Mobile Network Slicing under Demand Uncertainty: A Stochastic Programming Approach

Anousheh Gholami, Nariman Torkzaban, and John S. Baras
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD, 20742, USA
Email: {anousheh, narimant, baras}@umd.edu

2023 IEEE 9th International Conference on Network Softwarization (NetSoft) | 979-8-3503-9980-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/NetSoft57336.2023.10175453

***Abstract*—Constant temporospatial variations in the user demand complicate the end-to-end (E2E) network slice (NS) resource provisioning beyond the limits of the existing best-effort schemes that are only effective under accurate demand forecasts for all NSs. This paper proposes a practical two-time-scale resource allocation framework for E2E network slicing under demand uncertainty. At each macro-scale instance, we assume that only the spatial probability distribution of the NS demands is available. We formulate the NSs resource allocation problem as a stochastic mixed integer program (SMIP) with the objective of minimizing the total CN and RAN resource costs. At each micro-scale instance, given the exact NSs demand profiles known at operation time, a linear program is solved to jointly minimize the unsupported traffic and RAN cost. We verify the effectiveness of our resource allocation scheme through numerical experiments.**

***Index Terms*—Network slicing, end-to-end resource provisioning, demand uncertainty, stochastic programming**

I. Introduction

In the paradigm of 5G enhanced by *network slicing*, mobile network operators (MNOs) manage and set up network slices (NSs) and provide service providers (SPs) with an on-demand and scalable delivery of network services [1]. The SPs, a.k.a. *tenants*, seamlessly dedicate NSs to customers with various QoS requirements. A mobile NS spans across multiple domains including the radio access network (RAN), the core network (CN), and the transport network (TN), forming an end-to-end (E2E) sub-network. In contrast to the traditional static resoruce allocation schemes [2], 5G RAN slicing introduces the capability of sharing physical network infrastructure among mobile virtual network operators (MVNOs). Consequently, the core and radio resources are reserved on the fly according to end-users demand. Designing E2E NSs requires resource provisioning across heterogeneous physical and virtual network infrastructures each having specific technical constraints. Despite the desirable impacts of *network slicing* on the agility and flexibility of next-generation mobile networks (NGMN), practical NS deployment faces key challenges among which the *demand uncertainty* stands out. While the existing resource provisioning methods for *network slicing* are typically performed in a best-effort manner [3], the shared network resources must be dynamically and efficiently allocated to logical NSs based on changing user demands. Besides the dynamicity of the user demand profiles, the variation in the infrastructure resource availability status may degrade the slice QoS compared to the service level agreement (SLA) promised by the SPs [4]. In this paper, we propose a novel approach to optimize the E2E resource provisioning for *network slicing* under *demand uncertainty*. Stochastic programming is a powerful tool to address optimization under uncertainty. We consider the joint resource allocation of next-generation RAN (NG-RAN) and 5G core (5GC) for different NSs. In our proposed solution, the RAN slicing is triggered more frequently than the CN segment, due to the existence of more dynamic parameters in the RAN such as user mobility and varying channel condition. Therefore, our algorithm operates at two time scales. At each macro-scale instance, we assume that only the spatial probability distribution of the slice demands is available. We formulate the NSs resource provisioning problem as a stochastic mixed integer program (SMIP) with the objective of minimizing the total resource cost. At each micro-scale instance, utilizing the realized exact NS demands, a linear program is solved to jointly minimize the unsupported traffic and the RAN resource cost by adjusting the RAN slices and scaling E2E resource allocation.

The remainder of the paper is organized as follows. Section II describes the system model. The problem formulation and proposed solution are provided in section III. The numerical results and conclusion are provided in section IV.

II. System Model

A. Substrate Network Model

We consider a mobile network infrastructure (a.k.a. substrate network) that is comprised of next-generation NodeBs (gNBs) and CN nodes hosting 5GC components. Let $G = (V, E)$ denote the substrate network graph, where V and E represents the set of substrate nodes and links, respectively. We assume that $V = V_{gNB} \cup V_{-gNB}$ where V_{gNB} and V_{-gNB} are the substrate gNB and non-gNB nodes. Each gNB is characterized by its maximum supported traffic that is computed based on its available resource blocks (RBs) and antenna configuration. Non-gNB nodes are general-purpose servers providing essential capabilities to run core VNFs. A substrate node is characterized by its residual CPU, storage, and RAM resources. Let $W_j = (W_j^\nu, \nu \in \mathcal{T})$ represent the residual capacity of the substrate node $j \in V$, the set of resources defined as $\mathcal{T} = \{CPU, STO, RAM\}$. Similar to the RAN slicing model considered by [5], we assume that the gNB $j \in V_{gNB}$ has W_j^r RBs available to be allocated to different NSs. Furthermore, we assume that $E = E_{FH} \cup E_{BH}$ where E_{FH} and E_{BH} stand for the set of gNB-CN links and the remaining links, respectively. Each substrate link $e \in E$ is characterized by its available bandwidth, W_e^{BW}, and propagation delay, τ_e^{PRO}. Moreover, let $\mathcal{P}$ denote the set

of substrate paths used for traffic routing. Define $\mathcal{P}(i \rightarrow j)$ to be the set of substrate paths between nodes i and j. Therefore, $\mathcal{P} = \cup_{i,j \in V, i \neq j} \mathcal{P}(i \rightarrow j)$. Let $\mathcal{U}$ denote the set of UEs distributed across the geographical area. We assume that each UE is served by a gNB according to nearest association rule.

B. Network Slice Model

We assume that a NS consists of one or multiple SFCs, each comprised of a number of VNFs (e.g. gNB, AFM, UPF, SMF) and virtual links (VLs) between them. The set of NSs is denoted by $\mathcal{K} = \{1, ..., K\}$. Let $G'_k = (V'_k, E'_k)$ represent the kth slice SFC modeled as an undirected graph. V'_k, $V'_{k,gNB}$, $V'_{k,-gNB}$, E'_k, $E'_{k,FH}$, and $E'_{k,BH}$ are defined similar to the corresponding components of the substrate network. We assume that each instance of the SFC corresponding to NS k has QoS requirements expressed as network-level UE throughput and maximum E2E tolerable latency. In order to guarantee the requirement of different NSs, the network-level performance requirement are translated to cell-level radio resource requirement [6]. Let $\underline{R}_j^k$ denote the average number of RBs required for a UE requesting NS k served by gNB j. The value of $\underline{R}_j^k$ depends on the overall cell load, antenna configuration, channel condition, modulation and coding scheme (MCS), and NS QoS requirements. The details of translation mechanisms such as [7] are beyond the scope of this article.

We denote the set of substrate paths considered for slice k by $\mathcal{P}_k$. Let $\mathcal{P}'_k$ represent the set of all paths in G'_k. Each path of G'_k corresponds to either CP or UP data flows. For instance, the path gNB-AMF-SMF of a 5G NS is a CP path while gNB-UPF is a UP path. Let $\mathcal{P}'_{k,UP}$ and $\mathcal{P}'_{k,CP}$ denote the set of UP and CP paths in G'_k. We assume that the QoS requirement of each NS is given as the maximum tolerable UP (CP) latency denoted by $d_k^{UP}(d_k^{CP})$. Let $\mathcal{U}_k \subseteq \mathcal{U}$ denote the set of UEs requesting NS k. We define $u_j^{k,i}$ to be equal to 1 if $i \in \mathcal{U}_k$ and 0, otherwise. In order to support the demands of all UEs requesting a NS, multiple instances of the NS may need to be deployed. Let I_k be the number of instances of slice k deployed to support the load for slice k. Let $R_{j'} = (R_{j'}^{\nu}, \nu \in \mathcal{T})$ denote the required per-unit CPU, STO and RAM for VNF j'. Similarly, each VL $e' \in E'_k$ is characterized by its per-unit bandwidth requirement $R_{e'}^{BW}$ to meet the demand for data transmission between the two end-point VNFs of e'. The per-unit resource requirements must be scaled according to the traffic load and resource-sharing factor. We define $\chi_{k,j'}^{\nu}$ to be the scaling factor of resource ν, $\nu \in \mathcal{T}$ for VNFs of slice k. Moreover, $\chi_{k,e'}^{BW}$ stands for the BW scaling factor of VLs of slice k. The cost of running a VNF on the substrate node j is composed of two parts: (i) a fixed cost denoted by C_j, (ii) a variable cost that increases linearly with respect to the consumed amount of resources by that VNF. Let $C_j^{\nu}, \nu \in \mathcal{T}$ denote the per-unit cost of using resource ν of node j. The per-unit bandwidth cost of the infrastructure link e is C_e^{BW}.

III. Problem Formulation and Proposed Solution

In this section, we formulate the E2E network slicing problem under demand uncertainty as a two-stage SMIP. A *stochastic linear programs* (SLP) is a linear program where some data is uncertain, represented as random variables with given probability measures. The value of these random variables are known only after a random experiment. Thus, a SLP variables are divided into two groups: first-stage variables, determined before the experiment, and second-stage variables, determined after the experiment results are known. We refer the interested readers to [8] for further details.

A. Stochastic Mobile Network Slicing Optimization Model

We start by defining the decision variables for the optimization model as follows:

- $\boldsymbol{x}_k$: set of binary variables where $x_j^{k,j'}$ equals 1 if the VNF j' of NS k is placed on the substrate node j.
- $\boldsymbol{y}_k$: set of binary variables where $y_p^{k,e'}$ equals 1 if the VL e' of NS k is mapped to substrate path $p \in \mathcal{P}_k$.
- $\boldsymbol{z}_k$: set of continuous variables where z_j^k represents the fraction of gNB j spectrum allocated to NS k.

We model the resource provisioning problem for NSs as a two-stage SMIP. We refer to this problem as stochastic mobile network slicing (*SMNS*), explained in the following.

1) First-Stage Problem: The first-stage objective of *SMNS* is to minimize the total provisioning cost which consists of node and link deployment costs for all slices. Let $\mathcal{C}_k^N$ and $\mathcal{C}_k^L$ denote the node and link costs corresponding to NS k. Consequently, we have:

$$\mathcal{C}_k^N(\boldsymbol{x}_k) = \sum_{j \in V} \sum_{j' \in V'_k} C_j^k x_j^k + \sum_{j \in V} \sum_{j' \in V'_k} \sum_{\nu \in \mathcal{T}} \gamma_\nu R_{j'}^{\nu} C_j^{\nu} x_j^{k,j'} \tag{1}$$

$$\mathcal{C}_k^L(\boldsymbol{y}_k) = \sum_{p \in \mathcal{P}} \sum_{e \in p} \sum_{e' \in E'_k} \gamma_{BW} R_{e'}^{BW} C_e^{BW} y_p^{k,e'} \tag{2}$$

where

$$x_j^k = \begin{cases} 1 & \text{if } \sum_{j' \in V'_k} x_j^{k,j'} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Hence, the first-stage objective is:

$$\mathcal{C}_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k \in \mathcal{K}} \mathcal{C}_k^N(\boldsymbol{x}_k) + \sum_{k \in \mathcal{K}} \mathcal{C}_k^L(\boldsymbol{y}_k) \tag{4}$$

where $\gamma_\nu, \nu \in \mathcal{T} \cup \{BW\}$ are the weights used to balance the objective terms of (4) corresponding to different resources. Since each gNB(non-gNB) VNF should be placed at substrate gNB(non-gNB) nodes only, a valid slicing strategy satisfies:

$$\sum_{j \in V_{-gNB}} x_j^{k,j'} = 1, \ \forall k \in \mathcal{K}, j' \in V'_{k,-gNB} \tag{5}$$

$$x_j^{k,j'} = 0, \ \forall k \in \mathcal{K}, j \in V_{-gNB}, j' \in V'_{k,gNB} \tag{6}$$

The flow conservation is guaranteed by constraints (7) and (8):

$$\sum_{p \in \mathcal{P}_k(i \rightarrow j), q \in \mathcal{P}_k(j \rightarrow i), j \in V} y_p^{k,e'} - y_q^{k,e'} = x_i^{k,i'} - x_i^{k,j'},$$
$$\forall k \in \mathcal{K}, e' \in E'_{k,BH}, src(e') = i', dst(e') = j' \tag{7}$$

$$\sum_{p \in \mathcal{P}_k(i \rightarrow j), q \in \mathcal{P}_k(j \rightarrow i), j \in V} y_p^{k,e'} - y_q^{k,e'} = (x_i^{k,i'} - x_i^{k,j'}) I_k,$$
$$\forall k \in \mathcal{K}, i \in V_{-gNB}, e' \in E'_{k,FH}, src(e') = i', dst(e') = j' \tag{8}$$

The domain constraints are as follows,

$$x_j^{k,j'}, y_p^{k,e'} \in \{0, 1\} \ \forall k \in \mathcal{K}, j \in V, j' \in V'_k, p \in \mathcal{P}_k, e' \in E'_k \tag{9}$$

2) Second-Stage Problem: The objective of the second-stage problem is minimizing the cost of gNBs resources:

$$\mathcal{C}_2(\boldsymbol{z}) = \sum_{k\in\mathcal{K}} \sum_{j\in V_{gNB}} C_j^k z_j^k W_j^r \tag{10}$$

The processing delay of VNF $j' \in V'_k$ on a substrate node is represented by $\tau^{k,j'}$. The UP(CP) latency of slice k instances is guaranteed to be lower than the delay budget $d_k^{UP}(d_k^{CP})$ by:

$$\sum_{e''\in p'} \left[\sum_{p\in\mathcal{P}_k} y_p^{k,e''} \sum_{e\in p} \left[\sum_{k\in\mathcal{K}} \sum_{\substack{p\in\mathcal{P}_k \\ |e\in p}} \sum_{e'\in E'_k} y_p^{k,e'} \frac{\chi_k^{BW} R_{e'}^{BW}}{W_e^{BW}} \right] \right]$$
$$+ \sum_{j'\in p'} \left[\sum_{j\in V} x_j^{k,j'} \tau^{k,j'} \right] \leq d_k^{UP/CP}, \forall p' \in \mathcal{P}'_{k,UP/CP}, \forall k$$

The above constraint ensures that the maximum latency of each NS is less than its tolerable latency by enforcing the inequality for all paths of G'_k. We linearize it by introducing a set of additional continuous variables $\eta_p \geq 0$ and $\eta_p^{k,e'} \geq 0$, defined as the latency of path $p \in \mathcal{P}$ and the latency of VL $e' \in E'_k$ on path p, respectively. Therefore, we have:

$$\sum_{e\in q} \left[\sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k|e\in p} \sum_{e'\in E'_k} y_p^{k,e'} \frac{\chi_k^{BW} R_{e'}^{BW}}{W_e^{BW}} \right] = \eta_q, \ \forall q \in \mathcal{P} \tag{11}$$

$$\zeta y_p^{k,e'} + \eta_p - \eta_p^{k,e'} \leq \zeta, \ \forall p \in \mathcal{P}_k, e' \in E'_k, k \in \mathcal{K} \tag{12}$$

$$\sum_{e'\in p'} \sum_{p\in\mathcal{P}_k} \eta_p^{k,e'} + \sum_{j'\in p'} \sum_{j\in V} x_j^{k,j'} \tau^{k,j'} \leq d_k^{UP/CP},$$
$$\forall p' \in \mathcal{P}'_{k,UP/CP}, k \in \mathcal{K} \tag{13}$$

where ζ is a large constant. We also add the term $\epsilon\mathcal{D}$ to the objective function where $\mathcal{D}(\boldsymbol{\eta}) = \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k} \sum_{e'\in E'_k} \eta_p^{k,e'}$ and ϵ is a very small value in order to make sure that the main objective of minimizing the resource provisioning cost is not affected by adding $\mathcal{D}$. Given the network-level performance translation to cell-level metric $\underline{R}_j^k$, the RAN NS is:

$$\sum_{i=1}^{U_k} u_j^{k,i} \underline{R}_j^k \leq x_j^{k,j'} z_j^k W_j^r, \ \forall k, j \in V_{gNB}, j' \in V'_{k,gNB} \tag{14}$$

Using the big-M method and convert (14) to the following:

$$\sum_{i=1}^{U_k} u_j^{k,i} \underline{R}_j^k \leq z_j^k W_j^r + (1 - x_j^{k,j'})M, \ \forall k \in \mathcal{K},$$
$$j \in V_{gNB}, j' \in V'_{k,gNB} \tag{15}$$

In order to satisfy the slice isolation constraint and given the maximum number of RBs allowed for a NS at each gNB ($\overline{R}_j^k$), we formulate the slice isolation constraint using the following inequality:

$$z_j^k W_j^r \leq \overline{R}_j^k x_j^{k,j'}, \ \forall k \in \mathcal{K}, j \in V_{gNB}, j' \in V'_{k,gNB} \tag{16}$$

For the substrate nodes and links, the capacity constraints are:

$$\sum_{k\in\mathcal{K}} \sum_{j'\in V'_k} x_j^{k,j'} R_{j'}^{\nu} \chi_k^{\nu} \leq W_j^{\nu}, \ \nu \in \mathcal{T}, \forall j \in V \tag{17}$$

$$\sum_{k\in\mathcal{K}} \sum_{\substack{p\in\mathcal{P}_k \\ |e\in p}} \sum_{e'\in E'_k} y_p^{k,e'} R_{e'}^{BW} \chi_k^{BW} \leq W_e^{BW}, \ \forall e \in E \tag{18}$$

The capacity and placement constraints of gNBs enforced by:

$$\sum_{k=1}^{K} z_j^k = x_j^{k,j'}, \ \forall j \in V_{gNB}, j' \in V'_{k,gNB} \tag{19}$$

$$x_j^{k,j'} \geq \mathbb{I}\{\sum_{i=1}^{U_k} u_j^{k,i} \geq 1\}, \ \forall k \in \mathcal{K}, j \in V_{gNB}, j' \in V'_{k,gNB} \tag{20}$$

The domain constraints of the second-stage problem are:

$$z_j^k \in [0,1], \ \forall k \in \mathcal{K}, j \in V_{gNB}$$
$$\eta_p^{k,e'}, \eta_p \geq 0, \ \forall k \in \mathcal{K}, e' \in E'_k, p \in \mathcal{P}_k \tag{21}$$

Thus, the E2E network slicing problem with demand uncertainty is formulated as a two-stage SMIP presented below:

$$min \ \phi_1 \mathcal{C}_1(\boldsymbol{x}, \boldsymbol{y}) + \phi_2 \ \mathbb{E}_{\boldsymbol{\xi}}[min \ \mathcal{C}_2(\boldsymbol{z}) + \epsilon\mathcal{D}(\boldsymbol{\eta})] \tag{22}$$
$$\text{s.t.} \qquad (5) - (21)$$

The objective (22) minimizes the summation of the CN slice and the expectation of the RAN slice provisioning cost. The weights ϕ_1, ϕ_2 determine the balance between the objectives of the first and second stage problems. An interpretation of the *SMNS* is as follows:

- The MNO provisions the CN resources for different NSs, represented by the decision variables $\boldsymbol{x}, \boldsymbol{y}$, before the actual value of the random variable vector $\boldsymbol{\xi} = (\mathcal{U}_1, \ldots, \mathcal{U}_K, u_1^{1,1}, \ldots, u_1^{1,U_1}, \ldots, u_{|V_{gNB}|}^{K,\mathcal{U}_K}, \chi_1^{\nu}, \ldots, \chi_K^{\nu}, \nu \in \mathcal{T} \cup \{BW\})$ is realized.
- Once $\boldsymbol{\xi}$ realized, the RAN resource provisioning and delay decision variables, denoted by $\boldsymbol{z}, \boldsymbol{\eta}$ are determined.

The expectation term in (22) requires an integration over the high-dimensional random vector $\boldsymbol{\xi}$. To tackle this challenge, we use the sample average approximation (SAA) technique and replace the expectation in (22) with its SAA.

B. Deterministic Equivalent Reformulation

The SAA method is an approach for solving stochastic optimization problems by using Monte Carlo simulation. Suppose that H i.i.d observations of the random variables $u_j^{k,i}$, U_k, and χ_k^{ν} are available, denoted by $\tilde{u}_{j,h}^{k,i}$,$\tilde{U}_{k,h}$, $\tilde{\chi}_{k,h}^{\nu}$, $h = 1, \ldots, H$. For each realization, we define a separate set of second-stage decision variables $z_{j,h}^k, \eta_{p,h}, \eta_{p,h}^{k,e'}$. Thus, we convert *SMNS* to

its sampled deterministic equivalent problem:

$$min \quad \phi_1 \mathcal{C}_1(\boldsymbol{x}, \boldsymbol{y}) + \phi_2 \sum_{h=1}^{H} \frac{1}{H} \left(\mathcal{C}_2(\boldsymbol{z}_h) + \epsilon \mathcal{D}(\boldsymbol{\eta}_h) \right) \tag{23}$$

$$subject\ to: \qquad (5) - (9)$$

$$\sum_{e \in q} \left[\sum_{k \in \mathcal{K}} \sum_{\substack{p \in \mathcal{P}_k \\ |e \in p}} \sum_{e' \in E'_k} y_p^{k,e'} \frac{\tilde{\chi}_{k,h}^{BW} R_{e'}^{BW}}{W_e^{BW}} \right] = \eta_{q,h}, \forall q, h \tag{24}$$

$$\zeta y_p^{k,e'} + \eta_{p,h} - \eta_{p,h}^{k,e'} \leq \zeta, \ \forall p \in \mathcal{P}_k, e' \in E'_k, k \in \mathcal{K}, h \in \mathcal{H} \tag{25}$$

$$\sum_{e' \in p'} \sum_{p \in \mathcal{P}_k} \eta_{p,h}^{k,e'} + \sum_{j' \in p'} \sum_{j \in V} x_j^{k,j'} \tau^{k,j'} \leq d_k^{UP/CP},$$
$$\forall p' \in \mathcal{P}'_{k,UP/CP}, k \in \mathcal{K}, h \in \mathcal{H} \tag{26}$$

$$\sum_{k \in \mathcal{K}} \sum_{j' \in V'_k} x_j^{k,j'} R_{j'}^{\nu} \tilde{\chi}_{k,h}^{\nu} \leq W_j^{\nu}, \forall \nu \in \mathcal{T}, j \in V, h \in \mathcal{H} \tag{27}$$

$$\sum_{k \in \mathcal{K}} \sum_{\substack{p \in \mathcal{P}_k \\ |e \in p}} \sum_{e' \in E'_k} y_p^{k,e'} R_{e'}^{BW} \tilde{\chi}_{k,h}^{BW} \leq W_e^{BW}, \forall e \in E, h \in \mathcal{H} \tag{28}$$

$$\sum_{i=1}^{\tilde{U}_{k,h}} \tilde{u}_{j,h}^{k,i} \underline{R}_j^k \leq z_{j,h}^k W_j^r + (1 - x_j^{k,j'}) M,$$
$$\forall k \in \mathcal{K}, j \in V_{gNB}, j' \in V'_{k,gNB}, h \in \mathcal{H} \tag{29}$$

$$z_{j,h}^k W_j^r \leq \overline{R}_j^k x_j^{k,j'}, \forall k \in \mathcal{K}, j \in V_{gNB}, j' \in V'_{k,gNB}, h \in \mathcal{H} \tag{30}$$

$$\sum_{k=1}^{K} z_{j,h}^k \leq 1, \forall j \in V_{gNB}, h \in \mathcal{H} \tag{31}$$

$$x_j^{k,j'} \geq \mathbb{1}\{ \sum_{i=1}^{\tilde{U}_{k,h}} \tilde{u}_{j,h}^{k,i} \geq 1 \}, \forall k \in \mathcal{K}, j \in V_{gNB}, j' \in V'_{k,gNB}, h \tag{32}$$

$$z_{j,h}^k \in [0,1], \eta_{p,h}, \eta_{p,h}^{k,e'} \geq 0, \ \forall k \in \mathcal{K}, j \in V_{gNB}, h \in \mathcal{H} \tag{33}$$

We refer to the above problem as *DET-SMNS*($\mathcal{H}$). It is shown in [9] that as the sample size H increases and under certain mild conditions, the solution to *DET-SMNS*($\mathcal{H}$) converges to that of the original *SMNS* problem. To solve the E2E network slicing problem under demand uncertainty, we propose a two-timescale resource allocation scheme that utilizes *DET-SMNS*. In this approach, the E2E resource allocation problem of NSs is addressed through long and short time slots. We assume that the lifecycle of an NS is divided into a number of long time slots (macro-scale) indexed by T. At each macro-scale instance, the spatial probability distribution of the demand for NSs is known. Let $\mathcal{F}^T = \{\mathcal{F}_1^T, \ldots, \mathcal{F}_K^T\}$, where $\mathcal{F}_k(.,.) : \mathbb{R}^2 \rightarrow [0,1]$ denote the spatial density function of the kth NS demand across the considered geographical area in the macro-scale instance T. Given $\mathcal{F}^T$, we solve the resource provisioning problem *DET-SMNS*($\mathcal{H}$). In order to adjust the provisioned resources to address the actual demand, we further divide each macro-slot into N_T short time slots (micro-scale). At each micro-scale instance t, the actual demand for the duration of the micro-slots is observed and $\boldsymbol{\xi}$ becomes known, denoted by $\hat{\boldsymbol{\xi}}$. The demand of a NS is served if supported by the allocated resources in RAN and CN. We define $\sigma_j^k, k \in \mathcal{K}, j \in V_{gNB}$ to be the fraction of total demand requested for slice k in gNB j that is not supported. Given the solution of the *DET-SMNS*($\mathcal{H}$) and $\boldsymbol{\xi}$, we solve the following LP (*RNSR*($\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{\xi}}$)) with the objective of minimizing RAN resource allocation cost and total unsupported traffic.

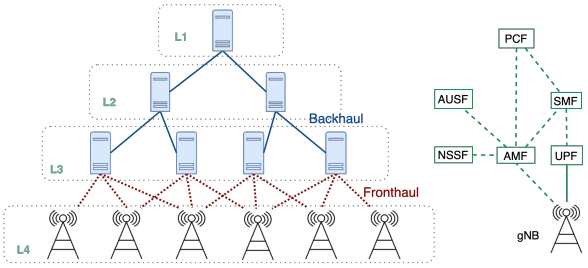


Fig. 1: Example substrate (left) and NS (right)

$$minimize \quad \sum_{k \in \mathcal{K}} \sum_{j \in V_{gNB}} \theta(\mathcal{C}_2(\boldsymbol{z})) + (1 - \theta)\sigma_j^k \tag{34}$$

$$(1 - \sigma_j^k)(\sum_{i=1}^{\hat{U}_k} \hat{u}_j^{k,i}) \underline{R}_j^k \leq z_j^k W_j^r + (1 - x_j^{k,j'}) M,$$
$$\forall k \in \mathcal{K}, j \in V_{gNB}, j' \in V'_{k,gNB} \tag{35}$$

$$\frac{\sum_{j \in V_{gNB}} (1 - \sigma_j^k) \sum_{i=1}^{\hat{U}_k} \hat{u}_j^{k,i}}{\hat{U}_k} \hat{\chi}_k^{\nu} \leq \chi_{k,prov}^{\nu},$$
$$\forall k \in \mathcal{K}, \nu \in \mathcal{T} \cup \{BW\} \tag{36}$$

$$z_j^k W_j^r \leq \overline{R}_j^k x_j^{k,j'}, \ \forall k \in \mathcal{K}, j \in V_{gNB}, j' \in V'_{k,gNB} \tag{37}$$

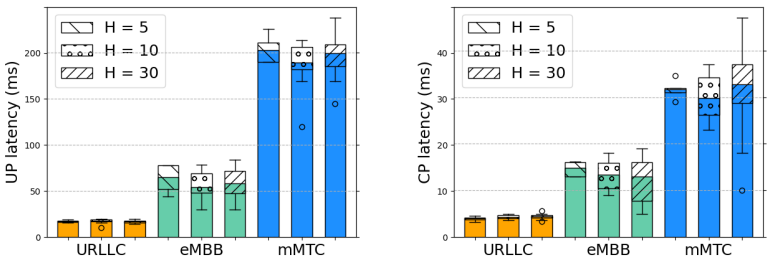
$$\sum_{k \in \mathcal{K}} z_j^k \leq 1, \ \forall j \in V_{gNB} \tag{38}$$

$$z_j^k, \sigma_j^k \in [0,1], \ \forall k \in \mathcal{K}, j \in V_{gNB} \tag{39}$$

IV. Performance Evaluation and Discussion

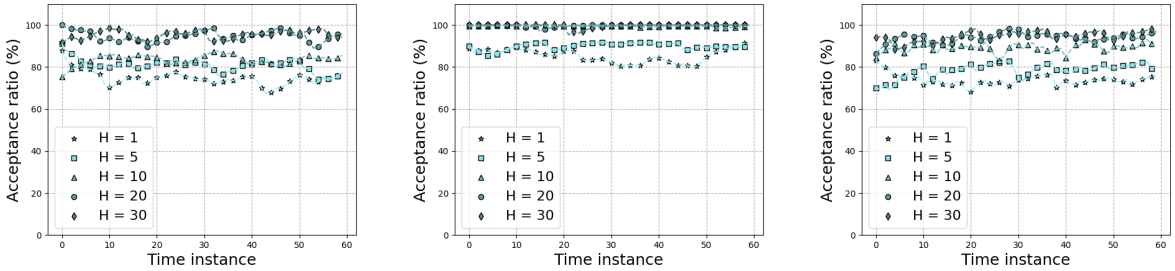
A. Simulation Setup

We consider a substrate network consisting of seven general-purpose servers at four levels of hierarchy and 6 gNB nodes connected to the third-level servers as illustrated in Fig. 1. L1, L2, L3, and L4 servers have CPU (cycle/s), storage (GB), and RAM (GB) capacities of $(72, 144, 288)$, $(36, 72, 144)$, $(18, 36, 72)$, and $(6, 12, 24)$, respectively. The backhaul links have a capacity of 2Gbps while the fronthaul links can support 1Gbps of traffic. For the sake of simplicity, the resource scaling factor χ_k^{ν} is assumed to be the same for all resource types $\nu \in \mathcal{T}$ and is equal to the number of users requesting slice k. Each VNF requires $(0.1\chi_k, 0.2\chi_k, 0.4\chi_k)$ units of CPU, STO, and RAM resources. The simulation environment is implemented in Java and we use the CPLEX commercial solver for solving the *DET-SMNS* and *RNSR* models. We assume that the demand density functions $(\mathcal{F}^T)$ change hourly, i.e. the value of the macro-slot is one hour. Moreover, the duration of the micro-slot is set to one minute, i.e. $N_T = 60$. We consider three slices, namely, eMBB, URLLC, and mMTC with the SFC graph as depicted in Fig. 1. The NS simulation parameters are given in Table I. We use the 3-shortest path algorithm to construct the set $\mathcal{P}(i \rightarrow j)$.



(a) UP latency for different network slices (b) CP latency for different network slices

Fig. 2: UP and CP latency for URLLC, eMBB, and mMTC slices



(a) eMBB slice (b) URLLC slice (c) mMTC slice

Fig. 3: Average request acceptance ratio for URLLC, eMBB, and mMTC slices

TABLE I: Simulation Parameters

Slice Type	d_k^{UP}/d_k^{CP}	mean value of $\underline{R}_j^k$	$\overline{R}_j^k$
eMBB (k=1)	100/20 ms	50	0.6
URLLC (k=2)	25/5 ms	10	0.2
mMTC (k=3)	300/60 ms	5	0.2

B. Numerical Results

Figure 4 illustrates the UP and CP latency for the URLLC, eMBB, and mMTC slices. We consider the results of 5 and depict both the average latency and the profile of latency values for different NSs, in the cases of $H = 5, 10, 30$. We observe that in all cases, the obtained UP and CP latencies are below the corresponding maximum tolerable latency given in Table I. Thus, the proposed two-time-scale resource provisioning algorithm provides solutions that meet the NSs' QoS requirement in terms of E2E UP and CP latency. In Fig. 5, the average acceptance ratio of different slices averaged for the duration of 60 micro-slots is illustrated. The percentage of accepted requests is obtained from the supported traffic solution of the *RNSR* problem, and averaged over different gNBs, i.e. Avg. acceptance ratio of slice $k = 100 * \sum_{j \in V_{gNB}} \sigma_j^k / |V_{gNB}|$. Figure 3a, 3b, and 3c denote the average acceptance ratio for the eMBB, URLLC, and mMTC slices, respectively. In this experiment, we change the number of realizations from 1 to 30. It is observed that as the value of H increases (more realizations are considered as an input to *DET-SMNS*($\mathcal{H}$)), the acceptance ratio enhances for all slices. This is due to the fact that as the number of realizations increases, the solution of the *DET-SMNS*($\mathcal{H}$) problem is a better approximation of the original *SMNS* problem modeled as a SMIP as explained in Section III-A. In summary, our proposed algorithm for the E2E network slicing under demand uncertainty operates in two phases with long and short time scales corresponding to the E2E resource provisioning followed by the RAN slice adjustments.

REFERENCES

[1] S. Zhang, "An overview of network slicing for 5g," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, 2019.

[2] Y. Shi, Y. E. Sagduyu, and T. Erpek, "Reinforcement learning for dynamic resource optimization in 5g radio access network slicing," in *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Comm. Links and Networks (CAMAD)*, 2020, pp. 1–6.

[3] R. Su, D. Zhang, R. Venkatesan, Z. Gong, C. Li, F. Ding, F. Jiang, and Z. Zhu, "Resource allocation for network slicing in 5g telecommunication networks: A survey of principles and models," *IEEE Network*, vol. 33, no. 6, pp. 172–179, 2019.

[4] Q.-T. Luu, S. Kerboeuf, and M. Kieffer, "Uncertainty-aware resource provisioning for network slicing," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 79–93, 2021.

[5] A. Papa, A. Jano, S. Ayvaşık, O. Ayan, H. M. Gürsu, and W. Kellerer, "User-based quality of service aware multi-cell radio access network slicing," *IEEE Transactions on Network and Service Management*, 2021.

[6] T. Guo and A. Suárez, "Enabling 5g ran slicing with edf slice scheduling," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, 2019.

[7] A. Oliveira and T. Vazao, "Mapping network performance to radio resources," in *2022 International Conference on Information Networking (ICOIN)*. IEEE, 2022, pp. 298–303.

[8] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[9] A. Shapiro, "Simulation-based optimization convergence analysis and statistical inference," *Stochastic Models*, vol. 12, pp. 425–454, 1996.