A Coverage-Aware Resource Provisioning Method for Network Slicing

Quang-Trung Luu, Sylvaine Kerboeuf, Alexandre Mouradian, Michel Kieffer

To cite this version:

Quang-Trung Luu, Sylvaine Kerboeuf, Alexandre Mouradian, Michel Kieffer. A Coverage-Aware Resource Provisioning Method for Network Slicing. IEEE/ACM Transactions on Networking, 2020, 28 (6), pp.2393-2406. 10.1109/TNET.2020.3019098 . hal-03097001

HAL Id: hal-03097001

https://centralesupelec.hal.science/hal-03097001

Submitted on 5 Jan 2021

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*—With network slicing in 5G networks, Mobile Network Operators can create various slices for Service Providers (SPs) to accommodate customized services. Usually, the various Service Function Chains (SFCs) belonging to a slice are deployed on a best-effort basis. Nothing ensures that the Infrastructure Provider (InP) will be able to allocate enough resources to cope with the increasing demands of some SP. 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.**

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 SFCs within a slice are aggregated within a graph of Slice Resource Demands (SRD). Infrastructure nodes and links have then to be provisioned so as to satisfy all SRDs. This problem leads to a Mixed Integer Linear Programming formulation. A two-step approach is considered, with several variants, depending on whether the constraints of each slice to be provisioned 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 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 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. The network slicing concept has thus emerged in 5G networks [3–5]. 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 infrastructures in a multi-tenancy context. As in [10], our work focuses on the problem of slice resource provisioning, *i.e.*, reservation. By provisioning we ensure that enough resource is reserved for further SFC deployment while satisfying coverage constraints for mobile end-users of the slice services. A two-step method is proposed for efficient slice deployment: The resource provisioning process is followed by the SFC embedding process. For the latter any state-of-the-art deployment approach may be employed on a simplified infrastructure network reduced to the nodes and links which have provisioned resources. The SFC embedding time may then be much smaller. We extend preliminary results obtained in [10], by accounting for radio coverage constraints. This requires the introduction of a radio propagation model in the radio resource provisioning phase. Moreover, coverage requirements impose several additional constraints on the SFCs to be deployed

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

Q.-T. Luu is with Nokia Bell Labs and the 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 L2S, CNRS-CentraleSupélec-Univ Paris-Sud-Univ Paris-Saclay, France, e-mail: {michel.kieffer, alexandre.mouradian}@l2s.centralesupelec.fr.

within the network infrastructure.

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. An optimal and four suboptimal variants of a coverage-aware slice resource provisioning algorithm are provided in Section V. Numerical results are presented in Section VI. Finally, Section VII 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 Providers* (InPs) own and manage the wireless and wired infrastructure such as the cell sites, the fronthaul and backhaul networks, and cloud data centers.

The *Mobile Network Operator* (MNO) leases resources from InPs to setup and manage the slices. The *Service Providers* (SPs) exploit the slices supplied by the MNO, and provide to their customers the required services that are running within the slices. Service needs are forwarded by the SP 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*. Due to user mobility, these characteristics may be time-varying. The MNO translates the SP 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.

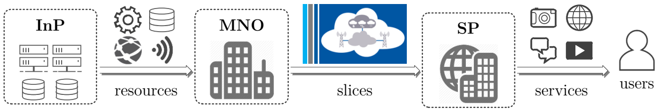


Fig. 1. System architecture

In this paper, one considers an infrastructure owned by a single InP. 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*) Base Stations (BS) providing radio resources to mobile users so as to satisfy coverage constraints, *ii*) the placement of the VNFs on the data center nodes, and *iii*) 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). Updates may be necessary when the service characteristics have changed significantly.

Our aim, with resource provisioning is to reserve, somewhat in advance, enough infrastructure resources to ensure that the MNO will have access to properly located radio resources and be able to deploy the set of SFCs with characteristics as stated in the SLA. The time scale at which provisioning is performed is much larger than that at which SFCs are deployed and adapted to meet actual time-varying user demands. One focuses on a time interval over which resources will be provisioned so as to be compliant with the variations of user demands within a slice. The duration of this time interval results from a compromise between the need to update the provisioning and the level of conservatism in the amount of provisioned resources required to satisfy fast fluctuating user demands.

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 [11]. The BBU pool hosts multiple virtual BBUs and handles 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 [12]. 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.

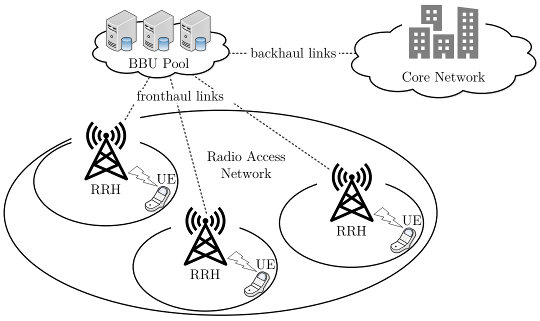


Fig. 2. General architecture of C-RAN

B. Related Work

Early results on assigning infrastructure network resources to virtual network components may be found, *e.g.*, in [13, 14]. Due to its capability of sharing efficiently network resources in 5G networks, the concept of network virtualization has gained renewed attention in the literature [5, 15–17] 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 [18, 19], 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) [19–21] or as a *Mixed Integer Linear Programming* (MILP) problem [14, 22], which are known to be NP-hard [23]. In [24], 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.*, [18–20]. For example, [18] introduced a heuristic based on the search of shortest paths to sequentially embed the SFCs. In [19], 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 [25]. With CG, 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 [25] or [26], CG has been used to relax ILP-based SFC embedding or reconfiguration problems. Specifically, in [25], the SFC embedding problem is addressed. Only core capacity and bandwidth resources for infrastructure nodes and links are considered. In [26], 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 resource 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 [27], 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, whose aim is to find, with a reduced complexity, the optimum matching between the SFC graph and the infrastructure graph. In [25], [26], and [27], 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 [28]. 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, whose objective is to maximize the total slice utility. Nevertheless, as pointed out in [28], such centralized solution 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. The results provided by all entities determines 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 [28] has been introduced in our previous work [10], 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 users 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 C-RAN 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 resources associated to the processing within the BBUs. Such resources are 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. The coverage area of RRH is considered, but the possible coverage constraints required by the slices are not taken into account.

C. Main Contributions

Compared to previous works, this paper considers 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 [10], and adapts it to the joint radio and network infrastructure resource provisioning. Constraints related to the infrastructure network considered in [10, 18, 19, 28] are combined with coverage and radio resource constraints introduced in [29–32]. The coverage constraints are very important to satisfy mobile service requirements. The amount of radio resources required depends on the location of users. A radio propagation model is thus introduced in the provisioning phase. The coverage constraints reduce the flexibility to select the nodes on which SFCs are deployed.

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 SP. 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 [18, 19], where each VNF is deployed on a single node. In [18, 19], 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 whose aim is to provide different services, indexed by $\sigma = 1, \dots, |\mathcal{S}|$, to mobile users. The geographical area under study is denoted by $\mathcal{A}$ and the subarea over which service σ has to be made available is denoted by $\mathcal{A}^{\sigma}$. For that purpose, each SP forwards his service requirements to the MNO, whose aim is to design a network slice able to satisfy these requirements. Figure 3 illustrates three typical geographical subareas over which three different services have to be deployed. 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 whose resource demands have been described by the SRD graph.

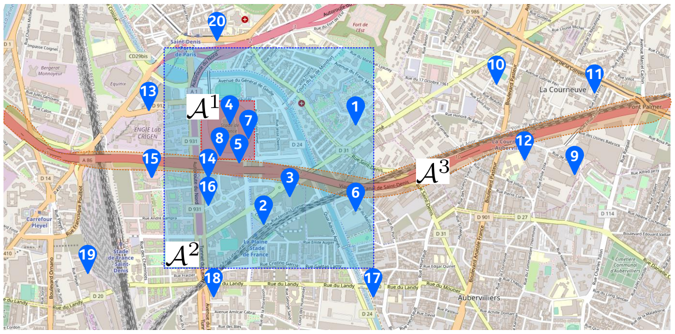


Fig. 3. The considered metropolitan area including the Stade de France (covered by the red rectangle representing $\mathcal{A}^1$), its surrounding (blue rectangle representing $\mathcal{A}^2$), and part of the A86 highway (orange shape representing $\mathcal{A}^3$). Blue markers show the location of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

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 InP. This network is represented by a directed graph $\mathcal{G}_{\text{I}} = (\mathcal{N}_{\text{I}}, \mathcal{E}_{\text{I}})$, where $\mathcal{N}_{\text{I}}$ is the set of infrastructure nodes and $\mathcal{E}_{\text{I}}$ 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 $i \in \mathcal{N}_{\text{I}}$ is characterized by a given amount of computing and storage resources, denoted as $a_{\text{c}}(i)$ and $a_{\text{s}}(i)$, which may be allocated to network slices. Radio resources are exclusively provided by a subset $\mathcal{N}_{\text{Ir}} \subset \mathcal{N}_{\text{I}}$ of RRH nodes, whose location in some Cartesian frame attached to $\mathcal{A}$ is denoted by x_i^{r}. The cost associated to the use of an infrastructure node i consists of a fixed part $c_{\text{f}}(i)$ for node disposal (paid by each slice using node i) and a variable part $c_{\text{c}}(i)$, $c_{\text{s}}(i)$, and $c_{\text{r}}(i)$, which depend linearly on the amount of computing, storage, and radio resources provided by that node.

Each infrastructure link $ij \in \mathcal{E}_{\text{I}}$ connecting node i to j has a bandwidth $a_{\text{b}}(ij)$, and an associated per-unit bandwidth cost $c_{\mathrm{b}}(ij)$. 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 $c_{\mathrm{b}}\left(ii\right)$.

B. SRD Model

An SRD is defined on the basis of an SLA between an SP and the MNO. The SLA may consider several time intervals over each of which the service characteristics and constraints are assumed constant, but may vary from one interval to the next one. These time intervals translate, *e.g.*, day and night variations of user demands. They last between tens of minutes and hours. It is of the responsibility of the SP and MNO to properly scale the requirements expressed in the SLA, by considering, for example, similar services deployed in the past.

In this paper, one considers a given time interval specified in the SLA. The SLA is also expressed in terms of supported service type and 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 of each client. The geographical distribution function $\rho^{\sigma}(x)$, with $x \in \mathcal{A}$, describes the *maximum* user/device density to be served around x within the considered time interval.

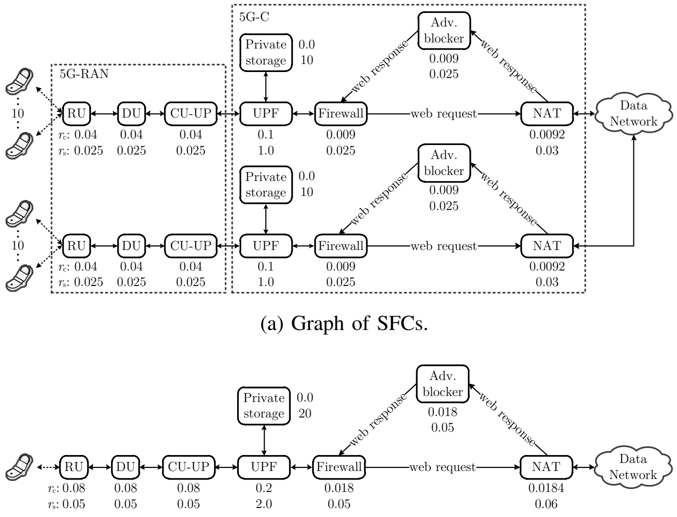
One assumes that the resource requirements for a slice can be represented by an SRD graph that mimics the graph of SFCs. The SRD graph for slice σ is an oriented graph denoted by $\mathcal{G}_{\mathrm{V}}^{\sigma} = (\mathcal{N}_{\mathrm{V}}^{\sigma}, \mathcal{E}_{\mathrm{V}}^{\sigma})$, where $\mathcal{N}_{\mathrm{V}}^{\sigma}$ 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 as $r_{\mathrm{c}}(v)$ and $r_{\mathrm{s}}(v)$ to sustain the aggregated demand for all instances of a given VNF in the slice. The minimum resources to deploy a single VNF instance are denoted as $\underline{r}_{\mathrm{c}}(v)$ and $\underline{r}_{\mathrm{s}}(v)$. 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 at least one VNF.

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

$$
\begin{aligned}
r_{\mathrm{u}}\left(v_{\mathrm{r}}\right) &= \underline{R}_{\mathrm{u}}^{\sigma} \int_{\mathcal{A}^{\sigma}} \rho^{\sigma}\left(x\right) \mathrm{d}x, \\
r_{\mathrm{d}}\left(v_{\mathrm{r}}\right) &= \underline{R}_{\mathrm{d}}^{\sigma} \int_{\mathcal{A}^{\sigma}} \rho^{\sigma}\left(x\right) \mathrm{d}x. \qquad (1)
\end{aligned}
$$

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 RU to handle RF operations, a DU, and a *Centralized Unit for User-Plane* (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) function. 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 whose 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. More details are provided in Section VI. 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.



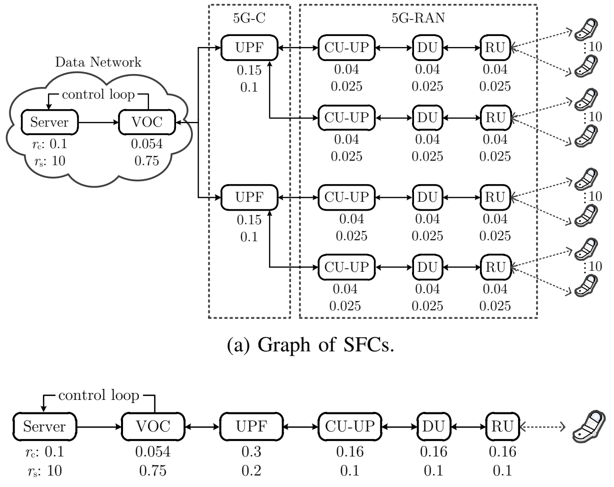
(a) Graph of SFCs.

(b) Corresponding SRD graph.

Fig. 4. SFCs and their required computing (in CPUs) and storage (in GBytes) resources for the deployment of a secured web browsing service with advertisement removal and their associated SRD graph.

When it is possible to reserve enough resources, the MNO will be ensured to be able to deploy a collection of SFCs needed to satisfy the SLA over its time interval of validity. When, for example, the user density over some subarea is larger than stated in the SLA, some users may not be served. Nevertheless, from the perspective of the InP, the SLA is still satisfied. On contrary, when the user density/requirements are less than the maximum specified in the SLA, some provisioned resources may remain unused, but this is the price to pay when provisioning resources.

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



(a) Graph of SFCs.

(b) Corresponding SRD graph.

Fig. 5. SFCs and their required computing (in CPUs) and storage (in GBytes) resources for the deployment of an adaptive wireless video streaming service and their associated SRD graph.

TABLE I
INFRASTRUCTURE NETWORK AND SLICE PARAMETERS.

Node resource type: n	
n	computing (c), storage (s), and radio (r)
Infrastructure network graph: $\mathcal{G}_{\mathrm{I}} = (\mathcal{N}_{\mathrm{I}}, \mathcal{E}_{\mathrm{I}})$	
$\mathcal{N}_{\mathrm{I}}$	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

IV. PROBLEM FORMULATION

The 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 that 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 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 physical 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.

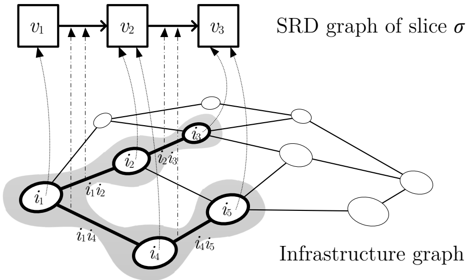


Fig. 6. Provisioning of infrastructure resource to an SRD graph: Resources from the infrastructure node i_1 is provisioned for SRD node v_1; Resources from i_2 and i_4 are provisioned for SRD node v_2; and resources from i_3 and i_5 are provisioned for SRD node v_3. Correspondingly, the infrastructure links i_1i_2 and i_1i_4 are provisioned for SRD link v_1v_2 and resources from links i_2i_3 and i_4i_5 are provisioned for SRD link v_2v_3.

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 $\mathcal{A}_q^{\sigma}$, $q \in \mathcal{Q}^{\sigma} = \{1, \dots, Q^{\sigma}\}$. 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 by $\eta_{\mathrm{u}}^{\sigma}(i,q) \in [0,1]$ and $\eta_{\mathrm{d}}^{\sigma}(i,q) \in [0,1]$ for uplink and downlink traffic, respectively. These quantities represent average proportions of RBs available during some typical interval of time and provisioned by RRH i. The time interval may be, *e.g.*, of one second[1]. The summed proportions of RBs provided by a given RRH i must be less than one

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

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

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

$$\sum_{i \in \mathcal{N}_{\mathrm{Ir}}} \eta_{\mathrm{d}}^{\sigma}(i,q)\, a_{\mathrm{r}}(i)\, b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) \geqslant \underline{R}_{\mathrm{d}}^{\sigma} \int_{\mathcal{A}_q^{\sigma}} \rho^{\sigma}(x)\, \mathrm{d}x, \qquad (4)$$

[1] Since $\eta_{\mathrm{u}}^{\sigma}(i,q)$ and $\eta_{\mathrm{d}}^{\sigma}(i,q)$ are averages, they may be accurately represented by real numbers in the interval $[0,1]$, even if in reality both quantities should be rational numbers.

which correspond to the satisfaction of the geographical coverage constraints for uplink and downlink traffic. 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 $\int_{\mathcal{A}_q^\sigma}\rho^\sigma\left(x\right)\mathrm{d}x$ 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 VI-A3. Summing (3) over all $q\in\mathcal{Q}^\sigma$ and using (1), one gets

$$\sum_{q\in\mathcal{Q}^\sigma}\sum_{i\in\mathcal{N}_{\mathrm{Ir}}}\eta_{\mathrm{u}}^\sigma\left(i,q\right)a_{\mathrm{r}}\left(i\right)b_{\mathrm{u}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^\sigma\right)\geqslant r_{\mathrm{u}}\left(v_{\mathrm{r}}\right), \tag{5}$$

$$\sum_{q\in\mathcal{Q}^\sigma}\sum_{i\in\mathcal{N}_{\mathrm{Ir}}}\eta_{\mathrm{d}}^\sigma\left(i,q\right)a_{\mathrm{r}}\left(i\right)b_{\mathrm{d}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^\sigma\right)\geqslant r_{\mathrm{d}}\left(v_{\mathrm{r}}\right), \tag{6}$$

which ensure, 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 $r_{\mathrm{u}}\left(v_{\mathrm{r}}\right)$ and $r_{\mathrm{d}}\left(v_{\mathrm{r}}\right)$. 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 $q\in\mathcal{Q}^\sigma$

$$\frac{\eta_{\mathrm{u}}^\sigma\left(i,q\right)a_{\mathrm{r}}\left(i\right)b_{\mathrm{u}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^\sigma\right)}{r_{\mathrm{u}}\left(v_{\mathrm{r}}\right)}=\frac{\eta_{\mathrm{d}}^\sigma\left(i,q\right)a_{\mathrm{r}}\left(i\right)b_{\mathrm{d}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^\sigma\right)}{r_{\mathrm{d}}\left(v_{\mathrm{r}}\right)}. \tag{7}$$

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 $\widetilde{\eta}^\sigma\left(i\right)\in\{0,1\}$, with $\widetilde{\eta}^\sigma\left(i\right)=1$ if $\sum_{q\in\mathcal{Q}^\sigma}\eta^\sigma\left(i,q\right)>0$, and $\widetilde{\eta}^\sigma\left(i\right)=0$ otherwise. The variables $\eta^\sigma\left(i,q\right)$ and $\widetilde{\eta}^\sigma\left(i\right)$ are gathered in the sets $\boldsymbol{\eta}^\sigma=\{\eta^\sigma\left(i,q\right)\}_{i\in\mathcal{N}_{\mathrm{Ir}},q\in\mathcal{Q}^\sigma}$ and $\widetilde{\boldsymbol{\eta}}^\sigma=\{\widetilde{\eta}^\sigma\left(i\right)\}_{i\in\mathcal{N}_{\mathrm{Ir}}}$, which are components of the sets $\boldsymbol{\eta}=\{\boldsymbol{\eta}^\sigma\}_{\sigma\in\mathcal{S}}$ and $\widetilde{\boldsymbol{\eta}}=\{\widetilde{\boldsymbol{\eta}}^\sigma\}_{\sigma\in\mathcal{S}}$. The relation between $\eta^\sigma\left(i,q\right)$ and $\widetilde{\eta}^\sigma\left(i\right)$ is nonlinear. Nevertheless, both quantities can be linked with the following linear constraints, $\forall\sigma\in\mathcal{S}$, $\forall i\in\mathcal{N}_{\mathrm{Ir}}$,

$$0\leq\widetilde{\eta}^\sigma\left(i\right)-\sum_{q\in\mathcal{Q}^\sigma}\eta^\sigma\left(i,q\right)<1, \tag{8}$$

with

$$\eta^\sigma\left(i,q\right)=\eta_{\mathrm{u}}^\sigma\left(i,q\right)+\eta_{\mathrm{d}}^\sigma\left(i,q\right). \tag{9}$$

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

$$c_{\mathrm{rr}}\left(\boldsymbol{\eta},\widetilde{\boldsymbol{\eta}}\right)=\sum_{\sigma\in\mathcal{S}}c_{\mathrm{rr}}^\sigma\left(\boldsymbol{\eta}^\sigma,\widetilde{\boldsymbol{\eta}}^\sigma\right), \tag{10}$$

where

$$\begin{aligned}c_{\mathrm{rr}}^\sigma\left(\boldsymbol{\eta}^\sigma,\widetilde{\boldsymbol{\eta}}^\sigma\right)&=\sum_{i\in\mathcal{N}_{\mathrm{Ir}}}c_{\mathrm{f}}\left(i\right)\widetilde{\eta}^\sigma\left(i\right)\\&+\sum_{\sigma\in\mathcal{S}}\sum_{i\in\mathcal{N}_{\mathrm{Ir}}}\sum_{q\in\mathcal{Q}^\sigma}\left[c_{\mathrm{r}}\left(i\right)-\lambda b_{\mathrm{u}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^\sigma\right)\right]a_{\mathrm{r}}\left(i\right)\eta_{\mathrm{u}}^\sigma\left(i,q\right)\\&+\sum_{\sigma\in\mathcal{S}}\sum_{i\in\mathcal{N}_{\mathrm{Ir}}}\sum_{q\in\mathcal{Q}^\sigma}\left[c_{\mathrm{r}}\left(i\right)-\lambda b_{\mathrm{d}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^\sigma\right)\right]a_{\mathrm{r}}\left(i\right)\eta_{\mathrm{d}}^\sigma\left(i,q\right)\end{aligned} \tag{11}$$

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 $\mathbf{\Phi}_{\mathrm{n}}^\sigma=\{\phi_n^\sigma(i,v)\}_{i\in\mathcal{N}_{\mathrm{I}},v\in\mathcal{N}_{\mathrm{V}}^\sigma,n\in\{\mathrm{c},\mathrm{s}\}}$, 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 $\mathbf{\Phi}_{\mathrm{b}}^\sigma=\{\phi_{\mathrm{b}}^\sigma(ij,vw)\}_{ij\in\mathcal{E}_{\mathrm{I}},vw\in\mathcal{E}_{\mathrm{V}}^\sigma}$, 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. The sets $\mathbf{\Phi}_{\mathrm{n}}=\{\mathbf{\Phi}_{\mathrm{n}}^\sigma\}_{\sigma\in\mathcal{S}}$ and $\mathbf{\Phi}_{\mathrm{b}}=\{\mathbf{\Phi}_{\mathrm{b}}^\sigma\}_{\sigma\in\mathcal{S}}$ 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 $\forall\sigma\in\mathcal{S}$ to

$$\sum_{i\in\mathcal{N}_{\mathrm{I}}}a_n\left(i\right)\phi_n^\sigma(i,v)\geq r_n(v),\ \forall n\in\{\mathrm{c},\mathrm{s}\},\forall v\in\mathcal{N}_{\mathrm{V}}^\sigma. \tag{12}$$

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

$$\sum_{\sigma\in\mathcal{S}}\sum_{v\in\mathcal{N}_{\mathrm{V}}^\sigma}\phi_n^\sigma(i,v)\leq 1,\ \forall n\in\{\mathrm{c},\mathrm{s}\},\forall i\in\mathcal{N}_{\mathrm{I}}. \tag{13}$$

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

$$\sum_{\sigma\in\mathcal{S}}\sum_{vw\in\mathcal{E}_{\mathrm{V}}^\sigma}\phi_{\mathrm{b}}^\sigma(ij,vw)\leq 1,\ \forall ij\in\mathcal{E}_{\mathrm{I}}. \tag{14}$$

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 $\underline{r}_n\left(v\right)$ for a VNF associated to the SRD node v

$$\begin{aligned}&a_n(i)\phi_n^\sigma(i,v)=\underline{r}_n(v)\kappa_n^\sigma(i,v),\\&\forall i\in\mathcal{N}_{\mathrm{I}},\forall v\in\mathcal{N}_{\mathrm{V}}^\sigma,\forall n\in\{\mathrm{c},\mathrm{s}\},\end{aligned} \tag{15}$$

where $\kappa_n^\sigma(i,v)$ is a positive integer belonging to the set of variables of the optimization problem. This ensures that enough resources are provisioned by an infrastructure node i to be able to deploy an integer number $\kappa_n^\sigma(i,v)$ of VNF instances associated to the SRD node v.

It is usually difficult, if not impossible, when deploying a given VNF, to benefit from the storage of one infrastructure node and from the computing resources of an other infrastructure node. Consequently, resources of each type have to be provisioned in a balanced way by an infrastructure node for an SRD node, consistently with the requirements of the SRD node. This ensures to be able to deploy a VNF on a single infrastructure node. For example, if an infrastructure node provides 10% of the computing demand of a given SRD node, it should also provide 10% of its storage demand. This translates into the following resource provisioning proportionality constraints $\forall \sigma \in \mathcal{S}$,

$$\frac{a_{\mathrm{c}}(i)}{r_{\mathrm{c}}(v)}\phi_{\mathrm{c}}^{\sigma}(i,v) = \frac{a_{\mathrm{s}}(i)}{r_{\mathrm{s}}(v)}\phi_{\mathrm{s}}^{\sigma}(i,v),\ \forall i \in \mathcal{N}_{\mathrm{I}}, \forall v \in \mathcal{N}_{\mathrm{V}}^{\sigma}. \tag{16}$$

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, $\forall \sigma \in \mathcal{S}$ and $\forall i \in \mathcal{N}_{\mathrm{Ir}}$,

$$\begin{aligned}\frac{a_{\mathrm{c}}(i)}{r_{\mathrm{c}}(v_{\mathrm{r}})}\phi_{\mathrm{c}}^{\sigma}(i,v_{\mathrm{r}}) = \frac{a_{\mathrm{s}}(i)}{r_{\mathrm{s}}(v_{\mathrm{r}})}\phi_{\mathrm{s}}^{\sigma}(i,v_{\mathrm{r}}) = \frac{a_{\mathrm{r}}(i)}{r_{\mathrm{r}}(v_{\mathrm{r}})}\\ \times \sum_{q\in\mathcal{Q}^{\sigma}}\left(\eta_{\mathrm{u}}^{\sigma}(i,q)\,b_{\mathrm{u}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^{\sigma}\right)+\eta_{\mathrm{d}}^{\sigma}(i,q)\,b_{\mathrm{d}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^{\sigma}\right)\right).\end{aligned} \tag{17}$$

The constraints (16) and (17) ensure a balanced resource provisioning by infrastructure nodes. In (17), $r_{\mathrm{r}}(v_{\mathrm{r}})$ 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 $\forall \sigma \in \mathcal{S}$, $\forall j \in \mathcal{N}_{\mathrm{Ir}}$, $\forall v v_{\mathrm{r}} \in \mathcal{E}_{\mathrm{V}}^{\sigma}$,

$$\begin{aligned}\sum_{i\in\mathcal{N}_{\mathrm{I}}\backslash\mathcal{N}_{\mathrm{Ir}}}\frac{a_{\mathrm{b}}(ij)}{r_{\mathrm{b}}(vv_{\mathrm{r}})}\phi_{\mathrm{b}}^{\sigma}(ij,vv_{\mathrm{r}}) = \left(\frac{r_{\mathrm{b}}(vv_{\mathrm{r}})}{\sum_{uv_{\mathrm{r}}\in\mathcal{E}_{\mathrm{V}}^{\sigma}} r_{\mathrm{b}}(uv_{\mathrm{r}})}\right)\frac{a_{\mathrm{r}}(j)}{r_{\mathrm{d}}(v_{\mathrm{r}})}\\ \times\sum_{q\in\mathcal{Q}^{\sigma}}\eta_{\mathrm{d}}^{\sigma}(j,q)\,b_{\mathrm{d}}\left(x_j^{\mathrm{r}},\mathcal{A}_q^{\sigma}\right)\end{aligned} \tag{18}$$

In (18), 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 (18) 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 (18), represents the proportion of the data traffic that *is provided* by all infrastructure links ij, $i \in \mathcal{N}_{\mathrm{I}}\backslash\mathcal{N}_{\mathrm{Ir}}$ for the SRD link vv_{r}. Both terms have thus to be equal.

For uplink traffic (links with RRH as ingress), one has, $\forall \sigma \in \mathcal{S}$, $\forall i \in \mathcal{N}_{\mathrm{Ir}}$, $\forall v_{\mathrm{r}} v \in \mathcal{E}_{\mathrm{V}}^{\sigma}$,

$$\begin{aligned}\sum_{j\in\mathcal{N}_{\mathrm{I}}\backslash\mathcal{N}_{\mathrm{Ir}}}\frac{a_{\mathrm{b}}(ij)}{r_{\mathrm{b}}(v_{\mathrm{r}}v)}\phi_{\mathrm{b}}^{\sigma}(ij,v_{\mathrm{r}}v) = \left(\frac{r_{\mathrm{b}}(v_{\mathrm{r}}v)}{\sum_{v_{\mathrm{r}}u\in\mathcal{E}_{\mathrm{V}}^{\sigma}} r_{\mathrm{b}}(v_{\mathrm{r}}u)}\right)\frac{a_{\mathrm{r}}(i)}{r_{\mathrm{u}}(v_{\mathrm{r}})}\\ \times\sum_{q\in\mathcal{Q}^{\sigma}}\eta_{\mathrm{u}}^{\sigma}(i,q)\,b_{\mathrm{u}}\left(x_i^{\mathrm{r}},\mathcal{A}_q^{\sigma}\right)\end{aligned} \tag{19}$$

In (19), 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 (19) 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 (19), 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. Combined with (6), the constraints (18) and (19) impose that the total radio resources provisioned by the RRHs are above the required resources $r_{\mathrm{d}}(v_{\mathrm{r}})$ and $r_{\mathrm{u}}(v_{\mathrm{r}})$.

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 $\forall \sigma \in \mathcal{S}$, $\forall i \in \mathcal{N}_{\mathrm{I}}$, $\forall vw \in \mathcal{E}_{\mathrm{V}}^{\sigma}$,

$$\frac{a_{\mathrm{b}}(ii)}{r_{\mathrm{b}}(vw)}\phi_{\mathrm{b}}^{\sigma}(ii,vw) = \left(\frac{r_{\mathrm{b}}(vw)}{\sum_{vu\in\mathcal{E}_{\mathrm{V}}^{\sigma}} r_{\mathrm{b}}(vu)}\right)\frac{a_{\mathrm{c}}(i)}{r_{\mathrm{c}}(v)}\phi_{\mathrm{c}}^{\sigma}(i,v) \tag{20}$$

$$= \left(\frac{r_{\mathrm{b}}(vw)}{\sum_{uw\in\mathcal{E}_{\mathrm{V}}^{\sigma}} r_{\mathrm{b}}(uw)}\right)\frac{a_{\mathrm{c}}(i)}{r_{\mathrm{c}}(w)}\phi_{\mathrm{c}}^{\sigma}(i,w), \tag{21}$$

In (20), the term $a_{\mathrm{c}}(i)\phi_{\mathrm{c}}^{\sigma}(i,v)/r_{\mathrm{c}}(v)$ represents the proportion of computing resource 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 (20) 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 (19) represents the proportion of the data traffic that *is provided* by the internal link ii for the SRD link vw. The constraint (21) 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 $\forall \sigma \in \mathcal{S}$

$$\sum_{j\in\mathcal{N}_{\mathrm{I}}}\frac{a_{\mathrm{b}}(ij)}{r_{\mathrm{b}}(vw)}\phi_{\mathrm{b}}^{\sigma}(ij,vw) = \left(\frac{r_{\mathrm{b}}(vw)}{\sum_{vu\in\mathcal{E}_{\mathrm{V}}^{\sigma}} r_{\mathrm{b}}(vu)}\right)\frac{a_{\mathrm{c}}(i)}{r_{\mathrm{c}}(v)}\phi_{\mathrm{c}}^{\sigma}(i,v), \tag{22}$$

The right-hand side of (22) is the same as that of (20). The left-hand side of (22) 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 $\forall \sigma \in \mathcal{S}$

$$\sum_{j\in\mathcal{N}_\mathrm{I}} \frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ji,vw) = \left(\frac{r_\mathrm{b}(vw)}{\sum_{uw\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(uw)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(w)}\phi_\mathrm{c}^\sigma(i,w) \tag{23}$$

The right-hand side of (23) is the same as that of (21). The left-hand side of (23) 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

$$\sum_{j\in\mathcal{N}_\mathrm{I}} \left[\frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ij,vw) - \frac{a_\mathrm{b}(ji)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ji,vw)\right] = 0 \tag{24}$$

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

The constraints (22-24) can be gathered in the following single constraint, which should be valid, $\forall \sigma \in \mathcal{S}$, $\forall i \in \mathcal{N}_\mathrm{I}$, $\forall vw \in \mathcal{E}_\mathrm{V}^\sigma$,

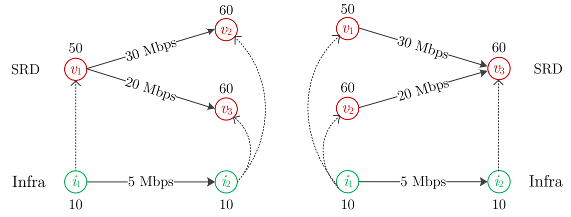
$$\begin{aligned}
&\sum_{j\in\mathcal{N}_\mathrm{I}} \left[\frac{a_\mathrm{b}(ij)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ij,vw) - \frac{a_\mathrm{b}(ji)}{r_\mathrm{b}(vw)}\phi_\mathrm{b}^\sigma(ji,vw)\right] \\
&= \left(\frac{r_\mathrm{b}(vw)}{\sum_{vu\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(vu)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(v)}\phi_\mathrm{c}^\sigma(i,v) \\
&- \left(\frac{r_\mathrm{b}(vw)}{\sum_{uw\in\mathcal{E}_\mathrm{V}^\sigma} r_\mathrm{b}(uw)}\right)\frac{a_\mathrm{c}(i)}{r_\mathrm{c}(w)}\phi_\mathrm{c}^\sigma(i,w),
\end{aligned} \tag{25}$$

In (20), (21), and (25), the consistency with the other provisioned resources is ensured by (16).

Note that the flow conservation constraints (25) imposes a relation between the $\phi_n^\sigma(i,v)$s for different i and v. Since $\phi_n^\sigma(i,v)$ and $\kappa_n^\sigma(i,v)$ are proportional, see (15), the relations between $\kappa_n^\sigma(i,v)$ for different i and v are also imposed without specifying any additional constraint.

Figure 7 illustrates 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 $i_1 i_2$ is mapped onto the pair $(v_1 v_2, v_1 v_3)$. Considering the infrastructure node i_1 and the SRD link $v_1 v_2$, the constraint (25) leads to $\frac{5}{30}\phi_\mathrm{b}(i_1 i_2, v_1 v_2) - 0 = \frac{30}{50}\frac{10}{50}\phi_n(i_1, v_1) - 0$, hence $\phi_\mathrm{b}(i_1 i_2, v_1 v_2) = \frac{18}{25}\phi_n(i_1, v_1)$. Similarly, considering i_1 and $v_1 v_3$, one gets $\phi_\mathrm{b}(i_1 i_2, v_1 v_3) = \frac{8}{25}\phi_n(i_1, v_1)$. 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 $\phi_\mathrm{b}(i_1 i_2, v_1 v_2) = \frac{8}{26}$, and $\phi_\mathrm{b}(i_1 i_2, v_1 v_3) = \frac{18}{26}$. Considering $\phi_n(i_1, v_1) = 1$ would lead to $\phi_\mathrm{b}(i_1 i_2, v_1 v_2) + \phi_\mathrm{b}(i_1 i_2, v_1 v_3) = \frac{26}{25} > 1$, which is not consistent with (14).

In Figure 7b, an SRD graph with a merge topology is depicted. Through similar calculations, one gets $\phi_n(i_1, v_1) = \frac{1}{2}$, $\phi_n(i_1, v_2) = \frac{2}{5}$, $\phi_n(i_2, v_3) = 1$, $\phi_\mathrm{b}(i_1 i_2, v_1 v_3) = \frac{9}{15}$, and $\phi_\mathrm{b}(i_1 i_2, v_2 v_3) = \frac{4}{15}$. 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 (25) 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 of slice σ, one introduces the sets of binary variables $\widetilde{\boldsymbol{\Phi}}^\sigma = \left\{\widetilde{\phi}^\sigma(i)\right\}_{i\in\mathcal{N}_\mathrm{I}}$ and $\widetilde{\boldsymbol{\Phi}} = \left\{\widetilde{\boldsymbol{\Phi}}^\sigma\right\}_{\sigma\in\mathcal{S}}$, which represents node mapping indicators, *i.e.*, $\widetilde{\phi}^\sigma(i) = 1$ if at least one of the elements of $\{\phi_\mathrm{c}^\sigma(i,v), \phi_\mathrm{s}^\sigma(i,v)\}_{v\in\mathcal{N}_\mathrm{V}^\sigma}$ is strictly positive, and $\widetilde{\phi}^\sigma(i) = 0$ otherwise. The relation between $\phi_n^\sigma(i,v)$ and $\widetilde{\phi}^\sigma(i)$ is again nonlinear. As in (8), both quantities may be linearly related as follows, $\forall i \in \mathcal{N}_\mathrm{I}, \forall \sigma \in \mathcal{S}$,

$$\sum_{v\in\mathcal{N}_\mathrm{V}^\sigma}\sum_{n\in\{\mathrm{c,s}\}} \frac{\phi_n^\sigma(i,v)}{2\left|\mathcal{N}_\mathrm{V}^\sigma\right|} \leq \widetilde{\phi}^\sigma(i) < \sum_{v\in\mathcal{N}_\mathrm{V}^\sigma}\sum_{n\in\{\mathrm{c,s}\}} \frac{\phi_n^\sigma(i,v)}{2\left|\mathcal{N}_\mathrm{V}^\sigma\right|} + 1. \tag{26}$$

The leasing 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

$$c_\mathrm{wr}\left(\boldsymbol{\Phi}_\mathrm{n}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_\mathrm{b}\right) = \sum_{\sigma\in\mathcal{S}} c_\mathrm{wr}^\sigma\left(\boldsymbol{\Phi}_\mathrm{n}^\sigma, \widetilde{\boldsymbol{\Phi}}^\sigma, \boldsymbol{\Phi}_\mathrm{b}^\sigma\right), \tag{27}$$

with

$$\begin{aligned}
c_\mathrm{wr}^\sigma\left(\boldsymbol{\Phi}_\mathrm{n}^\sigma, \widetilde{\boldsymbol{\Phi}}^\sigma, \boldsymbol{\Phi}_\mathrm{b}^\sigma\right) = &\sum_{i\in\mathcal{N}_\mathrm{I}\backslash\mathcal{N}_\mathrm{Ir}} \widetilde{\phi}^\sigma(i) c_\mathrm{f}(i) \\
&+ \sum_{i\in\mathcal{N}_\mathrm{I}}\sum_{v\in\mathcal{N}_\mathrm{V}^\sigma}\sum_{n\in\{\mathrm{c,s}\}} a_n(i)\phi_n^\sigma(i,v)c_n(i) \\
&+ \sum_{ij\in\mathcal{E}_\mathrm{I}}\sum_{vw\in\mathcal{E}_\mathrm{V}^\sigma} a_\mathrm{b}(ij)\phi_\mathrm{b}^\sigma(ij,vw)c_\mathrm{b}(ij),
\end{aligned} \tag{28}$$

where the first term represents the cost for deploying VNFs in infrastructure nodes, while the second and the third term indicate the total cost for leasing resources from infrastructure nodes and links. In the first term, the fixed infrastructure node disposal cost related to RRH nodes is not considered, since it has already been taken into account in (10).

V. 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 (10) and (27)

$$c_\mathrm{tot}\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}, \boldsymbol{\Phi}_\mathrm{n}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_\mathrm{b}\right) = c_\mathrm{rr}\left(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}\right) + c_\mathrm{wr}\left(\boldsymbol{\Phi}_\mathrm{n}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_\mathrm{b}\right) \tag{29}$$

with the constraints introduced in Sections IV-A and IV-B. The provisioning algorithm minimizing (29) and considering all slices jointly is denoted as JRN (Joint Radio and Network provisioning).

When the number of variables in $(\boldsymbol{\Phi}_{\text{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\text{b}})$ and $(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ increases, the problem may become intractable. Therefore, a two-step provisioning algorithm, denoted as CARP (Coverage-Aware Resource Provisioning), see Algorithm 1, is introduced where both terms of (29) are minimized separately. The *Radio resource Provisioning* problem, denoted by RP, involving the radio coverage constrains introduced in Section IV-A, is solved first. Then, the *Network resource Provisioning*, denoted by NP, is solved using using the solution of the RP problem and considering the other resource constraints introduced in Section IV-B.

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. Let SR and JR denote the sequential and joint RP, and similarly SN and JN denote the sequential and joint NP.

Algorithm 1: Coverage-Aware Resource Provisioning

```
Input: $\mathcal{G}_{\text{I}} = (\mathcal{N}_{\text{I}}, \mathcal{E}_{\text{I}}), \mathcal{S}, \{\mathcal{G}_{\text{V}}^{\sigma}, \sigma \in \mathcal{S}\}, \{\mathcal{A}^{\sigma}, \sigma \in \mathcal{S}\}$
Output: $(\widehat{\boldsymbol{\eta}}, \widehat{\widetilde{\boldsymbol{\eta}}})$ and $(\widehat{\boldsymbol{\Phi}}_{\text{n}}, \widehat{\widetilde{\boldsymbol{\Phi}}}, \widehat{\boldsymbol{\Phi}}_{\text{e}})$
1  # Radio resource provisioning - JR variant
2  Evaluate $(\widehat{\boldsymbol{\eta}}, \widehat{\widetilde{\boldsymbol{\eta}}}) = \arg\min_{\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}}} c_{\text{rr}}(\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}})$,
3  subject to: (2),
4          (3)-(4), (7), $\forall \sigma \in \mathcal{S}, \forall q \in \mathcal{Q}^{\sigma}$,
5          (5)-(6), $\forall \sigma \in \mathcal{S}$,
6          (8), $\forall i \in \mathcal{N}_{\text{Ir}}$, $\forall \sigma \in \mathcal{S}$.
7  # Radio resource provisioning - SR variant
8  for $\sigma \in \mathcal{S}$ do
9      Evaluate $(\widehat{\boldsymbol{\eta}}^{\sigma}, \widehat{\widetilde{\boldsymbol{\eta}}}^{\sigma}) = \arg\min_{\boldsymbol{\eta}^{\sigma}, \tilde{\boldsymbol{\eta}}^{\sigma}} c_{\text{rr}}^{\sigma}(\boldsymbol{\eta}^{\sigma}, \tilde{\boldsymbol{\eta}}^{\sigma})$,
10     subject to:
11         $\sum_{\sigma' \leqslant \sigma} \sum_{q \in \mathcal{Q}^{\sigma'}} \left( \eta_{\text{u}}^{\sigma'}(i, q) + \eta_{\text{d}}^{\sigma'}(i, q) \right) \leqslant 1, \forall i \in \mathcal{N}_{\text{Ir}}$,
12         (3)-(4), (7), $\forall q \in \mathcal{Q}^{\sigma}$,
13         (5)-(6),
14         (8), $\forall i \in \mathcal{N}_{\text{Ir}}$.
15 # Other network resource provisioning - JN variant
16 Evaluate $(\widehat{\boldsymbol{\Phi}}_{\text{n}}, \widehat{\widetilde{\boldsymbol{\Phi}}}, \widehat{\boldsymbol{\Phi}}_{\text{b}}) = \arg\min_{\boldsymbol{\Phi}_{\text{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\text{b}}} c_{\text{wr}}(\boldsymbol{\Phi}_{\text{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\text{b}})$
17 subject to:
18         (12), $\forall \sigma \in \mathcal{S}$
19         (13), (14),
20         (15)-(21), (25), $\forall \sigma \in \mathcal{S}$,
21         (26), $\forall i \in \mathcal{N}_{\text{I}}$, $\forall \sigma \in \mathcal{S}$.
22 # Other network resource provisioning - SN variant
23 for $\sigma \in \mathcal{S}$ do
24     Evaluate
           $(\widehat{\boldsymbol{\Phi}}_{\text{n}}^{\sigma}, \widehat{\widetilde{\boldsymbol{\Phi}}}^{\sigma}, \widehat{\boldsymbol{\Phi}}_{\text{b}}^{\sigma}) = \arg\min_{\boldsymbol{\Phi}_{\text{n}}^{\sigma}, \widetilde{\boldsymbol{\Phi}}^{\sigma}, \boldsymbol{\Phi}_{\text{b}}^{\sigma}} c_{\text{wr}}^{\sigma}(\boldsymbol{\Phi}_{\text{n}}^{\sigma}, \widetilde{\boldsymbol{\Phi}}^{\sigma}, \boldsymbol{\Phi}_{\text{b}}^{\sigma})$
25     subject to: (12),
26         $\sum_{\sigma' \leqslant \sigma} \sum_{v \in \mathcal{N}_{\text{V}}^{\sigma}} \phi_{n}^{\sigma'}(i, v) \leq 1, \forall n \in \{\text{c}, \text{s}\}, \forall i \in \mathcal{N}_{\text{I}}$,
27         $\sum_{\sigma' \leqslant \sigma} \sum_{vw \in \mathcal{E}_{\text{V}}^{\sigma'}} \phi_{\text{b}}^{\sigma'}(ij, vw) \leq 1, \forall ij \in \mathcal{E}_{\text{I}}$,
28         (15)-(21), (25),
29         (26), $\forall i \in \mathcal{N}_{\text{I}}$.
```

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 $\mathcal{A}_{q}^{\sigma}$, $q \in \mathcal{Q}^{\sigma} = \{1, \dots, Q^{\sigma}\}$.

In Step 1 (Lines 1-6 (for JR) or Lines 7-14 (for SR) of Algorithm 1), the values of $\boldsymbol{\eta}$ and $\widetilde{\boldsymbol{\eta}}$ minimizing $c_{\text{rr}}(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ while satisfying all constraints related to radio provisioning (2)-(9) are evaluated.

In Step 2 (Line 15-21 (for JN) or Lines 22-29 (for SN) of Algorithm 1), the values of $\boldsymbol{\Phi}_{\text{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\text{b}}$ minimizing $c_{\text{wr}}\left(\boldsymbol{\Phi}_{\text{n}}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_{\text{b}}\right)$, subject to the constraints (12)-(26) are evaluated. The constraints (17), (18), (19) are evaluated with the help of $\boldsymbol{\eta}$ and $\widetilde{\boldsymbol{\eta}}$ obtained at Step 1.

Combining these methods gives four variants of the CARP 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. The complexity of the single-step JRN algorithm, performing a simultaneous joint radio and network provisioning for all slices is provided as a reference. In Table II, the variables $\kappa_{n}^{\sigma}(i, v)$ introduced in (15) 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 $|\mathcal{S}|$ optimization problems, but with $|\mathcal{S}|$ less variables compared to the joint variants (JR and JN). Since each problem is NP-hard, the sequential variants may obviously be solved faster than the joint variants. Section VI-A compares these variants on simulations.

When the amount of available infrastructure resources is not sufficient to accommodate all slices, the proposed joint approaches return no solution. In the sequential approach, the provisioning is performed slice-by-slice. The first processed requests are likely to be satisfied. Next requests may only be satisfied when resources are released. This solution works on a first-arrived-first-served strategy, and has thus some fairness. The main drawback is the suboptimality of the sequential approach, which will be discussed in the next section.

Alternatively, in the joint approach, one may renegotiate the SLAs of all slices to provide some fairness by deploying a part of the services. This may be done by provisioning resources so as to satisfy only a fixed proportion $\delta \in]0, 1]$ of demands of each slice. The search for δ may be done by dichotomy.

TABLE II
VARIANTS OF THE PROVISIONING ALGORITHM.

Variant	*#problems*	*#variables/problem*
JRN	1	$\lvert\mathcal{S}\rvert(\lvert\mathcal{N}_{\text{Ir}}\rvert\,(1 + \lvert\mathcal{Q}^{\sigma}\rvert) +$ $2\lvert\mathcal{N}_{\text{I}}\rvert\lvert\mathcal{N}_{\text{V}}^{\sigma}\rvert + \lvert\mathcal{N}_{\text{I}}\rvert + \lvert\mathcal{E}_{\text{I}}\rvert\lvert\mathcal{E}_{\text{V}}^{\sigma}\rvert)$
SR-SN	$\lvert\mathcal{S}\rvert$ RP	$\lvert\mathcal{N}_{\text{Ir}}\rvert\,(1 + \lvert\mathcal{Q}^{\sigma}\rvert)$
	$\lvert\mathcal{S}\rvert$ NP	$2\lvert\mathcal{N}_{\text{I}}\rvert\lvert\mathcal{N}_{\text{V}}^{\sigma}\rvert + \lvert\mathcal{E}_{\text{I}}\rvert\lvert\mathcal{E}_{\text{V}}^{\sigma}\rvert$
SR-JN	$\lvert\mathcal{S}\rvert$ RP	$\lvert\mathcal{N}_{\text{Ir}}\rvert\,(1 + \lvert\mathcal{Q}^{\sigma}\rvert)$
	1 NP	$\lvert\mathcal{S}\rvert\,\left(2\lvert\mathcal{N}_{\text{I}}\rvert\lvert\mathcal{N}_{\text{V}}^{\sigma}\rvert + \lvert\mathcal{E}_{\text{I}}\rvert\lvert\mathcal{E}_{\text{V}}^{\sigma}\rvert\right)$
JR-SN	1 RP	$\lvert\mathcal{S}\rvert\lvert\mathcal{N}_{\text{Ir}}\rvert\,(1 + \lvert\mathcal{Q}^{\sigma}\rvert)$
	$\lvert\mathcal{S}\rvert$ NP	$2\lvert\mathcal{N}_{\text{I}}\rvert\lvert\mathcal{N}_{\text{V}}^{\sigma}\rvert + \lvert\mathcal{E}_{\text{I}}\rvert\lvert\mathcal{E}_{\text{V}}^{\sigma}\rvert$
JR-JN	1 RP	$\lvert\mathcal{S}\rvert\lvert\mathcal{N}_{\text{Ir}}\rvert\,(1 + \lvert\mathcal{Q}^{\sigma}\rvert)$
	1 NP	$\lvert\mathcal{S}\rvert\,\left(2\lvert\mathcal{N}_{\text{I}}\rvert\lvert\mathcal{N}_{\text{V}}^{\sigma}\rvert + \lvert\mathcal{E}_{\text{I}}\rvert\lvert\mathcal{E}_{\text{V}}^{\sigma}\rvert\right)$

VI. EVALUATION

In this section, one evaluates via simulations the performance of the proposed provisioning algorithms. The simulation set-up is described in Section VI-A. The variants of the provisioning algorithm introduced in Section V are first compared in Section VI-B. Then, Section VI-C illustrates the benefits of provisioning prior to SFC embedding compared to direct SFC embedding. All simulations are performed with the CPLEX MILP solver interfaced with MATLAB.

A. Simulation Conditions

1) Infrastructure Topology: Consider the 1.43 km×4.95 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[2].

For the wired part of the infrastructure network, as in [18, 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 $(a_\text{c}, a_\text{s}, a_\text{r})$ expressed in available number of CPUs, Gbytes of storage, and available RBs at each RRH, depending on the level they belong to.

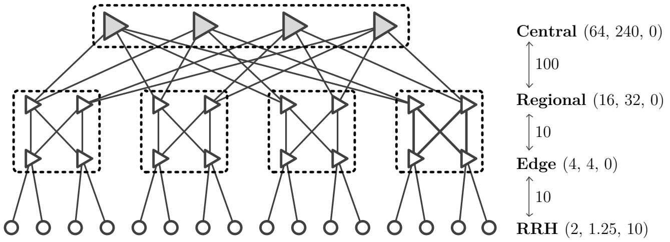


Fig. 8. Description of the k-ary fat-tree infrastructure network in case $k = 4$; Nodes provide a given amount of computing a_c, storage a_s, and radio resource a_r measured in number of used CPUs, Gbytes, and RBs respectively; Links are assigned with a given amount of bandwidth a_b measured in Gbps.

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. The leasing costs of each resource of the infrastructure network is detailed in Table III.

TABLE III
INFRASTRUCTURE COST

Node	$c_\text{f}(i)$	$c_\text{r}(i)$	$c_\text{c}(i)$	$c_\text{s}(i)$
$i \in \mathcal{N}_\text{I} \backslash \mathcal{N}_\text{Ir}$	20	–	1	1
$i \in \mathcal{N}_\text{Ir}$	25	0.05	1	1

2) Slice Resource Demand (SRD): Three types of slices are considered.

- Slices of type 1 cover the *Stade de France* and aim to provide an HD video streaming service at 4 Mbps for at most 200 VIP users within the stadium (downlink traffic);
- Slices of type 2 are dedicated to provide an SD video streaming service at 0.5 Mbps, and cover the blue-highlighted area in Figure 3 (downlink traffic);
- Slices of type 3 aim to provide a video surveillance and traffic monitoring service at 1 Mbps for 50 cameras installed on the A86 highway (uplink traffic).

The first two slice types 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 type consists of five virtual functions: a vBBU, a vGW, a virtual Traffic Monitor (vTM), a vVOC, and a virtual Intrusion Detection Prevention System (vIDPS)..

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 V. Numerical values in Table V have been adapted from [35].

In the following, different scenarios are considered with an increasing number of slices whose distribution among each type is given in Table IV. This represents, *e.g.,* situations where slices of the same type are provided by different SPs.

TABLE IV
NUMBER OF SLICES OF EACH TYPE AS A FUNCTION OF $|\mathcal{S}|$

$\lvert\mathcal{S}\rvert$	4	6	8
Type 1	2	2	4
Type 2	1	2	1
Type 3	1	2	3

The coverage area $\mathcal{A}^\sigma$ associated to each slice type is partitioned into rectangular subareas $\mathcal{A}^\sigma_q$ of 90 m×103 m.

Functional structure and resource requirements for the three slice types are described in Table V.

TABLE V
SLICE RESOURCE DEMAND.

Slice 1: HD video streaming						
Node	r_c	$\underline{r}_\text{c}$	r_s	$\underline{r}_\text{s}$	*Link*	r_b
vVOC	1.35	0.14	3.75	0.38	vVOC→vGW	1.0
vGW	0.23	0.02	0.13	0.01	vGW→vBBU	1.0
vBBU	1.00	0.10	0.13	0.01		
Slice 2: SD video streaming						
Node	r_c	$\underline{r}_\text{c}$	r_s	$\underline{r}_\text{s}$	*Link*	r_b
vVOC	1.08	0.11	1.88	0.19	vVOC→vGW	0.5
vGW	0.18	0.02	0.06	0.01	vGW→vBBU	0.5
vBBU	4.00	0.40	0.06	0.01		
Slice 3: Video surveillance and traffic monitoring						
Node	r_c	$\underline{r}_\text{c}$	r_s	$\underline{r}_\text{s}$	*Link*	r_b
vIDPS	0.535	0.054	0.006	0.001	vIDPS→vVOC	0.05
vVOC	0.270	0.027	0.188	0.019	vVOC→vTM	0.05
vTM	0.665	0.067	0.006	0.001	vTM→vGW	0.05
vGW	0.045	0.005	0.006	0.001	vGW→vBBU	0.05
vBBU	0.200	0.020	0.006	0.001		

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 an RB for a user located in $\mathcal{A}^\sigma_q$ and served by an RRH located in x^r_i are now considered.

[2] L'Agence nationale des fréquences (ANFR): https://data.anfr.fr/

Let $d\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$ be the distance between x_i^{r} and the center of each rectangle $\mathcal{A}_q^{\sigma}$. Focusing on downlink traffic, according to [38], one assumes that

$$b_{\mathrm{d}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right) = W_i \log_2\left(1 + \frac{P_{\mathrm{rx,d}}\left(d\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)\right)}{P_{\mathrm{n}}}\right), \qquad (30)$$

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

$$P_{\mathrm{rx,d}}(d) = P_{\mathrm{tx,d}} + G_{\mathrm{tx,d}} + G_{\mathrm{rx,d}} - PL(d), \qquad (31)$$

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

$$PL(d) = 10\alpha \log_{10}(d) + \beta + 10\gamma \log_{10}(f_i), \qquad (32)$$

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 (30) may be derived for $b_{\mathrm{u}}\left(x_i^{\mathrm{r}}, \mathcal{A}_q^{\sigma}\right)$.

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

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

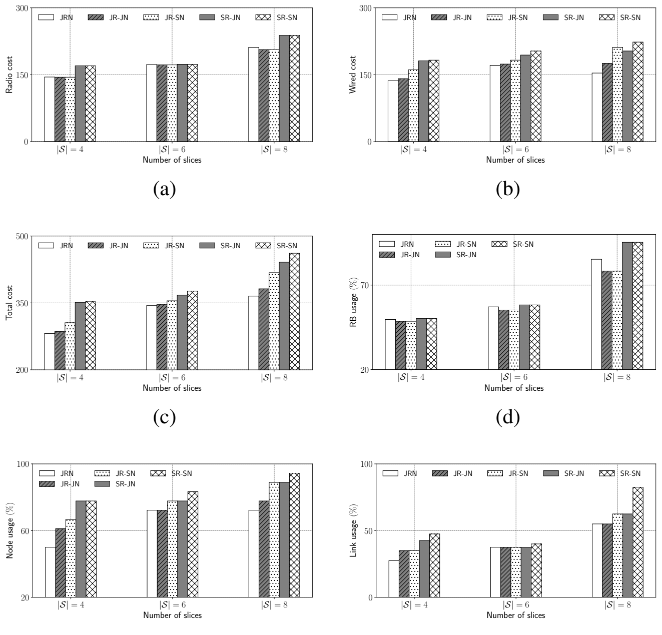
Parameter	*Definition*	*Value*
$a_{\mathrm{r}}(i)$	Number of RBs available at RRH i	100
f_i	Carrier frequency of RRH i	2.6 GHz
W_i	Bandwidth of a RB of RRH i	0.2 MHz
$P_{\mathrm{tx,d}}$	Antenna transmit power of each RRH	43 dBm
$G_{\mathrm{tx,d}}$	Antenna gain of each RRH	15 dBi
$P_{\mathrm{tx,u}}$	Antenna 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	$(3.6, 7.6, 2)$

B. *Comparison of Provisioning Algorithms*

This section illustrates the performance of the JRN joint approach and of the four variants of the CARP two-step provisioning algorithm described in Table II when four, six, and eight slices of different types have to be deployed, see Table IV.

Figure 9a illustrates the radio provisioning costs obtained with the various approaches. One observes that the joint RP schemes (JRN, JR-SN, and JR-JN) yield a smaller cost whatever the NP allocation method. Note that the JRN scheme provides a wireless provisioning cost slightly larger than that of the JR-SN or JR-JN approaches.

Figure 9b illustrates the cost related to the wired part of the infrastructure network. The JRN scheme provides the best results and is always able to compensate for the somewhat larger radio provisioning cost, as illustrated in Figure 9c, which shows the total provisioning costs. Considering the suboptimal approaches, Figures 9b and 9c show that the JR-JN scheme performs better than the other approaches and SR-SN provides always the largest costs, as expected.



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

Fig. 9. 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.

To explain these results, one may consider first the use of radio resource blocks detailed in Figure 9d. The results are consistent with those in Figure 9a: the joint RP approaches (JRN, JR-SN, and JR-JN) outperform the sequential approaches (SR-JN and SR-SN), since the joint RP aims at finding the optimal wireless provisioning for all the slices, while the sequential method only accounts for the constraints of each slice sequentially. The JRN approach does not select the best RRHs for the radio resource provisioning, as compared to the JR-JN or the JR-SN approach, but rather selects the RRHs so as to facilitate the wired network resource provisioning. This leads to a slightly higher utilization of RBs and radio cost (see Figures 9d and 9a), but lower utilization of infrastructure nodes and links (see Figures 9e and 9f), and finally allows the JRN approach to obtain the lowest total cost.

For the suboptimal approaches, the joint RP approach also leads to an efficient utilization of infrastructure nodes and links when solving the NP problem, as shown in Figures 9e and 9f.

The difference in performance of these two sets of methods (JR-SN and JR-JN versus SR-JN and SR-SN) becomes more significant when the number of slices increases. For instance, with six slices, a difference of 11.11% in terms of link utilization is observed in favor of the JR-JN approach, see Figure 9f, whereas with eight slices, the difference is 16.67%. Overall, the JR-JN approach provides the best performance in terms of provisioning costs among the four suboptimal methods.

As expected, the methods involving sequential provisioning (SR and SN) perform better than the joint approaches (JR, JN, and JRN) in terms of computing time. Increasing the number of slices leads to an increase of the cardinality of the sets of variables $(\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}})$ and $\left(\boldsymbol{\Phi}_\mathrm{n}, \widetilde{\boldsymbol{\Phi}}, \boldsymbol{\Phi}_\mathrm{b}\right)$ and therefore increases the computing time. In sequential provisioning, slices are considered successively. Therefore, among the four suboptimal methods, the SR-SN approach and the JR-JN approach are respectively the least and most time-consuming, as shown in Figure 10. The computing time of the optimal JRN is up to 4 times larger than that of the SR-SN approach. Moreover, it increases faster than the other approaches when the number of slices increases. Figure 11 illustrates the way RBs are provisioned by the various RRHs for each slice, when $|\mathcal{N}_\mathrm{Ir}| = 8$ and $|\mathcal{S}| = 8$. Thanks to the rate-related discount introduced in the objective function, RRHs that are close to the coverage area of each slice are chosen in priority. For instance, with the JRN and JR approach, Slice 1, which covers the stadium, has its radio resource demand provisioned by RRH 5 and RRH 7. With the SR approach, radio resource demand of Slice 1 is provisioned by RRH 4 and RRH 7. These three RRHs are both close to the stadium.

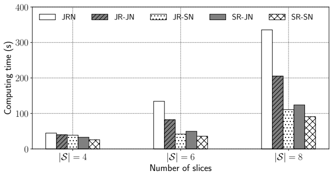


Fig. 10. Computing time of the 4 proposed provisioning variants

The advantage of the JRN and JR over the SR approach can be observed: with the SR approach, all RRHs are required to provision resources, whereas with the JRN or JR approach, only seven RRHs are needed.

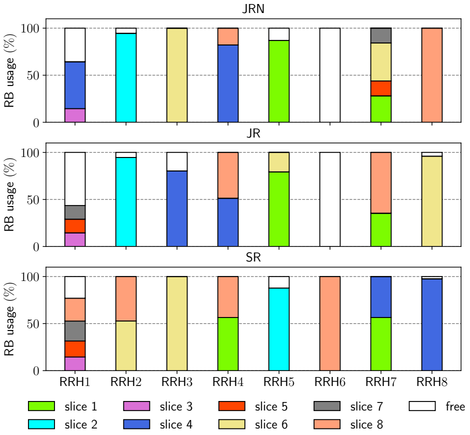


Fig. 11. Provisioned RBs by RRHs for each slice considering the JRN (top), the JR (middle), and the SR (bottom) approaches.

Finally, Figure 12 focuses on the RP problem and shows the maximum 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 from users, *i.e.*, $\sum_{\sigma \in \mathcal{S}} u^\sigma \underline{R}^\sigma$, where u^σ is the number of users in σ, when $|\mathcal{N}_\mathrm{Ir}| = 8$ and 3 slices of type 1, 2, and 3 have to be deployed. $\underline{R}^\sigma$ 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 JR approach, a larger aggregated data rate is supported: provisioning of slices with more users is then possible.

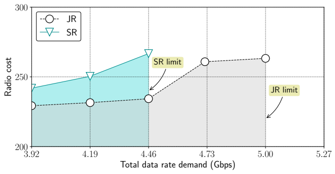


Fig. 12. Maximum supported data rate associated to the SR and JR provisioning approaches when 3 slices of type 1, 2, and 3 have to be deployed.

C. Resource Provisioning vs Direct Embedding

In this section, one assumes that the radio provisioning step has been performed and one focuses on the wired part of the provisioning problem.

To evaluate the benefits of a provisioning approach prior to SFC embedding, the latter is compared to a direct SFC embedding approach. A single slice of type 1 is considered.

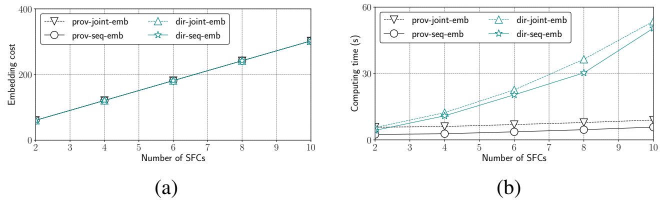
For the SFC deployment, the ILP-based SFC embedding algorithm is adapted from [18]. Specifically, the objective function in [18] is modified to allow the simultaneous embedding of multiple SFCs. Both sequential and joint SFC embedding schemes are performed. The proposed methods, where provisioning is done before a joint and sequential SFC embedding, are denoted respectively as `prov-joint-emb` and `prov-seq-emb`. Direct joint and sequential SFC embedding are denoted as `dir-joint-emb` and `dir-seq-emb`, where prior provisioning is not considered.

The k-ary fat-tree infrastructure topology considered in Section VI-A is used here again. The amount of network infrastructure resource available at each node and link of the infrastructure remains the same.

Figures 13a and 13b show respectively the cost and the required computing time for different number of SFCs belonging to Slice of type 1 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 methods, *i.e.*, `prov-joint-emb` and `prov-seq-emb`, , have similar cost performance as that of the direct embedding, *i.e.*, `dir-joint-emb` and `dir-seq-emb`. Nevertheless, as depicted in Figure 13b, the proposed approach is faster than a direct embedding, when either performing in a joint or sequential fashion. The difference increases with the number of SFCs to embed. Note that in the proposed approach (*i.e.*, `prov-joint-emb` or `prov-seq-emb`), the computing time for the provisioning step has been taken into account.

VII. 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



(a)

(b)

Fig. 13. (a) Embedding costs and (b) computing time of `prov-joint-emb`, `prov-seq-emb`, `dir-joint-emb`, and `dir-seq-emb` approaches as a function of the number of SFCs to embed.

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 on the basis of the SLA between an SP and the MNO. This graph describes the aggregated resource requirements of the SFCs that will be deployed by the MNO for a given slice

Adopting the point of view of the InP, one tries to minimize the cost related to the usage of the network infrastructure, in particular the radio access network, while satisfying radio coverage constraints, to ensure a minimum data rate for users in the geographical areas where services have to be made available. This problem is cast in the framework of MILP problem.

A two-step approach is proposed to address the complexity of this problem. Radio resources on RRH are provisioned first to ensure the satisfaction of the coverage constraints. Other 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 simulations, 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. When both coverage and infrastructure network constraints have to be taken into account simultaneously, *i.e.*, the `JRN` approach, a minimum provisioning cost could be achieved, but this approach requires a much larger time complexity than the four variants of the suboptimal `CARP`.

Once resources have been provisioned, the approach introduced in [18, 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.

Only static provisioning is considered in this paper. Resource provisioning was done for a given time interval specified in the SLA over which the service characteristics and constraints are assumed constant and compliant with the variations of user demands within a slice. A level of conservatism in the amount of provisioned resources is then required to satisfy fast fluctuating user demands. One could imagine adaptive SLAs to meet more closely the actual demands. The SLA may consider several time intervals over each of which the service characteristics and constraints are assumed constant, but may vary from one interval to the next one. On the other hand, one could imagine that already allocated SFCs may be updated during the lifetime of the slice. Adaptive SLAs and dynamic provisioning techniques will be considered in future work.

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 *ACM AllThingsCellular*, 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 Commun. Mag.*, 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 VTC*, 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," *Comput. Commun.*, 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 Comput.*, vol. 21, no. 5, pp. 20–27, 2017.

[9] A. Kaloxylos, "A Survey and an Analysis of Network Slicing in 5G Networks," *IEEE Commun. Std. Mag.*, vol. 2, no. 1, pp. 60–65, 2018.

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

[11] 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. IEEE ICAC*, 2017, pp. 221–226.

[12] ITU-T, "GSTR-TN5G: Transport Network Support of IMT-2020/5G," *ITU Technical Report*, 2018.

[13] Y. Zhu and M. Ammar, "Algorithms for Assigning Substrate Network Resources to Virtual Network Components," in *Proc. IEEE INFOCOM*, 2006.

[14] M. Chowdhury, M. R. Rahman, and R. Boutaba, "ViNEYard: Virtual Network Embedding Algorithms," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 206–219, 2012.

[15] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, 2017.

[16] 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," *J. Inf. Process.*, vol. 25, pp. 153–163, 2017.

[17] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 2018.

[18] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling Wireless Virtual Networks Functions," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 2, pp. 240–252, 2016.

[19] P. Vizarreta, M. Condoluci, C. M. Machuca, T. Mahmoodi, and W. Kellerer, "QoS-driven Function Placement Reducing Expenditures in NFV Deployments," in *Proc. IEEE ICC*, 2017.

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

[21] 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*, 2016, pp. 243–250.

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

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

[24] 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 Trans. Netw. Service Manag.*, no. July, pp. 1–20, 2018.

[25] N. Huin, B. Jaumard, and F. Giroire, "Optimization of Network Service Chain Provisioning," in *Proc. IEEE ICC*, 2017.

[26] J. Liu, W. Lu, F. Zhou, P. Lu, and Z. Zhu, "On Dynamic Service Function Chain Deployment and Readjustment," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 3, pp. 543–553, 2017.

[27] M. Mechtri, C. Ghribi, and D. Zeghlache, "A Scalable Algorithm for the Placement of Service Function Chains," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 533–546, 2016.

[28] H. Halabian, "Distributed Resource Allocation Optimization in 5G Virtualized Networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 627–642, 2019.

[29] S. Chatterjee, M. J. Abdel-rahman, and A. B. Mackenzie, "Virtualization Framework for Cellular Networks with Downlink Rate Coverage Probability Constraints," in *Proc. IEEE 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. 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. 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 Trans. Netw.*, 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 Comput.*, 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," *Comput. Netw.*, 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 NFV-SDN*, 2016, pp. 191–197.

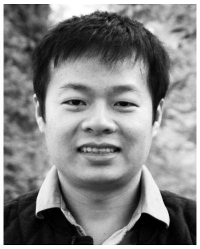
[36] Y. Shi and Y. T. Hou, "Approximation Algorithm for Base Station Placement in Wireless Sensor Networks," in *Proc. IEEE SeCON*, 2007, pp. 512–519.

[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 Trans. Netw. Service Manag.*, vol. 14, no. 2, pp. 317–331, 2017.

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

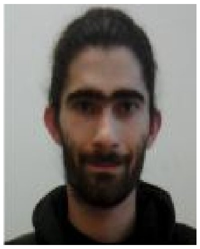
[39] S. Sun, T. S. Rappaport, S. Rangan, T. A. Thomas, A. Ghosh, I. Z. Kovacs, I. Rodriguez, O. Koymen, A. Partyka, and J. Jarvelainen, "Propagation Path Loss Models for 5G Urban Micro- and Macro-Cellular Scenarios," in *Proc. IEEE VTC*, 2016, pp. 1–6.

[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.



Quang-Trung Luu is currently a Ph.D student at Nokia Bell Labs and the Signal and Systems Laboratory (L2S) of the University of Paris-Sud and CentraleSupélec, France. His research focuses on the dynamic control and optimization of wireless virtual networks, in particular on key enable technologies for 5G such as network slicing. He obtained two master's degrees, one in antennas and telecommunications, and one another in multimedia networking from the University of Paris-Sud and Télécom Paris, France in 2016 and 2017, respectively.

Sylvaine Kerboeuf received the M.S. degree in physics and the Ph.D. degree in solid-state physics from the University of Paris-Sud, Orsay France, in 1991 and 1994, respectively, and the Ph.D. degree in superconductivity from the Centre National d'Etude des Télécommunications, France Telecom, Paris, France. She joined the Research and Innovation Department, Alcatel-Lucent Bell Laboratories, Nozay, France, where she was involved in research projects on optoelectronics for several years. In 2004, she joined a project involved in radio access networks and focusing on fourth generation discontinuous networks and on caching technology. She is currently a Senior Researcher in the Wireless Program with Nokia Bell Laboratories. Her current research interests include end-to-end video delivery over wireless networks, video transport protocols, video quality of experience optimization, and software defined networking for 5G.



Alexandre Mouradian is currently an Assistant Professor (maître de conférences) at the University of Paris-Sud, L2S laboratory. Previously, he was temporary assistant professor/ATER (2014) and PhD student (2010-2013) at CITI lab and INSA de Lyon. His research focuses on MAC and routing in multi-hop radio networks with a special interest in timeliness (real-time properties). In the previous works, he has especially been interested in MAC and routing protocol design for WSNs and in the use of formal methods in order to verify time properties in such networks.

Michel Kieffer (M'02, SM'07) received the Ph.D. degree in control theory from the University of Paris XI, Orsay, France, in 1999. He is a Full Professor in signal processing for communications with the University of Paris-Sud and a Researcher with the Laboratoire des Signaux et Systèmes (L2S), Gif-sur-Yvette, France. Since 2009, he is a part-time Invited Professor with the Laboratoire Traitement et Communication de l'Information, Télécom Paris-Tech, Paris, France. He is coauthor of more than 150 contributions in journals, conference proceedings, or books. He is one of the coauthors of the books *Applied Interval Analysis* (Springer-Verlag, 2001) (this book was translated in Russian in 2005) and *Joint Source-Channel Decoding: A Cross-Layer Perspective With Applications in Video Broadcasting* (Academic, 2009). His research interests are in signal processing for multimedia, communications, and networking; distributed source coding; network coding; joint source-channel coding and decoding techniques; and joint source-network coding. Applications are mainly in the reliable delivery of multimedia contents over wireless channels. He is also interested in guaranteed and robust parameter and state bounding for systems described by nonlinear models in a bounded-error context. Prof. Kieffer became a junior member of the Institut Universitaire de France in 2011. He has served as an Associate Editor of SIGNAL PROCESSING since 2008 and of the IEEE TRANSACTIONS ON COMMUNICATIONS since 2012.