Placement of 5G RAN Slices in Multi-tier O-RAN 5G Networks with Flexible Functional Splits

Egemen Sarikaya
Department of Computer Engineering
METU, Ankara, Turkey
e203614@metu.edu.tr

Ertan Onur
Department of Computer Engineering
METU, Ankara, Turkey
eronur@metu.edu.tr

***Abstract*—The network slicing concept has gained much attention with the development of software-defined network and network function virtualization technologies, enabling logically isolated networks for different purposes in the same network infrastructure. The virtualization of network functions enables the functional split of radio access network functions to fulfill different 5G radio access network requirements. Functional split can be expressed as deciding the distribution of radio access network functionalities between the radio unit placed at the edge of the network, the distributed unit placed close to the users, and the central unit placed centrally. The open radio access network concept is utilized to achieve a virtualized, interoperable radio access network among multiple vendors. In this work, we considered the placement of radio access network slices in a multi-tier 5G open radio access network architecture and formulated the problem as a Mixed-Integer Quadratically Constrained Programming, considering different functional split options for each network slice separately. Our results showed that flexible functional split utilizes physical network resources and satisfies different network slice requirements better than fixed functional split options.**

***Index Terms*—5G, Open RAN, network slicing, RAN slicing, VNF placement, baseband function placement, flexible functional split**

I. Introduction

As mobile networks get more and more complex, dedicated machines providing specific network functions are not flexible enough to satisfy the diverse requirements of different 5G use cases. Dedicated proprietary infrastructures are replaced by general-purpose systems that can host various Virtual Network Functions (VNFs) thanks to the software-defined network (SDN) [1] and network function virtualization (NFV) [2] technologies.

The virtualization of the network functions leads to the network slicing concept, one of the prominent paradigms to answer the diverse needs of 5G networks. The requirements of network slices in a network can differentiate from each other. One network slice may require low latency, while another network slice may be latency tolerant but requires high capacity. A network slice consists of a set of VNFs, and a traffic flow through these VNFs. To satisfy the needs of different network slices within a network infrastructure, the optimal placement of the VNFs of network slices onto the network infrastructure is a crucial challenge.

The virtualization of the network functions brings new opportunities and challenges, such as where to place these functions considering different network and user aspects. Placing VNFs onto the network infrastructure most efficiently and profitably considering different requirements of these functions is widely tackled in the literature, named the VNF placement problem. VNF placement and Virtual Network Embedding (VNE) are similar to each other. VNF placement problem concerns how to place VNFs to a particular physical infrastructure, whereas the VNE problem concerns the optimal mapping of virtual nodes and links satisfying specific requirements to substrate nodes and links.

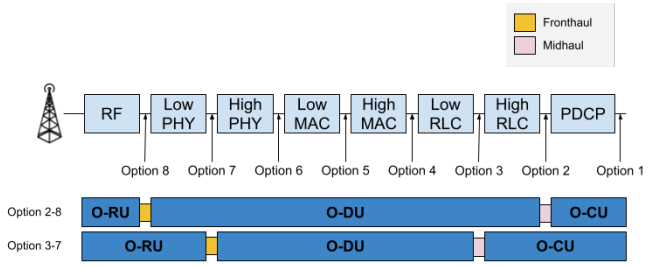


Fig. 1: Functional split options proposed by 3GPP [3] and two example mapping of dual functional split options to O-RU/O-DU/O-CU.

New 5G services have a wide range of requirements, such as low latency, high throughput, and high connection density. 5G Radio access network (RAN) has to accommodate these diverse requirements. In previous mobile generation networks, Centralized-RAN (C-RAN) architecture decouples the base station (BS) functions into two entities, the baseband unit (BBU) and the remote radio head (RRH). RRHs are placed at the edge, and BBUs are centralized into BBU pools to achieve low cost and low energy consumption. However, C-RAN requires a tremendous amount of fronthaul capacity from aggregation sites to operate. In 5G, different functional splits are proposed [3] further to decouple the baseband unit into a distributed unit (DU), a centralized unit (CU), and RRU in C-RAN becomes the radio unit (RU). In the context of Open RAN, these elements are named O-RAN distributed unit (O-DU), O-RAN centralized unit (O-CU), and O-RAN radio unit (O-RU) [4]. Adopting Open RAN has several benefits. One of the essential benefits is interoperability between multi-vendor devices. The interoperability aspect prevents vendor lock-ins

978-3-903176-36-2 ©2021 IFIP

and provides flexibility while designing the network and replacing or upgrading the network with products from different vendors. The functional split architecture we considered, dual split, splits the baseband functionalities with a high layer split (HLS) option and a lower layer split (LLS) option, which are illustrated in Fig 1.

We considered the optimal placement of network slices in multi-tier 5G radio access networks with a flexible functional split option for each network slice in this work. The placement of a network slice is formulated as a virtual network function placement problem, considering a network slice is a set of virtualized baseband functions, O-RU, O-DU, and O-CU, chained under bandwidth capacity and latency requirements.

Our contributions in this paper are as follows:

1) Placement of a 3-layer RAN slice (O-RU, O-DU, O-CU) and flexible dual split options are jointly considered. Most of the works in the literature tackle these problems separately.
2) We used different high layer split (HLS) and low layer split (LLS) options for each RAN slice as a variable, allowing the selection of optimum functional split for different slices. Each RAN slice has its own HLS and LLS option to find the optimal use of limited network resources to accommodate the slice requirements.
3) We consider eMBB, URLLC, and mMTC network slices, three main network slice types, and the impacts of admitting different types of network slices.
4) Multi-tier aggregation sites network topology is considered to reflect real-life scenarios.
5) As can be seen in Section IV, we compared the flexible functional split with the fixed functional split options regarding the utilization of physical network resources and the admission ratio of RAN slice requests.

The rest of the paper is structured as follows. Section II covers the related work done in the literature. Our optimization framework FlexNSP and network structure are covered in Section III. In Section IV, the experiment parameters are presented and explained, the experiment results are discussed. Lastly, Section V concludes the paper, presents the final remarks, and discusses the future work.

II. Related Work

In [5], Xiao et al. proposed a fine-grained functional split architecture alternative to standardized 3GPP functional split options [3] and present a quantitative model for the computational complexity of different functions. However, they do not consider the latency requirement between each fine-grained unit (FU), which corresponds to the latency consideration of different functional splits.

In [6], Guan et al. worked on a modified version of the VNE problem. They proposed an end-to-end network slice placement algorithm linking virtual resources to physical resources with the optimization objective of minimizing physical network resource usage. In their model, each network slice is a virtual network instance that needs to be mapped to the physical network. They abstracted the infrastructure network as an undirected weighted graph.

In [7], Alleg et al. worked on the Placement and Chaining of Virtual Network Functions (PC-VNF) problem and formulated the problem as a Mixed-Integer Quadratically Constrained program (MIQCP). Their main difference from the literature is the relation between allocated resources and the processing delay of a VNF. A linear relation is assumed and considered in the mathematical formulation.

The placement algorithm maps virtual network functions to physical resources, then selects the links connecting these functions. They considered three types of slices with different requirements, eMBB slices, mMTC slices, and uRLLC slices. Three different heuristic algorithms are proposed for these slice types separately. In [8], Harutyunyan et al. proposed a virtual network embedding (VNE) problem with different functional splits. Later, in [9], they expanded the work they done in [8] by considering a star-shaped topology. This work has similar aspects to our work as their embedding model includes selecting functional splits flexibly. However, the authors select functional split options for substrate small cells rather than virtual cells, which takes away the flexibility of selecting a different functional split option for each request. Specifically, PHY-RF (option 8), PHY (option 7), and MAC (option 5) splits are considered.

In [10], Coelho et al. illustrated the effects of different functional split options with different network sharing policies. They focused on end-to-end network slicing problems and interactions between slices, isolating each slice or some functions depending on network sharing policies.

In [11], Alba et al. proposed a dynamic functional split design where a RAN can change selected functional split option during run-time depending on the current traffic. In their implementation, switching between two functional split options is illustrated.

VNE placement problem with functional splitting is formulated as a Mixed Integer Quadratically Constrained Programming (MIQCP), and a MaxSAT problem to minimize the physical network resource usage minimization of network latency in [12]. They suggested different functional split options for different types of network slices. However, placing different types of network slices on the same network structure is not considered, affecting the performance of individual network slices and the whole network. The topology they used is similar to the one we use, except they do not consider more resources available at more centralized entities. Scarce resource distribution on the edge of the network should be considered because it can affect the quality of results.

In [13], Zhang et al. worked on maximizing the energy efficiency of a BBU pool in a C-RAN architecture. In their model, they dynamically turn on and off the BBU cards in the BBU pool. They formulate the problem as a version of the 2D bin-packing problem.

In [14], Yu et al. considered the placement of DU and CU functions over optical aggregation networks. They proposed an Integer Linear Programming (ILP) model for baseband

function placement. Functional split options 2 and 7 are fixed as fronthaul and midhaul interfaces respectively, flexible function split options are not considered.

In [15], isolation-aware 5G RAN network slice mapping problem in WDM metro-aggregation networks is investigated. Their network structure consists of cell sites (CSs) at the edge, access central offices (COs), main COs, and core COs. They proposed a heuristic algorithm to minimize the number of used central offices and wavelength channels while satisfying the requirements of different types of slices (URLLC, eMBB). They claimed that they proposed a flexible functional split concept, but the flexible functional split is not considered in their work; fronthaul and midhaul interfaces are fixed to split options 2 and 7.

In [16], the authors introduced a CU placement problem for mobile networks as an ILP problem while considering the impact of different functional split options. Since ILP is not scalable to vast networks, they proposed a heuristic algorithm to solve the mentioned problem. They considered PHY-RF split, PHY split, MAC split, and PDCP-RLC split, corresponding to options 8,7,5, and 2 in Fig. 1. Their problem definition has virtual DUs, CUs, and links mapped to the substrate DUs, CUs, and links. Unlike this work, our problem definition allows a site to host different DUs and CUs simultaneously, leading to more flexible placements.

Zorello et al. proposed an optimization framework considering functional split to minimize power usage of links and nodes in [17]. They choose split options 2 (PDCP-RLC) and 6 (MAC-PHY) and compute these options' bandwidth, process, and latency requirements.

In [18], De Domenico et al. considered the VNF placement problem on a hybrid cloud infrastructure. The network infrastructure is divided into remote radio heads (RRHs), edge cloud, and central cloud. Their hybrid cloud infrastructure is similar to our considerations of multi-tiered network infrastructure. They proposed a mixed-integer problem (MIP) to optimize the placement of network slices. In order to scale their framework, they proposed a heuristic solution to use when mathematical optimization has too many variables and takes too much time to find the optimal solution. However, both their MIP solution and heuristic solution fail to address the bandwidth capacity requirements of functional split options, which is a crucial aspect of the functional split concept. The main reason for proposing the functional split concept in 5G is that split option eight used in C-RAN requires a high amount of link capacity, and this requirement limits the flexibility of the network.

[11], [10] and [17] consider the effects of functional split. [7], [6], [13], [5] and [15] work on the placement of network functions. [8], [9], [12], [14], [16], and [18] work on the placement of functions considering different functional splits, but none of them used functional split as a decision variable in their problem definition. To observe the effects of a selected functional split of a slice on other slices in the network, and optimally utilize the physical network infrastructure, flexible functional split per RAN slice should be considered. We considered different types of slices with a unique functional split option assigned to each slice coexisting in the same physical network structure.

III. Flexible Network Slice Placement (FlexNSP)

There are three main types of slices. URLLC slices require low latency and high reliability; eMBB slices require high spectral efficiency, high user experienced data rate, and mMTC requiring high connection density (for example, 1 million devices/km^2 according to ITU [19]).

According to the requirements of different user groups of network slices, each network slice fits into one of the main slice types, while the requirements of slices of the same type can differ to some degree. That is, not all the eMBB slices are identical in terms of capacity and latency requirements. There may be more than one suitable functional split option for each slice, satisfying the slice requirements and fitting with the network infrastructure. In our model, we select the optimal functional split option of a slice from the perspective of Infrastructure Providers (InPs), meaning that the selected functional split option maximizes the network's total financial value while satisfying the needs of that network slice.

This section presents the Flexible Network Slice Placement problem (FlexNSP), a VNF placement problem considering the placement of network slices consisting of virtualized O-RU, O-DU, and O-CU, considering different HLS and LLS options per slice on multi-tier 5G O-RAN.

A. Network Structure

In [20], multi-tier 5G RAN illustrated, which can be seen in Fig. 2. In this example structure, cell sites are connected to tier-1 sites, tier-1 sites are connected to tier-2 sites, and tier-2 sites are connected to a tier-3 site. Note that there can be more than three tiers in a multi-tier structure; the three-layered structure in Fig. 2 is just an example for illustration purposes. From another perspective, the hierarchy can be considered as tiers in one or more data centers.

There are T site levels in the network such that tier 1 sites are the closest sites to the cell sites; tier T sites are the closest sites to the external networks. $\mathcal{N}$ is the set of sites in the network. $\mathcal{N}_t$ represents the set of tier t sites. There are also cell sites, $\mathcal{N}_c$, connected to $\mathcal{N}_1$ sites. It is convenient to have more tier sites closer to the edge because all cell sites hosting RU functions should connect to a physically close tier site to accommodate high bandwidth and low latency requirements of fronthaul communication. Therefore, central tier sites have fewer instances compared to the sites that close to the edge,

$$|\mathcal{N}_t| \geq |\mathcal{N}_{t+1}|, t = 1, ..., T-1.$$

Cell sites and tier sites have limited computational capacity in terms of giga operations per second (GOPS). $\mathcal{C}$ is the set of resource capacities of sites,

$$\mathcal{C} = \{C_1, ..., C_{|\mathcal{N}|}\},$$

Where C_n represents computational capacity of a site $n \in \mathcal{N}$. High-capacity network resources at the edge are

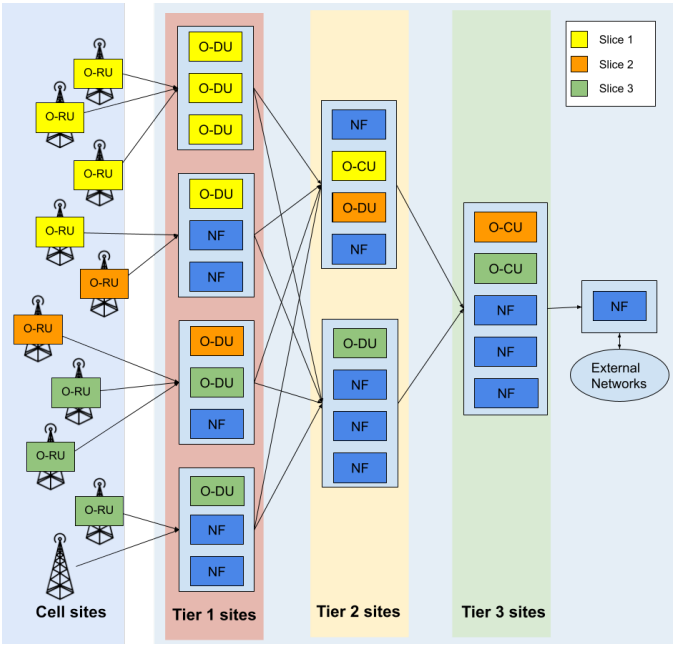


Fig. 2: Example placement on a multi-tiered 5G RAN structure.

not feasible because deployment and operational costs are high due to the high number of instances at the edge; thus, centralized sites have more capacity than peripheral sites.

$\mathcal{L}$ is the set of links that connect any pair of sites. l_{n_1,n_2} represents a link that connects sites n_1 and n_2,

$$l_{n_1,n_2}, l \in \mathcal{L}_t, n_1 \in \mathcal{N}_t, n_2 \in \mathcal{N}_{t+1}.$$

$\mathcal{Q}$ and $\mathcal{E}$ are the set of capacities and latencies of links in $\mathcal{L}$ respectively. q_l and e_l denote the capacity and latency of link $l \in \mathcal{L}$.

$\mathcal{P}$ is the set of all paths in the network between two tier sites. $p_{n_1,n_2} \in \mathcal{P}$ represents a path between servers n_1 and n_2 and consists of a set of links in $\mathcal{L}$.

B. Computational Cost Calculation

When placing network slices to the network infrastructure, the computational cost of O-RU must be satisfied by the cell site that it is placed. On top of that, the tier sites that host O-DU and O-CU of a slice must have sufficient resources to satisfy the computational costs of O-DU and O-CU.

To calculate the computational requirements of O-RU, O-DU, and O-CU of a network slice according to each functional split in terms of GOPS, a computational complexity estimation model is derived from [5],

$$D_s^F = \alpha_{s,F}(3A + A^2 + \frac{MCL}{3})\frac{\Gamma}{5},$$

Where F denotes the network slice function O-RU, O-DU or O-CU that we are calculating the computational requirements, A denotes the number of antennas, Γ denotes the number of resource blocks required by the network slice, M denotes modulation bits, C denotes coding rates, L denotes the number of MIMO layers of each radio unit. $\alpha_{s,F}$ is a scaling factor of the network slice function F depending on the selected functional split option of slice s. In other words, α is the ratio of functionalities in F to all functionalities in the network slice.

When the scaling factor α of the model equals one, the model calculates the total computational requirement of the network slice. Scaling factor α distributes the computational cost to the network slice functions proportional to the baseband functionality (RF, PHY, MAC, RLC, and PCDP) processed in the network slice function. In higher functional split options like option 2, the scaling factor of O-CU is close to one, whereas lower layer split options like option 8 have a lower scaling factor value of O-CU and a higher scaling factor value of O-RU.

TABLE I: Parameters and variables.

Parameter	Explanation	Unit
T	The set of tiers in the multi-tier network	
$\mathcal{N}$	The set of all sites	
$\mathcal{N}_t$	The set of sites in tier t	
$\mathcal{N}_c$	The set of cell sites	
C_n	The capacity of a site $n \in \mathcal{N}$	GOPS
$\mathcal{L}$	The set of all links	
$l_{(n_1,n_2)} \in \mathcal{L}$	The link that connects site $n_1, n_2 \in \mathcal{N}$	
e_l	Link latency on link $l \in \mathcal{L}$	ms
q_l	Link capacity on link l	MHz
$\mathcal{P}$	The set of all paths consist of a subset of links in $\mathcal{L}$	
$p_{(n_1,n_2)} \in \mathcal{P}$	Denoted the path that connects n_1 and n_2	
$\mathcal{S}$	The set of RAN slices.	
R_s^e	Denotes the end-to-end latency requirement of a slice $s \in \mathcal{S}$.	ms
R_s^{CS}	Denotes the set of requested cell sites of a slice $s \in \mathcal{S}$.	
$A_{s,j}^{HLS/LLS}$	Indicated whether HLS/LLS option j is selected for slice s.	
$J^{HLS/LLS}$	The set of all HLS/LLS options	
$\mathcal{D}$	The set of server capacity demands depending on the function split options	
D_s^{DU}	Denotes the cost of a DU function of slice s.	GOPS
$H_s^{mid/front}$	Denotes the midhaul/fronthaul requirement of slice s	
$m_{s,n}^f$	The binary variable to indicate if VNF $f \in F_s$ (RU,DU or CU) is placed on server n or not.	
x_s	The binary variable to indicate if slice s is admitted or not.	
$k_s^{r,d}$	The binary variable to indicate if $RU_{s,r}$ is connected to $DU_{s,d}$.	
$y_{s,p}^{mid,d}$	The binary variable to indicate if path p is connecting DU $d \in F_s^{DU}$ and CU_s.	
$y_{s,p}^{front,r}$	The binary variable to indicate if path p is selected for RU $r \in F_s^{RU}$ to connect it to a DU of slice s	
F_s^f	The set of network functions $f \in$ {O-RU, O-DU, O-CU} of a slice s	
ϵ_s	Centralization factor of the slice s	
v_s	Potential financial value when the slice is admitted	

C. Formulation of FlexNSP

A RAN slice can be seen as a set of virtual RAN functions (O-RU, O-DU, O-CU) that together constitutes a complete radio access network. $\mathcal{S}$ is the set of RAN slice requests that can be placed in the network,

$$\mathcal{R} = \{R_1, ..., R_{|\mathcal{S}|}\}.$$

$\mathcal{R}$ is the set of RAN slice requirements that need to be satisfied if the slice is placed. R_s can be expanded as different types of requirements of a RAN slice. R_s^e denotes the latency requirement of RAN slice s. R_s^{mid} and R_s^{front} corresponds to the midhaul and fronthaul capacity requirements of slice s respectively,

$$R_s = \{R_s^e, R_s^{front}, R_s^{mid}\}, s \in \mathcal{S}.$$

The distribution of RAN functions between O-RU, O-DU and O-CU is vital to meet slices' requirements. In Fig. 1, different function split options proposed by 3GPP [3] and some example mappings of baseband functionalities to O-RU/O-DU/O-CU can be seen. It is important to note that not all mapping possibilities (Options 2-8, 2-7, 2-6, 3-8, 3-7, 3-6 for our problem) are represented in Fig. 1 to avoid cluttering.

Baseband functions on the left of the LLS option are assigned to O-RU, and functions on the right of the HLS option are assigned to O-CU. The functions between the LLS option and the HLS option are assigned to O-DU.

The placement of these functions on the mentioned multi-tier aggregation 5G O-RAN is a crucial aspect of network slicing, as placing functions closer to the edge provides benefits like low latency required by some specific types of slices. However, as functions are moved to the edge, some techniques that require centralization of the functions, such as Coordinated Multipoint (CoMP), cannot be utilized in the network [21].

Our objective is to maximize the profit of InP. We considered that we have a set of network slice requests gathered from different customers in our problem. From the perspective of Infrastructure Provider (InP), each network slice request $s \in \mathcal{S}$ has a value v_s demonstrating the financial gain InP acquires if the network slice is admitted to the network infrastructure. Our problem's objective is to maximize the total financial value while satisfying the requirements of admitted slices.

While assigning financial values to the network slice requests, analysis of financial values considering individual network slice requests is out of our work's scope. We employ a simple strategy to assign values to network slice requests, a random value from a defined range of values depending on the type of the network slice. As we can see from the ranges of admittance value (financial value) in Table IV, URLLC slices are more valuable than other types because they are generally mission-critical slices for customers; deployment of these slices should have a priority over other types. Also, low latency requirements of URLLC slices cause them to be placed closer to the edge, consuming infrastructure's scarce edge resources rather than residing in more central sites.

Considering the functional splits in Fig. 1, we considered functional split options 2 and 3 as high layer split options; functional split options 6, 7 and 8 as low layer split options,

$$J^{HLS} = \{j_2, j_3\},$$
$$J^{LLS} = \{j_6, j_7, j_8\}.$$

J^{HLS} is the set of higher layer functional split options. J^{LLS} is the set of lower layer functional split options.

It is important to note that functional split option 8 is the split used in the C-RAN architecture. In option 8, all baseband processing is centralized and fronthaul requirements are very demanding. The reason for excluding options 1, 4, 5 is that these functional split options are not feasible to select in any scenario, bandwidth and latency requirements of these functional splits can be too demanding for the benefits they present [3].

$\mathcal{A}$ is the set of functional split variables,

$$\mathcal{A} = \{A_1, ..., A_{|\mathcal{S}|}\}.$$

A_s^{HLS} and A_s^{LLS} are the functional split option variables for slice $s \in \mathcal{S}$,

$$A_s^{HLS} = \{A_{s,j}^{HLS} | j \in J^{HLS}\},$$
$$A_s^{LLS} = \{A_{s,j}^{LLS} | j \in J^{LLS}\},$$

where $A_s^{HLS^T}$ and $A_s^{LLS^T}$ denote the transposes of A_s^{HLS}, A_s^{LLS} respectively. For a RAN slice s, the selected HLS and LLS options dictate latency and midhaul requirements.

The computational cost of O-RU, O-DU, and O-CU depends on which HLS and LLS options are used. As mentioned in subsection III-B, $\mathcal{D}$ denotes the set of processing demands of O-RU, O-DU, and O-CU functions according to different functional splits. $\mathcal{D}_s^F$ is the computational cost of $F \in \{$O-RU, O-DU, O-CU$\}$ of slice s.

$\mathcal{H}^{mid}$ and $\mathcal{H}^{front}$ denote the set of midhaul and fronthaul capacity requirements according to different functional splits respectively,

$$\mathcal{H}^{mid} = \{H_j^{mid} | j \in J^{HLS}\},$$
$$\mathcal{H}^{front} = \{H_j^{front} | j \in J^{LLS}\}.$$

For each network slice, midhaul and fronthaul capacity requirements depend on the selected functional split options. Selected midhaul and fronthaul paths should satisfy these requirements to admit the network slice in the network. R_s^{mid} and R_s^{front} denote the midhaul and fronthaul capacity requirements of slice s,

$$R_s^{mid} = \mathcal{H}^{mid} A_s^{HLS^T},$$
$$R_s^{front} = \mathcal{H}^{front} A_s^{LLS^T}.$$

Each network slice have different end-to-end maximum latency tolerances according to their use case requirements. Therefore, midhaul and fronthaul path latencies should be considered jointly as transport latency, and transport latency

should comply with the requirements of the slice. $\mathcal{G}^{mid}$ and $\mathcal{G}^{front}$ denote the sets of transport latency requirements, meaning maximum allowed latency of midhaul and fronthaul paths,

$$\mathcal{G}^{mid} = \{G_j^{mid} | j \in J^{HLS}\},$$
$$\mathcal{G}^{front} = \{G_j^{front} | j \in J^{LLS}\}.$$

G_s^{mid} and G_s^{front} denote the HLS and LLS latency requirement of slice s,

$$G_s^{mid} = \mathcal{G}^{mid} A_s^{HLS^T}, s \in \mathcal{S},$$
$$G_s^{front} = \mathcal{G}^{front} A_s^{LLS^T}, s \in \mathcal{S}.$$

Mathematical Formulation of FlexNSP

In the light of these definitions, the mathematical formulation of FlexNSP is as follows:

$$\max_{x} \quad \sum_{s \in \mathcal{S}} x_s \upsilon_s \epsilon_s \tag{1.1}$$

$$\text{s.t.} \quad \sum_{j \in J^{HLS}} A_{s,j}^{HLS} = x_s, s \in \mathcal{S} \tag{1.2}$$

$$\sum_{j \in J^{LLS}} A_{s,j}^{LLS} = x_s, s \in \mathcal{S} \tag{1.3}$$

$$\sum_{n \in N} m_{s,n}^{CU} = x_s, s \in \mathcal{S} \tag{1.4}$$

$$\sum_{n \in N} m_{s,n}^{DU,d} = x_s, s \in \mathcal{S}, d \in F_s^{DU} \tag{1.5}$$

$$\sum_{p \in \mathcal{P}} y_{s,p}^{front,r} = x_s, s \in \mathcal{S}, r \in R_s^{CS} \tag{1.6}$$

$$\sum_{p \in \mathcal{P}} y_{s,p}^{mid,d} = x_s, s \in \mathcal{S}, d \in F_s^{DU} \tag{1.7}$$

$$y_{s,p_{n_1,n_2}}^{front,r} \leq m_{s,n_1}^{RU,r}, s \in \mathcal{S}, p \in \mathcal{P}, r \in R_s^{CS} \tag{1.8}$$

$$y_{s,p_{n_1,n_2}}^{front,r} \leq m_{s,n_2}^{DU,r}, s \in \mathcal{S}, p \in \mathcal{P}, r \in R_s^{CS} \tag{1.9}$$

$$y_{s,p_{n_1,n_2}}^{mid,d} \leq m_{s,n_1}^{DU,d}, s \in \mathcal{S}, p \in \mathcal{P}, d \in F_s^{DU} \tag{1.10}$$

$$y_{s,p_{n_1,n_2}}^{mid,d} \leq m_{s,n_2}^{CU}, s \in \mathcal{S}, p \in \mathcal{P}, d \in F_s^{DU} \tag{1.11}$$

$$\sum_{s \in \mathcal{S}} \sum_{d \in F_s^{DU}} m_{s,n}^{CU} D_s^{CU} + m_{s,n}^{DU,d} D_s^{DU} \leq C_n, n \in \mathcal{N} \tag{1.12}$$

$$\sum_{p \ni l, p \in \mathcal{P}} \sum_{s \in \mathcal{S}} y_{s,p}^{mid,r} R_s^{mid} + y_{s,p}^{front,r} R_s^{front} \leq q_l, l \in \mathcal{L} \tag{1.13}$$

$$\sum_{p \in P} \sum_{l \in p} y_{s,p}^{mid} e_l \leq G_s^{mid}, s \in \mathcal{S} \tag{1.14}$$

$$\sum_{p \in P} \sum_{l \in p} y_{s,p}^{front} e_l \leq G_s^{front}, s \in \mathcal{S} \tag{1.15}$$

$$\sum_{p \in P} \sum_{l \in p} y_{s,p}^{front} e_l + y_{s,p}^{mid} e_l \leq R_s^e, s \in \mathcal{S} \tag{1.16}$$

The objective expressed in (1.1) indicates the maximization of the total value of admitted RAN slices. Centralization factor ϵ_s is a multiplier in the equation, increasing the slice value as more central function split options are selected. We want to locate our functions as central as possible because the centralization of a function enables centralization benefits such as Coordinated Multipoint (CoMP) and low operational expenditures (OpEx). Each RAN slice has a centralization factor ϵ_s correlated to the selected functional split of slice s to reflect the centralization gains to our objective. Lower functional split options have a lower centralization factor; higher functional split options have a high centralization factor to reflect that centralization of a slice increases the value of the slice by increasing efficiency and decreasing the cost of managing the slice.

x_s is a decision variable indicating whether slice s is admitted or not. $m_{s,n}^{CU}$ is a decision variable indicating if CU function of slice s is placed on site n. Similarly, $m_{s,n}^{DU,d}$ indicates if DU function d of slice s is placed on site n. Since a RAN slice consists of a single CU function and multiple DU functions, each DU function has a separate decision variable. However, $m_{s,n}^{RU,r}$ is not a decision variable because placements of RU functions depend on the desired location, which is given information with the slice request. In our problem, a functional split option is selected for each slice to accommodate the slice's requirements. Constraints (1.2), (1.3) ensure that each slice selects one HLS option and one LLS option if the slice is admitted, respectively. Functions of a slice should be placed in the network if and only if the slice is admitted. Constraint (1.4) ensures CU function of slice s is placed if slice s is admitted. Constraints (1.5) ensures a DU function of slice s is placed for each RU if slice s is admitted.

Note that the number of DU functions of slice s, $|F_s^{DU}|$, is equal to the number of RU functions of the same slice s, $|R_s^{CS}|$, because each RU function has to connect a DU function.

After placing the functions to tier sites, the connectivity between them must be ensured. From all possible paths in the network, appropriate ones should be assigned to connect slice functions. Constraints (1.6) and (1.7) ensure that the number of fronthaul and midhaul paths selected is equal to the number of RU_s and DU_s functions placed. A path p_{n_1,n_2} connects a lower tier server n_1 to a higher tier server n_2,

$$t_1 < t_2, t_1 \in T, t_2 \in T, n_1 \in \mathcal{N}_{t_1}, n_2 \in \mathcal{N}_{t_2}, p_{n_1,n_2} \in \mathcal{P}.$$

F_s is the set of functions that composes the slice s. $F_s^{RU} \in F_s$ is the set of RU functions in the slice s. Constraints (1.8) and (1.9) ensure that a path p_{n_1,n_2} can be selected as a fronthaul path of O-RU function $r \in R_s^{CS}$ if and only if O-RU function is placed at cell site n_1 and O-DU function is placed at tier site n_2. Constraints (1.10) and (1.11) ties the selected midhaul paths to the placement of DU functions and the CU function of slice s.

The capacity of physical network resources, tier sites, and links should satisfy the requirements of all placed slices and slice functions. Constraint (1.12) ensures that total resource demand in a site $n \in \mathcal{N}$ cannot exceed the total capacity of the site C_n. Constraint (1.13) ensures that the the capacity of the link l, q_l, is not exceeded.

Constraints (1.14), (1.15) ensure that selected midhaul and fronthaul path p for a slice s can satisfy the midhaul and fronthaul functional split latency requirement of slice s, G_s^{mid}, G_s^{front} respectively. Constraint (1.16) dictates that the total one-way latency between the user and slice functions should be lower than the slice latency requirement.

TABLE II: Functional Split Requirements [22].

Functional Split Option	Bandwidth Requirement	Latency Requirement	Centralization Factor
Option 2	4.016 Gb/s	10 ms	1
Option 3	4 Gb/s	6 ms	1.1
Option 6	4 Gb/s	8 ms	1.1
Option 7	38 Gb/s	5 ms	1.2
Option 8	157.3 Gb/s	3 ms	1.3

IV. Evaluation and Discussion

In this section, we discuss the solutions to our Mixed-Integer Quadratically Constrained Programming (MIQCP) problem. We solved the MIQCP problem defined in (1.1), ..., (1.16), with different optimization parameters, using Gurobi v9.0 solver , which is a commercial mathematical optimization solver [23]. For conducting the experiments, we used a computer with Intel® Core™ i5-8250U Processor and 12 GB RAM. Different functional split requirements derived from [22] can be seen in Table II.

TABLE III: Radio parameters used in the experiments.

Parameter	Value
Number of antennas	4
Number of MIMO layers	2
Modulation	64-QAM
Coding	5/6

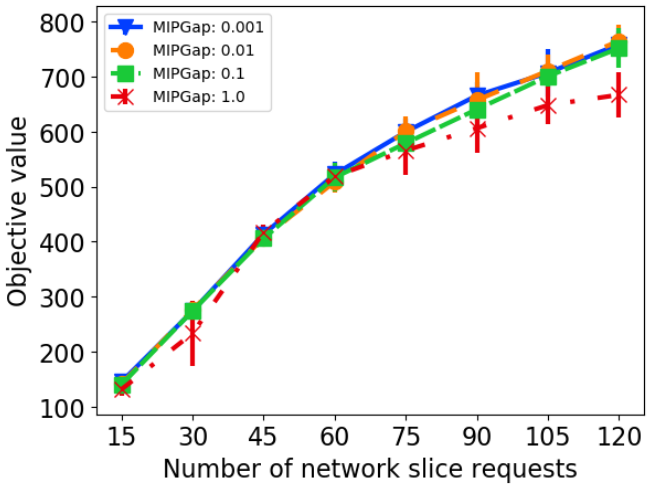
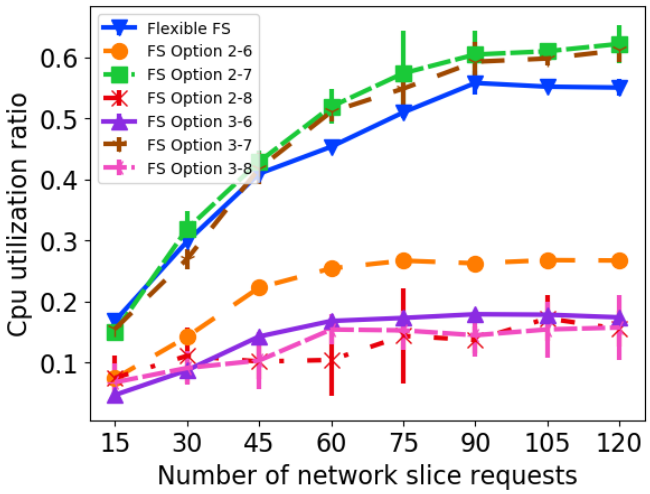
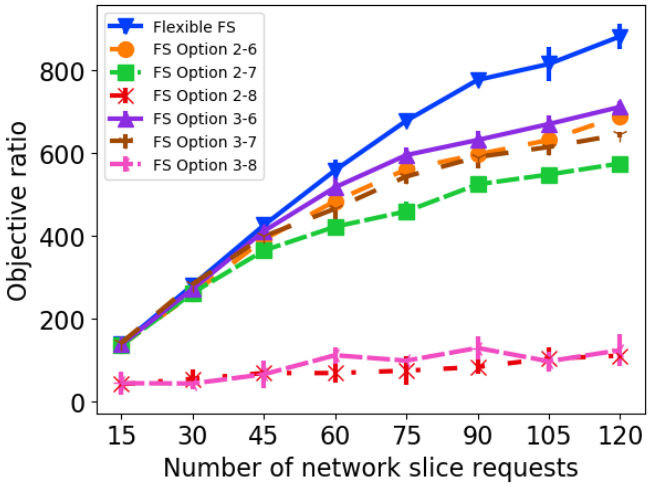


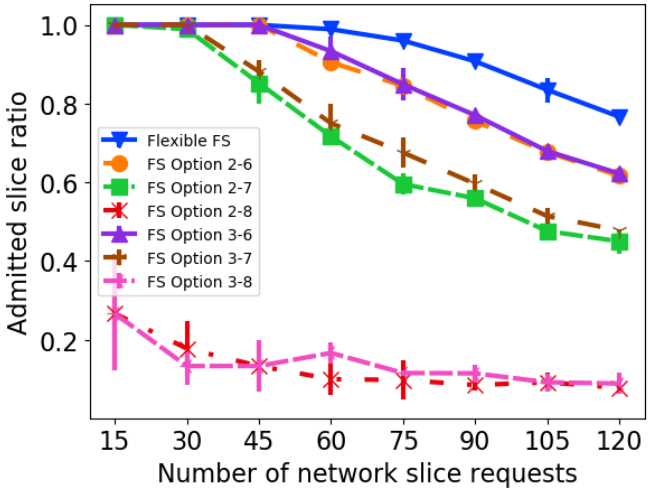
Fig. 3: Performances of Flexible Functional Split for different MIPGap values.



(a) Utilization of site resources of different functional split options.

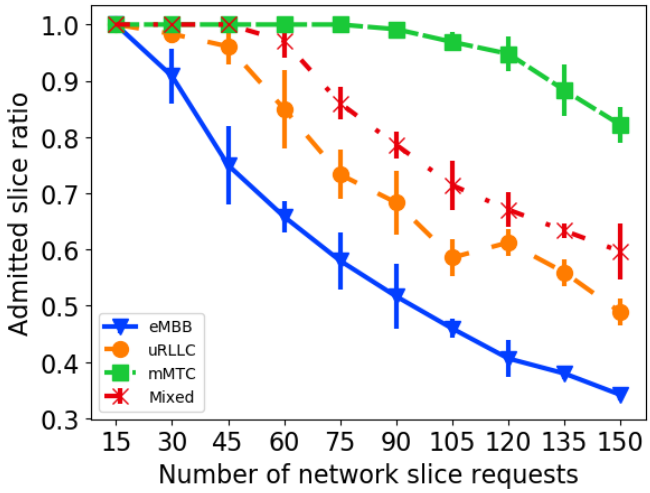


(b) Determined optimization objective performances of different functional split options.

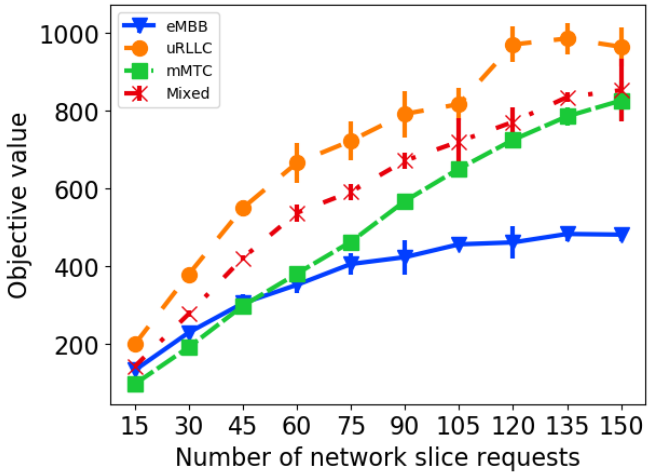


(c) The ratio of admitted RAN slice requests to all RAN slice requests.

Fig. 4: Objective value and network slice acceptance results.



(a) Admission rate of network slices of different network slice types.



(b) Determined optimization objective performances of different network slice types.

Fig. 5: Results of different network slice types.

A. Experimental Settings

The experiment topology has a similar structure to Fig. 2. Processing requirements of O-RU, O-DU, and O-CU are calculated using the model mentioned in subsection III-B, with the radio parameters listed in Table III. In each experiment, experiment values are set between mentioned ranges randomly. Since we adopted O-RAN architecture, our network infrastructure can have sites with different GOPS capacities from different vendors. To simulate the heterogeneous aspect of the infrastructure, we assigned random GOPS capacities to the sites within a range depending on the tier of the site. In tier-3, there is one site with [1600, 2000] GOPS capacity available for VNFs. In tier-2, there are five sites, each containing [800, 1000] GOPS capacity available. In tier-1, 25 sites with [400, 500] GOPS capacity are located. The capacity of each link between tier-2 sites and tier-3 sites is [400, 800] Gbps. Between tier-1 sites and tier-2 sites, it is [200, 300] Gbps.

B. Results and Discussion

We consider three types of network slices with different requirements, which can be seen in Table IV.

In our results, mathematical solutions with different MIPGap values are presented. MIPGap is the Mixed Integer Program (MIP) gap, representing the gap between the upper and lower objective bound. When we use lower MIPGap values, the solution we found is closer to the optimal solution. However, using too low MIPGap values increases the time consumption of the optimization process to unreasonable levels, whereas using too high MIPGap values generates poor sub-optimal results.

Fig. 3 illustrates the effects of different MIPGap values on the quality of the solution Gurobi finds. As expected, better results are achieved with smaller MIPGap values. It is important to mention that it took $\sim$ 50000 seconds (more than 13 hours) to find the solution at 0.001 MIPGap value when there are 120 RAN slice requests, although the achieved objective value is not significantly higher than 0.01 MIPGap at any point of the experiments, which took $\sim$ 900 seconds to compute.

For our other experiments, we chose the 0.01 MIPGap value because it is a good compromise between the quality of the solution and the time consumption of optimization.

In Fig. 4a, 4b and 4c, we compared flexible functional split per RAN slice with fixed functional splits. Fig. 4a illustrates the utilization of resource tier sites that hosts at least one virtual network function. As the number of slice requests increases, the difference between flexible functional split and fixed split options can be seen easily. Flexible functional split utilizes the physical network better than any fixed functional split option. The objective value represents the value of our problem's objective function that can be seen in (1.1). The objective ratio is the ratio between the best objective value found by solving the model and the maximum theoretical objective value achieved if we have infinite resources and admit all requests with maximum centralization. For different functional split options, Fig. 4b illustrates the objective ratio achieved with different number of RAN slice requests. From the figure, we can see that flexible functional split achieves better objective ratios, especially at high numbers of slice requests. The decrease of ratio for all options is expected because as the number of requests increases, we need to choose the most profitable requests and reject low-value requests to maximize our profit from the limited physical network resources. The ratio of admitted RAN slice requests to all requests can be seen in Fig. 4c. Since our physical network resources are not changing, increasing the number of candidate network slices lowers the admission ratio. As shown in the figure, the flexible functional split has the highest admission ratio because the flexible functional split can satisfy the diverse RAN slice request requirements. Although it is beneficial to have a high admission ratio, we do not try to achieve the highest admission ratio since each request has a value; that is, admitting a highly valued request over low valued multiple

TABLE IV: Network slice parameters.

Slice Type	Number of O-RUs	Number of RBs per O-RU	Maximum latency tolerance	Admittance value
eMBB	[2, 5]	50	[4, 10]	[3, 5]
uRLLC	[1, 10]	25	[1, 2]	[5, 7]
mMTC	[6, 10]	5	[10, 30]	[2, 4]

requests can be beneficial for maximum profit.

In Fig. 5a and 5b, effects of different network slice types are compared. In these two experiments, network slice types are evenly distributed in the mixed case, while all network slices have the same type in the other cases. Fig. 5a indicates that when all network slices are mMTC slices, we achieve the highest admission rate despite requiring the highest number of O-RUs because each O-RU requires significantly fewer resource blocks compared to other types of network slices. eMBB slices have the lowest admission rate because of the higher resource block requirements than other types. Total objective value is presented in Fig. 5b shows that uRLLC slices maximizes profit because uRLLC slices are higher value than other types of slices with stringent requirements.

These results showed that jointly considering functional split and network slice placement yields better results than selecting a functional split across the whole network. Also, selecting the proper optimization parameters is essential as getting the best result in a reasonable time requires careful selection.

V. Conclusion

We introduced FlexNSP, a network slice placement problem with flexible functional splitting in multi-tier 5G open RAN, formulated it as a mathematical problem, and compared the results with fixed functional split options solved with Gurobi solver. Our results have shown that flexible functional splits enable infrastructure providers to utilize their limited resources better and satisfy the needs of more mobile network operators to increase both sides' profits compared to fixed functional split options. As future work, we plan to implement different heuristic algorithms and experiment on different network topologies because mathematical solutions are not scalable for vast network topologies and compare our results with state-of-art algorithms.

VI. Acknowledgement

This work was supported by Vodafone within the 5G and Beyond Joint Graduate Support Program, which is run by the BTK.

References

[1] D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2014.

[2] G. ETSI, "Etsi gs nfv 001 v1. 1.1 (2013-10) network functions virtualisation (nfv); use cases," *oct-2013*, 2013.

[3] "Technical specification group radio access network; study on new radio access technology, radio access architecture and interfaces (release 14)," Tech. Rep. 3GPP TR 38.801, 3GPP, April 2017.

[4] Open RAN Alliance, "O-RAN: towards an open and smart ran," *White Paper*, 2018.

[5] 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.

[6] W. Guan, X. Wen, L. Wang, Z. Lu, and Y. Shen, "A service-oriented deployment policy of end-to-end network slicing based on complex network theory," *IEEE Access*, vol. 6, pp. 19691–19701, 2018.

[7] A. Alleg, T. Ahmed, M. Mosbah, R. Riggio, and R. Boutaba, "Delay-aware vnf placement and chaining based on a flexible resource allocation approach," in *2017 13th International Conference on Network and Service Management (CNSM)*, pp. 1–7, 2017.

[8] D. Harutyunyan and R. Riggio, "Flexible functional split in 5G networks," in *2017 13th International Conference on Network and Service Management (CNSM)*, pp. 1–9, 2017.

[9] 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.

[10] W. da Silva Coelho, A. Benhamiche, N. Perrot, and S. Secci, "On the impact of novel function mappings, sharing policies, and split settings in network slice design," in *2020 16th International Conference on Network and Service Management (CNSM)*, pp. 1–9, 2020.

[11] A. M. Alba, J. H. G. Velásquez, and W. Kellerer, "An adaptive functional split in 5G networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 410–416, IEEE, 2019.

[12] J. Yusupov, A. Ksentini, G. Marchetto, and R. Sisto, "Multi-objective function splitting and placement of network slices in 5G mobile networks," in *2018 IEEE Conference on Standards for Communications and Networking (CSCN)*, pp. 1–6, IEEE, 2018.

[13] J. Zhang, Y. Ji, X. Xu, H. Li, Y. Zhao, and J. Zhang, "Energy efficient baseband unit aggregation in cloud radio and optical access networks," *Journal of Optical Communications and Networking*, vol. 8, no. 11, pp. 893–901, 2016.

[14] H. Yu, F. Musumeci, J. Zhang, Y. Xiao, M. Tornatore, and Y. Ji, "Du/cu placement for c-ran over optical metro-aggregation networks," in *International IFIP Conference on Optical Network Design and Modeling*, pp. 82–93, Springer, 2019.

[15] 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*, vol. 38, no. 6, pp. 1125–1137, 2020.

[16] D. Harutyunyan, R. Riggio, S. Kuklinski, and T. Ahmed, "Cu placement over a reconfigurable wireless fronthaul in 5G networks with functional splits," *International Journal of Network Management*, vol. 30, no. 1, p. e2086, 2020.

[17] L. M. M. Zorello, S. Troia, M. Quagliotti, and G. Maier, "Power-aware optimization of baseband-function placement in cloud radio access networks," in *2020 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–6, IEEE, 2020.

[18] A. De Domenico, Y.-F. Liu, and W. Yu, "Optimal virtual network function deployment for 5G network slicing in a hybrid cloud infrastructure," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 7942–7956, 2020.

[19] "Minimum requirements related to technical performance for imt-2020 radio interface(s)," Tech. Rep. ITU-R M.2410-0, Radiocommunication Sector of International Telecommunication Union, November 2017.

[20] "NGMN Overview on 5G RAN Functional Decomposition," tech. rep., Next Generation Mobile Networks Alliance, 2018.

[21] L. M. 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, 2018.

[22] "Small cell virtualization functional splits and use cases," Tech. Rep. SCF159.07.02, Small Cell Forum, January 2016.

[23] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2020.