdel-rahman, and A. B. Mackenzie, "Virtualization Framework for Cellular Networks with Downlink Rate Coverage Probability Constraints," in *Proc. 2018 IEEE Global Communications Conference (GLOBECOM)*, 2018.

[30] K. Teague, M. J. Abdel-rahman, and A. B. Mackenzie, "Joint Base Station Selection and Adaptive Slicing in Virtualized Wireless Networks: A Stochastic Optimization Framework," in *Proc. 2018 International Conference on Computing, Networking and Communications (ICNC)*, 2018.

[31] Y. L. Lee, J. Loo, and T. C. Chuah, "A New Network Slicing Framework for Multi-Tenant Heterogeneous Cloud Radio Access Networks," in *Proc. 2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES)*, 2016, pp. 414–420.

[32] S. D'Oro, F. Restuccia, T. Melodia, S. Member, S. Palazzo, and S. Member, "Low-Complexity Distributed Radio Access Network Slicing: Algorithms and Experimental Results," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2815 – 2828, 2018.

[33] J. Wang, K. L. Wright, and K. Gopalan, "XenLoop: A Transparent High Performance Inter-VM Network Loopback," *Cluster Computing*, vol. 12, no. 2 SPEC. ISS., pp. 141–152, 2009.

[34] I. Cerrato, F. Risso, R. Bonafiglia, K. Pentikousis, G. Pongrácz, and H. Woesner, "COMPOSER : A Compact Open-Source Service Platform," *Computer Networks*, vol. 139, pp. 151–174, 2018.

[35] M. Savi, M. Tornatore, and G. Verticale, "Impact of Processing-Resource Sharing on the Placement of Chained Virtual Network Functions," in *Proc. IEEE Conference on Network Function Virtualization and Software Defined Network (NFV-SDN)*, 2016, pp. 191–197.

[36] Y. Shi and Y. T. Hou, "Approximation Algorithm for Base Station Placement in Wireless Sensor Networks," in *Proc. 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SeCON)*, 2007, pp. 512–519.

[37] N. Bouten, R. Mijumbi, J. Serrat, J. Famaey, S. Latre, and F. De Turck, "Semantically Enhanced Mapping Algorithm for Affinity-Constrained Service Function Chain Requests," *IEEE Transactions on Network and Service Management*, vol. 14, no. 2, pp. 317–331, 2017.

[38] D. Tse and V. Pramod, *Fundamentals of Wireless Communication*, 2004.

[39] S. Sun, T. S. Rappaport, S. Rangan, T. A. Thomas, A. Ghosh, I. Z. Kovacs, I. Rodriguez, O. Koymen, A. Partyka, and J. Jarvelainen, "Propagation Path Loss Models for 5G Urban Micro- and Macro-Cellular Scenarios," in *Proc. IEEE Vehicular Technology Conference (VTC)*, 2016, pp. 1–6.

[40] ETSI, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception," *Tecnical Specification - ETSI TS 136 101 V10.21.0 (2016-04)*, 2016.