Isolation Aware CU-DU Mapping for Multi-Tenant 5G O-RAN Slices

Nabhasmita Sen
Department of CSE, IIT Hyderabad
cs17resch11001@iith.ac.in

Antony Franklin A
Department of CSE, IIT Hyderabad
antony.franklin@cse.iith.ac.in

***Abstract*—Resource efficiency is critical for 5G and beyond 5G (B5G) mobile networks. Leveraging different Radio Access Network (RAN) functional splits, baseband functions in 5G are disaggregated into three main components- Radio Unit (RU), Distributed Unit (DU), and Centralized Unit (CU). These disaggregated components can be placed in different geographical locations, leading to a higher flexibility and efficiency in RAN. However, such disaggregation makes the placement of baseband functions challenging due to the constraints imposed by the requirements of network slices, baseband functions, limited capacity in processing nodes and transport links, etc. It becomes even more challenging when slices of multiple tenants have different isolation requirements regarding the baseband functions. In this work, we address the problem of resource-efficient baseband function placement for multi-tenant slices in 5G Open RAN (O-RAN). We formulate the problem as an Integer Linear Programming (ILP) based optimization model to minimize the cost of processing and bandwidth resources. We consider the various requirements of delay, data rate, and sharing policies of multi-tenant slices, as well as limited resource capacity in the network while placing the functions. We perform extensive simulations to analyze the behavior of our model and show that it incurs lesser cost than baselines while placing the functions in the network. To deal with the high computational complexity of ILP, we also propose a low-complexity heuristic algorithm to achieve reasonable performance in significantly less time.**

***Index Terms*—Baseband function, Multi-tenancy, Network Slice, Open Radio Access Network.**

I. INTRODUCTION

5G and Beyond 5G (B5G) mobile networks are becoming ubiquitous for providing a broad range of services to a large number of users. Efficient management of infrastructure resources is pivotal to bring down the Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) of 5G Mobile Network Operators (MNO). In the traditional Radio Access Network (RAN), all the functions of a base station (also known as baseband functions) are processed in dedicated hardware at the cell sites. However, due to temporal load variations, most of the infrastructure resources remain idle, leading to a higher cost for the operators. With the introduction of functional split in RAN, baseband functions can be disaggregated and placed separately in geographically distributed locations by virtualizing them. According to Open RAN (O-RAN), a base station is disaggregated into three main components- a) Radio Unit (RU), b) Distributed Unit (DU), and c) Centralized Unit (CU). Each of these components performs specific functionalities of a base station. Such disaggregation and virtualization help in increasing the flexibility of the network. However, on the downside, this makes the placement of baseband functions challenging due to various constraints imposed by requirements of slices, baseband functions, and available capacity in the network. Besides, multiple tenants need to coexist over the same infrastructure with different isolation requirements. This makes the efficient placement of the baseband functions even more challenging. There are various factors that can influence the placement of baseband functions, described as follows.

1) The disaggregated functions have different requirements in terms of delay and data rate. For example, the lower layer functions have stringent delay and more bandwidth requirements than higher layer functions [1].
2) 5G and B5G networks will support services with different delay and data-rate requirements, such as a) enhanced Mobile Broadband (eMBB), b) Ultra Reliable Low Latency Communication (URLLC), and c) massive Machine Type Communication (mMTC). While placing baseband functions of a specific slice type, we must ensure its requirements are fulfilled.
3) Slices of different tenants can have different isolation requirements that must be taken care of during the placement of the baseband functions.
4) The capacity of processing nodes and transport links in different locations is limited. So, all deployment options for baseband functions may not be feasible.

Recently, some of the works have considered baseband function placement [2]–[5] in 5G RAN. However, only a few of them consider all these factors together. Most of the previous works do not consider different delay requirements of slices. Moreover, only a few works consider different isolation requirements of multi-tenant slices [6], [7]. In this paper, we consider all the above-mentioned factors for the cost-efficient placement of baseband functions. The main contributions of this work are as follows:

- We design an Integer Linear Program (ILP) based optimization model for placing the virtualized baseband functions (CU and DU) of multi-tenant 5G O-RAN slices to minimize their deployment cost. While placing the functions we consider the delay, data rate, and isolation requirement of baseband functions as well as different slices, availability of processing and transport resources, their respective costs, etc.

- We compare the results of our proposed solution with baseline approaches and show its efficiency in providing a cost-efficient baseband function placement solution.
- To deal with the high complexity of the optimization model, we propose a polynomial time heuristic algorithm for solving the problem.

II. RELATED WORKS

Some of the research works that consider baseband function placement in O-RAN are as follows. In [8], the authors propose a gradient-based strategy for minimizing the delay of baseband function placement in O-RAN. In [9], the authors propose an MILP and a heuristic solution to maximize the user admittance ratio by properly placing the O-RAN baseband functions. In [10], the authors present an orchestration framework to provide practical solutions for challenges in O-RAN such as meeting different intents of network operators using data-driven algorithms. In [11], the authors propose a two-level RAN slicing approach for allocating the communication and computation resources in O-RAN. Authors of [12] address the service allocation scheduling (SAS) problem by proposing various algorithms for slices in O-RAN. In [13], the authors propose a Reinforcement Learning (RL) based solution to place the VNFs in O-RAN to minimize downtime in the network. Authors of [14], [15] propose an RL-based solution for selecting functional split for slices in the O-RAN to minimize energy consumption. In [16], the authors propose a heuristic and MILP-based solution for deploying O-RAN slices.

However, the above mentioned works don't consider all the factors mentioned in section I. Such as, very few works consider isolation requirement of multi-tenant slices [6], [7]. In this work, we propose an ILP and a heuristic that considers the requirements of baseband functions, different types of slices, their sharing policies, capacity constraints of processing nodes and midhaul links to minimize the deployment cost of slices.

III. SYSTEM MODEL

A base station performs a series of functions known as baseband functions. Leveraging different functional splits, the Open RAN (O-RAN) alliance has disaggregated a 5G base station into three main components: a) Radio Unit (RU), b) Distributed Unit (DU), and c) Centralized Unit (CU). RU processes the lower physical layer (Low-PHY) and has antennas for the transmission and reception of radio signals. The locations of RUs are fixed, and they also denote the cell sites. The DU performs higher Physical layer (High-PHY), Medium Access Control (MAC) layer, and Radio Link Control (RLC) layer. The CU performs upper-layer functionalities, namely Radio Resource Control (RRC) layer and Physical Data Convergence Protocol (PDCP) layer.

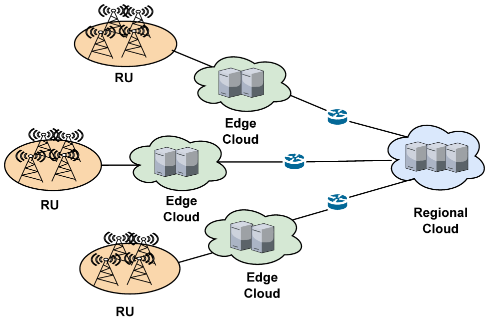


Fig. 1. RAN System Model.

We consider a hybrid cloud architecture (shown in Fig. 1) as our system model, which conforms to the O-RAN deployment scenario [17]. Multiple RUs are connected to their corresponding edge clouds and the edge clouds are further connected to the regional cloud. The links between a RU and an edge cloud, between an edge cloud and the regional cloud, and between the regional cloud and core network are known as fronthaul, midhaul, and backhaul, respectively. The edge clouds and the regional cloud consist of multiple processing nodes where the CUs and DUs can be placed as Virtualized Network Functions (VNF) in different Virtual Machines (VM) [18]. As the DUs have high data rate and low latency requirements, they are placed at the edge. The CUs, on the other hand, have relaxed latency and data requirements and hence they can be placed either in the edge cloud or in the regional cloud as long as the requirements of the slices are met [19].

IV. PROBLEM FORMULATION

In this section, we propose an Integer linear programming (ILP) based optimization model for cost-efficient placement of virtualized baseband functions (CU and DU) for multi-tenant O-RAN slices. The notations used in the formulation are listed in Table I.

1) **Decision Variables**: We consider the following decision variables in our formulation.
   (i) Binary variable x_{sf}^{m} denotes if slice s uses VM m for processing its function f or not.
   (ii) Binary variable y_{vm} denotes if VM m runs VNF v or not.
   (iii) Binary variable z_m denotes if VM m is being used for processing any function or not.

2) **Objective Function**: The objective of this optimization model is to minimize the total cost of baseband function placement. We consider the following costs for calculating the total cost.
   (i) Cost of using a VM: This is the operational cost of an active VM which is defined as,

$$C_n = \sum_{m \in M} z_m c_m \tag{1}$$

   , where c_m is the cost of using VM m.
   (ii) Cost of routing the traffic: The cost of routing the traffic in the midhaul for placing the baseband functions in regional cloud is expressed as follows,

$$C_r = \sum_{m \in \mathrm{RM}} \sum_{s \in S} \sum_{f \in F} x_{sf}^{m} tr_s c_t \tag{2}$$

where tr_s denotes the traffic requirement of slice s and c_t is the cost of routing unit traffic in the midhaul. All the costs are expressed in any Monetary Unit (MU). The total cost is expressed as,

$$Cost = C_n + C_r \tag{3}$$

Hence, the objective is finally expressed as,

$$Minimize : Cost \tag{4}$$

3) **Constraints**: The constraints for the proposed optimization model are described as follows.

a) Capacity constraint of processing nodes: The total processing done in any VM should not exceed the maximum capacity of the VM.

$$\sum_{s \in S} \sum_{f \in F} x^m_{sf} d_{sf} \leq \Gamma_m, \forall m \in M \tag{5}$$

where d_{sf} denote the processing requirement of function f of slice s.

b) Capacity constraint of transport links: Total traffic routed through a midhaul link should not exceed the capacity of that link.

$$\sum_{s \in S} \sum_{m \in \mathrm{RM}} x^m_{s1} \eta_{se} tr_s \leq K_e, \forall e \in EC \tag{6}$$

where η_{se} denotes if slice s belongs to edge e.

c) If any slice uses VM m for processing its function then m is considered to be activated.

$$z_m \geq x^m_{sf}, \forall m \in M, \forall s \in S, \forall f \in F \tag{7}$$

d) Each function of a slice can be placed on only one VM that it has a link to connect.

$$\sum_{m \in M} x^m_{sf} = 1, \forall s \in S, \forall f \in F \tag{8}$$

$$\sum_{m \in M} x^m_{sf} \psi_{sm} = 1, \forall s \in S, \forall f \in F \tag{9}$$

e) Each VM can run only one type of VNF.

$$\sum_{v \in V} y_{mv} \leq 1, \forall m \in M \tag{10}$$

f) A function can be placed in a VM if its type matches the VNF running in that VM.

$$\begin{gathered} y_{mv} \geq x^m_{sf} t^v_{sf}, \\ \forall s \in S, \forall f \in F, \forall m \in M, \forall v \in V \end{gathered} \tag{11}$$

g) All functions placed in a VM should be of the tenants that agree to share with each other.

$$\begin{gathered} x^m_{sf} x^m_{s'f'} \leq \mathrm{comp_{ss'}}, \\ \forall m \in M, \forall s, s' \neq s \in S, \forall f, f' \in F \end{gathered} \tag{12}$$

The quadratic term in this equation can be linearized as both x^m_{sf} and $x^m_{s'f'}$ are binary.

TABLE I

NOTATION AND DESCRIPTION

Notation	Description
Γ_m	Capacity of VM m
RM	Set of VMs in regional cloud
EM	Set of VMs in edge cloud
EC	Set of edge clouds
RU	Set of Radio Units
F	Set of baseband functions
S	Set of all slices
V	Set of VNFs
M	Set of all VMs
η_{se}	Adjacency matrix of an edge cloud and a slice
ψ_{sm}	Adjacency matrix of slice s and VM m
d_{sf}	Processing requirement of function f of slice s
δ_s	Delay budget of slice s
$\mathrm{lat_{e_s}}$	Delay of midhaul link of edge where slice s belongs
t^v_{sf}	Function f of slice s is of VNF type v or not
$\mathrm{comp_{s,s'}}$	Tenants of s and s' agree to share function or not
K_e	Capacity of midhaul link of edge e.

h) The DU of each slice must be placed in the edge cloud.

$$\sum_{m \in \mathrm{RM}} x^m_{s0} = 0, \forall s \in S \tag{13}$$

i) The delay of a midhaul link used by a slice for placing its CU in the regional cloud must not exceed the delay requirement of the slice.

$$x^m_{s1} \mathrm{lat_{e_s}} \leq \delta_s, \forall s \in S, \forall m \in \mathrm{RM} \tag{14}$$

V. PROPOSED HEURISTIC

Solving an ILP can be time-consuming for large-scale scenarios [20]. To tackle the scalability issue of ILP, in this section we propose a heuristic algorithm (shown in Algorithm 1). We consider the following factors for designing the heuristic. Energy consumed by active processing nodes incurs a higher cost than routing [21]. Hence, in the heuristic, we mainly focus on minimizing the cost of using Virtual Machines (VM). Moreover, a VM in the regional cloud is connected to more RUs than a VM in the edge cloud. Therefore, a VM in the regional cloud can provide a higher level of consolidation. However, as discussed in Section I, the DUs are always placed in the edge cloud, and the CUs of the delay-sensitive slices are also placed in the edge cloud. As a result, some VMs in the edge cloud must be activated to serve them. We try to utilize these activated VMs at the edge as much as possible so that a VM in the regional cloud is not activated unnecessarily. We assume enough capacity exists to serve all slices, and the cost of using a VM in the regional cloud is less than or equal to using a VM in the edge cloud [21].

We sort the slices based on their processing requirements. We first consider the delay-sensitive functions (DUs and CUs of delay-sensitive slices) for their placement. The reason for considering the delay-sensitive functions first (Lines 2-8 of Algorithm 1) is to use fewer VMs in the edge cloud. We follow the first-fit strategy while selecting a VM and check if it is

Algorithm 1: Proposed Heuristic Solution

```
   Data: Slices with their load, type, origin, and available
         network capacity.
   Result: Baseband function placement for RAN slices
1  $S' \leftarrow Sort(S)$ // Sort slices based on data
       rate requirements in decreasing order
   /* Part 1: Find the set of edge servers to
       be activated $\mathrm{EM}'$                          */
2  $\mathrm{EM}' = \phi$// Initialize $\mathrm{EM}'$
3  foreach slice $s$ in $S'$ do
4      foreach delay sensitive function of $s$ do
5          foreach VM $m$ in $\mathrm{EM}$ do
6              a. Check if the VM can support the function
                 regarding capacity, type and isolation
                 requirements.
7              b. If found, place the function and update
                 residual capacity and break.
8              c. If not found, activate another VM $m$, place
                 the function in a new VNF in that server and
                 add $m$ to $\mathrm{EM}'$, update residual capacity and
                 break.
   /* Part 2: Find the set of activated
       servers in regional cloud $\mathrm{RM}'$             */
9  $\mathrm{RM}' = \phi$// Initialize $\mathrm{RM}'$
10 foreach slice $s$ in $S'$ do
11     foreach VM $m$ in $\mathrm{EM}'$ do
12         Repeat Line 6-8
       // If not placed
13     foreach VM $m$ in $\mathrm{RM}$ do
14         Repeat Line 6-7
15         c. If not found, activate use another VM $m$, place
             the function there and add $m$ to $\mathrm{RM}'$, update
             residual capacity and break.
```

already in use and can support the current function regarding its VNF type, capacity, and isolation requirements. If such a VM is found, we place the function in that VM. Otherwise, we activate another VM, which can support the requirements of the function. At the end of this step, we get the set of VMs in the edge clouds EM' that need to be activated.

Next, we place the remaining functions (CUs of delay-tolerant slices) shown in Lines 9-16. For this, we first check the set of active VMs in the edge cloud (EM' from the previous step) if any of them can support the current function according to its requirements. This ensures that the active VMs at the edge are utilized as much as possible to prevent unnecessary activation of VMs in the regional cloud, eventually minimizing the objective function (Equation 4). If no active VM in the edge cloud can accommodate the current function, then a VM in the regional cloud is used. We check if any active VM in the regional cloud can accommodate the current function according to its VNF type and requirements. If such a VM is found, it is placed there. Otherwise, a new VM is activated in the regional cloud. After these two steps, we get the set of VMs in the edge and regional cloud i.e. EM' and RM', which will be used for placing the baseband functions.

The time complexity of the algorithm is $O(|S|log|S|) + O(|S| * |F| * |M|)$, where S is the set of network slices, F is the set of baseband functions and M is the set of VMs in the edge and regional cloud.

TABLE II
SIMULATION PARAMETERS

Simulation Parameters	Description
Number of edge clouds	3
Number of servers	13 servers
Number of server in regional cloud	4
Number of servers in each edge cloud	9
Number of VMs	65
Number of VMs in regional cloud	20
Number of VMs in each edge cloud	45
Number of tenants	4
Slice-type	eMBB and URLLC
URLLC data rate	40-50 Mbps
eMBB data rate	80-100 Mbps
URLLC and eMBB Delay	1 & 10 ms
Midhaul link delay	2-10 ms
Number of slices	12-60 slices
Server capacity	2000 GOPS
VM Capacity	400 GOPS
Normalized cost of using a VM	1 MU per VM
Normalized cost of bandwidth usage	0.0001 MU per Mbps

VI. SIMULATION AND RESULTS

The simulation parameters are shown in Table II. The network consists of 3 edge clouds and a regional cloud with 13 servers where 65 VMs are deployed. Four tenants exist in the network and slices of each tenant have different requirements of delay, data rate, and isolation. Two types of slices are there in the network- eMBB and URLLC. Slices of Tenant 1 and 2 are URLLC type and others are eMBB type. The server capacity is 2000 Gigabit Operations Per Seconds (GOPS) and the VM capacity is 400 GOPS. Considering 2x2 MIMO RUs with 20MHz bandwidth the processing and bandwidth requirements are approximated with the help of [22], [23]. The normalized cost for using a VM and bandwidth consumption per Mbps is considered to be 1 MU and 0.0001 MU, respectively. We run all simulations for 10 randomly generated input instances and report the results.

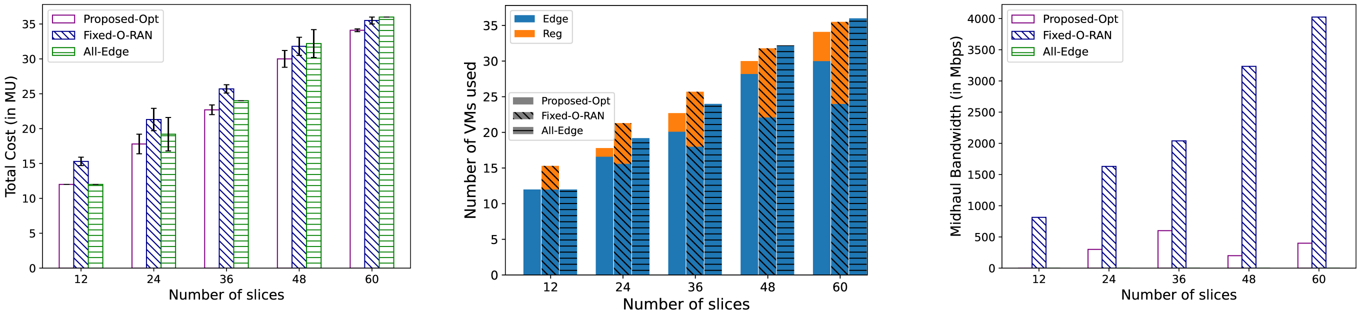
We consider the following baseline strategies to compare our proposed optimization model (Proposed-Opt).

- *All-Edge*: Baseband functions of all slices are placed in the edge cloud.
- *Fixed-O-RAN*: It follows fixed CU-DU deployment strategy. i.e., where both baseband functions of URLLC slices are placed at the edge clouds and the CUs of eMBB slices are always placed in the regional cloud [24].

We consider VNF sharing for both the baselines and that there are enough resources in the network to support all slices with different deployment options.

A. Comparison of baseline strategies

In this section, we compare the performance of the proposed optimization model with baselines. For this simulation we consider tenants 1 and 3 share their functions and tenants 2 and 4 agree to share their functions. From Fig. 2, we can observe that Fixed-O-RAN does not consider the underutilized VMs in the edge. Hence, it unnecessarily uses some VMs in the

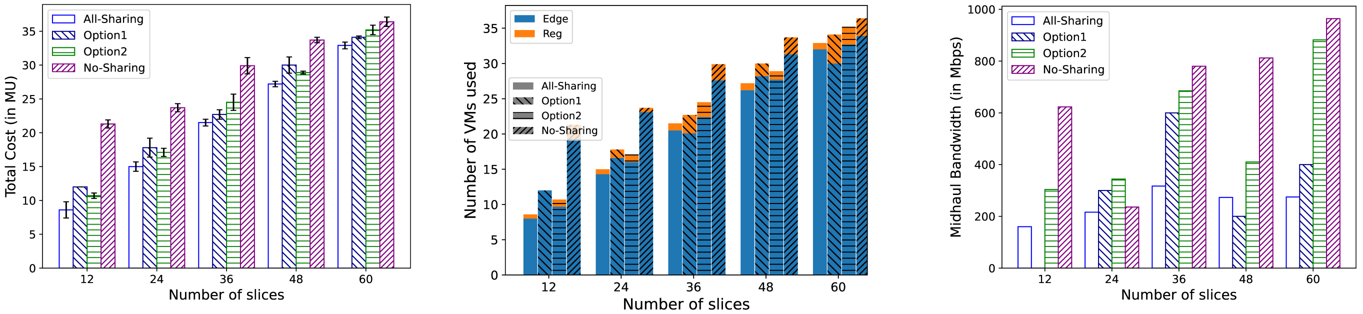


(a) Total cost

(b) Used VMs

(c) Midhaul bandwidth consumption

Fig. 2. Comparison of different baselines



(a) Total cost

(b) Used VMs

(c) Midhaul bandwidth (in Mbps)

Fig. 3. Comparison of different sharing policies

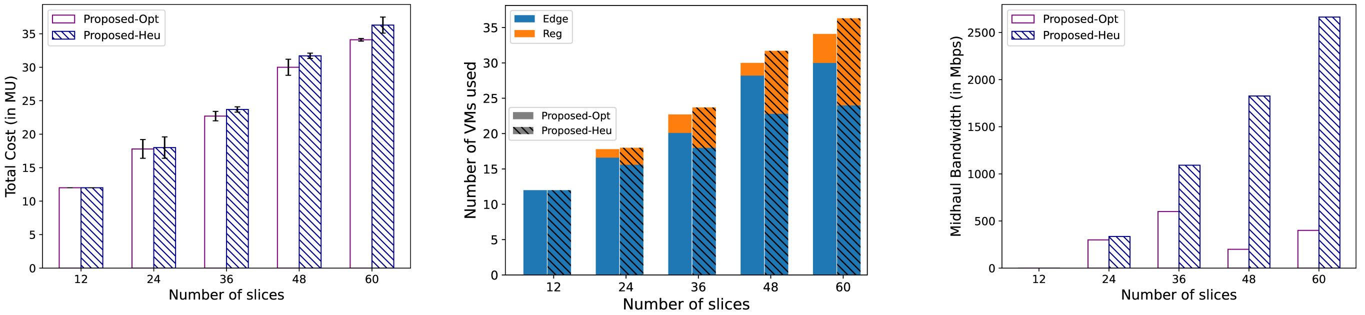
regional cloud, leading to higher costs. All-Edge places all the functions in the edge cloud. Hence, it does not use VMs in the regional cloud, which can provide a higher consolidation. As a result, it incurs higher costs by activating more VMs in the edge cloud. On the other hand, the proposed model places the functions of different slices in such a way that the overall cost is minimized. On average, the proposed model incurs around 6% and 10% lesser cost than All-Edge and Fixed-O-RAN, respectively. Fig. 2b shows the total number of activated VMs in the edge and regional cloud for all strategies, which verifies the cost shown in Fig. 2a. In Fig. 2c, we observe the midhaul bandwidth consumption in the network. As Fixed-O-RAN always places the CUs of eMBB slices in the regional cloud, it generates the highest amount of traffic in the midhaul. Whereas All-Edge places all functions in the edge, due to which it does not consume any midhaul bandwidth. The proposed model consumes some midhaul bandwidth as it places the functions in a way to minimize their deployment cost.

B. *Impact of different sharing policies*

In this simulation, we consider different sharing policies among slices of four different tenants and observe the behavior of the proposed optimization model. We observe the result for the two extreme scenarios- a) No-sharing is considered among tenants, and b) All-sharing, where all the tenants trust each other. We also consider when only some of the tenants agree/disagree to share their baseband functions (Option 1 and 2). From Fig. 3, we can observe that in the case of All-Sharing, the total cost is minimum in the network as VNFs of all the tenants can be shared. On the other hand, in the case of No-sharing, it incurs the highest amount of cost as it activates more VMs due to the isolation constraint from different tenants. In other instances where some tenants trust other specific tenants to share their VNFs, the cost is between the two extremes. This is because it can allow the usage of fewer VMs in the network by sharing some of the VNFs. In a real deployment scenario, slices of multiple tenants may have different types of isolation requirements. Hence, this should be kept in check while deciding the mapping of the functions in the network. Fig. 3b and Fig. 3c show the number of used VMs in the edge and regional cloud and the midhaul bandwidth consumption, respectively, which also describes the cost shown in Fig. 3a.

C. *Comparison with the heuristic*

In this section, we analyze the performance of our proposed heuristic algorithm. Fig. 4 shows the comparison between the optimization model and the heuristic. We observe that the heuristic algorithm performs comparably with the ILP in case of less number of slices in the network (Fig. 4a). As the number of slices increases, the optimization model outperforms the heuristic. However, in Fig. 5, we can see that the optimization model takes exponential time as the input size increases. On the other hand, the heuristic takes significantly less time as the number of slices increases, making it suitable in large-scale deployment scenarios.



(a) Total Cost (b) Used VMs (c) Midhaul bandiwdth consumption (in Mbps)

Fig. 4. Comparison of heuristic and the optimal

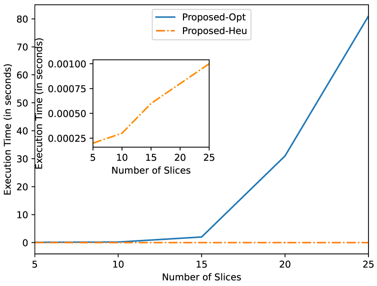


Fig. 5. Comparison of heuristic and the optimal.

VII. CONCLUSION AND FUTURE WORK

In this paper, we address the problem of placing baseband functions (CU and DU) for 5G multi-tenant O-RAN slices to minimize processing and bandwidth costs, considering delay, data rate requirements, tenant isolation, and network capacity limitations. We present an Integer Linear Programming (ILP) model, comparing its performance with baseline strategies. To handle the complexity of ILP, we propose a heuristic-based solution for large-scale scenarios. Simulations demonstrate the efficiency of our approach in achieving cost-efficient placements for baseband functions. Moreover the heuristic can efficiently place the functions in a reasonable time frame. Our future work will involve exploring machine learning (ML)-based solutions to handle diverse tenant requirements and implementing them in an experimental testbed.

REFERENCES

[1] L. M. P. Larsen, A. Checko, and H. L. Christiansen, “A survey of the functional splits proposed for 5g mobile crosshaul networks,” *IEEE Communications Surveys Tutorials*, 2019.

[2] F. Z. Morais and et al., “Placeran: optimal placement of virtualized network functions in beyond 5g radio access networks,” *IEEE Transactions on Mobile Computing*, 2022.

[3] A. M. et al., “Efficient management of flexible functional split through software defined 5g converged access,” in *2018 IEEE International Conference on Communications (ICC)*, 2018.

[4] N. Sen and A. F. A, “Towards energy efficient functional split and baseband function placement for 5g ran,” in *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, 2023.

[5] M. Zorello and et al., “Power-efficient baseband-function placement in latency-constrained 5g metro access,” *IEEE Transactions on Green Communications and Networking*, 2022.

[6] H. Yu, F. Musumeci, J. Zhang, M. Tornatore, and Y. Ji, “Isolation-aware 5g ran slice mapping over wdm metro-aggregation networks,” *Journal of Lightwave Technology*, 2020.

[7] W. da Silva Coelho, A. Benhamiche, N. Perrot, and S. Secci, “Function splitting, isolation, and placement trade-offs in network slicing,” *IEEE Transactions on Network and Service Management*, 2022.

[8] N. Kazemifard and V. Shah-Mansouri, “Minimum delay function placement and resource allocation for open ran (o-ran) 5g networks,” *Computer Networks*, 2021.

[9] H. Hojeij and et al., “Dynamic Placement of O-CU and O-DU Functionalities in Open-RAN Architecture,” in *IEEE International Conference on Sensing, Communication, and Networking (SECON)*, Sep. 2023.

[10] S. D’Oro and et al., “Orchestran: Network automation through orchestrated intelligence in the open ran,” in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022.

[11] A. Filali and et al., “Communication and computation o-ran resource slicing for urllc services using deep reinforcement learning,” *IEEE Communications Standards Magazine*, 2023.

[12] S. Ravindran and et al., “Efficient service allocation scheduling algorithms for 5g user equipments in slice-in-slice networks,” in *IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 2021.

[13] I. Tamim, A. Saci, M. Jammal, and A. Shami, “Downtime-aware o-ran vnf deployment strategy for optimized self-healing in the o-cloud,” in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021.

[14] T. Pamuklu, M. Erol-Kantarci, and C. Ersoy, “Reinforcement learning based dynamic function splitting in disaggregated green open rans,” in *ICC 2021 - IEEE International Conference on Communications*, 2021.

[15] E. Amiri, N. Wang, M. Shojafar, and R. Tafazolli, “Energy-aware dynamic vnf splitting in o-ran using deep reinforcement learning,” *IEEE Wireless Communications Letters*, 2023.

[16] S. Mondal and M. Ruffini, “Optical front/mid-haul with open access-edge server deployment framework for sliced o-ran,” *IEEE Transactions on Network and Service Management*, 2022.

[17] “O-ran use cases and deployment scenarios,” O-RAN Alliance, Tech. Rep., 2020.

[18] F. Malandrino, C. F. Chiasserini, G. Einziger, and G. Scalosub, “Reducing service deployment cost through vnf sharing,” *IEEE/ACM Transactions on Networking*, 2019.

[19] “O-ran.wg1.slicing-architecture-v06.00,” O-RAN Alliance, Tech. Rep., 2022.

[20] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. USA: W. H. Freeman & Co., 1990.

[21] Y. L. et al., “End-to-end energy models for edge cloud-based iot platforms: Application to data stream analysis in iot,” *Future Generation Computer Systems*, 2018.

[22] A. Garcia-Saavedra, G. Iosifidis, X. Costa-Perez, and D. J. Leith, “Joint optimization of edge computing architectures and radio access networks,” *IEEE Journal on Selected Areas in Communications*, 2018.

[23] S. Matoussi, I. Fajjari, S. Costanzo, N. Aitsaadi, and R. Langar, “A user centric virtual network function orchestration for agile 5g cloud-ran,” in *2018 IEEE International Conference on Communications (ICC)*, 2018.

[24] N. Sen and A. F. A, “Intelligent admission and placement of o-ran slices using deep reinforcement learning,” in *2022 IEEE 8th International Conference on Network Softwarization (NetSoft)*, 2022.