Slice-aware Open Radio Access Network planning and dimensioning

Parisa Foroughi, Philippe Martins, Patrice Nivaggioli, Jean-Louis Rougier

Slice-aware Open Radio Access Network planning and dimensioning

Parisa Foroughi[†‡], Philippe Martins[†], Patrice Nivaggioli[‡], Jean-Louis Rougier[†]
[†] Télécom Paris - Department of Networks and Computer Science – [‡] Cisco Systems France
Email: parisa.foroughi, martins, rougier@telecom-paris.fr; pforough, pnivaggi@cisco.com

2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall) | 978-1-6654-5468-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/VTC2022-Fall57202.2022.10012946

***Abstract*—The fifth-generation (5G) of mobile networks and beyond is to host a variety of services for industry verticals with a diverse range of requirements. Network slicing (NS) is considered to be the fundamental enabling technology to address legacy networks' shortcoming, by tailoring logical virtual networks, called network slices, over the same infrastructure. Adopting the concepts of virtualization and open interfaces, Virtual radio access networks (vRAN) and Open RAN (ORAN) are two of the most promising architectures proposed for slicing radio access networks. To realize the efficient deployment (i.e. increase flexibility, scalability ,and decreased CAPEX and OPEX) of these architectures, a proper network planning approach is essential. This paper introduces a novel approach to planning and design of the ORAN architecture that takes into account QoS, CAPEX, OPEX and the transport network, simultaneously. The ORAN slice planning and design is formulated as a multi-objective optimization with binary variables and solved by simulated annealing.This paper provides a comprehensive discussion of the results. The proposed approach can be used in designing 5G ORAN network slices but also can be used as a transition network solution to integrate the 4G tier together with 5G for enabling a smooth and less costly transition.**

***Index Terms*—optimization, RAN planning, network slicing, transition network.**

I. Introduction

Next generation of mobile networks are expected to host diverse services and applications ranging from massive internet of things (IoT) to autonomous driving, augment reality, etc. These services impose a wider range of performance and cost requirements onto the legacy networks [1]. The three main categories of services, i.e, Enhanced Mobile Broadband Connectivity (eMBB), Massive Machine Type Communications (mMTC) and Ultra-Reliable Low Latency Communication (URLLC), prioritize high-speed and capacity, high density support and reliability, and low latency for service provisioning, respectively [2].

The diverse range of requirements, imposed by new set of services from vertical industries, do not integrate well with the traditional one-size-fits-all approach of network planning, design and operation. In the traditional design approach, all services are provisioned over the same network [1]. In addition, the existing network infrastructures incorporate a variety of proprietary and monolithic devices that makes the integration of new services more difficult, costly and less flexible. Thus, new technologies, deployment and design solutions are expected to be cost-efficient and flexible in both infrastructure and operation management. Network slicing is the key enabler technology proposed to address the above mentioned challenges.

Network slicing is the concept of sharing the same physical infrastructure for building several logical networks known as network slices. Each Network slice is designed and tailored to requirements of a specific service. The virtual RAN (vRAN) and open RAN (ORAN) alliances are the 2 main next-generation RAN (NG-RAN) architectures proposed for integrating NS. The vRAN architecture decomposes RAN to 8 virtual functions and aims at deploying them on common-off-the-shelf (COTS) hardware to decrease the capital and operation expenditure (CAPEX and OPEX) [3]. The ORAN alliance packages the virtual functions into 3 main units called radio, distributed/data and central units (RU, DU and CU). Thus, it considers low level splits which connects RU and DU via fronthaul. The high-level splits connect DU and CU via midhaul. It also splits the core network into the user and control plane based on the software defined networking (SDN) concept. However, a proper planning and deployment approach is essential to fully meet the goals of the above architectures. Moreover, to realize the requirements of some of the services, it is envisioned that the components (RU, DU and CU) are to be placed in multi-access edge computing [2] clouds/servers distributed closer to the end user. To realize the ORAN architecture's future vision, an adequate design and deployment and management plan is essential. Thus, the research community has proposed several optimization approaches [4]–[11] to address the radio resource managements (RRM) challenges. On the other hand, only a few works tackle the design and deployment of ORAN architecture.

The addition of the network slicing concept to radio access networks results in the introduction of new challenges, such as split option selection, assignment of the radio resources to several services, and the location options of different components in vRAN and ORAN architecture. There exist only a few works that address the packaging, as well as the placement of the functions in a, RAN equipped with distributed MEC servers. Morais et al. [3] propose an exact model for placing virtual functions in a vRAN architecture for a given set of RUs by minimizing the computation resources and maximizing the aggregation of functions. Murti et al. [12] propose an optimal deployment approach for placing DU and CU radio functions. Unlike our proposed solution, neither of the above articles considers the radio coverage of different slice types or the integration of existing base stations into their

formulation. Garcia et al. [13] propose another virtual function placement approach called FluidRAN that explores the trade-off of different split options, network cost, and base station load. However, none of the above works take into account the geographical coverage of the area (i.e. coverage holes) or consider the integration of already existing eNBs when deploying different slice types. This paper's contribution can be summarized as follows:

- Introducing an optimization model for 5G ORAN planning considering geographical coverage constraints
- Empowering the proposed model from the designing phase to be compatible with the network slicing concept and to enable the support of different slice types
- Enabling a simultaneous trade-off evaluation of different network operational costs, crosshaul delay, and the transport network
- Providing a comprehensive analysis of the results and providing guidelines to generalize the model for any mobile network

In this paper, a novel approach to the planning of radio units in ORAN architectures is proposed. The proposed approach considers the assignment of slice types to the next-generation of base stations(gNBs) while providing radio coverage over the given area. The proposed approach not only alleviates the complexity of the RRM approaches for spectrum sharing by reducing the number of the variables but also models and optimizes the CAPEX, and OPEX. Moreover, it takes into consideration the fronthaul and the delay constraints, simultaneously. The gNBs which are only assigned a certain type of slice may be substituted with an eNB for reducing CAPEX. The planning model proposed in this paper is unlike the isolated hierarchical traditional approaches in RAN planning that take weeks to provide a solution. The proposed model is capable of finding a practical optimal solution in less than 3 hours for a network consisting of 50 nodes.

The remainder of the paper is organized as follows: Section II sheds light on slice aware planning, the system model and the problem formulation. Section III includes the optimization algorithm (simulated annealing) methodology. Section IV presents the numerical results and compares the impact of the different cost components in the overall design objective. Section V includes the conclusion and future works.

II. ORAN PLANNING

With the prevalence of virtualization and the need for placements of different components closer to the end user, new approaches for 5G network planning and design are necessary. The common approach in network planning is to first collect the general requirements in terms of user, service, the operational environment, and business goals. The network operator (NO) then designs a slice(i.e. selecting the split option, estimating the necessary virtual infrastructure resources, and the assignment of radio resources) based on these requirements. The next step is to map the virtual resources to the physical infrastructure. This last step wraps up the initial deployment of the network. However, to keep the network in check and obtain run-time insurance, dynamic management of resources (optical wavelengths, resource blocks, etc.) is required. Therefore, resource adjustments are made based on service demand, the number of active users, policy adaptation, and mobility management [14]. To the authors' knowledge, all existing design and planning approaches for ORAN and vRAN architectures are based on the full implementation of a 5G tier network, with the support of all slice types on every gNB. The network planning approaches have isolated hierarchical steps which could take up to weeks to design. This paper presents a novel practical approach to the slice-aware design of ORAN radio units, which can provide a reasonable solution in only a few hours. The proposed approach determines the number of slice instances and the location and assignment of slice types to the BSs. The model also provides a rough approximation of the spectrum shares required on gNBs that support more than one slice type.

A. Network model

In this work, a slice instance is defined as the combination of an RU, DU, and CU of a particular type of service. Therefore, thorough planning requires the placement of all 3 components. In this paper, the planning problem is split into two parts: i) RU placement and ii) DU and CU placement. This paper only addresses the first half of the planning problem due to space constraints. The RU placement problem includes finding an adequate number of slice instances and their types as well as their assignment to geographical locations so that radio coverage constraints are met. To provide radio coverage for a given type of service over a given geographical area, several of the slice instances are necessary. Fig. 1 depicts a simple scenario.

In this paper, the infrastructure network is modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. There exists A antenna locations in set $\mathcal{A} = \{1, ..., A\}$ which are candidate RUs for different slice instances. The candidate locations could be already existing antenna polls or only candidate locations. There are N edge nodes in the set $\mathcal{N} = \{A+1, ..., A+N\}$ which are candidate nodes with or without computation resources for placing DUs and CUs. The set $\mathcal{R} = \{A+N+1, ..., V\}$ consists of R routers with no extra compute resources which are only responsible for the data transmission and are not suitable for resource placement. The index 0 is reserved for the Evolved packet core (EPC), which in this work, is where the core network is located. Thus, the set of vertices is $\mathcal{V} = \mathcal{A} \cup \mathcal{N} \cup \mathcal{R} \cup 0$. Each edge/link of e_{ij} which connects node i to j has the capacity c_{ij}. The set $\mathcal{T} = \{t_1, t_2, t_3\}$ are the 3 types of the slices i.e., URLLC, eMMBC and mMTC, respectively. The parameter ^{t} is defined as the expected number of users to be covered by slice type t in the given area. The variables r_a^t is a binary variable which is set to 1 when node $a \in A$ is set to support the slice type t.

B. Problem formulation

Providing radio coverage for users of a given area is one of the most pivotal factors for operators when building and

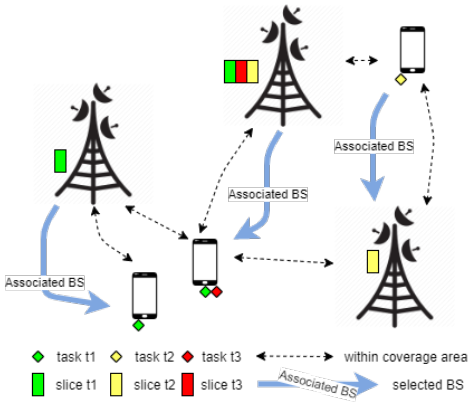

Fig. 1. Illustration of potential slice assignment to BSs and user association (note: they may not connect to their BEST choice of BS)

designing their virtual networks (slices). Every gNB in the 5G tier network is expected to support all slice types and provision their spectrum by the Bandwidth parts (BWPs) concepts [8]. The BWPs concept divides the available bandwidth between the slices and assign each particular portion to one slice with potentially different sub-carrier spacing and radio configurations. Although hosting every slice on every gNB is a straight forward way to provide radio coverage, it introduces additional complexities to the DUs and CUs as well as the radio resource scheduling (spectrum and resource block). The RU planning problem is thus formulated as follows:

The binary variable $r_a^t = 1$ denotes the association of BS a to slice type t. The parameter ξ is the maximum number of slice types to be hosted on a BS, which in this paper is set to 3. Eq. 1 ensures correct association of BSs and slice types.

$$\forall a \in \mathcal{A}, \quad \sum_{t \in \mathcal{T}} r_a^t \leq \xi \tag{1}$$

To ensure the radio coverage of every slice, this paper utilizes the simplicial topologies and Betti numbers to model the radio coverage conditions [8]. Simplicial topology is a part of algebraic topology that uses simplexes to summarize the information in a topology. A simplicial complex is a combination of vertices (i.e. a combination of base stations) that intersect each other. The number of elements in the combination gives the dimension of the simplicial complex. A k dimension simplicial complex is called a k-simplex. Thus, a vertex is a 0-simplex, an edge is a 1-simplex, a triangle a 2-simplex, a tetrahedron a 3-simplex, etc. Betti numbers (noted β_0, β_1,...,β_k) are the dimensions of each homology group. Their geometric meaning is defined to be the number of k-dimension holes in the network. Thus, β_0 represents the number of related/connected components in the network, β_1 represents the number of coverage holes, β_2 represents the number of zones where there is no 2-connectivity, etc. Eq. 2 and 3 impose the absence of any coverage holes and the connectivity of BSs in the network for every slice type, respectively.

$$\forall t \in \mathcal{T}, \quad \beta_1(a, a \in \mathcal{A} \quad where \quad r_a^t = 1) = 0 \tag{2}$$

$$\forall t \in \mathcal{T}, \quad \beta_0(a, a \in \mathcal{A} \quad where \quad r_a^t = 1) = 1 \tag{3}$$

Another condition to factor in when planning radio networks is the number of users to be covered by the area. In this paper, the spectrum available at BSs is assumed to provide Λ^t number of users for slice t if it were to provide all its BW to that slice. To simplify the model the whole available BW load of a BS is considered to be 1 and thus, the following conditions will ensure that the spectrum resources of the active BSs are enough to provide for the planning demand in terms of user numbers.

$$\forall a \in \mathcal{A}, \quad \sum_{t \in \mathcal{T}} \rho_{gnb}^t \leq 1 \tag{4}$$

The notation $\rho^t \in [0, 1]$ is the portion of spectrum resources on gNBs for slice t to realize the user demand λ^t when the users are divided equally between the gNBs. Note that in this calculation, the eNBs are considered to be at a full load of their spectrum$\rho_{enb}^t = 1$. Thus, the condition to fulfill the user demand is as follows:

$$\forall t \in \mathcal{T}, \quad \sum_{a \in \mathcal{A}} r_a^t \rho^t \Lambda^t \geq \lambda^t \tag{5}$$

Although in this work, the RU selection and placement are treated as separate problems from the DU and CU placement, the location of RU and its link capacity and delay to other nodes plays an important part in the placement of the DU and CU. Therefore, to factor in the transport network in the cost function the following cost is defined:

$$\forall t \in \mathcal{T}, \quad \gamma^{fh,t} = |1 - \frac{\sum_{a \in A} C_a r_a^t}{C_{FH} \times \frac{\sum_{a \in \mathcal{A}} r_a^t}{\sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}} r_a^t}}| \tag{6}$$

where C_{FH} and C_a are the total available fronthaul capacity and the node fronthaul capacity of node a, respectively which are calculated as follows:

$$C_{FH} = \sum_{a \in A} \sum_{j \in N \cup R} I_{aj} c_{aj}$$

$$\forall a \in \mathcal{A}, \quad C_a = \sum_{j \in R \cup N} I_{aj} c_{aj}$$

The binary parameter I_{aj} is 1 when the physical link e_{aj} exists between node a and node j. Note that Eq. 6 is an indication of overall fornthaul link capacity available to a slice type compared to their supposed share of fronthaul capacity based on their assigned number of BSs in the network. The condition above, though not perfect, is a step towards considering the transport network in the selection of the RUs. Another important factor in RU placement is the capital expenditure (CAPEX) of BSs which is denoted as γ^a. The cost of BSs, existing or not formulated as Eq. 7 follows:

$$\forall a \in \mathcal{A}, \quad \gamma^{node,a} = (\sum_{t \in T} r_a^t) \cdot \gamma^{comp} + \gamma^{setup} \cdot \gamma_a^{ru} \tag{7}$$

where γ^{comp} is the complexity cost (in terms of resource allocation and management) of placing different slice types in one node, γ^{setup} is the cost of prepping up the node if it already exists, and γ_a^{ru} factors in whether the node is existing or just a candidate. If existing $\gamma_a^{ru} = 1$. Otherwise, the candidate node will be considered to have a link to the closest edge nodes and routers with a default link capacity value. Therefore, γ_a^{ru} already encapsulates the node and link setup costs for candidate nodes. The cost of transport is assumed to be proportional to the distance of fiber required for the Ethernet connection.

To integrate the difference in the quality of service provided by eNB and gNBs, the cost function γ^{delay} is introduced. Eq. 8 integrates the effects of using larger sub-carrier spacing in gNBs which will potentially result in less user experience delay. $f(r_a^t)$ is a factor defined based on sub-carrier spacing. In this paper, $f(r_a^t)$ of the gNBs (60 kHz) is set to 0.25 and eNBs (15 kHz) is 1. $\tau^{fh,t}$ is the delay budget of slice type t for split option 7.2.

$$\forall t \in \mathcal{T}, \quad \gamma^{delay,t} = \tau^{fh,t} \times f_1(r_a^t) \tag{8}$$

The final term of the cost function is the slice-awareness factor. The cost function in Eq. 9 provides a bias in placing the more *sensetive* slice instances closer to the edge nodes. Note that the priority can be given to any of the slice types based on policies by changing the parameter values as long as placing the slice closer to the edge nodes is desired. $f_2(t)$ is a function of slice type that prioritize the slice types. In this paper, $f_2(t) = b_1\delta(t-t_1) + b_2\delta(t-t_2) + b_3\delta(t-t_3)$ where $\delta(t)$ is a the Dirac function.

$$\forall a \in \mathcal{A}, \quad \gamma^{sa,a} = \sum_{t\in\mathcal{T}} r_a^t\big(f_2(t) \times e_a + (1-e_a) \times b_0\big) \tag{9}$$

$e_a = 1$ if an edge node is not in direct neighbor nodes of the antenna node a. Moreover, $b_0 < \{b_1, b_2, b_3\}$ to emphasize the advantage of having an edge node as neighbor. The total cost function is thus as defined in Eq. 10. Note that to be able to add the different components, the cost functions are scaled with their maximum possible value. Note that some of the cost functions may never reach their maximum value throughout the optimization based on the topology thus, there might be differences in the amplitude of the cost functions.

$$\Gamma^{Tot} = w_1 \underbrace{\sum_{t\in\mathcal{T}} \gamma^{fh,t}}_{\gamma^{fh}} + w_2 \underbrace{\sum_{a\in\mathcal{A}} \gamma^{SA,a}}_{\gamma^{sa}} + w_3 \underbrace{\sum_{a\in\mathcal{A}} \gamma^{node,a}}_{\gamma^{node}} + w_4 \underbrace{\sum_{t\in\mathcal{T}} \gamma^{delay,t}}_{\gamma^{delay}} \tag{10}$$

The optimization is, therefore, formulated as follows:

$$\min_{r_a^t,\rho^t} \ \Gamma^{Tot}$$
$$subject\ to; \quad (1),(2),(3),(4),(5)$$

III. Methodology: simulated annealing

This section provides the algorithm used to solve the optimization problem presented in Section II-B. The optimization approach used in this paper is simulated annealing which is often used in global optimization of large solution spaces [15].

In general, simulated annealing methods work as follows. The algorithm starts from a positive temperature value. The temperature decreases at each step and is proportional to the probability of accepting worse solutions throughout the optimization. The higher the temperature, the more the possibility of accepting worse solutions. At each iteration/time step, a random solution close to the current one is selected and evaluated against the previous valid solution. The solution is accepted or rejected based on the temperature of the system at that step. The optimization is often stopped after a certain number of steps.

In this paper, the temperature (T) is set to cool down around the 300^{th} step thus $T_0 = 1/300$. The minimum acceptable temperature is set to 0.01. The acceptance probability is defined as presented in [15]. However, the following minor adjustments are made to the algorithm's general form.

- **Random neighbor generation:** the solutions are not generated in a completely random manner. At each iteration, one type of slice is randomly selected and is then removed from a randomly selected BS, which already supports the mentioned slice type. The selected slice type instance is only removed if the radio coverage of that slice type remains intact. The algorithm proposed to generate neighbor solutions is presented in Algorithm 1.For simplicity, the set of $r_a^t, \rho_a^t \ \forall a \in \mathcal{A}, \ \forall t \in \mathcal{T}$ is replaced with the matrix $\pi \in R^{\mathcal{A}\times\mathcal{T}}$ and matrix $\varrho \in R^{\mathcal{A}\times\mathcal{T}}$ notations, respectively. π_i is the solution in iteration/step i of the optimization.
- **Reset mechanism:** in cases where the algorithm fails to find a valid new neighbor 20 times in a row, the core solution is changed to one of the previous valid ones. This process allows for exploring more of the solution space.

Note that since the objective function is the sum of several cost functions there are several feasible solutions. In this paper, all feasible solutions are stored in a database but only the selected ones according to the simulated annealing method are depicted in the results.

IV. Results analysis

In this section, a comprehensive discussion of the simulations is presented. The network graph $\mathcal{G}$ is randomly generated over a $5.5km \times 5.5\ km$ area. In the generated network $V = 50, A = 19, R = 12$ and $N = 19$. The link capacities are randomly assigned from $4Gb/s$ to $16Gb/s$. The fronthaul delay budget is considered to be $100, 300$ and 5000 microseconds for *URLLC*, *eMBB* and *mMTC*, respectively. The user coverage demand is 1000, 80 and 20000 mobile users for *URLLC*, *eMBB* and *mMTC*, respectively. The maximum iteration is set to 5000. Moreover, for simplicity, the weights in Eq. 10 are presented as $W = [w_1, w_2, w_3, w_4]$ in this section.

Algorithm 1: Random neighbor solution generator in step i

Input: System parameters $\mathcal{G} = (\mathcal{V}, \mathcal{E}), \pi_{i-1}$, Algorithm parameters T

Output: π_i that is the neighbor solution to π_{i-1}

```
Select a random active BS k from the solution π_{i-1};
ζ ← Get neighbors of the BS k in terms of radio coverage;
flag ← True ;
t ← random slice type from 𝒯
while flag do
    π_i ← π_{i-1}
    for bs in ζ do
        if r^t_bs = 1 ∈ π_{i-1} then
            π_i ← update π_{i-1} by setting r^t_bs ← 0
            flag ← False
            break
    flag ← False
```

The first step in a weighted multi-objective optimization is to assign the proper weights. Although the cost functions are normalized appropriately, some of the cost functions are dependent on the network topology and the value calculated in the first reference step is not necessarily the maximum cost value of that cost function. Thus, the scaling applied in this paper can be done by division of an approximation or ideal expectation for the cost function.

Fig. 2 shows the values of different cost components in Γ^{Total} when the same weights,$[0.25, 0.25, 0.25, 0.25]$, are assigned to all components. It can be seen that the node and delay cost have maximum values of 1 while the other two cost functions have values less than 0.5 throughout the optimization. Therefore, the weight vector $[0.25, 0.25, 0.25, 0.25]$ is effectively similar to applying $[0.25, 0.5, 1, 1]$. For instance, when node cost can have an impact of up to 1 in the cost function, the slice-awareness cost can only go as far as changing the total cost up to 0.5. Thus, the effective weights are calculated if the cost functions were to have the same range of values. The reason behind defining this concept is that the range of the cost function values can slightly change in different networks and so the effective weight can help in generalizing the explanation. In simple terms, to have a balanced cost function and avoid missing the impact of different cost components, proper scaling of the cost functions should be done. This normalization not only allows for a fair comparison of different costs but also fair optimization progress. To avoid any confusion, the weights in the remainder of the paper are translated to effective weights as described above.

Note that although often the total sum of weights is assumed to be a constant value, this assumption is solely for the sake of making the total cost comparable. However, in this paper, the goal is not to find the optimal weights for the optimization but rather to compare the impact of the different cost functions as each of them has its practical importance. Thus, instead of having a constant total of weights to allow for a total cost comparison, the total cost is presented as a percentage relative to the cost function of a reference deployment (The reference deployment is considered to have all BSs support all slice types) when each weight set is applied. It is noteworthy to mention that, though finding the best weights configuration is not the ultimate goal of this study, the best option for this network after several experiments based on the optimization progress and the cost values is found to be $W1$.

Fig. 2 also shows that γ^{node} has a step-like reduction during optimization while γ^{delay} and γ^{sa} has a gradual, almost linear decrease in their values. Finally, γ^{fh} has the least changes and does not appear to follow a particular pattern and solely depends on the selection of nodes and their assignment to different slice types. Based on the above observations, each cost function's weight is elevated to facilitate exploring the effect on the total cost and optimization progress.

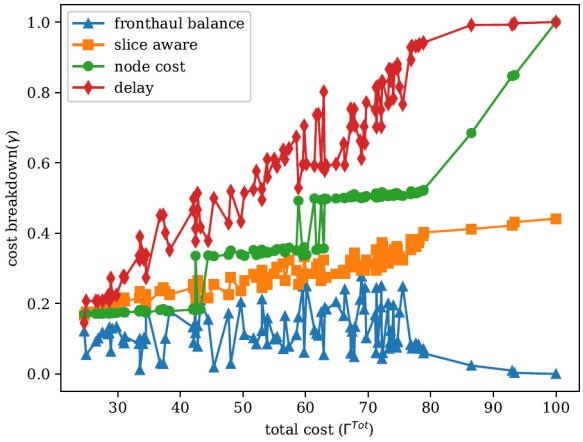

Fig. 2. Normalized total cost in percentage

Fig. 3 shows the total normalized cost throughout the optimization. The optimal solution found for the weights $W1, W2, W3, W4, W5$ is at the 24%, 38%, 40%, 22% and 26% of the full 5G deployment costs, respectively. It is important to understand that, often, elevating the weights of a certain cost function mostly depends on the priorities of the operators. While one operator may have limited access to transport network resources (thereby using elevated w_1), another might be prioritizing the cost of BS deployment (preferring elevating w_3). It can be seen from the results that the proposed model and architecture can reduce the cost by 70% on average in the given network in this paper. It can also be seen that the first 500 steps of the optimization explore numerous options in the solution space. The solution space will later be further explored thanks to the reset mechanism explained in Section III while maintaining the convergence of the algorithm. It is seen that different weights have different convergence speeds. For instance, $W4$ requires 4232 steps to converge while $W2$ only requires 985 iterations.

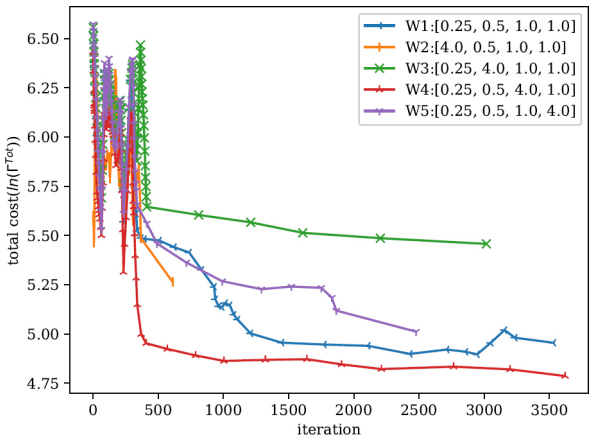

Fig. 3. The ln of the normalized total cost by the iteration number

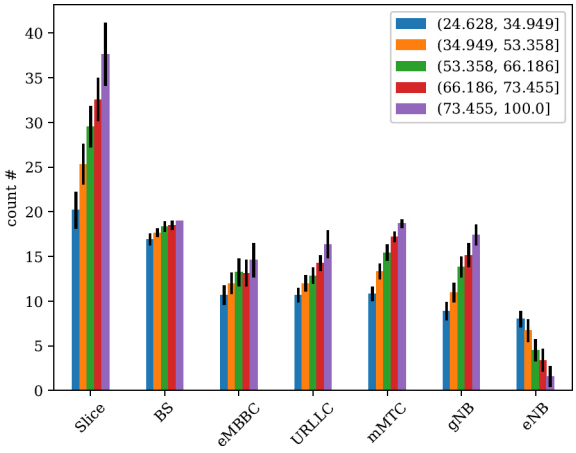

Fig. 4. Different parameters relation to the different quantized cost ranges throughout one optimization for weight configuration W1

Fig. 4 shows the relationship between the different more tangible factors in planning (the number of active BSs, slice instances, and their distribution between different types) when Γ^{tot} is quantized to 5 buckets. The 5 buckets ranges can be seen in the legend and is differentiated by different colors in the figure. In the first column, it can be seen that the fewer the number of slice instances, the less the cost function will be. The same conclusion applies to the number of BSs and different types of slices in the next two columns. Whereas, the more eNBs in the network, the less the cost function as can be seen in the last column. Yet, the most interesting point to mention is found when comparing the error lines. When the 2 error lines of two bars is overlapping, it means that the 2 bars does have similarity in their solution with respect to that column. The means, at two cost ranges we can have solutions that have similar slice number, BS number etc. In other words, for a given cost there could be several different solutions that have a different number of slice instances, active BSs, and different associations of slice types to BSs and that is because the error bars overlap in some cases. It can also be seen in the figure that the mMTC experiences the least number of variations in the same range of costs and thus have almost similar number of slice instances which is due to the huge number of users it can cover by utilizing only one BSs full spectrum. Since a fixed minimum number of BSs is required to provide coverage for each slice type, often the requested number of users are easily covered by the already associated BSs to this slice types. Therefore, the solutions in fixed cost ranges experience less variation.

In multi-objective optimizations it is important to understand the reasons behind different cost function behaviors, to find make the best decision for optimizing a certain network. Fig. 5 breaks down the multi-objective cost function to its building blocks. The γ^{fh} values always will have their least value in the beginning of the optimization. That is because all BSs are sharing the same resources and host the same slice types thus there is no imbalance in their accessible fronthaul resources. It is also seen that it starts to decrease from a certain total cost. Often, it is expected to have oscillations in this cost function since no particular measure is taken in the neighbor generation, which directly results in the decrease of fronthaul capacity gap. However, note that elevating its weight to 4 time the other cost functions certainly imposes the total cost to follow its behavior (see $W2$), though it does not necessarily result in the best values for the other cost functions. Another different scenario is the case of elevated w_3 to 4 times the max value. It is seen that it results in step-like behavior in all the cost functions but γ^{fh}. That is mostly due to the huge impact of γ^{node}. The node cost function is directly impacted by the number of active BS, as well as the state of BS (existing or only candidate). Moreover, it also has a smaller component connected to the number of slice instances on each BS, which represents the complexity of DUs and schedulers. Due to its impact on the number of slice instances, the γ^{sa} and γ^{delay} is also affected. The reason they do not follow the exact patterns is that they integrate some other details in their cost function, which are the location of different slice instances and the type (eNB, gNB) of active BSs, respectively. Note that the other combinations of weights will often result in a decreasing trend in all cost functions but γ^{fh}, which is expected.

Fig. 6 shows the number of extra users, compared to the given user demand in the explored valid solutions. The solutions are presented by their number of eNBs and gNBs. It can be seen that generally more gNBs (see columns) will result in more extra users but it is more costly. It is also seen that the more eNBs in the network, the more users can be covered. However, deployment of more BSs introduces more variables to the radio resource (i.e. BS spectrum) optimization and makes it more complex. Note that there are still cases where the above conclusion does not necessarily hold. For instance, in the case of 9 gNBs (see the row with gNB value of 9), it can be seen that the number of extra users in eMBB and

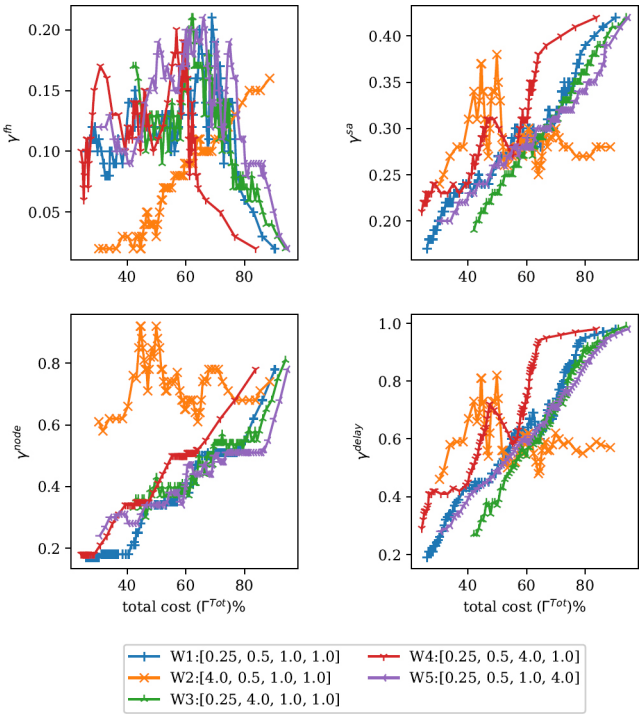

Fig. 5. The cost breakdown of different weights by the total cost

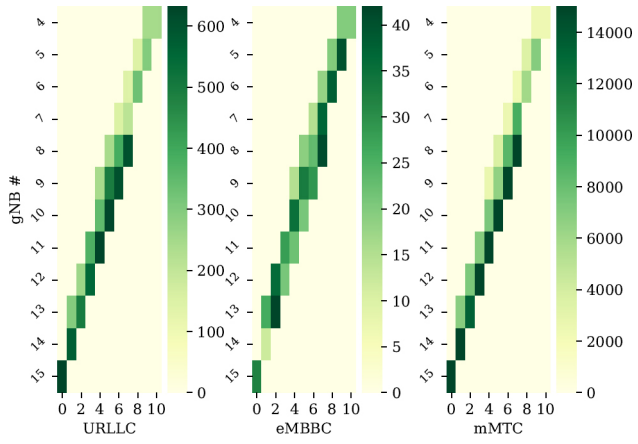

Fig. 6. The number of extra 5G users for $W1$ case

URRLC, when accompanied with 5 eNBs, is more than when it considers 6 eNBs with the same 9 gNB. This difference is due to the number of slice instances in the solution. In the case of 9 gNB and 5 eNB the number of mMTTC slice instances is more than the other solutions with the same gNB numbers.

V. CONCLUSIONS

This paper proposes a novel a slice-aware optimization model for slice design and planning of the radio access networks for the ORAN architecture. The proposed model optimizes the CAPEX and integrates elements in its cost functions to optimize OPEX by the right association of slice instances concerning available transport network capacity and edge nodes. It also integrates the quality of service of users by considering the fronthaul delay in the optimization function. The paper provides a comprehensive analysis of the results and the impact of different cost functions. The proposed model resulted in an average of 70% reductions in cost for the given network. The proposed model can be utilized in planning full 5G tier networks as well as designing transition networks by integrating 4G base stations to further reduce the CAPEX.

Future works include analysis of the impact of different coverage demands on optimization, as well as complementing the model with a placement strategy for the DU and CUs.

REFERENCES

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

[2] P. Foroughi, H. Beyranvand, M. Gagnaire, and S. Al Zahr, "User association in hybrid uav-cellular networks for massive real-time iot applications," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020, pp. 243–248.

[3] F. Z. Morais, G. M. F. De Almeida, L. L. Pinto, K. Cardoso, L. M. Contreras, R. da Rosa Righi, and C. B. Both, "Placeran: optimal placement of virtualized network functions in beyond 5g radio access networks," *IEEE Transactions on Mobile Computing*, 2022.

[4] B. Khodapanah, A. Awada, I. Viering, J. Francis, M. Simsek, and G. P. Fettweis, "Radio resource management in context of network slicing: What is missing in existing mechanisms?" in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–7.

[5] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí, "On the configuration of radio resource management in a sliced ran," in *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2018, pp. 1–6.

[6] H. M. Soliman and A. Leon-Garcia, "Qos-aware frequency-space network slicing and admission control for virtual wireless networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2016, pp. 1–6.

[7] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On radio access network slicing from a radio resource management perspective," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166–174, 2017.

[8] A. de Javel, J.-S. Gomez, P. Martins, J.-L. Rougier, and P. Nivaggioli, "Slice-aware energy saving algorithm for 5g networks based on simplicial homology." in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. IEEE, 2021, pp. 1–5.

[9] P. Caballero, A. Banchs, G. De Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6419–6432, 2018.

[10] K. Abbas, M. Afaq, T. Ahmed Khan, A. Rafiq, and W.-C. Song, "Slicing the core network and radio access network domains through intent-based networking for 5g networks," *Electronics*, vol. 9, no. 10, p. 1710, 2020.

[11] T. Wang and S. Wang, "Online convex optimization for efficient and robust inter-slice radio resource management," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 6050–6062, 2021.

[12] F. W. Murti, J. A. Ayala-Romero, A. Garcia-Saavedra, X. Costa-Pérez, and G. Iosifidis, "An optimal deployment framework for multi-cloud virtualized radio access networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2251–2265, 2020.

[13] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith, and G. Iosifidis, "Fluidran: Optimized vran/mec orchestration," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 2366–2374.

[14] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain, "Network slicing for 5g: Challenges and opportunities," *IEEE Internet Computing*, vol. 21, no. 5, pp. 20–27, 2017.

[15] P. Serafini, "Simulated annealing for multi objective optimization problems," in *Multiple criteria decision making*. Springer, 1994, pp. 283–292.