Sliced-RAN: Joint Slicing and Functional Split in Future 5G Radio Access Networks

Behnam Ojaghi, Ferran Adelantado
Wireless Networks Research Lab (WINE)
Universitat Oberta de Catalunya (UOC)
Castelldefels, Spain
{bojaghi,ferranadelantado}@uoc.edu

Elli Kartsakli
Iquadrat Informatica S.L.
Barcelona, Spain
ellik@iquadrat.com

Angelos Antonopoulos, Christos Verikoukis
Centre Tecnologic de Telecomunicacions de Catalunya (CTTC/CERCA)
Castelldefels, Spain
{aantonopoulos, cveri}@cttc.es

***Abstract*—The unprecedented surge in mobile data traffic, along with the wide range of services and the corresponding different performance requirements, has raised the need for a new mobile network architecture. In that sense, 5G has been conceived as a software defined network able to provide service-tailored connectivity. Network slicing is a key mechanism to serve efficiently the diversified service requirements. In this paper, we formulate the joint RAN slicing and functional split with optimization of centralization degree (CD) and throughput. Our results show that even though in terms of CD we have more costs, we can better meet the service requirements and also have more throughput in the network.**

***Index Terms*—5G network, functional split, slicing**

I. Introduction

The unprecedented surge in data traffic experienced over the last decade has stretched the telecommunications networks to their capacity. According to Cisco's forecast [1], global IP traffic will have increased 127-fold from 2005 to 2021. This demand rise will be exacerbated for wireless and mobile traffic, which is expected to account for more than 60% of the total IP traffic by 2021, growing twice as fast as fixed IP traffic. It is precisely in this context of traffic explosion in which the requirements for 5G have been defined [2]. In a nutshell, 5G must support high data rates, low latency targeting about 1 ms round-trip time, a reduction of the cost and energy consumption, and massive connectivity [3].

The diversity of performance requirements, ranging from 1Gbps peak rate to 1ms end-to-end latency, renders a classical static network architecture unfeasible for 5G. Thus, 5G is conceived as a network able to provide service-tailored connectivity, by leveraging network virtualization techniques based on Software Defined Networking (SDN) [4] which provides the network with enough flexibility, mainly to create isolated slices on-demand on top of the physical network.

The virtualization and the slicing process required to adapt the Radio Access Network (RAN) to users' performance requirements pose significant challenges in 5G networks. In the future, RAN will be composed of a Central Unit (CU) and a set of geographically distributed Remote Radio Heads (RRH) connected through a packet-based network, as proposed by 3GPP in [5] and in [6][1]. Although the optimal distribution of layers PDCP/RLC/MAC/PHY between the CU and the RRHs (known as *functional split*) has attracted the attention of the research community, it still remains as an open research problem [7]–[10]. In this paper, we propose a joint analysis of the functional split along with RAN slicing to meet the requirements of diverse traffic slices simultaneously.

The rest of the paper is organized as follows. In Section II, an overview of the current state of the art is provided, and the main contributions of the paper are stated. Section III describes the problem. In Section IV, we formulate the optimization of computational cost and the throughput. Section V explores the performance analysis of the system model and shows the corresponding results. Finally, Section VI presents the conclusion for this work.

II. RAN Virtualization and Slicing

Virtualization and slicing have become two major concepts of future 5G networks. Given the importance of SDN as an enabler for both virtualization [11] and slicing [12], some recent works have been focused on RAN virtualization platforms and slicing designs. Foukas et al. propose FlexRAN in [13], a flexible and programmable software-defined RAN (SD-RAN) platform. The platform is composed of a centralized controller and one agent per evolved Node B (eNB) that separates control and data planes and allows a flexible control plane design. This architecture enables the dynamic allocation of control functions between the centralized controller and the decentralized agents, thus tailoring the RAN to meet the performance requirements. However, despite making a step forward in the direction of the virtualization of the RAN, the proposal still lacks a slicing design. The same authors in [14] proposed Orion, which is a RAN slicing design running on the FlexRAN platform, while guaranteeing the functional isolation among slices. Isolation of functions among slices is of paramount importance, since it allows a slice-custom functional split within a single shared eNB. In other words, two different slices sharing the same physical node can be configured with different functional splits.For instance, in a specific eNB, a slice serving

[1]The packet-based network connecting the CU and RRHs is an integrated backhaul/fronthaul network, also known as *cross-haul* in [6].

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

a high speed user better suits a centralized functional split, so that the coordination among neighboring cells is tighter and the handover performance can be simplified. Conversely, the slice serving a low latency user requires more decentralized functional splits to reduce the HARQ delay. This is the main weakness of [7], [8], [15] where, for simplicity, either no slicing is considered or all slices are assumed to suit the same functional split. Two notable recent works, [6] and [16], address the optimization of the functional split. Specifically, WizHaul is proposed in [6] as a joint routing and functional split optimization to achieve maximum centralization. Similarly, FluidRAN follows the same rationale but targeting the monetary cost minimization [16]. However, none of them consider the slicing of the RAN.

In this context, this paper is aimed to design a joint routing (from user to CU) and functional split optimization while considering different slices. As shown schematically in Fig. 1, the slicing of the RAN allows a customized functional split deployment per slice, thus optimizing the available resources, e.g. transport network capacity and RRH or CU computational capacity. Fig. 1 conveys how different control functions are allocated either at the RRH or at the CU for each slice.

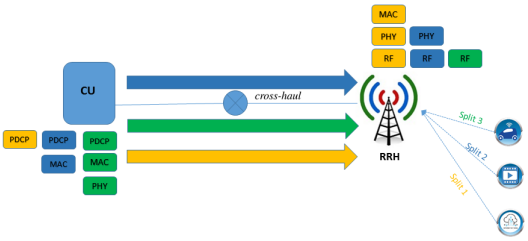

Fig. 1. Slicing based on service requirements.

Specifically, the contributions of this paper are the following:

- We propose a joint slicing and functional split RAN optimization. As described in [5] and implemented in [14], functional isolation is assumed at the eNB. This means that each slice of an eNB can have a different functional split.
- The slice creation is extended up to the user. In general, the slicing only considers the RAN, [6], [16]. Instead, this paper aims to exploit the high density of RRHs by jointly analysing routing in the RAN and user association.
- Not all services support the whole range of possible functional splits. For instance, high transmission rates usually require a high degree of centralization to implement efficient Coordinated Multipoint (CoMP). We take this into account in the functional split optimization.

III. System Model

As thoroughly described in Section II, the RAN can be modeled as a CU, referred to as node 0, and a set of RRHs, $\mathcal{R} = \{1, \ldots, R\}$. Connecting the RRHs and the CU, there is a transport network (fronthaul/backhaul) composed of a set of forwarding nodes, $\mathcal{Q} = \{1, \ldots, Q\}$, connected among them and with the CU and the RRHs through a set of links $\mathcal{L} = \{l_{i,j} : i, j \in \mathcal{R} \cup \mathcal{Q} \cup \{0\}\}$. Each link $l_{i,j} \in \mathcal{L}$ has a capacity equal to $\omega_{i,j} \geq 0$ (in bps) and a delay $d_{i,j} \geq 0$ (in sec). The connection between the CU and an RRH can be realized through multiple paths [16]. Let us then define the ith path $P^{r,i}$ from RRH r to the CU as the set of links between them. Given that there might exist multiple paths between RRH r and CU, all possible paths from RRH r to CU is denoted by P^r. Additionally, the set of users served by the network is denoted by $\mathcal{U}$ and are characterized by their required data rate λ_u^s, where $u \in \mathcal{U}$ and $s \in \mathcal{S}$ is the service type of the user.

3GPP has proposed in [5] a wide range of possible granularities of the functional split, from the coarsest granularity (the functional split is determined by the computational capacity of the RRH and CU computational capacities and the transport network capacity) to the finest granularity (the functional split is decided on a user, bearer or slice basis). Without precluding any of the granularity levels proposed by 3GPP, in the sequel we focus on the traffic type based functional split. As shown in [5], the control functions of layers PDCP, RLC (high and low), MAC (high and low) and PHY (high and low) can be allocated either in the CU or in the RRH. Accordingly, a specific allocation (i.e., functional split) can be determined per slice. Given that slicing will be done on a service type basis, hereafter service and slice concepts will be interchangeable. In the sequel we will assume a set of four RAN functions, denoted as $\mathcal{F} = \{f_0, f_1, f_2, f_3\}$, where f_0 is the low layer network function (RF, signal and analog processing, etc.) which is always placed in the RRH and f_3 is the high layer network function (e.g., PDCP and above layers). Depending on the functional split, these functions will be allocated either at the RRH or at the CU. That is, a completely centralized functional split will allocate f_0 at the RRH and the rest of functions at the CU. Conversely, a completely decentralized functional split will accommodate all functions at the RRH. Initially, f_0 is always placed in RRH and f_3 is in CU, no matter which split we use. Based on the aforementioned definitions, three different functional splits are considered in the following, as also assumed in [6]. In split 1, ϕ_1, f_1 and f_2 are located in RRH. This split is useful to serve users with low latency requirements. Split 2, ϕ_2, which enables a better utilization of hardware, only allocates f_1 in RRH while f_2 is moved to CU. In split 3, ϕ_3, all functions are moved to CU (known as Centralized-RAN, C-RAN) which allows a higher degree of coordination among eNBs.

Each functional split imposes different maximum latency and minimum throughput constraints to the fronthaul/backhaul. For instance, in functional split 1, the decentralization reduces traffic overhead and the required backhaul capacity can be approximated by the aggregate users' traffic. Instead, for functional split 3, only RF function is located at the RRH, thus transmitting IQ samples through the fronthaul. In this case, samples are

usually encapsulated with CPRI [17] and the required fronthaul capacity depends on the bandwidth of the eNB, the number of antennas, etc. That is, fronthaul capacity requirement does not depend on the users' traffic for ϕ_3. This can be observed in Table I, where fronthaul/backhaul bandwidth and latency requirements for each split are shown as described in [16].

TABLE I
BANDWIDTH AND LATENCY REQUIREMENTS OF FUNCTIONAL SPLITS (λ_u^s IS THE DATA RATE OF USER u WITH SERVICE TYPE s)

Split	Bandwidth (bps)	Latency (ms)	Functions allocated at the RRH
ϕ_1	λ_u^s	30	f_0, f_1, f_2
ϕ_2	$1.02\lambda_u^s$	2	f_0, f_1
ϕ_3	$2.5 \cdot 10^9$	0.25	f_0

Given the described scenario, the network can create different slices on top of the physical RAN, each one with a specific functional split tailored to provide the required QoS, while guaranteeing the data rate and latency constraints of each functional split.

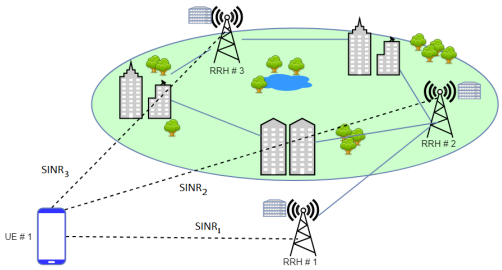

Fig. 2. Illustration of user associations with SINR

A. Optimal Slicing and Functional Split

The management and operation of the dynamic functional split in the RAN poses significant challenges with several trade-offs. Not only reduction on the transport networks load can be achieved by locating RAN functions at RRHs, but also offloading the RAN functions and pooling at the CU benefits from a computing costs reduction and offers centralized control that can improve the networks performance. While some splits have very tight delay constraints and create high fronthaul traffic, in other splits the CU might not have enough computation power to accommodate all RAN functions. In this context, a convenient network slicing algorithm provides the network with a higher degree of flexibility, enabling the adaptation of the functional split of each slice to traffic requirements and network limitations (RRH and/or CU computing capacity, transport network capacity and delay, etc). Thereby, we aim to find the optimal joint functional split and network slicing for future 5G networks.

B. User Associations

As mentioned earlier, each user's service type can determine a minimum degree of centralization (i.e., the set of possible functional splits). For instance, a high speed moving user requires a high degree of centralization to better coordinate the handover process or, in other words, it can only be supported by split ϕ_3. On the contrary, an Ultra-Reliable Low Latency (URLLC) user higher decentralization to guarantee low delay. Therefore, not all services can be supported with all functional splits.

Let us define the spectrum allocated to RRH r as W_r. The spectrum is divided in Physical Resources Blocks (PRB), each one with a bandwidth equal to w (in Hz). Based on the definitions, if the user u with type s and transmission rate λ_u^s (in bps), requires a number of PRBs per subframe equal to $\rho_u^{r,s}$ when served by RRH r, the user will be served by the RRH with the highest SINR (i.e., the smaller $\rho_u^{r,s}$) as long as there are enough resources (See Fig. 2). Thus, the user association must hold the condition $\sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}} x_u^{r,s}\rho_u^{r,s} \leq \rho^r$, where $x_u^{r,s} \in \{0,1\}$ is a binary variable to check whether user u with type s is connected to RRH r or not, and $\rho^r = W_r/w$ is the number of available PRBs in that RRH.

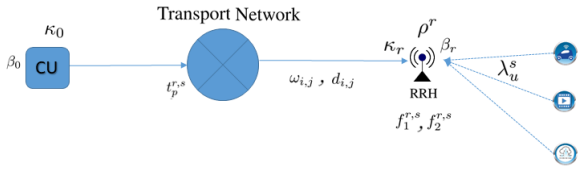

Fig. 3. System model scheme

C. Routing and Delay Constraints

Connected users inject traffic (bps) into the network, which is transmitted through the transport network over path $p \in P^{r,i}$ between CU and RRH r. This is shown in Fig. 3. Each RRH has connections to CU via sliced paths. These paths are used based on different demands of users traffic. Indeed, each RRH is responsible of transmitting flows of connected users to CU based on their service types. Hence, for each RRH there are different paths from CU. We define the traffic of service type s served by RRH r through path p as $t_p^{r,s}$, which must hold $\sum_{p\in P^r} t_p^{r,s} = T^{r,s}$, where $T^{r,s}$ is the total traffic of service type s served by RRH r (in bps). Based on Table I and the results of [16]:

$$T^{r,s} = f_1^{r,s}Q_1 - f_2^{r,s}Q_2 + (1 - f_1^{r,s})Q_3 \quad (1)$$

where $f_1^{r,s}$ and $f_2^{r,s}$ are equal to 1 when function f_1 and f_2, respectively, are located at the RRH r for service s and they are equal to 0 otherwise, and

$$\begin{cases} Q_1 = \sum_{u\in U} 1.02\lambda_u^s.x_u^{r,s} + 1.5 \\ Q_2 = \sum_{u\in U} 0.2\lambda_u^s.x_u^{r,s} + 1.5 \\ Q_3 = 2500 \end{cases}$$

Equation (1) provides the bandwidth requirements for a given functional split. Thus, when functional split ϕ_1 is used, $f_1^{r,s} = f_2^{r,s} = 1$, and so $T^{r,s} = Q_1 - Q_2$. Conversely, for functional split ϕ_3, $f_1^{r,s} = f_2^{r,s} = 0$ and consequently $T^{r,s} = Q_3$. Therefore, the set of paths from CU to RRH r must be able to forward a traffic of type s of at least $T^{r,s}$.

Routing also takes into account the delay of each link. In particular, slices with a functional split ϕ_n, with $n = \{1, 2, 3\}$, can only forward traffic over paths with an aggregate delay below the constraints shown in Table I.

D. Computational Cost Model

The deployment of network functions either at the RRH or at the CU incur a computational burden or cost. In the following the computational cost in the RRH and in the CU are stated. Thus, the computational cost required at the RRH r is given by:

$$C_r = \beta_r \sum_{u \in U} \sum_{s \in \mathcal{S}} \lambda_u^s . x_u^{r,s} \left(f_1^{r,s} . c_1^{r,s} + f_2^{r,s} . c_2^{r,s} \right) \quad (2)$$

where $c_1^{r,s}$ and $c_2^{r,s}$ are the computational cost of each function located at the RRH in CPU operations per bit per second, and λ_u^s is the user transmission rate. We also use (monetary units per cycle) which is the the average cost for serving each request; hence, for RRHs we set $\beta_r = 0.017$ and for CU $\beta_0 = 1$ as detailed in [16]. As for the CU,

$$C_0 = \beta_0 \sum_{r \in R} \sum_{s \in S} \sum_{u \in U} \lambda_u^s . x_u^{r,s} (c_1^{r,s}(1 - f_1^{r,s}) + c_2^{r,s}(1 - f_2^{r,s})) \quad (3)$$

IV. Problem Formulation

A. Optimization of Computational Cost

The main objective of this optimization model is to maximize the Centralization Degree (CD), which is defined as the inverse of the computational cost, $CD = (\sum_{r \in \mathcal{R}} C_r + C_0)^{-1}$. As discussed in [16], the computational costs of RRHs and CU can not be directly compared. In general, deploying additional capacity in a RRH is more costly than doing it in the CU. This effect is considered in (2) and (3) by assuming $\beta_r > \beta_0$. Therefore, the minimization of the total cost is equivalent to the maximization of the CD. This is the objective function of the optimization problem, as shown in (4) as long as it holds the capacity constraints in the integrated fronthual/backhaul. With regard to constraints, constraint (5) ensures the computation capacity needed to process $f_1^{r,s}$ and $f_2^{r,s}$ is less than the available capacity in RRH r (κ_r).

Equation (6) is used to bound the maximum computational capacity supported by the CU (κ_0). Equation (7) is explained in Section III-C. Constraint (8) states that the flow from each RRH r to CU is bounded by the capacity of the links of the paths, denoted as $\omega_{i,j}$ (in bps). The binary variable $y_{i,j}^p$ is used to check if the path p includes link $l_{i,j}$ or not. Equations (9), (10), (11) as described in Section III-C, are used for routing paths. Indeed each path is considered for the specific slices, and these paths have constraints in terms of delay requirements (ms). Note that M is a large positive number, to zeroize the paths with unacceptable delays for each configuration. Constraint (12) is explained in Section III-B, and constraint (13) is ensuring that each user is allowed to be connected to only one RRH. As discussed, not all paths are feasible solutions for a given functional split. In particular, only paths with an aggregate delay below the maximum delay supported by the functional split will be considered. Thus, based on Table I, we define for each RRH r the set of paths with a delay above 30 ms (the maximum delay allowed by split ϕ_1) as $P_r^{\phi_1}$. Analogously, $P_r^{\phi_2}$ as the set of paths with a delay larger than 2 ms and $P_r^{\phi_3}$ as the set of paths with a delay larger than 0.25 ms.

$$\max\ CD = \left(\sum_{r \in R} C_r + C_0 \right)^{-1} \quad (4)$$

subject to:

$$\sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}} x_u^{r,s} \lambda_u^s \left(f_1^{r,s} c_1^{r,s} + f_2^{r,s} c_2^{r,s} \right) \leq \kappa_r,\ \forall r \in \mathcal{R} \quad (5)$$

$$\sum_{r \in R} \sum_{s \in S} \sum_{u \in U} \sum_{n=1}^{2} \lambda_u^s . x_u^{r,s} c_n^{r,s} (1 - f_n^{r,s}) \leq \kappa_0 \quad (6)$$

$$\sum_{p \in P^r} t_p^{r,s} = T^{r,s},\ \forall r \in \mathcal{R}, \forall s \in \mathcal{S} \quad (7)$$

$$\sum_{r \in R} \sum_{s \in S} \sum_{p \in P^r} t_p^{r,s} . y_{i,j}^p \leq \omega_{i,j},\ \forall j \neq i \in \mathcal{Q} \quad (8)$$

$$\sum_{p \in P_r^{\phi_1}} t_p^{r,s} \leq M(2 - f_1^{r,s} - f_2^{r,s}),\ \forall r \in \mathcal{R}, \forall s \in \mathcal{S} \quad (9)$$

$$\sum_{p \in P_r^{\phi_2}} t_p^{r,s} \leq M(1 - f_1^{r,s} + f_2^{r,s}),\ \forall r \in \mathcal{R}, \forall s \in \mathcal{S} \quad (10)$$

$$\sum_{p \in P_r^{\phi_3}} t_p^{r,s} \leq M(f_1^{r,s} + f_2^{r,s}),\ \forall r \in \mathcal{R}, \forall s \in \mathcal{S} \quad (11)$$

$$\sum_{s \in S} \sum_{u \in U} x_u^{r,s} \rho_u^{r,s} \leq \rho^r,\ \forall r \in \mathcal{R} \quad (12)$$

$$\sum_{r \in R} x_u^{r,s} = 1,\ \forall u \in \mathcal{U}, \forall s \in \mathcal{S} \quad (13)$$

$$f_1^{r,s}, f_2^{r,s} \in \{0, 1\},\ \forall r \in \mathcal{R}, \forall s \in \mathcal{S} \quad (14)$$

B. Optimization of Throughput

As explained in Section III-C, the aggregation of the traffic in each RRH defines the throughput of the whole network. The problem defined in Section IV-A can be concerted from the maximization of the CD into the maximization of the throughput by changing the objective function (4) for constraint (7).

C. Linearization of the Problem

In general, the method to linearize non-linear problems, where constraints include multiplication of variables, depends on the kinds of variables involved in the constraints. For the case of binary variables, for instance, if we have multiplication of binary variables, say x, y, it can be solved by adding another binary variable, z, that holds: $z \geq 0, z \geq x + y - 1$ and $z \leq x, z \leq y$. Since our optimization problem is a Mixed integer Non Linear Programming (MINLP) problem, it can be concerted into a linear problem (i.e, MIP), by defining new

binary variables $z_1^{u,r,s}$ and $z_2^{u,r,s}$ to model linearly the multiplication of binary variables of $f_1^{r,s}.x_u^{r,s}$, $f_2^{r,s}.x_u^{r,s}$ stated in (1), (5), and (6). This formulation extends naturally to more indices and it must hold the following constraints:

$$x_u^{r,s}, f_1^{r,s}, f_2^{r,s}, z_1^{u,r,s}, z_2^{u,r,s} \in \{0,1\} \tag{15}$$

$$\begin{cases} z_1^{u,r,s} \leq x_u^{r,s} \\ z_1^{u,r,s} \leq f_1^{r,s} \\ z_1^{u,r,s} \geq x_u^{r,s} + f_1^{r,s} - 1 \end{cases} \tag{16}$$

$$\begin{cases} z_2^{u,r,s} \leq x_u^{r,s} \\ z_2^{u,r,s} \leq f_2^{r,s} \\ z_2^{u,r,s} \geq x_u^{r,s} + f_2^{r,s} - 1 \end{cases} \tag{17}$$

V. Performance Analysis

In this section, we explore joint optimization of functional split and slicing with maximization of centralization degree (i.e., minimization of computational cost). For the numerical analysis, Monte carlo simulations have been run for the averages of 100 times to get statistical significance. In order to highlight the impact of the joint slicing and functional split optimization, users of three different services have been considered. The first service has a data rate uniformly distributed in the range 10-1000 kbps. This service can only be supported with functional split ϕ_1. The second service has a data rate in the range 10-50 Mbps and can be supported by any of the three functional splits. Finally, the last service has a data rate within the range 70-200 Mbps and it can only be served with functional split ϕ_3. The scenario is composed of a single CU connected to a set of 10 RRHs. As for the computational capacity, we utilized the values used in [16] , with $\kappa_0 = 100$, $\kappa_r = 1$ CPU Reference Core per Gbps. Regarding the computational cost, $c_1^{r,s} = 3.25$ and $c_2^{r,s} = 0.75$ CPU reference core per Gbps. We obtain results after running 100 times for two metrics of minimizing the computation cost (i.e., maximizing the centralization degree) and maximizing the throughput. Results are compared with where no slicing is considered. Therefore, a single functional split is applied to each RRH. In the sequel, the optimization proposed in this paper is denoted by $SlicedRAN$ and the average cost and the average throughput for the first, second, and third splits, when no slicing is considered, are denoted by $split_1$, $split_2$, $split_3$, respectively. It is worth mentioning that in each scenario we set all RRHs to a single functional split (i.e., $split_1$, $split_2$, $split_3$) in order to compare them with $SlicedRAN$.

In Fig. 4, we first present the computational cost. The proposed algorithm presents a higher computational cost than $split_1$, $split_2$, $split_3$ (i.e., a lower centralization degree). This increase of the computational cost can be justified with the inherent flexibility of $SlicedRAN$. In particular, $SlicedRAN$ aims to centralize network functions as long as networks constraints (backhaul/fronthaul capacity and delay, and available computational capacity in RRHs and CU) permit it. This decision is made on a slice basis. On the contrary, each of $split_1$, $split_2$, $split_3$, is configured for RRHs. With this, users that can not be served with the functional split deployed in the RRH will have to be dropped. For $split_3$, it is expected to have the computation cost lower than $SlicedRAN$ as shown in Fig. 4. However, for $split_1$ and $split_2$, for all of the numerical evaluations, we found that both splits give close results and with big difference with $SlicedRAN$. The existing difference between the computation cost of these splits with $SlicedRAN$ is due to dropping more users for $split_1$ and $split_2$.

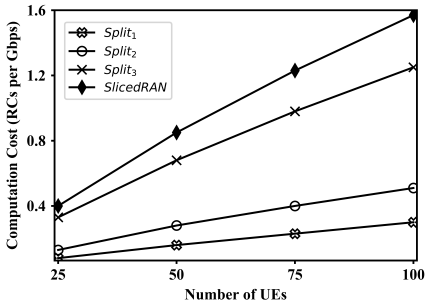

Fig. 4. Computation Cost analysis as a function of the number of users

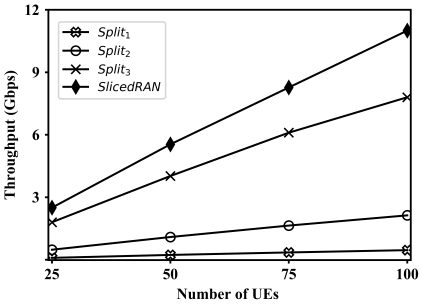

Fig. 5. Throughput analysis as a function of the number of users

The behavior explained above is translated into a decrease of the throughput in the $split_1$, $split_2$, $split_3$, as it can be observed in Fig. 5. It shows that with $SlicedRAN$ more throughput is achieved due to serving more users with different type of services. Finally, Fig. 6 shows the throughput/computation cost for 30, 60, 90,

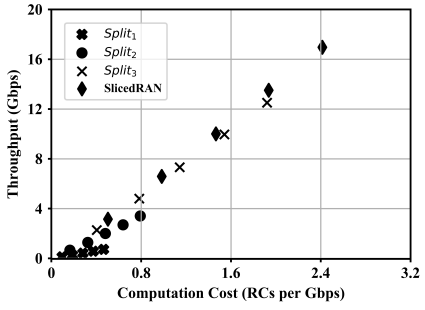

Fig. 6. Throughput vs Computation Cost trade-off for 30, 60, 90, 120 and 150 users respectively

120 and 150 users. In particular, for a given algorithm (i.e., $SlicedRAN$, $split_1$, $split_2$, and $split_3$), the number of users increases from left to right. This figure shows how the proposed algorithm can achieve different trade-offs between the computation cost (i.e., the inverse of centralization degree) and the throughput when the range of services covers completely different Quality of Service requirements. The trade-off between these two metrics can be clearly seen: as expected, as the number of users increase, the computation cost also increases. Likewise, throughput can only be improved at the expense of additional computational cost (decentralization). This improvement is directly related with the type of services of users which means the traffic type with higher amount, the more throughput can be achieved. As it is shown in this figure, the slope of $SlicedRAN$ is much higher than the slope of the $split_1$, $split_2$, $split_3$. This means that the computation cost increase required to increase the throughput is smaller for $SlicedRAN$. For example, for $split_2$, the increase in throughput is about 400 (Mbps) (i.e., 0.4 Gbps) whereas this increase is about 12 (Gbps) for $SlicedRAN$. Hence, $SlicedRAN$ outperforms the increase in throughput when compared with single functional split cases (i.e., $split_1$, $split_2$, $split_3$).

VI. Conclusions

RAN slicing along with functional split are two major concepts of future 5G networks. In this paper, we proposed $SlicedRAN$, which is a joint slicing and functional split optimization framework for 5G not yet fully investigated. We formulated this problem as a Mixed Integer Programming (MIP) to jointly optimize the centralization degree and throughput. Our results show that there is a trade-off between the centralization degree and the throughput. According to our results, computational cost in $SlicedRAN$ is higher than the one in $split_1$, $split_2$, $split_3$, which means that the centralization degree is lower than in $split_1$, $split_2$, $split_3$. However, this is compensated by the increase in the throughput. Furthermore, in $SlicedRAN$ the throughput increase is much higher and needs less increase in the computation cost. In the future work, we aim at extending this work by assessment of the system performance through various experimental results, and considering the other metrics in combination of the explored ones in the proposed approach.

Acknowledgment

This work has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No. 722788 (5G-Spotlight project) and SPOT5G (TEC2017-87456-P), and from AGAUR (2017 SGR 891), and is partially supported by the Spanish Ministry of Economy and the FEDER regional development fund under SINERGIA project (TEC2015-71303-R) and AGAUR (2017 SGR 60).

References

[1] CISCO, "Cisco Visual Networking Index: Forecast and Methodology, 20162021," Tech. Rep., 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

[2] NGMN Alliance, "NGMN 5G White Paper," Tech. Rep., 2015.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[4] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless network virtualization with sdn and c-ran for 5g networks: Requirements, opportunities, and challenges," *IEEE Access*, vol. 5, pp. 19 099–19 115, 2017.

[5] 3GPP, "TR 38.801. Technical Specification Group Radio Access Network; Study on new radio access technology: Radio access architecture and interfaces (Release 14)," 2017.

[6] 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*, p. 1, 2018.

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

[8] C. Y. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos, and C. Bonnet, "Impact of packetization and functional split on C-RAN fronthaul performance," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–7.

[9] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G Backhaul Challenges and Emerging Research Directions: A Survey," *IEEE Access*, vol. 4, pp. 1743–1766, 2016.

[10] Small Cell Forum, "Small cell virtualization functional splits and use cases," 2016.

[11] X. Wang, C. Cavdar, L. Wang, M. Tornatore, H. S. Chung, H. H. Lee, S. M. Park, and B. Mukherjee, "Virtualized Cloud Radio Access Network for 5G Transport," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 202–209, 2017.

[12] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, 2018.

[13] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks," in *Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies*, ser. CoNEXT '16. New York, NY, USA: ACM, 2016, pp. 427–441. [Online]. Available: http://doi.acm.org/10.1145/2999572.2999599

[14] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN Slicing for a Flexible and Cost-Effective Multi-Service Mobile Network Architecture," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '17. New York, NY, USA: ACM, 2017, pp. 127–140. [Online]. Available: http://doi.acm.org/10.1145/3117811.3117831

[15] A. Maeder, M. Lalam, A. D. Domenico, E. Pateromichelakis, D. Webben, J. Bartelt, R. Fritzsche, and P. Rost, "Towards a flexible functional split for cloud-RAN networks," in *2014 European Conference on Networks and Communications (EuCNC)*, 2014, pp. 1–5.

[16] A. Garcia-Saavedra, X. Costa-Perez, D. Leith, and G. Iosifidis, "FluidRAN: Optimized vRAN/MEC Orchestration," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1–9.

[17] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 152–159, 2016.