Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Slice Aware Baseband Function Placement in 5G RAN using Functional and Traffic Split

NABHASMITA SEN and ANTONY FRANKLIN A. (Senior Member, IEEE)
Indian Institute of Technology Hyderabad, India (e-mail: cs17resch11001@iith.ac.in, antony.franklin@cse.iith.ac.in)
Corresponding author: Nabhasmita Sen (e-mail: cs17resch11001@iith.ac.in).

ABSTRACT 5G and Beyond 5G (B5G) are undergoing numerous architectural changes to enable higher flexibility and efficiency in mobile networks. Unlike traditional mobile networks, baseband functions in 5G are disaggregated into multiple components - Radio Unit (RU), Distributed Unit (DU), and Centralized Unit (CU). These components can be placed in different geographical locations based on the latency sensitivity and available capacity in the network. Processing baseband functions in a centralized location offer various advantages (known as centralization benefit in RAN) to mobile network operators, which have been a point of interest for several research works. However, achieving maximum centralization is challenging due to various factors such as limited capacity in the midhaul network, delay requirement of different functional splits and network slices, etc. In this work, we aim to address these challenges by proposing a slice-aware baseband function placement strategy. Our primary objective is to maximize the degree of centralization in the network by appropriate selection of functional split. To achieve this objective, we jointly consider functional split, traffic split, different placement options for baseband functions, and network slice-specific requirements. We also consider the minimization of active processing nodes in cloud infrastructure of different levels (edge and regional) to provide additional resource efficiency. To this end, we formulate an optimization model using Mixed Integer Linear Programming (MILP) and compare its performance with different baseline techniques. We show that the proposed model achieves 6.5% more degree of centralization than the state-of-the-art while placing baseband functions in the network. To tackle the high computational complexity of the MILP model, we also present a polynomial-time heuristic algorithm for solving the problem in large-scale scenarios. We show that although the optimization model achieves around 4% more degree of centralization than the heuristic, the heuristic solves the problem in a reasonable amount of time, making it suitable for real deployment scenarios.

INDEX TERMS Centralization benefit, Functional split, Network slice, RAN disaggregation, Resource efficiency.

I. INTRODUCTION

Mobile networks of 5G and beyond are going through several technological advancements to serve a massive number of users with a broad range of services. The introduction of Software Defined Networking (SDN) and Network Function Virtualization (NFV) has improved the flexibility and efficiency of mobile networks. In contrast to traditional Radio Access Network (RAN), baseband functions in 5G are disaggregated using different functional splits [1]. These disaggregated functions can be further virtualized and placed in shared infrastructures enabling higher flexibility in mobile networks. Processing the baseband functions in a centralized location has various advantages known as centralization benefit in RAN [1], [2]. Centralizing different layers of the baseband function protocol stack generates different degrees of centralization [3]. E.g., the centralization of the PDCP (Physical Data Convergence Protocol) layer provides a centralized over-the-air encryption facility and greater coordination for mobility-related handovers. A centralized RLC (Radio Link Control) layer can offer high reliability. Centralization of the MAC (Medium Access Control) layer offers centralized scheduling, joint transmission, and better interference management [4], [5]. The centralization of physical layer functionalities can benefit centralized scheduling, joint

transmission, and joint reception [1]. Hence, maximizing centralization is important, which can also profit the mobile network operators in specific scenarios.

To maximize the degree of centralization in RAN, selection of appropriate functional split is critical, which in turn depends on various factors described as follows.

1) 5G has to support a wide range of services with different service requirements. Based on their needs, these services are mainly divided into three categories - eMBB (enhanced Mobile Broadband), mMTC (massive Machine Type Communication), and URLLC (Ultra-Reliable Low Latency Communication), which have different delay and data rate requirements. These varied services can be efficiently provided to the users with the help of network slicing [6]. Due to different data rate and delay requirements of slices, all functional split options cannot support all slices. Hence, the selection of functional split must be made accordingly.
2) Different functional splits have different delay requirements. On the other hand, different paths in the midhaul network have different delays. To route the traffic of a slice using a specific functional split, the delay of the considered path should be less than the delay requirement of that functional split. Hence, functional split should be selected based on the available path characteristics.
3) Different functional splits have different bandwidth requirements. Hence, based on the available capacity in the midhaul network, appropriate functional split needs to be selected. On the other hand, if a single path cannot route the traffic from a slice due to its limited capacity, splitting the traffic among multiple paths can be helpful to push more functions in the regional cloud resulting in a higher degree of centralization.
4) The capacity of processing nodes in different clouds is limited. Therefore, all baseband function placement options may not be supported by the processing nodes.

Various works have focused on RAN centralization by considering one or many factors mentioned above. The authors of [3] and [7] maximize the degree of centralization by minimizing the computational cost for processing the baseband functions in different locations. Authors of [4] aim to minimize interference related issues by selecting functional split for base stations. Recent works like [8] and [5] assign different centralization values to different functional splits and maximize the centralization degree in the network by selecting proper functional split. However, further exploration is required on this topic as none of the previous works consider all the aforementioned factors together.

In this work, we jointly consider functional split, traffic split, network slice-specific requirements, and different baseband function placement options to maximize the degree of centralization in the network while minimizing the number of active processing nodes to place the baseband functions. The main contributions can be summarized as follows:

- We propose a Mixed Integer Linear Programming (MILP) based optimization model to maximize the degree of centralization in RAN while minimizing the number of active processing nodes in different levels of cloud (edge and regional).
- We consider the delay and data rate requirement of slices while selecting functional split, baseband function placement and paths to route the traffic. Moreover, the delay requirements for different functional splits are ensured while selecting the paths. We also consider traffic splitting to tackle the limited capacity in the midhaul network.
- We compare our proposed optimization model with different baseline strategies and show its superiority in selecting functional split and baseband function placement options for different RAN slices.
- To tackle the high computational complexity of MILP, we provide a low-complexity heuristic algorithm that can be applied in large-scale scenarios.

The organization of the rest of the paper is as follows. Section II contains the related works. The system model and its related concepts are described in Section III. Section IV describes the problem formulation. Section V and VI provide the simulation setup and results, respectively. Section VII presents a heuristic algorithm to address the high computational complexity of the proposed optimization model. Section VIII summarizes the paper and mentions the possible future works.

II. RELATED WORKS

In this section, we provide a literature survey on the selection of functional split in RAN that considers one or more factors described in Section I. Authors of [9] and [10] jointly minimize the energy and bandwidth consumption in a hybrid Cloud RAN (C-RAN) by selecting appropriate functional splits. Authors of [11] discuss Virtualized Network Embedding (VNE) algorithms for flexible selection of functional split for each small cell in 5G RAN. A user-centric functional split is considered in [12], where the functional split per user is selected to minimize the energy and bandwidth consumption. However, delay requirements of different functional splits and slices are not considered in the works mentioned till now. In [13], Virtual Network Function (VNF) deployment in RAN is considered while selecting functional splits. In [14] and [15], functional split and baseband function placement decisions are considered for base stations. Authors of [16] have proposed solutions to minimize the cost of a MEC-enabled RAN using functional and traffic split. Authors of [17] consider functional splits for base stations in a multi-cloud scenario. Nevertheless, slice-specific requirements are not taken into account in these works.

In [7] and [18], slice-centric functional split and user association are performed while considering functional and traffic splitting, though slice delay requirements are not considered. Moreover, these works also consider a single CU in the

network, due to which different baseband function placement options are not considered. Several works like [7], [17], [19] perform functional split selection for slices and base stations based on the two-tier architecture i.e., DU and CU. However, most standards have agreed to a three-tier architecture consisting of CU, DU, and RU, which can provide more flexibility in the network [20]. In [21] and [22], service-oriented CU and DU placement are done with the help of Reinforcement Learning. In [23], the baseband function placement strategy is proposed with the help of an optimization model and heuristic to minimize the power consumption in the network. However, in [21]–[23], only fixed functional split options are considered between CU and DU.

Centralization of baseband processing functions can offer several benefits to the mobile operators (described in Section I). To maximize the centralization in the network, the authors of [24] and [25] propose a solution for selecting functional split for base stations based on variation in the midhaul link traffic. The authors of [26] discuss the impact of split granularity in the centralization gain of RAN. In [3], the authors maximize the degree of centralization in the network by selecting appropriate functional splits. However, functional split specific to slices is not considered here. In our previous work [27], the impact of slice granularity in the centralization benefit of RAN is analyzed. Although, slice-specific delay requirements are not taken into account. In [8] and [5], functional split and baseband function placement decision is considered for RAN slices to maximize the degree of centralization without considering traffic splitting.

In contrast to the existing works, we jointly consider functional split, traffic split, slice-specific requirements, and different baseband function placement options while placing the functions. We explore the selection of functional split to maximize the centralization in the network in a capacity constrained midhaul network. To provide resource efficiency, the optimization model also minimizes the number of activated processing nodes in different clouds. To deal with the limited capacity in the midhaul, we further consider splitting the traffic among multiple paths, which helps to improve the degree of centralization in the network.

III. SYSTEM MODEL

In this section, we provide a detailed description of our system model. Before that, let us briefly introduce concepts related to the system model. A base station has to perform various functions known as Baseband Processing Functions (BPF) [12]. The protocol stack of baseband functions includes Radio Resource Control (RRC) layer, Physical Data Convergence Protocol (PDCP) layer, Radio Link Control (RLC) layer, Medium Access Control (MAC) layer, and a Physical (PHY) layer. The PHY layer is further divided into Higher Physical (High-PHY) layer and Lower Physical (Low-PHY) layer. To introduce better flexibility, this chain of baseband functions is split at different points, which are known as functional splits [1] in RAN. The functional splits are used to propose a three-tier architecture for 5G RAN

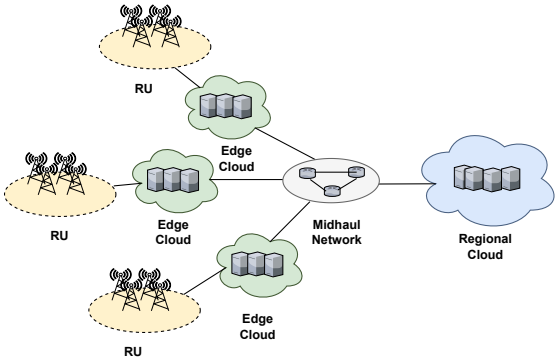


FIGURE 1: RAN System Model.

composed of Centralized Unit (CU), Distributed Unit (DU), and Radio Unit (RU). The RUs are placed at the cell site, have antennas to transmit and receive radio signals, and processing power to perform the Low-PHY layer functionalities. The rest of the layers are placed in the CU or DU depending on the functional split between them.

We consider a hybrid cloud architecture as our system model (shown in Fig. 1) along with the three-tier architecture for RAN, which conforms to the 5G RAN deployment scenario [28], [29]. The multiple RUs in the network denote the cell sites. Different clusters of RUs are connected to their corresponding edge clouds. The edge clouds are further connected to the regional cloud through the midhaul network. The midhaul network is considered to be a metro aggregation network composed of links of different latencies and capacities. The edge and regional cloud consist of multiple processing nodes where the baseband functions can be placed. The DUs are placed on the edge cloud (near the cell sites), whereas the CUs are placed in the regional cloud. Placing the baseband functions in the regional cloud can provide a higher centralization benefit than placing them in the edge cloud. This is because baseband functions from significantly more RUs can be placed together in the regional cloud. However, due to delay and capacity constraints in the midhaul network, all functions cannot be placed in the regional cloud. Hence, some functions must be placed at the edge clouds in such scenarios. Different functional splits have different latency and bandwidth requirements [1], whereas different network slices have different delay requirements, due to which some functional splits may not be applied to some slices. This way, multiple functional splits can be present in the same RU [30].

In this work, we consider that the High-PHY layer is always placed at the edge cloud due to its stringent delay and high bandwidth requirement. Other upper layers (RRC-PDCP, RLC, MAC) have less stringent delay requirements and have similar bandwidth requirements [31]. These layers are placed in the DU or CU based on the functional split resulting in four different functional splits. We consider RRC and PDCP together due to less processing requirement of RRC layer [28]. We assign split numbers 0-3, with 0 being the lowest split and 3 being the highest. In the lowest split, all

the layers of the baseband function protocol stack are placed in the DU at the edge cloud. Whereas, all the layers except High-PHY are placed in the CU at the regional cloud for the highest split.

IV. PROBLEM FORMULATION

Maximizing the degree of centralization is important because of various centralization benefits. Placing more functions in the regional cloud can increase the degree of centralization. However, all baseband functions cannot be placed in the regional cloud due to various delay and capacity constraints (as discussed in Section I). We define our problem (Split-RAN) as follows: *Given the data rate and delay requirement of RAN slices, slice origin and underlying network characteristics (node capacity, link capacity, path delay), select functional split, baseband function placement, and paths to route the traffic for the slices such that the degree of centralization in the network is maximized and the number of active processing nodes in edge and regional cloud is minimized.*

We formulate Split-RAN as an optimization model using Mixed Integer Linear Programming (MILP). The optimization model is mainly beneficial when there is a requirement to find the optimal solution, which can also act as a benchmark for evaluating other potential solutions to the Split-RAN problem. Table 1 shows the notations used for the problem formulation. The decision variables, constraints, and objective function of our proposed optimization model are defined as follows.

A. DECISION VARIABLES

We consider the following decision variables in our formulation.

(i) We define a binary variable $k_{s,f}$ to denote whether the functional split f is chosen for a slice s or not.

$$k_{s,f} = \begin{cases} 1, & \text{if slice } s \text{ selects functional split } f \\ 0, & \text{otherwise} \end{cases}$$

(ii) A binary variable $x_{s,m}$ indicates whether slice s is assigned to processing node m in the regional cloud for placing its CU.

$$x_{s,m} = \begin{cases} 1, & \text{if slice } s \text{ selects node } m \text{ for its CU} \\ 0, & \text{otherwise} \end{cases}$$

(iii) Binary variable $y_{s,n}$ denotes if slice s is assigned to processing node n in the edge cloud for placing its DU.

$$y_{s,n} = \begin{cases} 1, & \text{if slice } s \text{ selects node } n \text{ for its DU} \\ 0, & \text{otherwise} \end{cases}$$

(iv) A binary variable z_m to capture if a processing node m in the regional cloud is switched ON or not.

$$z_m = \begin{cases} 1, & \text{if regional cloud node } m \text{ is activated} \\ 0, & \text{otherwise} \end{cases}$$

TABLE 1: Notation and Description

Notation	Description
ES	Set of servers in edge clouds
RS	Set of servers in regional cloud
$\mathrm{cu}_{s,f}$	Processing required for CU of slice s using split f
$\mathrm{du}_{s,f}$	Processing required for DU of slice s using split f
$\lambda_{l,p}$	Link l belongs to path p or not
CAP_l	Capacity of link l
EC	Set of edge clouds
F	Set of functional splits
S	Set of all slices
P	Set of all paths
CE_n	Capacity of edge cloud server n
CR_m	Capacity of regional cloud server m
π_s^n	Connectivity of edge server n and slice s
Φ_s	Centralization benefit related to slice s
μ_f	Centralization factor of a functional split f
$\mathrm{PC}_{s,p}$	1 if slice s is connected to path p, else 0
$Q_{s,p}$	1 if path p satisfies delay requirement of slice s, else 0

(v) Binary variable w_n indicates whether a processing node n in edge cloud is switched ON or not.

$$w_n = \begin{cases} 1, & \text{if edge cloud node } n \text{ is activated} \\ 0, & \text{otherwise} \end{cases}$$

(vi) We consider a continuous variable $\zeta_{s,p}$ which denotes the amount of traffic from slice s going through path p.

B. OBJECTIVE FUNCTION

The objective of our model is to maximize the centralization of the network while minimizing the number of active processing nodes in both the regional and edge cloud.

The degree of centralization in the network is expressed as,

$$C = \sum_{s \in S} \sum_{f \in F} k_{s,f} \Phi_s \mu_f \tag{1}$$

where μ_f denote the centralization factor of a functional split f, Φ_s denotes the centralization benefit related to slice s, and $k_{s,f}$ indicates the functional split selected for slice s. The value of μ_f increases with the number of functions centralized in the regional cloud. For the four functional splits described in Section III, we set μ_f as 0.1, 0.2, 0.3, 0.4 respectively from the lowest to highest functional split. In this work, we consider the centralization benefit of slice s (Φ_s) is proportional to its data rate requirement.

The number of active processing nodes in the edge and regional cloud is expressed as,

$$A = \sum_{m \in \mathrm{RS}} z_m + \sum_{n \in \mathrm{ES}} w_n \tag{2}$$

Now, the final objective is defined as,

$$Maximize : \alpha \frac{C}{C'} - \beta \frac{A}{A'} \tag{3}$$

where α and β are the weighing factor used to adjust the weightage of C and A respectively. C' and A' are the normalization factors that denote the maximum value of centralization (C) and active processing nodes (A), respectively.

As our main goal is to maximize the centralization gain in the network, we adjust α and β such that the centralization is prioritized. However, these factors can be set by the infrastructure providers according to their requirements.

C. CONSTRAINTS

The constraints of the optimization model are defined as follows.

(i) Capacity constraint of processing nodes: The total processing performed in any processing node in the edge or regional cloud should not exceed the capacity of that node.

$$\sum_{s\in S}\sum_{f\in F} x_{s,m}k_{s,f}\mathrm{cu}_{s,f} \leq \mathrm{CR}_m, \forall m \in \mathrm{RS} \tag{4}$$

$$\sum_{s\in S}\sum_{f\in F} y_{s,n}k_{s,f}\mathrm{du}_{s,f} \leq \mathrm{CE}_n, \forall n \in \mathrm{ES} \tag{5}$$

(ii) Capacity constraint of transport links: The total traffic routed through a transport link should not exceed the capacity of that link.

$$\sum_{s\in S}\sum_{p\in P} \zeta_{s,p}\lambda_{l,p} \leq \mathrm{CAP}_l, \forall l \in L \tag{6}$$

(iii) Total traffic constraint: This constraint ensures that the total traffic from a slice for the selected split remains equal to the sum of all its traffic going through different paths.

$$\sum_{f\in F} k_{s,f}t_{s,f} = \sum_{p\in P} \zeta_{s,p}, \forall s \in S \tag{7}$$

where $t_{s,f}$ is the traffic for slice s when split f is used.

(iv) Activation of a processing node: A processing node is considered active if at least one function is placed on it.

$$z_m \geq x_{s,m}, \forall m \in \mathrm{RS}, \forall s \in S \tag{8}$$

$$w_n \geq y_{s,n}, \forall n \in \mathrm{ES}, \forall s \in S \tag{9}$$

(v) Each slice can use only one server in edge cloud only if it is connected to the edge server.

$$\sum_{n\in \mathrm{ES}} y_{s,n} = 1, \forall s \in S \tag{10}$$

$$\sum_{n\in \mathrm{ES}} y_{s,n}\pi_s^n = 1, \forall s \in S \tag{11}$$

(vi) Except for Split-0, each slice can use only one server in the regional cloud. For Split-0, no server is used in the regional cloud as all functions are placed at the edge.

$$\sum_{m\in \mathrm{RS}} x_{s,m} = 1 - k_{s,0}, \forall s \in S \tag{12}$$

(vii) Only one functional split can be selected for a particular slice.

$$\sum_{f\in F} k_{s,f} = 1, \forall s \in S, \tag{13}$$

(viii) The delay of a path should not exceed the delay requirement of a functional split. We consider $P_f \subseteq P$ denotes the set of paths whose delay is greater than the delay requirement of functional split f. Hence, no path $p \in P_f$ should carry any traffic from slice s when split f is used for that slice.

$$\sum_{p\in P_f} \zeta_{s,p} \leq M(1 - k_{s,f}), \forall s \in S, \forall f \in F \tag{14}$$

where M is a big integer that is used to ensure that the selected split must support the delay requirement.

(ix) A path can be used by a slice only if the slice is connected to the path through its edge cloud.

$$\zeta_{sp} \leq M{\cdot}\mathrm{PC}_{s,p}, \forall s \in S, \forall p \in P \tag{15}$$

where $\mathrm{PC}_{s,p}$ denotes if slice s is connected to path p through its edge.

(x) If the delay of a path is more than the required delay of a slice then that path cannot be used for the same slice.

$$\zeta_{sp} \leq M{\cdot}Q_{s,p}, \forall s \in S, \forall p \in P \tag{16}$$

where $Q_{s,p}$ denotes if path p satisfies delay requirement of slice s.

D. LINEARIZATION OF THE OPTIMIZATION MODEL

The optimization model described above has some non-linear terms due to the multiplication of two variables. We remove these non-linearities by introducing new variables and their related constraints. For example, the term $y_{s,n}k_{s,f}$ is replaced with a new variable $\mathrm{yk}_{s,n,f}$ and the related constraints are added as follows,

$$\mathrm{yk}_{s,n,f} \leq y_{s,n} \tag{17}$$

$$\mathrm{yk}_{s,n,f} \leq k_{s,f} \tag{18}$$

$$\mathrm{yk}_{s,n,f} \geq y_{s,n} + k_{s,f} - 1 \tag{19}$$

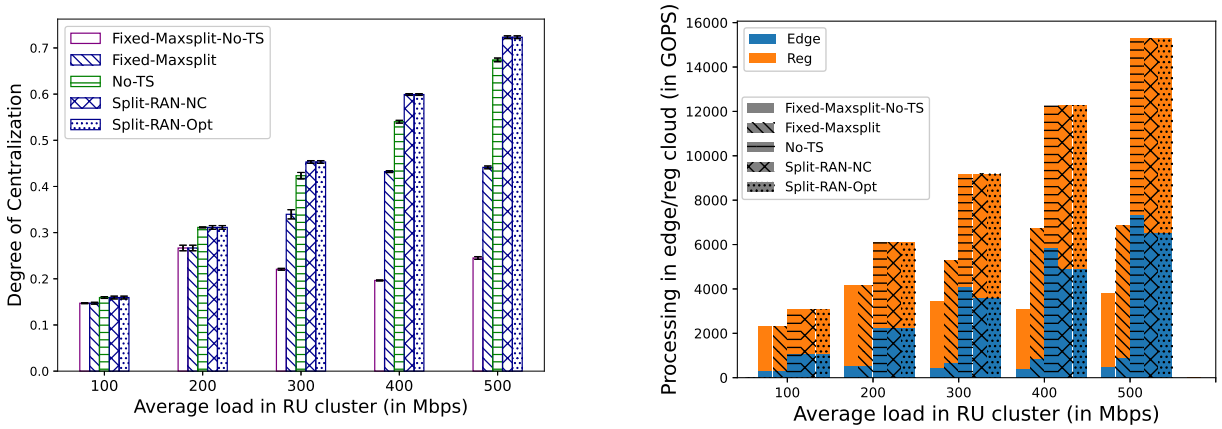
Similarly, the other quadratic terms are also linearized.

V. SIMULATION SETUP

This section provides the necessary details about our simulation environment and the baseline strategies. The simulation parameters shown in Table 2 are chosen from various references to simulate a real deployment scenario. We consider

TABLE 2: Simulation Parameters

Simulation Parameters	Description
Number of Clouds	11
Number of edge cloud	10
Number of regional cloud	1
Total number of servers	64 servers
Number of server in regional cloud	24
Number of servers in edge cloud	4 in each cloud
Slice-type	eMBB, URLLC, MMTC [6]
Load in each RU cluster	100-500 Mbps
eMBB, URLLC, mMTC load	50%,25%,25 %
eMBB, URLLC, mMTC Delay	10, 1, 10 ms [32]
Number of slices	3-30 slices
Server capacity	1200 GOPS
Number of paths in midhaul	50
Path delay	2-30 ms
Aggregated midhaul capacity	2-28 Gbps



(a) Degree of Centralization

(b) Processing in edge and regional cloud

FIGURE 2: Comparison of strategies for different load in the network.

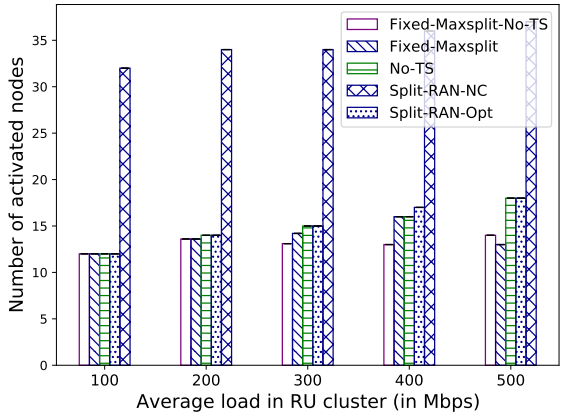


FIGURE 3: Number of active nodes corresponding to Fig. 2

ten edge clouds and a regional cloud in the network. Each edge cloud is connected to a RU cluster. We consider 64 servers (4 servers in each edge cloud and 24 servers in the regional cloud) with a capacity of 1200 Gigabit Operations Per Second (GOPS) [33], [34]. The edge clouds are connected to the regional cloud through the midhaul network. The midhaul network consists of multiple paths which have delays in the range of 2-30 ms and aggregated path capacities ranging from 2-28 Gbps [28], [33]. There are three slices in each RU cluster for eMBB, URLLC, and mMTC services, respectively [6]. The load in each RU cluster is varied from 100-500 Mbps, considering 4 RUs with 2×2 MIMO and 20 MHz bandwidth in each RU cluster [31]. As the eMBB slice has the highest data rate requirements, and mMTC and URLLC slices have similar data rate requirements [35], we consider that the eMBB, mMTC, and URLLC slices have 50%, 25%, and 25% of the total load, respectively. The delay requirement of eMBB, URLLC, and mMTC slices are set according to [32]. The processing and bandwidth requirements for different baseband functions corresponding to each slice are approximately calculated based on the models given in [34], [36] and [31]. In the objective function (Eqn. 3), we set $\alpha = 1$ and $\beta = 0.04$ such that the centralization is given higher weightage. For a given α, the value of β is calculated using the simulation parameter values in Table 2. We perform the simulations with 50 randomly generated data instances for different load conditions in each slice and report the results with a 95% confidence interval obtained for different strategies. We implement the optimization model using Gurobi solver [37] (version 9.5.0) with python interface and Python 3.8 environment. All the simulations are performed in an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz machine.

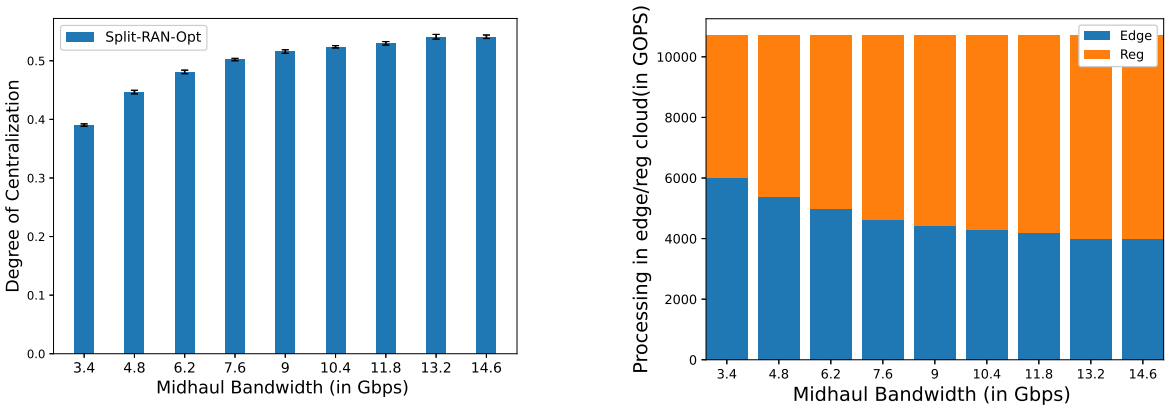
We name our proposed optimization model as *Split-RAN-Opt* and compare its performance with the following baseline strategies.

- *No-TS*: This strategy is based on [8] where no traffic splitting is considered while maximizing the centralization in the network.
- *Fixed-Max-Split*: In this strategy, the highest functional split (Split-3) is selected as the fixed functional split option as it has the highest centralization factor. Since fixed split option may not support some of the slices, few constraints in the optimization model are relaxed accordingly.
- *Fixed-Max-Split-No-TS*: It is same as Fixed-Max-Split without considering the traffic splitting.
- *Split-RAN-NC*: This strategy is a variation of Split-RAN-Opt where the only objective is to maximize the centralization, i.e., the second term in Eqn. 3 (involving the minimization of active nodes) is not considered. This baseline can be seen as a combination of [5] and [7].

VI. RESULTS AND ANALYSIS

A. COMPARISON WITH BASELINES

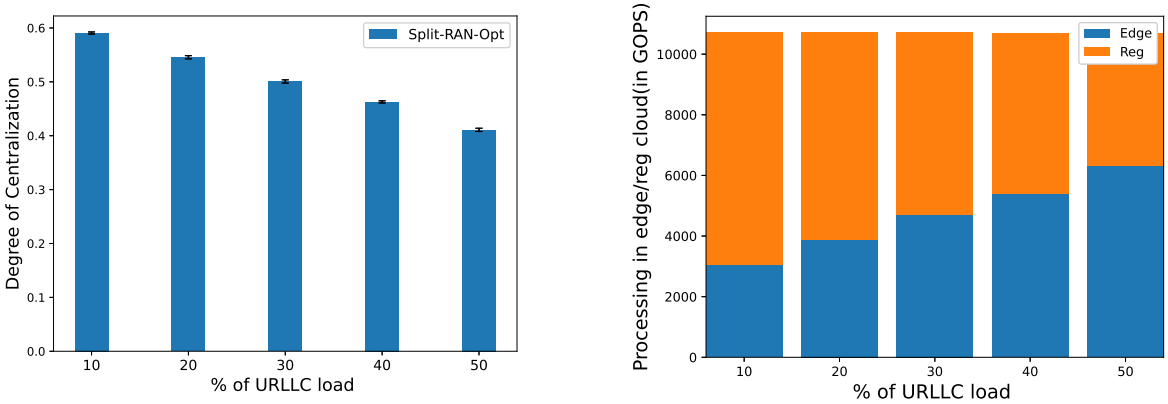
In this section, we compare the performance of our proposed model Split-RAN-Opt with the baseline strategies for different network loads while selecting functional split and



(a) Centralization Degree

(b) Processing in edge and regional cloud

FIGURE 4: Impact of midhaul capacity on Split-RAN-Opt.



(a) Centralization Degree

(b) Processing in edge and regional cloud

FIGURE 5: Impact of slice delay requirement on Split-RAN-Opt.

baseband function placement options. The observations from the simulation results are as follows.

1) Degree of Centralization

Fig. 2a shows the degree of centralization achieved by different strategies. In the case of low load (100-200 Mbps), Split-RAN-Opt and No-TS achieve a similar degree of centralization due to sufficient midhaul capacity. In high load, Split-RAN-Opt generates higher centralization than No-TS because of splitting the traffic among multiple paths to cope with the less capacity of the midhaul network. Overall, we observe that Split-RAN-Opt generates 6.5% more centralization than NO-TS. Split-RAN-NC achieves same centralization as Split-RAN-Opt since they consider the same factors for maximizing the degree of centralization. As discussed earlier, Fixed-Max-Split and Fixed-Max-Split-No-TS consider only Split-3 as the fixed functional split. However, Split-3 cannot support a delay sensitive slice when the delay between its corresponding edge cloud and the regional cloud is higher than its delay requirement. On the other hand, when sufficient midhaul capacity is unavailable, the fixed split strategies do not try to assign other possible split options. Hence, the fixed split strategies cannot support some of the slices, resulting in a lower degree of centralization than the other strategies.

2) Total Processing in the Edge and Regional Cloud

Fig. 2b shows the total processing in the edge and regional cloud for placing the baseband functions of different slices. This analysis aims to verify the degree of centralization shown in Fig. 2a. We observe that Split-RAN-Opt has the highest amount of processing in the regional cloud as it tries to place slices with higher demand (Φ) in the regional cloud to maximize the degree of centralization (Eqn. 1). Compared to Split-RAN-Opt, No-TS has less processing in the regional cloud in high load as it does not consider traffic splitting. Overall, Split-RAN-Opt places around 9% more amount of processing in the regional cloud than No-TS. Consequently, No-TS places more processing in the edge cloud than Split-RAN-Opt. The fixed split strategies do not support some of the slices due to capacity and delay constraints. As a result, the total processing in regional and edge cloud for these strategies is lesser than other strategies. Fixed-Max-Split has more processing in the edge and regional cloud

than Fixed-Max-Split-No-TS as it can support more slices due to considering traffic splitting. Split-RAN-NC has the same amount of processing in edge and regional cloud as Split-RAN-Opt due to considering the same factors for maximizing the centralization.

3) Number of activated nodes in different strategies

In Fig. 3, we analyze the number of active nodes while performing the simulation shown in Fig. 2. Firstly, Split-RAN-Opt activates fewer nodes than Split-RAN-NC even though they achieve similar centralization. This is because Split-RAN-Opt considers the minimization of the number of active processing nodes in its objective function. Secondly, fixed split strategies use fewer processing nodes than the other strategies as they do not support many slices (as discussed in Section VI-A1 and VI-A2). Thirdly, maximizing centralization does not always help in minimizing active nodes. Some of the processing nodes must be activated in the edge cloud to support the delay sensitive functions. This is why it is not always possible to minimize the total number of active nodes in edge and regional clouds when the primary objective is to maximize centralization. For instance in Fig. 3, we can observe that Split-RAN-Opt sometimes activates more nodes than No-TS, even though it attains higher centralization.

B. IMPACT OF MIDHAUL CAPACITY

In this subsection, we analyze the impact of midhaul network capacity on the performance of our proposed optimization model. In Fig. 4a, we observe that with the increase in the midhaul capacity, Split-RAN-Opt achieves better centralization as it can place more functions in the regional cloud. However, the delay sensitive functions must be placed in the edge cloud and cannot be centralized in the regional cloud. After reaching the maximum limit, the degree of centralization becomes fixed and does not increase even if the midhaul capacity increases. We can also verify this observation from Fig. 4b, where we show the processing in edge and regional cloud corresponding to Fig. 4a. In Fig. 4b, the amount of processing in the regional cloud increases with the increase in the midhaul capacity. However, after a certain limit, processing in the regional cloud does not increase even with the increase in the midhaul capacity as the remaining functions must be placed in the edge cloud to satisfy the delay constraints.

C. IMPACT OF SLICE DELAY REQUIREMENT

In this subsection, we observe the impact of slice delay requirement on the proposed optimization model. We keep the mMTC slice load fixed at 25% of the total load and divide the rest among eMBB and URLLC slices. We vary the percentage of URLLC slice load from 10-50% of the total load. URLLC slices are delay sensitive. This is why baseband functions of URLLC slices are mostly placed at the edge cloud to satisfy the delay constraint. Hence, in Fig. 5a, we can observe that the degree of centralization decreases with the increase in URLLC slice load. Fig. 5b also verifies the same, showing that the amount of processing in the regional cloud decreases as the URLLC load percentage increases due to placing more functions at the edge.

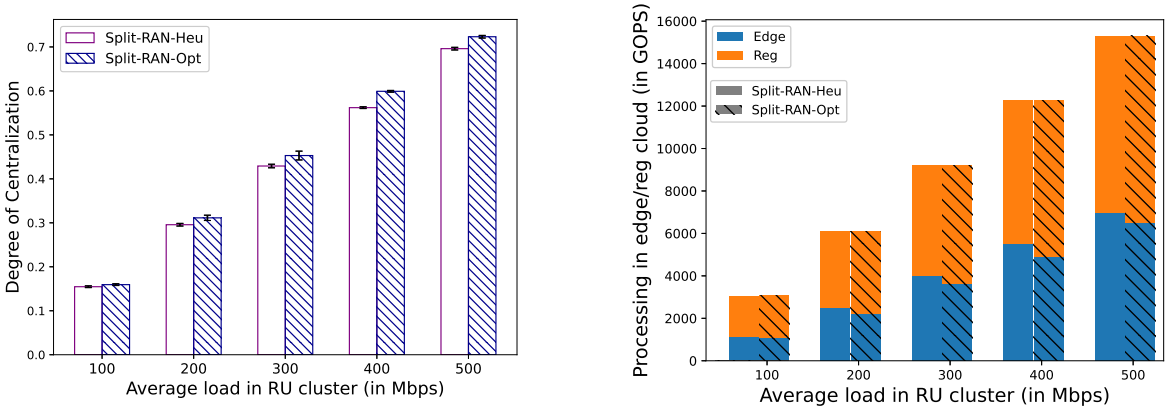
Algorithm 1: Heuristic Algorithm for Split-RAN

Data: Data rate and delay requirement of slices, slice origin, node capacity, link capacity, and path delay
Result: Selection of functional split, baseband function placement, paths for each slice

$S' \leftarrow sort(S, \Phi_s)$ // Sort slices based on their load in decreasing order
foreach *slice* s *in* S' **do**
  $f \leftarrow 3$ // Start from the highest split
  while $f \geq 0$ **do**
    // Find nodes to place CU and DU based on f
    if $f > 0$ **then**
      $not\_assigned \leftarrow True$
      foreach $r \in \mathrm{RS}$ **do**
        // Check available capacity for CU placement
        if $cap[r] >= \mathrm{cu}[s][f]$ **then**
          $cu\_select[s] \leftarrow r$
          $not\_assigned \leftarrow False$
        if $not\_assigned$ **then**
          goto 37
      $not\_assigned \leftarrow True$
      foreach $e \in \mathrm{ES}$ **do**
        // Check available capacity for DU placement
        if $cap[e] \geq \mathrm{du}[s][f]$ **then**
          $du\_select[s] \leftarrow e$
          $not\_assigned \leftarrow False$
      if $not\_assigned$ **then**
        goto 37
      // Find paths to route traffic
      if $f > 0$ **then**
        foreach $p \in P$ **do**
          if $\delta[p] \leq \delta[s]$ *and* $\delta[p] \leq \delta[f]$ **then**
            if $cap[p] \geq t[s][f]$ **then**
              $placed[s] \leftarrow 1$
              $split\_select[s] \leftarrow f$
              $rem\_cap[s] \leftarrow 0$
              Update remaining capacity of processing nodes, path p and its links
            else
              // Split the traffic
              $rem\_cap[s] \leftarrow rem\_cap[s] - cap[p]$
              Update remaining capacity of path p and its links
            if $placed[s] \leftarrow 1$ **then**
              break
        if $placed[s] \leftarrow 1$ **then**
          break
        else
          Revert back current changes in $cap, cu\_select, du\_select, split\_select$
      $f \leftarrow f - 1$

VII. HEURISTIC SOLUTION

Split-RAN is an NP-Hard problem (proof in Appendix A). In section IV, we proposed an optimization model which



(a) Degree of Centralization

(b) Processing in edge and regional cloud

FIGURE 6: Comparison of Split-RAN-Opt and Heuristic

requires solving MILP. It is known that solving MILP is NP-hard and the time complexity of MILP increases exponentially with the number of binary decision variables [38]. Hence, the complexity of our optimization model will also increase exponentially with the number of slices, available functional splits, and processing nodes in different clouds, making it inefficient for real deployment scenarios. This is why the optimization model will be mostly useful when the input instance is small and we need an optimal solution. To address this scalability issue, we propose a heuristic algorithm to solve Split-RAN that requires a time polynomial in the size of the input instance.

A. PROPOSED ALGORITHM

In this subsection, we propose a priority-based greedy heuristic to solve Split-RAN for large-scale scenarios. The algorithm takes information about slices such as data rate and delay requirement, its origin, node capacity, link capacity, and path delay as input and returns the selected functional split for each slice, their baseband function (CU and DU) placement options, and the paths to route their traffic as output. The proposed heuristic algorithm is shown in Algorithm 1.

Let S be the set of all slices. We first sort S based on the centralization benefit (Φ) of its slices to get S'. We also assume that sufficient capacity is there in the network to support all slices with at least the lowest functional split (Split-0). For each slice in S', we begin the assignment of functional split starting from the highest functional split i.e. Split-3. Doing so helps to maximize the degree of centralization as Split-3 has the highest centralization factor (μ_f). Now, for the given slice s and functional split f, we find the processing nodes for the placement of its CU and DU. We select a processing node in the regional cloud to place the CU from the list of switched-on servers in a first-fit strategy (shown in lines 6-12). If the switched-on servers are not able to place the CU, then a new server is activated. If none of the servers can place the CU for the current functional split, then the algorithm tries to do the same with the next functional split. For the lowest split (Split-0), all functions are placed in the DU. Therefore no server needs to be selected for the CU in the regional cloud. We then select a processing node in the edge cloud to place the DU in a similar fashion (shown in lines 13-19).

The next step of the algorithm (lines 20-32) is to find the paths to route the traffic among the CU and DU for the given slice s and functional split f. We consider the set of all available paths (P) from its corresponding edge cloud and regional cloud, which can satisfy the delay requirement of both slice s and functional split f. If a path $p \in P$ cannot accommodate the total traffic from slice s and functional split f, then the other paths are checked for routing the remaining traffic. This way, the traffic from a slice is routed among multiple paths enabling traffic splitting. If a set $P' \subseteq P$ can route the traffic from slice s, then the functional split and baseband function placement decisions are updated for that slice. The remaining capacity of the processing nodes, path, and its links are also updated accordingly. If no subset of paths can accommodate the traffic of slice s, the heuristic algorithm tries to place the slice using the next functional split with a lesser centralization factor. For the lowest functional split (Split-0), no paths need to be selected, as all the functions will be placed in the edge cloud. However, in this case, the slice achieves the lowest centralization gain.

The time complexity of Algorithm 1 is $O(|S|log|S| + |S|\cdot|F|(|\mathrm{ES}| + |\mathrm{RS}| + |P|\cdot|L|))$, where S is set of slices, F is set of functional splits, P is the set of paths and L is the set of links. ES and RS are the set of edge and regional cloud servers respectively.

B. COMPARISON WITH SPLIT-RAN-OPT

In this subsection, we compare the performance of the proposed heuristic algorithm (Split-RAN-Heu) and the optimization model (Split-RAN-Opt). In Fig. 6a, we can observe that the heuristic achieves similar centralization to Split-RAN-Opt in low load (100-200 Mbps). As the load increases, Split-RAN-Opt starts to outperform Split-RAN-Heu. Overall, Split-RAN-Opt achieves 4% more centralization than

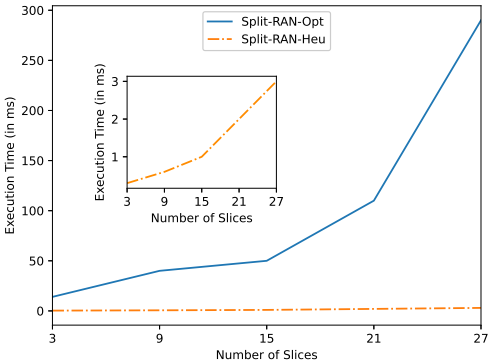


FIGURE 7: Comparison of execution time of Split-RAN-Opt and Split-RAN-Heu

Split-RAN-Heu. However, as discussed in Section VII-A, the heuristic algorithm tries to maximize the centralization with a priority-based greedy method and supports traffic splitting. Thus, even though Split-RAN-Heu falls behind Split-RAN-Opt, it can achieve a reasonable degree of centralization in significantly lesser time. The same observation can be verified from the amount of processing in edge and regional cloud shown in Fig. 6b. We can notice that the heuristic places as much processing in the regional cloud as possible. However, unlike the optimal solution, it does not explore all options of functional split and baseband function placement options. As a result, it places 6% less processing in the regional cloud than Split-RAN-Opt.

The main motivation for proposing the heuristic algorithm is to address the scalability issue of Split-RAN-Opt. Fig. 7 compares the execution time of both methods. We can observe that the execution time for Split-RAN-Opt rapidly increases with the number of slices, whereas Split-RAN-Heu can generate the solutions in a significantly lesser time. Thus, Split-RAN-Heu proves to be much more scalable compared to Split-RAN-Opt.

VIII. CONCLUSION

In this work, we discuss the usefulness of centralization for mobile network operators while placing the baseband functions in 5G RAN and analyze different factors that can maximize the degree of centralization. To address the limitations of the existing strategies, we propose to jointly consider functional split, traffic split, different placement options for baseband functions, and network slice-specific requirements. We formulate our problem as an MILP-based optimization model to maximize the degree of centralization. The objective function also includes the minimization of active processing nodes in different clouds to support resource efficiency. We show that the proposed optimization model outperforms the baseline strategies. We analyze the impact of midhaul capacity and delay requirements of slices on the performance of our optimization model. To deal with the high computational complexity of MILP, we propose a polynomial time heuristic algorithm. We show that the heuristic algorithm achieves a reasonable degree of centralization compared to the proposed optimization model in a significantly less amount of time and hence can be applied to large-scale real deployment scenarios. In future work, we want to explore how the degree of centralization impacts different factors such as interference mitigation, energy efficiency, spectral efficiency, etc. We also want to develop better heuristic algorithms leveraging advanced techniques.

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, vol. 21, no. 1, pp. 146–172, 2019.
[2] Z. Zhang and et al., "Statistical multiplexing gain analysis of processing resources in centralized radio access networks," IEEE Access, vol. 7, pp. 23 343–23 353, 2019.
[3] A. Garcia-Saavedra, J. X. Salvat, X. Li, and X. Costa-Perez, "Wizhaul: On the centralization degree of cloud ran next generation fronthaul," IEEE Transactions on Mobile Computing, vol. 17, no. 10, pp. 2452–2466, 2018.
[4] A. M. Alba, S. Janardhanan, and W. Kellerer, "Enabling dynamically centralized ran architectures in 5g and beyond," IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3509–3526, 2021.
[5] E. Sarikaya and E. Onur, "Placement of 5g ran slices in multi-tier o-ran 5g networks with flexible functional splits," in 2021 17th International Conference on Network and Service Management (CNSM), 2021, pp. 274–282.
[6] "Description of network slicing concept," NGMN, Tech. Rep., 2016.
[7] B. Ojaghi et al., "Sliced-ran: Joint slicing and functional split in future 5g radio access networks," in IEEE International Conference on Communications (ICC), 2019, pp. 1–6.
[8] C. C. Erazo-Agredo et al., "Joint route selection and split level management for 5g c-ran," IEEE Transactions on Network and Service Management, vol. 18, no. 4, pp. 4616–4638, 2021.
[9] X. Wang, A. Alabbasi, and C. Cavdar, "Interplay of energy and bandwidth consumption in cran with optimal function split," in 2017 IEEE International Conference on Communications (ICC), 2017, pp. 1–6.
[10] A. Alabbasi, X. Wang, and C. Cavdar, "Optimal processing allocation to minimize energy and bandwidth consumption in hybrid cran," IEEE Transactions on Green Communications and Networking, 2018.
[11] D. Harutyunyan and R. Riggio, "Flex5g: Flexible functional split in 5g networks," IEEE Transactions on Network and Service Management, vol. 15, no. 3, pp. 961–975, 2018.
[12] S. Matoussi, I. Fajjari, S. Costanzo, N. Aitsaadi, and R. Langar, "A user centric virtual network function orchestration for agile 5g cloud-ran," in IEEE International Conference on Communications (ICC), 2018, pp. 1–7.
[13] A. G. Dalla-Costa et al., "Orchestra: A customizable split-aware nfv orchestrator for dynamic cloud radio access networks," IEEE Journal on Selected Areas in Communications, vol. 38, no. 6, pp. 1014–1024, 2020.
[14] F. Z. Morais et al., "Placeran: optimal placement of virtualized network functions in beyond 5g radio access networks," IEEE Transactions on Mobile Computing, pp. 1–1, 2022.
[15] G. M. Almeida, L. d. L. Pinto, C. B. Both, and K. V. Cardoso, "Optimal joint functional split and network function placement in virtualized ran with splittable flows," IEEE Wireless Communications Letters, vol. 11, no. 8, pp. 1684–1688, 2022.
[16] 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, 2018, pp. 2366–2374.
[17] F. W. Murti, A. Garcia-Saavedra, X. Costa-Perez, and G. Iosifidis, "On the optimization of multi-cloud virtualized radio access networks," in IEEE International Conference on Communications (ICC), 2020, pp. 1–7.
[18] B. Ojaghi, F. Adelantado, A. Antonopoulos, and C. Verikoukis, "Slicedran: Service-aware network slicing framework for 5g radio access networks," IEEE Systems Journal, vol. 16, no. 2, pp. 2556–2567, 2022.
[19] F. W. Murti, S. Ali, and M. Latva-Aho, "Constrained deep reinforcement based functional split optimization in virtualized rans," IEEE Transactions on Wireless Communications, vol. 21, no. 11, pp. 9850–9864, 2022.
[20] Y. Xiao, J. Zhang, and Y. Ji, "Can fine-grained functional split benefit to the converged optical-wireless access networks in 5g and beyond?" IEEE Transactions on Network and Service Management, vol. 17, no. 3, pp. 1774–1787, 2020.

[21] Y. Xiao, J. Zhang, Z. Gao, and Y. Ji, "Service-oriented du-cu placement using reinforcement learning in 5g/b5g converged wireless-optical networks," in 2020 Optical Fiber Communications Conference and Exhibition (OFC), 2020, pp. 1–3.
[22] Z. Gao and et al., "Deep reinforcement learning for bbu placement and routing in c-ran," in 2019 Optical Fiber Communications Conference and Exhibition (OFC), 2019, pp. 1–3.
[23] L. M. Moreira Zorello et al., "Power-efficient baseband-function placement in latency-constrained 5g metro access," IEEE Transactions on Green Communications and Networking, vol. 6, no. 3, pp. 1683–1696, 2022.
[24] A. Marotta et al., "Efficient management of flexible functional split through software defined 5g converged access," in 2018 IEEE International Conference on Communications (ICC), 2018, pp. 1–6.
[25] A. Marotta and et al., "Exploiting flexible functional split in converged software defined access networks," Journal of Optical Communications and Networking, vol. 11, no. 11, pp. 536–546, 2019.
[26] Y. Xiao, J. Zhang, and Y. Ji, "Can fine-grained functional split benefit to the converged optical-wireless access networks in 5g and beyond?" IEEE Transactions on Network and Service Management, vol. 17, no. 3, pp. 1774–1787, 2020.
[27] N. Sen and A. A. Franklin, "Impact of slice granularity in centralization benefit of 5g radio access network," in 2020 6th IEEE Conference on Network Softwarization (NetSoft), 2020.
[28] R. Singh and et al., "Energy-efficient orchestration of metro-scale 5g radio access networks," in IEEE INFOCOM, 2021, pp. 1–10.
[29] "O-ran use cases and deployment scenarios," O-RAN Alliance, Tech. Rep., 2020.
[30] Y. Tsukamoto, R. K. Saha, S. Nanba, and K. Nishimura, "Experimental evaluation of ran slicing architecture with flexibly located functional components of base station according to diverse 5g services," IEEE Access, vol. 7, pp. 76 470–76 479, 2019.
[31] "Small cell virtualization functional splits and use cases," Small Cell Forum, Tech. Rep., 2016.
[32] A. Kayum, "5g transport requirement, a guiding tool for planning 5g transport network," 2020, telecommunication Engineering Centre (TEC), Government of India.
[33] M. Sulaiman et al., "Multi-agent deep reinforcement learning for slicing and admission control in 5g c-ran," in NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, 2022.
[34] 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.
[35] 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.
[36] 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, vol. 36, no. 11, pp. 2433–2443, 2018.
[37] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2022. [Online]. Available: https://www.gurobi.com
[38] M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences), first edition ed. W. H. Freeman, 1979.
[39] M. Mostofa Akbar, M. Sohel Rahman, M. Kaykobad, E. Manning, and G. Shoja, "Solving the multidimensional multiple-choice knapsack problem by constructing convex hulls," Computers & Operations Research, vol. 33, no. 5, pp. 1259–1273, 2006.

.

APPENDIX A SPLIT-RAN IS NP-HARD

Split-RAN can be shown to be NP-Hard using a polynomial time reduction from the Multi-dimensional Multiple-choice Knapsack Problem (MMKP) [39], which is known to be NP-Hard. In MMKP, there are n groups of items and m types of resources where each group i has l_i items. Each item j of group i has value $v_{i,j}$ and requires $r_{i,j,k}$ units of type-k resource. The objective of MMKP is to select one item from each group such that the value of collected items is maximized subject to the constraints for each resource.

Let us now consider a restricted case of Split-RAN where a) all the paths are eligible with respect to the delay and capacity constraints, b) each edge cloud and regional cloud consists of only one server, and c) the goal is to select one functional split for each slice to maximize the degree of centralization subject to the constraints for processing and bandwidth resources. We can transform an instance of MMKP into an instance of the restricted case of Split-RAN as follows: a) consider each group in MMKP as a slice, b) consider each item in a group as a functional split, c) consider selecting exactly one item from each group as the selection of one functional split for each slice, d) resource constraints of the knapsack as resource availability constraint for processing resources in edge and regional cloud, e) maximizing the value in MMKP as maximizing the degree of centralization. Since this transformation can be done in polynomial time of the input size, MMKP is polynomial time reducible to the restricted case of Split-RAN. Hence, Split-RAN is NP-Hard.

NABHASMITA SEN received her B.E. degree in Computer Science and Engineering from Shri Govindram Seksaria Institute of Technology and Science (SGSITS), Indore, India in 2014, and M.Tech. degree in Computer Science and Engineering from the Indian Institute of Technology Kharagpur (IIT KGP), India, in 2016. She is currently pursuing Ph.D. degree in Computer Science and Engineering at the Indian Institute of Technology Hyderabad (IITH), India. Her research interest includes 5G, Virtualized RAN, Network slicing, Energy efficiency and AI in mobile networks.

ANTONY FRANKLIN A received his B.E. degree in Electronics and Communication Engineering from Madurai Kamaraj University, India, in 2000, M.E. degree in Computer Science and Engineering from Anna University, India, in 2002, and a Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology Madras, India, in 2010. He is currently working as an Associate Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology Hyderabad (IITH), India. Before joining IITH, he worked as a Senior Engineer at DMC R&D Center, Samsung Electronics, South Korea between 2012 and 2015, and as a Research Engineer in Electronics and Telecommunications Research Institute (ETRI), South Korea between 2010 and 2012. His current research is on the development of next-generation mobile network architectures and protocols which include Cloud Radio Access Networks (C-RAN), Mobile Edge Computing (MEC), Multi-Radio Aggregation, Internet of Things (IoT), and SDN/NFV. He has published over 75 articles in refereed international journals and conferences and granted 14 US patents. He is a senior member of the IEEE and a member of the ACM. He has received Best Academic Demo Award at COMSNETS 2018 and 2nd Best Paper Award at IEEE ANTS 2017. He has served as TPC cochair for National Conference on Communications (NCC) 2018, COMSNETS (Posters) 2019, and ADCOM 2019 conferences.