Joint and dynamic optimization of functional split selection and slice configuration in vRAN

Neco Villegas, Sofía Pérez, Luis Diez, Ramón Agüero
Communication Engineering Department
Universidad de Cantabria
Santander, Spain
villegasn@unican.es, sofia.perezr@alumnos.unican.es, {ldiez, ramon}@tlmat.unican.es

***Abstract*—We jointly consider network slicing and functional split for 5G/6G Radio Access Network (RAN). Exploiting virtualization techniques, virtual Radio Access Network (vRAN) is an architectural shift that moves some of the functions that were traditionally performed by the base station to centralized nodes, deployed at cloud-based resources. By appropriately selecting the location of these nodes, they can effectively coordinate a number of radio access elements, thus bringing potential gains, for instance, in terms of interference reduction. One pivotal decision for vRAN operation is how to split the functions between these central elements and the corresponding distributed units. Deploying more functions at the cloud could potentially bring more benefits, due to the tighter cooperation between access elements. On the other hand, higher centralization poses more requirements, in terms of computational resources at the nodes where these functions are deployed. Along with the virtualization, RAN slicing enables the deployment of independent service verticals over the same physical infrastructure, and this could greatly benefit from the coordination achieved by exploiting the vRAN features. In this paper we propose a vRAN model to jointly consider functional split and RAN slicing, and we establish the corresponding configuration, exploiting Lyapunov's Theory to tackle the underlying problem. The results show that, in highly heterogeneous networks, the dynamic configuration of the functional split can reach the same performance of fully centralized networks, with substantially fewer computation resources.**

***Index Terms*—vRAN, 5G, functional split, RAN slice, optimization, Lyapunov**

I. Introduction

The deployment and roll-out of 5G technology enable new heterogeneous scenarios and novel services, with more ambitious requirements in terms of transmission speed, bandwidth, latency, and security, among others [1]. Some of these services, such as real-time HD video, demand high data speeds and extremely low delay, while others require highly reliable connectivity with lower error rates, such as remote control of sensitive equipment.

One fundamental aspect of 5G and 6G networks is the evolution of the RAN architecture. This shift, enabled by Software Defined Networking (SDN) and Network Function Virtualization (NFV) techniques, redefines the traditional functions associated with the base station. Under this new paradigm, some of these functions are relocated to central nodes, typically hosted at cloud servers, resulting in what is known as Cloud RAN (C-RAN) [2]. C-RAN is based on the separation of Baseband Unit (BBU) with respect to their corresponding Remote Radio Head (RRH), grouping them in a so-called BBU pool. Each RRH element is thus connected to its corresponding BBU pool via low-latency optical fronthaul links. Then, a backhaul link connects each BBU pool to the core network. The C-RAN architecture increases network scalability, improves spectral efficiency, reduces energy consumption, simplifies network management and maintenance, and facilitates load balancing [3]. Additionally, it achieves cost reductions compared to conventional mobile networks [4], and enables shared processing [5].

Despite the many advantages of the C-RAN architecture, it requires a low-latency, high-capacity fronthaul network, which may require large investments from operators [6]. To solve this problem, the concept of functional split, realized under the concept of vRAN, brings the possibility that not all processing functions are centralized, but only a set of them are selected to be implemented at central units, while the remaining are deployed at distributed units [7]. The introduction of the functional split concept brings an architectural transformation, in which the BBU is divided into Centralized Unit (CU) and Distributed Unit (DU) components, while the RRH is rebranded as Radio Unit (RU) [5], [8]. The virtualization of the entities that compose the base station, along with the isolation of radio resources, allow the definition of independent RAN slices, so that multiple virtual base stations can be deployed using the same physical access element.

The focus of this work lies in the joint management of radio resources and the functional division within vRAN architectures. Our main contributions are briefly summarized below:

1) Proposal of a network model that takes into account the interaction of the RAN slice definition and configuration of functional split with time performance constraints, considering temporal evolution.
2) Proposal of an adaptive solution to optimize the network operation under random uncontrolled conditions.
3) Validation of the algorithm operation in an heterogeneous canonical scenario and performance evaluation of the joint split and slice configuration compared to fixed functional split setups.

The rest of the paper is structured as follows: Section II discusses some related work, highlighting how this research differs from existing papers. In Section III we introduce the system model that addresses the joint resource allocation and functional split decision-making, and we propose an optimization problem to address it. Then, Section IV discusses the performance of the proposed scheme, after a thorough experiment campaign. The paper concludes in Section V, where we provide an outlook of our future work.

II. Related work

As mentioned earlier, 5G networks need to integrate a multitude of services with diverse performance demands, all of them within a unified physical network infrastructure. Furthermore, the ultimate goal would be to furnish each service with a tailored logical network [9]. Hence, there is a growing need to exploit the potential of dynamically adjusting the functional split based on specific network conditions and traffic loads. This flexible functional split, as it is often termed, has a direct impact on network operation and management. Nonetheless, existing works exploring the potential benefits of this approach and proposing performance assessment models are still relatively sparse.

Some works have studied the selection of functional splits and the optimization of the fronthaul network. In [10], the authors address the challenge of minimizing power consumption in vRAN, with a specific focus on two functional splits, while considering fronthaul configuration functionalities. Other strategies for selecting functional splits have been put forward to enhance various performance metrics, considering diverse scenario characteristics. For example, [11] considers the placement of virtual entities and focuses on maximizing data rates in the fronthaul by posing an Integer Linear Programming (ILP) problem that considers routing, bandwidth assignment, and latency. Similarly, Abdulrahman *et al.* [12] take into consideration the joint minimization of system power and transmission capacity consumption while ensuring end-to-end latency. Liumeng and Sheng propose in [13] a Markov Decision Process (MDP) based solution to maximize the throughput in the fronthaul under average rate constraints, considering radio units powered with renewable energy. Differently, the authors in [14] introduce a user-centric functional split orchestration approach, taking into account power, computational resource, and bandwidth usage. As can be observed, the scope of the aforementioned works combine the split selection with the optimization of the fronthaul network, while we focus on the functional split along with the management of radio resources.

Closer to our work, Alba and Kellerer propose in [15] a dynamic split selection algorithm to optimize the throughput perceived by users. However, they do not consider the resource allocation along with the split selection, neither they take into account performance guarantees. In [16] Wu *et al.* adopt a slice-centric approach to design a flexible RAN architecture that considers both processing and transmission resource saving, while ensuring compliance with the requirements of the slices. While this work shares some similarities with ours, it is not designed for time-varying scenarios, and service performance constraints are not included.

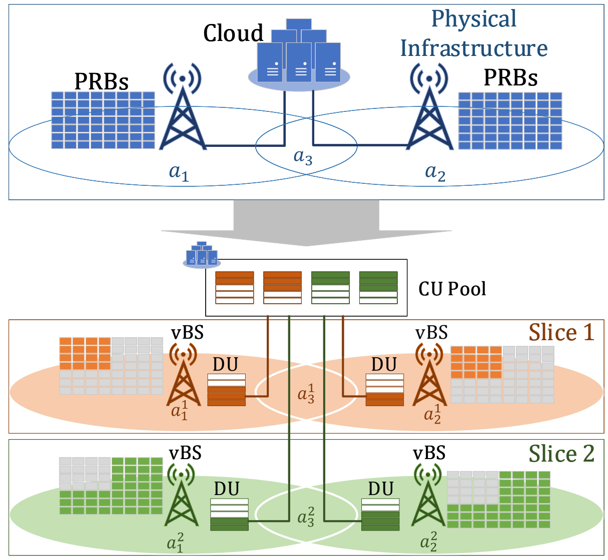


Fig. 1: Scenario example

It is also worth mentioning the work of Vajd *et al.* [17], where the authors focus on the optimization of the fronthaul network by the selection of functional split and configuration of optical transponders. Although it does not focus on radio resource management, the authors consider time-average constraints and they apply Lyapunov optimization to tackle the resulting optimization problem.

All in all, we believe that our work complements the existing literature, by jointly considering the functional split selection and the allocation of radio resources to RAN slices. To our best knowledge, this is the first time these two processes have been jointly tackled.

III. System model and problem formulation

In this section we present the system model for the joint resource allocation and functional-split decision under random conditions, and we pose an optimization problem to tackle it. Then, we introduce a simplified version of such problem for those cases in which the functional-split is fixed. This allows us comparing the benefits brought by dynamic functional-split under similar conditions.

We consider a scenario where services are provided by means of network slices. For simplicity, we assume that each slice bears a single service, but the model could be easily adapted for multi-service slices. We focus on the RAN slicing, where physical radio resources are to be distributed among the slices. It leads to instantiating virtual base stations, accounting for the different slices, on top of the physical access elements. In addition, different centralization levels can be configured at virtual base stations, according to the functional splits. In turn, the configuration levels enable coordination techniques, which would for instance improve the Signal to Interference plus Noise Ratio (SINR) perceived by users [6].

Figure 1 depicts a simple scenario comprising two physical base stations and two slices. As can be observed, the radio resources of the physical base stations, Physical Resource Block (PRB), are distributed in an orthogonal way between the virtual ones that are instantiated for each slice. Then, the centralization level is defined for each virtual base station. In this way, by exploiting the tighter coordination between virtual base stations of the same slice, the link quality can be improved at the overlapping areas.

Having that in mind, we aim to adapt to changes on wireless conditions by dynamically tuning both the centralization level and the distribution of PRBs among slices. In this sense, since the decisions (physical resources and split) affect all users in a certain area, we average the wireless conditions over all users in such zone. It is assumed that Medium Access Control (MAC) scheduling and Adaptive Modulation and Coding (AMC) would define the SINR perceived for each single user. Although the underlying problem formulation that is presented below is generic, the decision time scale is considered long enough, so that the time elapsed for the network reconfiguration is negligible.

A. System model

Let $\mathcal{B}$ be the set of physical base stations and $\mathcal{S}$ the set of slices. Each base station $b \in \mathcal{B}$ has a number of physical wireless resources η_b. In general, it is assumed that slices provide elastic services with minimum requirements. This is, the service corresponding to the slice $s \in \mathcal{S}$ has a minimum throughput requirement d_s^{min}, which needs to be satisfied at all times, and an target throughput d_s that should be provided in average over time. It may happen that d_s is not fulfilled at particular time instants, but it should nonetheless be satisfied in average.

In order to satisfy the demand of slices, we instantiate virtual base stations within the physical ones. Time is slotted, and t indicates a given slot time. In each time slot we decide the amount of resources for each virtual base station, as well as the centralization level. We have to take into account that the capacity demand needs to be satisfied in the whole area covered by the physical base station. In this sense, overlapping zones will be treated as separate areas. Let $\mathcal{A}$ be the set of areas, and $\mathcal{A}_s$ the subset of areas where slice s is deployed.

For each slice s and area $a \in \mathcal{A}_s$, we use the Shannon's formula to compute the average rate at a time instant t, $\rho_{s,a}(t)$, as follows:

$$\rho_{s,a}(t) = W_{s,a}(t) \cdot PRW \cdot \log_2\left(1 + \frac{P_{s,a}(t)}{\sigma^2 + I_{s,a}(t)}\right) \quad (1)$$

where $W_{s,a}(t)$ is the amount of PRBs used by the slice s at the area a, while $P_{s,a}(t)$ and $I_{s,a}(t)$ hold for the average signal and interference experienced by the devices of the slice in such area, respectively. The size, in frequency, of each PRB is denoted as PRW, and it is assumed to be a configuration parameter. In our future work we will consider the impact of the different numerologies defined for 5G on the system model. Finally, σ^2 represents the noise factor.

Let $\mathcal{F}$ be the set of possible functional splits that are available in the virtual base stations, and $f_s^b(t)$ the functional split selected for the virtual base station used for slice s in physical base station b. In addition, $\alpha_s^b(t)$ holds for the amount of physical resources allocated from base station $b \in \mathcal{B}$ to slice $s \in \mathcal{S}$.

It follows that the actual values of (1) may depend on both random and decision variables. If we consider a generic random variable $\omega(t)$, we can express the variables in (1) as functions [1]: $W_{s,a}(t) = \widehat{W}(\alpha_s^b(t))$; $I_{s,a}(t) = \widehat{I}(\omega(t), f_s^b(t))$; and $P_{s,a}(t) = \widehat{P}(\omega(t))$.

We assume that two base stations covering the same area coordinate between them to improve the spectral efficiency. In this sense, the coordination requires overlapping of the coverage areas, as assumed in [18]. Thus, the following constraint holds:

$$W_{s,a}(t) = \widehat{W}(\alpha_s^b(t)) = min(\{\alpha_s^b(t) \mid b \in \mathcal{B}_a\}) \quad (2)$$

where $\mathcal{B}_a$ is the subset of physical base stations covering the area $a \in \mathcal{A}_s$.

Finally, the functional splits require different computation resources in the nodes where DU and CU pools are deployed. The former is assumed to have enough capacity for all considered splits. On the other hand, the computation resources for the CU pool are shared between various virtual base stations, thus limiting the decision space. Let C_{CU} denote the shared computation capacity of the node hosting the CU pool, and $c_s^b(t)$ the computation resources required by the virtual base station instantiated for slice s in physical base station b, according to the selected split f_s^b, at time slot t.

B. Problem formulation

Having in mind the above definitions, we aim to jointly allocate resources to the different slices and select the split level so that the average amount of resources is minimized over time. The time average expectation of a variable $y(t)$ is defined as follows [19]: $\overline{y} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\{y(t)\}$. Altogether, we pose a time-average optimization problem:

$$\min_{f(t),\alpha(t)} \quad \sum_{s\in\mathcal{S},b\in\mathcal{B}} \overline{\alpha_s^b} \quad (3)$$

$$s.t.$$

$$\overline{\rho_{s,a}} \geq d_s \qquad \forall a \in \mathcal{A}_s, \forall s \in \mathcal{S} \quad (4)$$

$$\sum_{s\in\mathcal{S}} \alpha_s^b(t) \leq \eta_b \quad \forall b \in \mathcal{B}, \forall t \quad (5)$$

$$\rho_{s,a}(t) \geq d_s^{\min} \qquad \forall a \in \mathcal{A}_s, \forall s \in \mathcal{S}, \forall t \quad (6)$$

$$\sum_{s\in\mathcal{S},b\in\mathcal{B}} c_s^b(t) \leq C_{CU} \qquad \forall t \quad (7)$$

where (4) guarantees that the throughput achieved at each area (for each slice) reaches the target one, in average. Then, (5)

[1]Symbol $\widehat{f}$ denotes a function that yields the value of the variable f.

ensures that the amount of granted resources does not exceed the resources of the physical base stations. (6) imposes that the minimum throughput is satisfied for all the slices in every time slot. It is worth highlighting that the variable $\overline{\rho_{s,a}}$ depends on $W_{s,a}(t)$, which in turn is a function of both $f_s^b(t)$ and $\alpha_s^b(t)$. Finally, (7) ensures that the aggregated computation capacity required by the CU pool does not exceed its available capacity. For simplicity, we define the action space in each slot t as $\Lambda(t)$, and we replace the constraints that need to be satisfied every slot, (5), (6) and (7), by a single feasibility constraint: $\{\alpha_s^b(t), f_s^b(t)\}_{s\in\mathcal{S},b\in\mathcal{B}} \in \Lambda(t)\ \forall t$.

The original problem can be formulated in a more standard form by defining $E_{s,a}(t) = d_s - \rho_{s,a}(t)$, so that inequality (4) becomes $\overline{E_{s,a}} \leq 0 \quad \forall a \in \mathcal{A}_s, \forall s \in \mathcal{S}$.

This type of stochastic optimization problem can be tackled using the Lyapunov's theory. It can be solved using the framework defined in [19] to transform the original formulation into one based on stability conditions by changing time-average equality and inequality constraints to virtual queues. In our case, we define a virtual queue as follows: $Z_{s,a}(t+1) = \max\{Z_{s,a}(t) + E_{s,a}(t), 0\}$. The original time-average minimization problem can be solved by applying the drift-plus-penalty algorithm, which, in every slot t, observes both the queue state and the corresponding random events and takes the decision $(f(t), \alpha(t))$ that minimizes the following problem:

$$\min_{f(t),\alpha(t)} \quad V \cdot \sum_{s\in\mathcal{S},b\in\mathcal{B}} \alpha_s^b(t) + \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} Z_{s,a}(t)\cdot E_{s,a}(t) \tag{8}$$

$$s.t.$$

$$\{\alpha_s^b(t), f_s^b(t)\}_{s\in\mathcal{S},b\in\mathcal{B}} \in \Lambda(t)\ \forall t \tag{9}$$

C. Resource allocation only

In those cases where there is not functional split, or it is fixed, the problem can be greatly simplified. The spectral efficiency in each area would not then depend on the decision, but it could be considered constant at each time slot: $E_{\mathit{eff}_{s,a}} = \log_2\left(1 + \frac{P_{s,a}(t)}{\sigma^2 + I_{s,a}(t)}\right)$. Thus, we can simplify the objective function as follows:

$$\begin{aligned} & V \cdot \sum_{s\in\mathcal{S},b\in\mathcal{B}} \alpha_s^b(t) + \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} Z_{s,a}(t)\cdot E_{s,a}(t) \\ = & \sum_{s\in\mathcal{S}} \Bigg[V \cdot \sum_{b\in\mathcal{B}} \alpha_s^b(t) + \sum_{a\in\mathcal{A}_s} Z_{s,a}(t)\cdot d_s - \\ & \sum_{a\in\mathcal{A}_s} Z_{s,a}(t) W_{s,a}(t) E_{\mathit{eff}_{s,a}} \Bigg] \end{aligned} \tag{10}$$

In (10) we can remove the product of the virtual queue by the throughput requirement ($Z_{s,a}(t)\cdot d_s$), since it does not depend on the decision, and the objective function can be thus reformulated as: $\sum_{s\in\mathcal{S}}\left(V\cdot\sum_{b\in\mathcal{B}}\alpha_s^b(t) - \sum_{a\in\mathcal{A}_s} Z_{s,a}(t)W_{s,a}(t)E_{\mathit{eff}_{s,a}}\right)$. In addition, since we are minimizing the amount of allocated resources, any resource not contributing to the throughput will

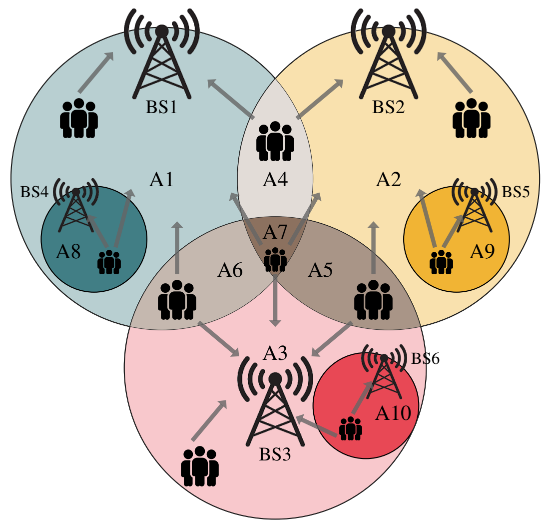


Fig. 2: Analysis scenario, heterogeneous network

not be granted. Thus, $W_{s,a}(t) = \alpha_s^b(t)\ \forall b \in \mathcal{B}_a$. Besides, at each area $a \in \mathcal{A}$, we only need to consider the resources granted by the base stations covering it, $\mathcal{B}_a$. Altogether, the objective function boils down to:

$$\begin{aligned} & \sum_{s\in\mathcal{S}}\left(V\cdot\sum_{a\in\mathcal{A}_s}\sum_{b\in\mathcal{B}_a} W_{s,a}(t) - \sum_{a\in\mathcal{A}_s} Z_{s,a}(t)E_{\mathit{eff}_{s,a}}W_{s,a}(t)\right) \\ = & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s} W_{s,a(t)}\left(V|\mathcal{B}_a| - Z_{s,a}(t)E_{\mathit{eff}_{s,a}}\right) \\ = & \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} K_{s,a}\cdot W_{s,a}(t) \end{aligned} \tag{11}$$

where $K_{s,a}(t) = V|\mathcal{B}_a| - Z_{s,a}(t)E_{\mathit{eff}_{s,a}}$. By replacing (8) by (11), the resulting problem is converted into a ILP one which can be solved with existing tools.

IV. Performance evaluation

This section discusses the results obtained by using the aforementioned system model for a heterogeneous scenario, where we apply the proposed algorithm. The main objective of the evaluation is to analyze the benefits of dynamically selecting the functional split, along with the resource allocation to RAN slices. Figure 2 depicts the evaluation setup, which comprises a layer of macro base stations and small cells, deployed within the coverage area of the macro cells. In this scenario, 2 slices are instantiated, the first one in all areas and the second one only in the areas shared by macro and small cells (areas numbered 8, 9, and 10).

The scenario configuration is described in Table I. All base stations have 100 PRB, and in each area the power and interference values are randomly selected from the corresponding ranges in each time slot. In general, we limit the spectral efficiency to 5 b/s/Hz (typical 4G maximum value), so that above a certain SINR, the spectral efficiency would not improve. All virtual base station can be configured with one of the following functional splits: C-RAN, MAC/PHY,

TABLE I: Configuration of the evaluation setup. Interference reduction factors and split computation requirements are based on [6] and [20], respectively

Access network	
# PRBs	100
PRW	180 KHz
Interference range	$[11, 15]$ mW
Power range	$[1, 81]$ mW
V parameter	$5e3$
CU pool capacity	$[15, 65]$ GOPS
Splits: C-RAN, MAC/PHY, RLC/MAC, PDCP/RLC	
Interference reduction factor	$\{0.01, 0.2, 0.6, 1\}$
CU computation requirement	$\{20.9, 10.7, 6.7, 4.7\}$ GOPS
Slices	
#Slices	2
Minimum throughput	$\{1, 1\}$ Mbps
Objective throughput.	$\{10, 10\}$ Mbps

RLC/MAC, PDCP/RLC. For each of them, we consider an interference reduction factor, obtained from [6], which is used to modulate (multiplying) the actual interference value. In this sense, if the PDCP/RLC split is selected, the interference is kept at the same value, while C-RAN configuration reduces the interference by a factor of $\times 100$. Altogether, if we consider the best scenario, where all the resources are granted, the highest centralization is configured and each PRB occupies 180 KHz, one area could reach more than 170 Mbps: $100 \cdot 180 \cdot \log_2 \left(1 + \frac{81}{11 \cdot 0.01}\right)$.

The evaluation tool consists of a proprietary system-level simulation, implemented in C++ that deploys the corresponding scenario and processes it to establish the corresponding optimization problem. The simulator also keeps track of the past decisions, to appropriately update the virtual queues defined in Section III-B. In particular, we use the GNU Linear Programming Kit (GLPK) framework [2] to solve problem instances that only solve resource allocation (see Section III-C). On the other hand, joint split selection and resource allocation problems are solved by iteratively instantiating resource-allocation problems, and calling GLPK. More efficient algorithms to solve the joint problem will be tackled in our future work, being our main objective herewith the evaluation of the potential gains brought by a joint decision. All the results shown in the following sections are obtained from experiments lasting $1e4$ slots.

A. Resource utilization

Figure 3 illustrates the split selection probability at those areas where slice 2 is deployed, while varying the computation capacity in the cloud node hosting the CUs. For each computational capacity value, we show the probability of selecting each split (see Table I) at each area, where the leftmost bar corresponds to PDCP/RLC (lowest centralization) and the rightmost to C-RAN (highest centralization). As can be observed, as the computational capacity increases, the

[2]https://www.gnu.org/software/glpk/

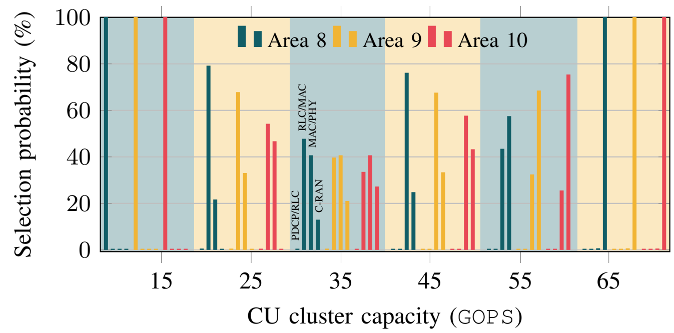


Fig. 3: Split selection probability

proposed algorithm tends to select higher centralization levels with higher probability. In this sense, when the computation capacity is 15 GOPS, the PDCP/RLC split is always selected at the three areas, while the highest centralization is chosen with the maximum computational capacity (65 GOPS). While these results could have been anticipated, they validate the operation of the decision algorithm under different circumstances.

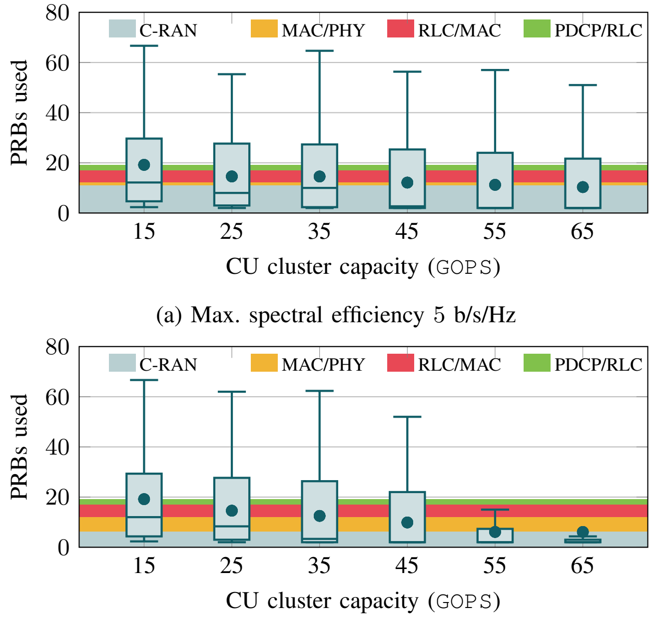
B. Split probability selection

We now study the impact of the joint decision over the average amount of PRBs used in the small cells ($W_{s,a}(t)$ in Section III). Figure 4 uses a boxplot to show the amount of resources used by the second slice, using the joint decision for different values of the computational capacity of the CU cloud. In addition, we indicate the resources required by the static split selection with colored bands. In Figure 4a we can observe that larger computation capacities yield a slight reduction of the amount of resources used by slice 2. However, the amount of computation resources required by the joint decision is lower than that needed by the static configuration. For instance, the joint solution achieves similar average resource usage (this is represented by a circle in the figure) to that brought by static C-RAN configuration when the CU cloud has 55 Giga Operations per Second (GOPS). On the other hand, the static configuration would require more than 60 GOPS (21×3). It is worth noting that there is a rather relevant difference between the results obtained when using C-RAN and MAC/PHY configurations, due to the saturation effect of the spectral efficiency gains: larger SINR values due to centralization would not yield additional efficiency gains.

In Figure 4b we increase the maximum spectral efficiency to 30 b/s/Hz, which captures a typical 5G maximum value. As can be seen, the benefits of the C-RAN configuration become much more clear in this case, leading to a notable reduction resource utilization. Again, we can observe that the joint solution achieves the same performance than the static C-RAN, but it requires fewer computational resources.

V. Conclusion

In this work, we have proposed a system model and an algorithm to tackle the reduction of radio resources allocated by base stations exploiting a dynamic selection of functional



(a) Max. spectral efficiency 5 b/s/Hz

(b) Max. spectral efficiency 30 b/s/Hz

Fig. 4: Distribution of PRBs used against computation capacity of CU cloud

split. We have assessed the potential benefits of a dynamic centralization level assignment, over a highly heterogeneous scenario. We exploit a proprietary tool, which integrates the GLPK framework to solve the corresponding optimization problems. After extensive simulations, the observed results evince that the proposed dynamic solution yields better performance, since it enables a significant reduction of the average number of resources to be allocated, compared to a more traditional configuration, where split levels are statically established.

This work has focused on the four most significant split configurations, characterizing their behavior in terms of resource allocation and average throughput. In our future work, we will broaden the analysis, to encompass the impact of considering all the centralization levels defined by 3GPP. Furthermore, we will also extend the analysis by considering more generic scenarios, as well as service requirements coming from particular use cases. As for the proposed model and algorithm, we will address the development of more efficient approaches to solve the resulting optimization problem. Finally, we will consider extensions, for instance including numerologies defined for 5G, which would modify the size of resource blocks.

Acknowledgment

This work has been supported by the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union – NextGenerationEU with the project SITED: Semantically-enabled Interoperable Trustworthy Enriched Data-spaces (PID2021-125725OB-I00), OPTIMAIX_OaaS (TSI-063000-2021-34) and 6GBLUR/joint (TSI-063000-2021-57). We have also received funds from the regional Cantabria Government, Fondo Europeo de Desarrollo Regional (FEDER), with the TCNIC program (2023/TCN/002).

References

[1] A. Gupta and R. K. Jha, “A survey of 5G network: Architecture and emerging technologies,” *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[2] M. Labana and W. Hamouda, “Advances in CRAN Performance Optimization,” *IEEE Network*, vol. 35, no. 3, pp. 140–146, 2021.

[3] H. Mrabet, “Performance Investigation of New Waveforms in CRAN Architecture for 5G Communication Systems,” in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, 2020, pp. 1–5.

[4] 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)*, 2019, pp. 410–416.

[5] Q. K. Ud Din Arshad, A. U. Kashif, and I. M. Quershi, “A review on the evolution of cellular technologies,” in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2019, pp. 989–993.

[6] 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.

[7] O. Arouk, T. Turletti, N. Nikaein, and K. Obraczka, “Cost Optimization of Cloud-RAN Planning and Provisioning for 5G Networks,” in *2018 IEEE International Conference on Communications (ICC)*, 05 2018, pp. 1–6.

[8] R. Gavrić, D. Ilišević, N. B. Ćurguz, and d. Budimir, “Comparison of basic characteristics of 4G/LTE and 5G NR technology,” in *2019 27th Telecommunications Forum (TELFOR)*, 2019, pp. 1–4.

[9] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain, “Network Slicing for 5G: Challenges and Opportunities,” *IEEE Internet Computing*, vol. 21, no. 5, pp. 20–27, 2017.

[10] Y. Zhou, J. Li, Y. Shi, and V. W. S. Wong, “Flexible Functional Split Design for Downlink C-RAN With Capacity-Constrained Fronthaul,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6050–6063, 2019.

[11] 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.

[12] A. Alabbasi, M. Berg, and C. Cavdar, “Delay Constrained Hybrid CRAN: A Functional Split Optimization Framework,” in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–7.

[13] L. Wang and S. Zhou, “Flexible Functional Split and Power Control for Energy Harvesting Cloud Radio Access Networks,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1535–1548, 2020.

[14] S. Matoussi, I. Fajjari, S. Costanzo, N. Aitsaadi, and R. Langar, “5G RAN: Functional Split Orchestration Optimization,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 7, pp. 1448–1463, 2020.

[15] A. Martinez Alba and W. Kellerer, “A Dynamic Functional Split in 5G Radio Access Networks,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[16] Y. Wu, L. Liang, Y. Jia, Z. Chen, and W. Wen, “Slicing Enabled Flexible Functional Split and Resource Provisioning in 5G-and-Beyond RAN,” in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.

[17] F. S. Vajd, M. Hadi, C. Bhar, M. R. Pakravan, and E. Agrell, “Dynamic joint functional split and resource allocation optimization in elastic optical fronthaul,” *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 4505–4515, 2022.

[18] L.-H. Shen, Y.-T. Huang, and K.-T. Feng, “CoMP-Enhanced Flexible Functional Split for Mixed Services in Beyond 5G Wireless Networks,” *IEEE Transactions on Communications*, vol. 71, no. 7, pp. 4133–4150, 2023.

[19] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Springer, 2010. [Online]. Available: http://dx.doi.org/10.2200/S00271ED1V01Y201006CNT007

[20] B. Debaillie, C. Desset, and F. Louagie, “A flexible and future-proof power model for cellular base stations,” in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, 2015, pp. 1–7.