User-Centric Slice Allocation Scheme in 5G Networks and Beyond

Salma Matoussi, Ilhem Fajjari, Nadjib Aitsaadi, *Member, IEEE*, and Rami Langar, *Member, IEEE*

***Abstract*—Network slicing is a key enabler in the Next Generation 5G Radio Access Network (RAN) to build the RAN-as-a-Service concept. Cloud-RAN, Network Function Virtualization, Software Defined Network and RAN functional splits are the main pillars expected to be integrated to provide the required flexibility. One of the major concerns is to efficiently allocate RAN resources for slices, while supporting multiple use-cases with heterogeneous Quality-of-Service (QoS) requirements. Current related work is adopting radio resource allocation scheme by considering a cell-centric deployment approach for slice embedding. However, to achieve greater flexibility and fine-grained tunable resource utilization, we believe that the deployment scheme should be integrated in the slice design. In this paper, we go a step further and propose a RAN slicing approach with customized deployment scheme on user basis. As the corresponding optimization problem is NP-Hard, we propose a low-cost and efficient heuristic algorithm for RAN Slice allocation based on the Particle Swarm Optimization approach. Our proposal jointly harnesses radio, processing and link resources at user level tailored to the QoS requirements, while customizing efficiently the underlying physical RAN resource usage.**

***Index Terms*—NG-RAN, slicing, 3GPP RAN functional split, radio resource allocation, particle swarm optimization.**

I. Introduction

NEXT Generation 5G Radio Access Network (NG-RAN) is expected to integrate major changes in the cellular communications beyond the new radio and wider spectrum. The objective is to build a flexible and cost efficient mobile network to convey services for enhanced Mobile BroadBand (eMBB), massive Machine Type Communications (mMTC) and ultra-Reliable Low Latency Communications (uRLLC) [1] use cases. Thereby, the network slicing concept [2] is proposed with the objective of enabling the network to deliver allocated resources (so called slice) as per service requirement. Within this perspective, multiple slices can be created on the same RAN infrastructure to convey services with heterogeneous requirements in terms of latency, reliability and throughput.

5G stakeholders make strong effort to redesign the RAN in aim of building service-oriented architecture [3]. 3GPP introduces in Release 15 [4], the NG-RAN architecture for a disaggregated RAN deployment, called RAN functional splits. Wherein, the new 5G eNodeB (so called gNB) is decoupled into i) Radio Unit (RU), ii) Distributed Unit (DU) and iii) Central Unit (CU). In doing so, the traditional BaseBand Unit (BBU) is henceforth disaggregated and deployed between DU and CU locations connected through a transport network (so called Midhaul). It is also foreseen that CUs of multiple *gNB*s will be centralized in a mobile edge cloud (Cloud-RAN) to achieve higher radio performance thanks to cell cooperation, while reducing CAPital EXpenditure (CAPEX) and OPerational EXpenditure (OPEX) budgets [5]. In addition, the Network Function Virtualisation (NFV) technique is adopted to separate the hardware from network software, while using the Software-Defined RAN (SD-RAN) for better programmability and customized control capabilities [2].

Cloud-RAN, NFV, Software Defined Network (SDN) and RAN functional splits are key enablers for slice management. Unlike slicing in the 5G core network, the RAN slicing still remains an open issue. Indeed, the latter should consider the radio resource nature and the RAN real time constraints. 3GPP has specified in [6] the RAN sharing concept, where multiple virtual operators can share the RAN infrastructure either with dedicated frequency band or with full spectrum sharing. These approaches lead to several works seeking how to efficiently manage the radio resource allocation with isolation and sharing capabilities [7], [8], [9], [10], [11]. However, the original scope of the network slicing is to consider all types of RAN resources. Current related work adopts a radio slice allocation scheme with a cell-centric deployment approach for slice embedding [12], [13], [14], [15], [16], [17], [18]. Leveraging RAN disaggregation through RAN functional splits, we believe that these deployment options should be integrated in the slice design. Within this context, a new challenge raises addressing the ability to fulfill the vastly use-case's Quality-of-Service (QoS) requirements, while considering heterogeneous resource types and multiple RAN functional split options in the physical infrastructure.

In this work, we put forward a user-centric RAN slicing scheme, providing suitable proportions of radio, link and computational resources for each User Equipment (UE). Our

Manuscript received 14 July 2022; revised 19 November 2022 and 23 March 2023; accepted 21 May 2023. Date of publication 8 June 2023; date of current version 12 December 2023. This work is partially supported by the ANR 5G-INSIGHT project (Grant no. ANR-20-CE25-0015), and by the Natural Sciences and Engineering Council of Canada, NSERC (Grant no. RGPIN-2022-03926). A preliminary version of this paper appeared in the 2020 IEEE Wireless Communications and Networking Conference (WCNC 2020) [36] [DOI: 10.1109/WCNC45663.2020.9120828]. The associate editor coordinating the review of this article and approving it for publication was R. Riggio. *(Corresponding author: Salma Matoussi.)*

Salma Matoussi is with the Research and Development Department, GANDI SAS, 75013 Paris, France (e-mail: Salma.Matoussi@gandi.net).

Ilhem Fajjari is with the Core Network, Automation, Security, E2E Services, Orange Innovation, 92320 Châtillon, France (e-mail: Ilhem.Fajjari@orange.com).

Nadjib Aitsaadi is with Université Paris-Saclay, UVSQ, DAVID, 78035 Versailles, France (e-mail: nadjib.aitsaadi@uvsq.fr).

Rami Langar is with the Software and Information Technology Engineering Department, Ecole de Technologie Supérieure de Montréal, Montreal, QC H3C 1K3, Canada, and also with the Computer Science Laboratory of Gaspard-Monge, University Gustave Eiffel, 77454 Marne-la-Vallée, France (e-mail: Rami.Langar@etsmtl.ca).

Digital Object Identifier 10.1109/TNSM.2023.3284206

approach is in compliance with the 3GPP slice vision, allowing more flexible deployment of NG-RAN slices. Specifically, our scheme fulfills each UE QoS requirement, while considering the underlying RAN infrastructure state. By enabling the selection of RAN functional split for each user, link and computational requirements become more tunable, which is a key to build cost effective RAN deployment solutions. The main contributions of our paper can be summarized as follows:

1) First, we design a RAN-as-a-Service orchestration framework compliant to the 3GPP standard [3], which enables on-demand RAN User-centric Slice Allocation, so called RAN-USA. Such a flexible RAN slicing system is tailored to temporal load variation, while managing a multi-sited RAN infrastructure with link aggregators in the transport network, as defined by 3GPP NG-RAN specification [4]. We refer to the latest advances of SD-RAN to monitor and control the network state and we consider the ETSI NFVI standard for orchestrating the RAN functional splits.
2) Second, we make use of an analytical model for quantifying the computational, link and latency resource requirements for each user split configuration. Then, we formulate the user-centric slicing allocation problem as an Integer Linear Problem (ILP) with multi-objective function. Two objectives are targeted: First, the maximization of the overall served throughput of users across the network through radio resource allocation. We exploit the regression linear method [19] to approximate the final served user throughput. Second, the minimization of the network deployment cost, while tuning the computational and link resource usage.
3) Considering that the formalized above optimization problem is NP-Hard, the resolution time becomes intractable in case of high density of user's traffic. In order to resolve our optimization problem in a polynomial time, we propose a low-cost and efficient heuristic algorithm based on the Particle Swarm Optimization approach [20]. The latter consists of creating initially a set of potential allocation solutions. Then, iteratively, the candidate solutions collaborate and evolve towards the best global allocation solution. The performance of RAN-USA is evaluated throughout extensive simulations on 3GPP eMBB and uRLLC traffic scenarios [1]. Obtained results highlight the effectiveness of our proposal in terms of scalability, QoS satisfaction and RAN deployment cost. We highlight the exploration and exploitation dilemma during the solution generation, which is ruled by the ε-greedy approach. Additionally, we evaluate the impact of user mobility on the performance and operational aspect of our proposed algorithm. We expect that the optimization process is triggered periodically to optimize the user slice allocation in a pro-active manner.

The rest of this paper is organized as follows. In Section II, we elaborate the state of the art addressing RAN slicing and related key enablers. Then, joint radio resource allocation with functional split problem in a multi-cell NG-RAN deployment is investigated. Section III sketches our proposed RAN-USA slice orchestration framework and its building blocks. We detail, in Section IV, the optimization user-centric slicing model. In Section V, we describe our proposed heuristic, while a description of our simulation and major results are presented in Section VI. Finally, Section VII concludes this paper.

II. Related Work

A. Towards Enabling RAN Slicing

As stated earlier, network slicing [2] is a concept that allows the network to deliver allocated resources (so called slice) as per service requirement. Unlike slicing in the 5G core network, the RAN slicing still remains an open issue. Till this moment, there are only proprietary solutions in MAC scheduler for service prioritization. Within this context, the main challenges that should be addressed in RAN slicing are how to enable on-demand provisioning and control of RAN resources across radio, computational and transport domains.

1) Radio Resource Slicing: The concept of spectrum sharing has been introduced to enable the partitioning of radio resources among operators as standardized by 3GPP in [6]. Accordingly, many works like in [7], [8], [9], [10], [11] seek on how to efficiently manage the radio resource allocation with isolation and sharing capabilities. In [21] and [22], the authors propose advanced mechanisms enabling tenants to reap the performance benefits of sharing, while retaining the ability to customize their own users' allocation. Other fundamental radio resources including transmission power and the cache space are still under discussion for slicing purpose.

From a technical perspective, radio resource slicing can be enabled through the Software-Defined RAN (SD-RAN) concept that decouples the Control Plane (CP) from the User Plane (UP) of radio functions. In doing so, access to the shared spectrum resources can be dynamically managed thanks to radio configuration functionalities located in CP. Authors in [23], [24], [25] focus on the different abstraction views of these radio control functions. Authors in [26] and [27] undertake the UP programmability and modularity aspect. FlexRAN is proposed in [28] to provide a general API for radio controlling with a custom RAN south-bound API, wherein RAN configuration can be enforced with a partial or full access to the allocated spectrum. In [29], 5G-EmPOWER is proposed as an open-source SDN platform with open protocol for radio resource control.

2) On-Demand RAN Function Placement: 5G stakeholders strive to redesign the RAN architecture in aim of building the service-oriented vision [3] for slice enabling. 3GPP introduces in Release 15 [4], the NG-RAN architecture with new terminologies, interfaces and functional modules. In this specification, the BBU is re-architected to be decomposed into a chained Processing Functions (PFs) that can be disaggregated at several conceivable points (functional splits). Wherein, the new 5G eNodeB (so called gNB) is decoupled into i) Radio Unit (RU), ii) Distributed Unit (DU) and iii) Central Unit (CU). The primary new F1 interface is defined between DU and CU, while the F2 interface is interconnecting the RU to DU. It is foreseen that CUs can be deployed on a cloud site

to take benefits from the cloud service infrastructure. The latter perspective was initially proposed under the Cloud RAN vision [5]. Besides, leveraging NFV [30], PFs can be virtualized to get rid of hardware dependency. Wherein, PFs constitute the NFV Infrastructure (NFVI) which is controlled by the Virtualized Infrastructure Manager (VIM) and managed by the MANagement and Orchestration (MANO) stack. The latter uses the Network Service Descriptor (NSD) as a template to instantiate a Network Service (slice instance) with customized network functions based on the end-user's requirements.

3) Slicing in the Transport Network: The disaggregated RAN deployment approach results in the definition of multiple functional split options. Accordingly, each of them is characterized by a dedicated interface interconnecting the CU and DU, with different guaranteed throughput, latency and reliability requirements. At the end, and by aggregating the traffic of multiple cells, the transport network will carry flows of data with heterogeneous requirements.

It is foreseen that a packet-based system with meshed connectivity between RAN locations would rather replace the CPRI fronthaul over wavelengths [31]. In addition to its lower cost compared to a fiber-based transport network, packet-based links are likely to create more programmable and flexible network. Authors in [32] implement an SD-RAN based platform for a hierarchical and programmable control and orchestration plane in a transport network. Authors in [33] leverage the SDN concept to support multi-tenancy in an efficient manner, where virtual networks can be allocated for distinct logical paths tailored to their service requirements.

It should be emphasized that the aforementioned works are implementing either a cell-centric strategy or service-oriented strategy for flow transport. Instead, we are implementing a fine-grained approach which is user-centric to carry all the 5G flows on the transport network.

B. RAN Slice Allocation Optimization

By enabling on the fly resource provisioning, RAN slices can be created and managed in a dynamic fashion, insuring the RAN-as-a-Service (RANaaS). One of the major concerns is how to meet the multitude use-case's requirements, while considering different designs in the physical infrastructure. In doing so, decisions on how much of radio resources to allocate and identify which network functions to place in DU or CU impose many challenges. Eventually, RAN slice allocation impacts directly the UE QoS performances and the operation cost, which is essential to design an orchestration solution able to rise these challenges. The network slicing approach that **jointly** optimizes the radio resource allocation and the functional split selection has motivated many research works.

In view of this, authors in [12] elaborate a joint functional split and BBU server scheduling problem to minimize the overall processing delay of downlink frames. The problem is formulated as a constrained shortest-path problem and resolved with a heuristic algorithm. However, in this work, the functional split selection is performed on a cell-basis. In [13], the authors elaborate a radio resource slicing scheme to fulfill the user Service Level Agreement (SLA), while insuring slice isolation. Although the approach is on user basis, the authors do not consider the RAN split in the slice allocation decision. In [15], a RAN runtime framework for slice control and orchestration is proposed. Then, a detailed approach on radio resource slicing with different levels of isolation and sharing is described. Although the disaggregated RAN scheme is integrated in the framework design, there is no problem modeling for functional split selection. In [16], a multi-tenant slicing scheme in Cloud-RAN is proposed taking into account tenant priority, computational resources, transport network capacity and interference levels. However, this work considers only a full centralized deployment scheme. In [17], the authors formulated a problem of the functional split selection, while considering the inter-cell interference level. A new heuristic is proposed to minimize jointly the inter-cell interference and the bandwidth utilization on the transport network. However, the functional split approach is performed at cell level. Authors in [18] propose a framework for slice management with functional split selection. They address the problem of joint radio allocation and split selection to meet the different use-case requirements. The approach is user-centric. However, the authors consider only one disaggregated scheme (i.e., RLC-MAC split). Therefore, our current study aims to fill the aforementioned gaps. Authors in [34] propose a joint slicing and functional split optimization framework for 5G. A Mixed Integer Programming model is formulated and then linearized. However, the adopted splits are not conform to the 3GPP specification. Subsequently, the splits' requirements are not realistic. In [35], a Functional Split Option-based Coordinated Multi-Point (CoMP) transmission for mixed eMBB-uRLLC services is proposed. The adopted scheme is resolved with the stochastic geometry approach. However, the model considers delay only for uRLLC users, which is not accurate, since eMBB users require an end-to-end delay of 8 ms, which exclude already a subset of splits.

C. Particle Swarm Optimization

Radio, computational and transport resource allocation impacts directly the end-user QoS and the deployment cost. This type of problem can be modeled as Multi Objective Combinatorial Optimization Problem (MOCOP). Then, the problem is decomposed into a set of single-objective subproblems using the weighted sum approach or others like the Tchebycheff approach, etc. When solved, MOCOPs are generally nondeterministic polynomial complete or nondeterministic polynomial hard [19]. Thus, we need to design an algorithm to solve it in a polynomial time by generating a near-optimal solution.

In this context, Particle Swarm Optimization (PSO) [20] approach is proposed as a population-based stochastic optimization algorithm inspired from birds foraging behavior. More specifically, PSO algorithm is characterized by an initial set of candidate solutions collaborating to find the global optimum of the optimization problem. In practice, swarm optimization algorithm for combinatorial optimization

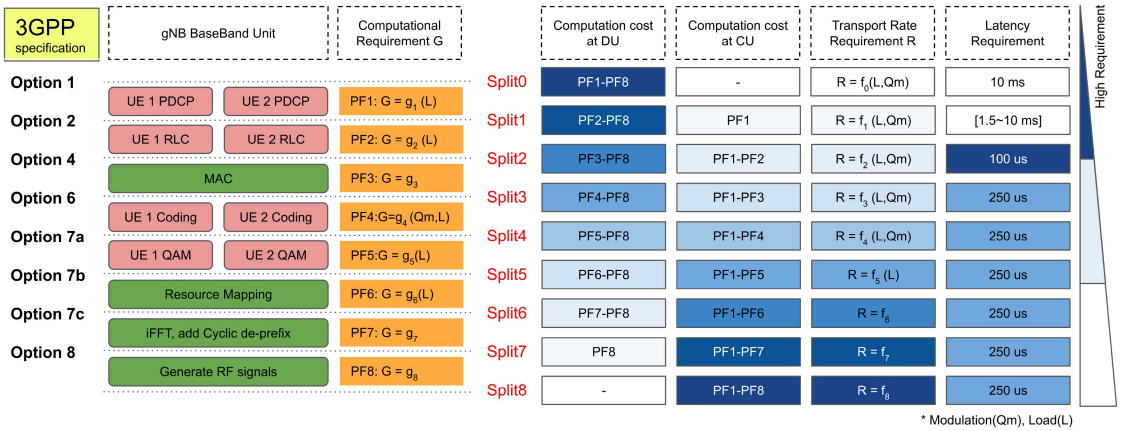


Fig. 1. Functional split requirements.

problem or what we call Set-based Particle Swarm Optimization S-PSO has been successfully applied in solving many problems like scheduling and vehicle routing problems [37], [38].

In order to solve our problem, we propose a set-based discrete particle swarm optimization based on MOCOP and S-PSO. Previously, we elaborated in [36], a model for RAN resource allocation that jointly optimizes the resource usage of radio, link and computational resources. In this paper, we propose to go a step further and optimize the radio resource allocation by means of the linear regression method [19] to approximate the final served user throughput. Our objective is to meet the user requirements in terms of throughput and latency, while considering multiple deployment approaches in the underlying physical infrastructure. In order to operate in a polynomial time, we propose a heuristic algorithm to handle the problem efficiently and pro-actively. As opposed to our previous work in [36], we present here a new model formulation for radio allocation based on the regression linear method to approximate the final served user throughput. We have also redesigned our heuristic by including a new convergence property that expresses the exploration and exploitation dilemma during the solution generation, which is ruled by the ε-greedy approach. In addition, we have extended our proposed slice allocation framework with the service-oriented vision of the 3GPP standard [3] by integrating our optimization scheme in a new flexible RAN orchestration framework, in compliance with the 3GPP RANaaS vision [3], taking benefits from the latest advances of SDN-RAN, ETSI NFVI and Cloud-RAN capabilities. Contrary to our previous work in [36], we have evaluated the dynamic and scaling aspect of our new approach in a multi-cell environment by varying the number of gNBs, UEs, uRLLC UEs and our solution parameters.

III. User-Centric RAN Slice Allocation Framework

In this section, we describe our proposed RAN User-centric Slice Allocation Framework, named RAN-USA implemented on top of a multi cell RAN infrastructure. The main idea behind our design is to ensure on-demand RAN resource provisioning with a fine-grained approach on user basis. Our goal is to fulfill end user's QoS requirements, while enabling customization of RAN resource usage.

A. Disaggregated RAN Model

Let us start by providing an overview of the functional split options in a disaggregated RAN architecture in compliance with 3GPP Specification [4]. Accordingly, the BBU layers are instantiated into containerized Processing Functions (PF) to perform either cell or user processing tasks. Hereafter, we detail the interface specification and requirement between each couple of *PFs* in DownLink (DL) as depicted in Fig. 1.

- $Split_0$ (3GPP Option 1), considered as a user split, which decentralizes all *PF*s of a given UE at the DU. This interface can transport IP UE packets on non ideal links with a latency up to 10 *ms* [39].
- $Split_1$ (3GPP Option 2), considered as a user split, which centralizes PF_1 of a given UE at CU site. PF_1 corresponds to the Packet Data Convergence Protocol (PDCP) receiving IP packets from higher layer and performing header compression and encryption operations. Accordingly, $Split_1$ interface requires a latency up to 10 *ms* to transport the IP UE packets.
- $Split_2$ (3GPP Option 4), considered as a UE split, that centralizes both PF_1 and PF_2 of a given UE. PF_2 performs Radio Link Control (RLC) that mainly ensures the unpacking and segmentation of PDCP flows. In this work, we do not consider the intra-RLC split (Option 3 in 3GPP [4]) as resource requirement model for RLC-high and RLC-low layers are still an ongoing work. $Split_2$ interface requires a tight latency of 100 *us* [39] to transport the processed radio bearer.
- $Split_3$ (3GPP Option 6), considered as a user split, which centralizes PF_1, PF_2 an PF_3 of a given UE. PF_3 is the Medium Access Control (MAC) function which performs the multiplexing of data from different radio bearers. It is responsible for building a transport block per UE based on the UE's context and its data buffer operating at user level. However, PF_3 contains a controller and a random access control entities, operating at a cell level, to maintain the cell's state for scheduling the radio transmission. Eventually, PF_3 is rather considered as a

cell-centric function. In this work, the intra-MAC split (3GPP-Option 5) is not considered as resource requirement models for MAC-high and MAC-low layers are still an ongoing work. $Split_3$ is often called the MAC-PHY split because it disaggregates the physical layer (layer 1) from MAC+RLC+PDCP (layer 2).

- $Split_4$ (3GPP Option 7a), considered as a user split, centralizing additionally PF_4. PF_4 performs encoding, forward error correction and rate matching at user level.
- $Split_5$ (3GPP Option 7b), considered as a user split, centralizes additionally PF_5. PF_5 mainly includes the Quadrature Amplitude Modulation (QAM) generating symbols for multi-antenna mapping.
- $Split_6$ (3GPP Option 7c), considered as a cell split because it centralizes all UE *PF*s from PF_1 to PF_5, PF_3 and PF_6. The latter corresponds to the resource mapping function, initiating the cell processing by mapping the symbols on resource elements.
- $Split_7$ (3GPP Option 8), considered as a cell split, which further centralizes PF_7 while keeping only PF_8 in DU. PF_7 adds the cyclic prefix and transforms symbols from frequency domain to time domain using iFFT Algorithm.
- $Split_8$ is considered as a cell split, centralizing all PFs at CU site. Accordingly, $Split_8$ interface manages in DL direction the traffic generated by PF_8 containing radio frequency signals with Parallel-to-Serial (P/S) conversion among others. $Split_8$ corresponds to the traditional full centralized Cloud-RAN architecture necessitating high bandwidth requirement in the midhaul. In addition, intra physical layer splits (i.e., $Split_3$, $Split_4$, $Split_5$, $Split_6$, $Split_7$, $Split_8$) require a tight latency of 250 *ms* [39].

It is worth noting that once RLC and PDCP layers are centralized, the latency requirement on the midhaul becomes stringent which satisfies broadly use cases with tight latency requirement. The aforementioned split options offer different design approaches for slice deployment that will be integrated in our proposed architecture for RAN resource orchestration, as detailed hereafter.

B. RAN-USA: RAN User-Centric Slice Allocation Framework Architecture

Fig. 2 depicts our proposed RAN-USA framework, which is a cloud-native C-RAN environment, using i) stateless architecture, ii) microservices and iii) containers technology. Such enablers will enhance the RAN processing functions (PF) development, while automating their service deployment and upgrade for better operational efficiency. Being packaged in containers instead of virtual machines, these *PF*s can be dynamically instantiated and destroyed within few micro seconds. Indeed, according to our experiments in [40], we have computed the average deployment time of a container-based *PF* to 1.8 us $\pm$ 0.2 us. The main objective of our solution is to ensure on-demand user-centric resources instantiation. To do so, one instantiated *PF* (i.e., container) is deployed to host several users' threads (i.e., light-weight process). Note that if the load of one container exceeds a given threshold, a new instance is created dynamically to host the increasing number (or load) of users.

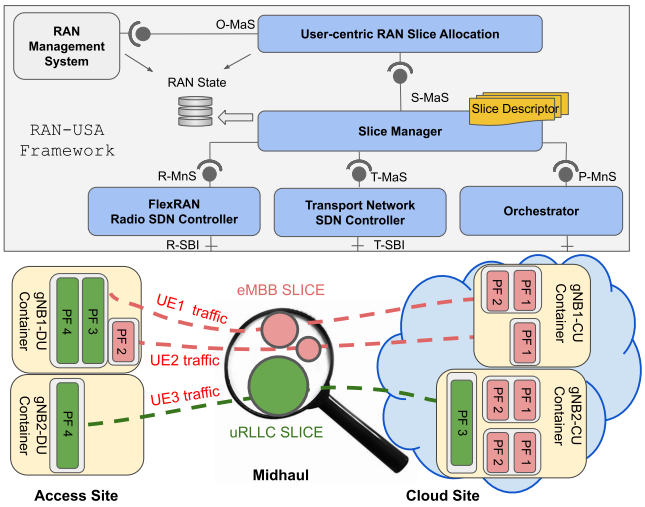


Fig. 2. RAN-USA Framework.

Our framework makes use of an optimization entity, which implements our user-centric RAN Slice Allocation solution that fetches RAN state information previously collected by the slice Manager via SDN controllers. This includes radio conditions (i.e., available spectrum, interference levels, *UE* radio channel estimation), link state (i.e., available bandwidth) and server capacities (i.e., processing power) of all RAN sites. The aim is to elaborate an optimized slice allocation decision that satisfies both UE throughput and latency demands, while keeping a cost effective RAN deployment. In doing so, and based on the aforementioned parameters, an optimized RAN slice allocation is formulated for each UE, by assigning the appropriate proportions of i) radio spectrum, ii) computational resources in DU site, iii) computational resources in CU site, and iv) bandwidth in the midhaul connecting the DUs to the CU. These information are registered in the slice descriptor resource requirements and then transferred to the Slice Manager entity through the Slice Management Service (S-MaS) interface.

The Slice Manager interacts with three resource management entities in order to deploy each user slice conforming to the slice descriptor specification. First, proportions of radio resources are allocated and configured by means of an SDN controller for radio resource management, namely FlexRAN [28]. The latter provides an API for radio controlling over multiple *gNB*s through its Radio SouthBound Interface (R-SBI). Second, the user processing resources are allocated both in DU and CU sites via the Processing Management Service interface (P-MaS). At this stage, DU and CU functions are instantiated into containerized network functions that can easily interact with each other and scale separately by mean of the Orchestrator. Third, the Slice Manager entity interconnects the *gNB*-CU and *gNB*-DU containers by programming the link bandwidth provisioning and latency control in the midhaul. This is handled by an SDN controller through a Transport Management Service interface (T-MaS) leveraging its centralized and abstract network view.

For instance, and as it can be seen in Fig. 2, our RAN slice allocation application instantiates dynamically three options of functional splits on top of the same physical infrastructure.

Our RAN-USA application chooses to centralize PF_1 and PF_2, while keeping the rest of *PFs* in the DU site for UE_1 generating an eMBB traffic. Then, only PF_1 is centralized in the cloud for UE_2 generating an important eMBB traffic. The aim is to reduce the data flow in the midhaul. Meanwhile, only PF_4 is kept in the DU site for UE_3 generating a uRLLC traffic as PF_3-PF_4 interface requires a stringent transport delay satisfying UE_3 latency requirement.

Our proposed framework offers an Optimization Management Service interface (O-MaS), through which the infrastructure provider (i.e., the operator) can customize the RAN resource usage, by providing optimal split decision. In doing so, the RAN Management System component subscribes to the RAN state entity through an event driven interface. Thus, it can be notified if a resource usage amount exceeds a given threshold (e.g., the amount of network traffic in the midhaul). In this case, the infrastructure provider chooses to tune the optimization entity by penalizing the allocation in the midhaul in order to reduce the link resource usage.

In the next section, we present our optimization model for joint Functional Split and Radio Resource Allocation. Our objective is to maximize the total offered throughput for the users across the network, while minimizing, at the same time, the total deployment cost. To achieve this, we propose to i) perform the cell attachment and radio resource block allocation, while considering the interference level of radio resource blocks, and at the same time, ii) choose the optimal functional split, for each user.

IV. Problem Formulation

In this section, we present the functional split model used in our analysis. Then, we present our joint Functional Split and Radio Resource Allocation (FSRRA) problem formulation, which can be divided into two correlated sub-problems: User-centric RAN slicing one and Radio Resource Allocation one.

A. Functional Split Model

In order to quantitatively study the computational resource requirement for each split in each RAN site, we refer to the conducted analysis in [41] and [42] expressing the amount of computational resources in Giga Operations Per Second (GOPS) consumed by a *PF*. We denote by g_k the computational requirement model of each PF_k in DL direction:

$$\mathrm{PF}_1 : \mathrm{g}_1(\mathrm{L_i}) = \mathrm{G}_1^{\mathrm{ref}} \frac{\mathrm{A}}{\mathrm{A_{ref}}} \times \boldsymbol{L_i} \quad (E1)$$

$$\mathrm{PF}_2 : \mathrm{g}_2(\mathrm{L_i}) = \mathrm{G}_2^{\mathrm{ref}} \frac{\mathrm{A}}{\mathrm{A_{ref}}} \times \boldsymbol{L_i} \quad (E2)$$

$$\mathrm{PF}_3 : \mathrm{g}_3 = \mathrm{G}_3^{\mathrm{ref}} \frac{\mathrm{A}}{\mathrm{A_{ref}}} \quad (E3)$$

$$\mathrm{PF}_4 : \mathrm{g}_4(\mathrm{Qm_i}, \mathrm{L_i}) = \mathrm{G}_4^{\mathrm{ref}} \frac{\mathrm{W}}{\mathrm{W_{ref}}} \frac{\mathrm{A}}{\mathrm{A_{ref}}} \frac{\boldsymbol{Qm_i}}{\mathrm{Qm_{ref}}} \times \boldsymbol{L_i} \quad (E4)$$

$$\mathrm{PF}_5 : \mathrm{g}_5(\mathrm{L_i}) = \mathrm{G}_5^{\mathrm{ref}} \frac{\mathrm{W}}{\mathrm{W_{ref}}} \left(\frac{\mathrm{A}}{\mathrm{A_{ref}}}\right)^2 \times \boldsymbol{L_i} \quad (E5)$$

$$\mathrm{PF}_6 : \mathrm{g}_6(\{\mathrm{L_i}\}) = \mathrm{G}_6^{\mathrm{ref}} \frac{\mathrm{W}}{\mathrm{W_{ref}}} \frac{\mathrm{A}}{\mathrm{A_{ref}}} \times \sum_{\mathrm{i}}^{\mathrm{UEs}} \boldsymbol{L_i} \quad (E6)$$

$$\mathrm{PF}_7 : \mathrm{g}_7 = \mathrm{G}_7^{\mathrm{ref}} \frac{\mathrm{W}}{\mathrm{W_{ref}}} \frac{\mathrm{A}}{\mathrm{A_{ref}}} \quad (E7)$$

$$\mathrm{PF}_8 : \mathrm{g}_8 = \mathrm{G}_8^{\mathrm{ref}} \frac{\mathrm{W}}{\mathrm{W_{ref}}} \frac{\mathrm{A}}{\mathrm{A_{ref}}} \quad (E8)$$

where G_k^{ref} refers to the PF_k's GOPS value in the reference scenario [41]. *W* is the carrier bandwidth, *A* is the number of antennas; *L* is the proportion of allocated Resource Blocks (RB) for UE_i and *Qm* is the QAM modulation. It is worth noting that, PF_7 and PF_8 are cell-centric for time-domain, while PF_3 corresponds to the platform control processing. Hence, their computational requirement is load independent. In contrast, PF_1, PF_2, PF_4 and PF_5 perform in frequency domain, i.e., take into account only frequency carriers having data signals which make them load dependent.

With reference to the model in [43], we quantitatively study the bandwidth requirement for each functional split in the midhaul. Accordingly, the generated traffic of each split interface in DL can be estimated as follows:

$$\mathrm{Split}_0 : \mathrm{f}_0(\mathrm{L_i}, \mathrm{Qm_i}) = \mathrm{c}_0(\boldsymbol{Qm_i}) \times \mathrm{A} \times \mathrm{B} \times \boldsymbol{L_i} \quad (E9)$$

$$\mathrm{Split}_1 : \mathrm{f}_1(\mathrm{L_i}, \mathrm{Qm_i}) = \mathrm{c}_1(\boldsymbol{Qm_i}) \times \mathrm{A} \times \mathrm{B} \times \boldsymbol{L_i} \quad (E10)$$

$$\mathrm{Split}_2 : \mathrm{f}_2(\mathrm{L_i}, \mathrm{Qm_i}) = \mathrm{c}_2(\boldsymbol{Qm_i}) \times \mathrm{A} \times \mathrm{B} \times \boldsymbol{L_i} \quad (E11)$$

$$\mathrm{Split}_3 : \mathrm{f}_3(\mathrm{L_i}, \mathrm{Qm_i}) = \mathrm{c}_3(\boldsymbol{Qm_i}) \times \mathrm{A} \times \mathrm{B} \times \boldsymbol{L_i} + \mathrm{c}_4 \quad (E12)$$

$$\mathrm{Split}_4 : \mathrm{f}_4(\mathrm{L_i}, \mathrm{Qm_i}) = \mathrm{A} \times \mathrm{B} \times (\mathrm{c}_5 + \mathrm{c}_6 \times \mathrm{A}) \times \boldsymbol{L_i} \times \boldsymbol{Qm_i} + \mathrm{c}_7 \quad (E13)$$

$$\mathrm{Split}_5 : \mathrm{f}_5(\mathrm{L_i}) = \mathrm{A} \times \mathrm{B} \times (\mathrm{c}_8 + \mathrm{c}_9 \times \mathrm{A}) \times \boldsymbol{L_i} + \mathrm{c}_{10} \times \mathrm{A} \quad (E14)$$

$$\mathrm{Split}_6 : \mathrm{f}_6 = \mathrm{c}_{11} \times \mathrm{A} \times \mathrm{B} \quad (E15)$$

$$\mathrm{Split}_7 : \mathrm{f}_7 = \mathrm{c}_{12} \times \mathrm{A} \times \mathrm{n_s} \quad (E16)$$

$$\mathrm{Split}_8 : \mathrm{f}_8 = \mathrm{c}_{13} \times \mathrm{A} \times \mathrm{n_s} \quad (E17)$$

where coefficients c_j, $\forall j \in \{1, 2, \ldots, 13\}$, are constants for the model [43]. *B* corresponds to the number of RBs and n_s refers to the sampling rate. It is straightforward to note that when the centralization level of *PF*s increases, the computational requirement in the cloud site increases accordingly, which rises the amount of the circulating data flow in the midhaul link.

B. User-Centric RAN Slicing Sub-Problem Formulation

Here, we consider a multi-cell RAN system with *M gNB*s. Each one *gNB* is characterized by a Distributed Unit (DU) located near the antenna radio unit and a Central Unit (CU) located at the cloud site. The computational capacity of one DU, (respectively one CU) is denoted by C^D_{MAX}, (respectively C^C_{MAX}) Giga Operation Per Second (GOPS). We assume that a set of *K* functional splits can be deployed for *N UE*s. Then, we consider the amount of GOPS consumed by UE_i in DU site (respectively in CU site) of gNB_m when split *k* is deployed, is denoted by C^D_{imk} (respectively C^C_{imk}). By aggregating all the computational requirements, we define C^D_m (respectively C^C_m) as the total amount of GOPS consumed at DU (respectively CU) of gNB_m. The connection between both DUs and CUs locations is maintained via an aggregated midhaul link with a capacity of R_{MAX} Mbps. Wherein, R_{imk} corresponds to

the amount of data flow generated for UE_i attached to gNB_m with split k. We define also R as the aggregated link bandwidth generated by all *UE*s in M *gNB*s. Formally, C^D, C^C and R are variables expressed as linear functions of *UE* loads L. We recall that *UE* load ld_{im} $(\forall i \in N, \forall m \in M)$ corresponds to the fraction of allocated RB for UE_i in gNB_m.

Our aim is to find the appropriate split k for each UE_i in gNB_m that minimizes the total deployment cost. Therefore, we define x_{im}^k as the binary variable, which is equal to 1 when split k is selected for UE_i in gNB_m and 0 otherwise. Then, we assume that the total available split options K can be divided into 3 subsets: K_c, K_{u_1} and K_{u_2}. K_c is the set of cell splits, namely splits $\{8, 7, 6\}$. K_{u_1} is the first set of user splits, namely $\{0, 1, 2\}$. Finally K_{u_2} is the second set of user splits, namely $\{3, 4, 5\}$, according to Section III-A. Let y_m^k be the binary variable, $\forall k \in \{0, .., K\}$ and $\forall m \in \{1, .., M\}$, that takes value 1 if the split k is activated for any *UE* in gNB_m and 0 otherwise. We also define the binary variable u_1^m (u_2^m respectively) that takes value 1 if a user split in subset $\{0, 1, 2\}$, ($\{3, 4, 5\}$ respectively) is activated in gNB_m and 0 otherwise. We model the attachment of UE_i to gNB_m with a binary variable t_{im}. The latter is equal to 1 when UE_i is attached to gNB_m and 0 otherwise.

In what follows, we present our model for RAN deployment cost minimization by optimizing the user split selection. Note that we make use of the Big-M modeling [44] to linearize different constraints, wherein, M_1 is an upper bound limit equal to N.

$$\mathcal{LP}_1\text{: Min } \alpha \sum_{m=1}^{M} \frac{C_m^D}{C_{MAX}^D} + \beta \sum_{m=1}^{M} \frac{C_m^C}{C_{MAX}^C} + \gamma \frac{R}{R_{MAX}}$$

$$\text{s.t.: } t_{im} = \sum_{k=0}^{K} x_{im}^k, \forall i \in N, \forall m \in M \quad (1)$$

$$x_{im}^k \nu_k \leq \nu_i, \forall i \in N, \forall m \in M, \forall k \in K \quad (2)$$

$$\sum_{i=1}^{N} x_{im}^k \leq M_1 y_m^k, \forall m \in M, \forall k \in K \quad (3)$$

$$y_m^k \leq \sum_{i=1}^{N} x_{im}^k, \forall m \in M, \forall k \in K \quad (4)$$

$$\sum_{i=1}^{N} x_{im}^{k'} \leq \sum_{i=1}^{N} t_{im} + M_1 \left(1 - y_m^{k'}\right), \forall m \in M, \forall k' \in K_c \quad (5)$$

$$\sum_{i=1}^{N} x_{im}^{k'} \geq \sum_{i=1}^{N} t_{im} - M_1 \left(1 - y_m^{k'}\right), \forall m \in M, \forall k' \in K_c \quad (6)$$

$$\sum_{k=1}^{K_{u_1}} y_m^k \leq |K_{u_1}| u_1^m, \forall m \in M \quad (7)$$

$$u_1^m \leq \sum_{k=1}^{K_{u_1}} y_m^k, \forall m \in M \quad (8)$$

$$\sum_{k=1}^{K_{u_2}} y_m^k \leq |K_{u_2}| u_2^m, \forall m \in M \quad (9)$$

$$u_2^m \leq \sum_{k=1}^{K_{u_2}} y_m^k, \forall m \in M \quad (10)$$

$$\sum_{k=1}^{K_c} y_m^k + u_1^m + u_2^m \leq 1, \forall m \in M \quad (11)$$

$$R = \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{k=0}^{K} x_{im}^k R_{imk} \leq R_{MAX} \quad (12)$$

$$C_m^D = \sum_{i=1}^{N} \sum_{k=0}^{K} x_{im}^k C_{imk}^D \leq C_{MAX}^D, \forall m \in M \quad (13)$$

$$C_m^C = \sum_{i=1}^{N} \sum_{k=0}^{K} x_{im}^k C_{imk}^C \leq C_{MAX}^C, \forall m \in M \quad (14)$$

$$x_{im}^k \in \{0, 1\}, \forall i \in N, \forall m \in M, \forall k \in K \quad (15)$$

$$y_m^k \in \{0, 1\}, \forall m \in M, \forall k \in K \quad (16)$$

$$u_1^m, u_2^m \in \{0, 1\}, \forall m \in M \quad (17)$$

The objective function in $\mathcal{LP}_1$ expresses the ability to tune the computational and link resource usage to minimize the RAN deployment cost, while considering the infrastructure capacity and *UE* latency constraints. This can be done by leveraging the user functional split that helps to find a trade-off between the centralization and decentralization levels of BBU functions. The first level of $\mathcal{LP}_1$ expresses the computational resource usage across DU sites, weighted by α. The second level is expressed as the computational resource usage across CU sites, weighted by β. Finally, the third level expresses the ongoing traffic in the aggregated midhaul links by calibrating the weighting factor γ. It is worth noting that we assume here that RUs and DUs are co-located in one site.

Constraint (1) expresses that attached UE_i in gNB_m can be assigned only one split k, $\forall k \in K$. Constraint (2) denotes that the latency generated by split k, ν_k, should satisfy the latency required by UE_i, ν_i. Constraint (3) activates the binary variable y_m^k when at least one user split k is activated in gNB_m. Constraint (4) expresses that when split k is deactivated for gNB_m, then no *UE* is assigned split k. Constraints (5) and (6) denote that the activation of one cell split k' in gNB_m, results in assigning split k' for all attached *UE*s. Constraints (7) and (8) activate the variable u_1^m when at least one split k in subset K_{u_1} is activated in gNB_m. Constraints (9) and (10) activate the variable u_2^m when at least one split k in subset K_{u_2} is activated in gNB_m. Constraint (11) denotes that for a given gNB_m, we may activate i) either one cell split k' in K_c or ii) a combination of user splits in K_{u_1} or iii) a combination of user splits in K_{u_2}. In (12), the total generated rate in the aggregated midhaul link should not exceed the link capacity R_{MAX}. Finally, constraints (13) and (14) express that the total allocated computational resources in DU, respectively CU, of gNB_m must not exceed the total computational capacity C_{MAX}^D, respectively C_{MAX}^C.

C. Radio Resource Allocation Sub-Problem Formulation

Hereafter, we consider N *UE*s statically located in a system of M *gNB*s with a frequency reuse factor of 1, i.e., the same set B of RBs are reused by each cell, which may induce interference on RB level. Each UE_i, $\forall i \in N$, generates a flow of throughput λ_i and maximum tolerated latency ν_i.

Considering the DL transmission direction, we calculate the Signal to Interference plus Noise Ratio (SINR) experienced by

UE_i from gNB_m, $\forall i \in N, \forall m \in M$, expressed as following:

$$\mathrm{SINR_{im}} = \frac{\overline{\mathrm{P_{gNB}}} \times \mathrm{h_{im}}}{\mathrm{I_{im}} + \sigma^2}, \ \mathrm{I_{im}} = \sum_{\mathrm{m'} \neq \mathrm{m}} \overline{\mathrm{P_{gNB}}} \times \mathrm{h_{im'}} \quad (E18)$$

where h_{im} denotes the channel gain between each gNB_m and UE_i, I_{im} stands for the interfering power received by the UE_i from other *gNB*s $m' \neq m$. σ^2 is defined as the noise power, while we assume that every *gNB* transmits a static amount of power, denoted by $\overline{P_{gNB}}$. Based on the SINR average estimation, we calculate the Channel Quality Indicator (CQI), the Modulation and Coding Scheme (MCS) and the Transport Block Size Index (I_{TBS}) between UE_i and gNB_m, $\forall i \in N$, $\forall m \in M$.

In order to proceed to radio resource allocation, we rely on following binary variables *t*, *w* and *L*. We recall that t_{im} is a binary variable, which is equal to 1 if UE_i is attached to gNB_m and 0 otherwise. w_{imb} is equal to 1 if the RB b of gNB_m is allocated to UE_i and 0 otherwise. L_{im} expresses the radio load of UE_i in gNB_m corresponding to the resulting fraction of total allocated RBs assigned to UE_i in gNB_m, with radio power equal to $\overline{p_{RB}}$: $L_{im} = \frac{\sum_{b=1}^{B} w_{imb}}{B}$.

For sake of simplicity, radio power allocation is not performed at this stage. Our aim is to find a *gNB* with an optimized radio load; i.e., an appropriate amount of allocated RBs with less interference level. The aim is to fulfill the throughput requirement λ_i of each UE_i. By means of the linear regression method [19], we propose an approximation of the served Transport Block Size, denoted by $\widetilde{TBS}$. With reference to [45, Table 7.1.7.2.1-1], we calculate the linear approximation $\widetilde{TBS}$ according to each TBSI value as following:

$$\widetilde{\mathrm{TBS}}(\mathrm{TBSI}) = \mathrm{TBS^L}(\mathrm{TBSI}) \times \mathrm{B} \times \boldsymbol{L_{im}} + \mathrm{TBS^o}(\mathrm{TBSI}) \quad (E19)$$

where $B \times \boldsymbol{L_{im}}$ is the supposed number of allocated RBs for UE_i in gNB_m. $TTBS^L(TBSI) \times B \times \boldsymbol{L_{im}}$ is called the response that depends from user load and $TBS^o(TBSI)$ is called the predictor which is independent from user load. Consequently, we express the approximation of the final served throughput $\widetilde{r}_{im}$ for UE_i in gNB_m, $\forall i \in N$, $\forall m \in M$ as function of the linear approximation of the Transport Block Size $\widetilde{TBS}_{im}$ with a multiplication factor c_{34} for the conversion from bytes to bits per second:

$$\widetilde{r}_{im} = c_{14} \times \widetilde{TBS}_{im} \quad (E20)$$

Hence, we define a second objective function $\mathcal{LP}_2$ that aims at maximizing the overall served user throughput across the network. This is achieved by finding for each UE_i, i) the best attached *gNB* m, $t^{\star}_{im}$ and ii) the best set of RBs $w^{\star}_{imb}$, while keeping a low interference level. We make use of the Big-M modeling [44] to linear different constraints with upper bound limits M_2 and M_3, as follows.

$$\mathcal{LP}_2: \ \mathrm{Max} \quad \sum_{i=1}^{N} \frac{\sum_{m=1}^{M} \widetilde{r}_{im}}{\lambda_i}$$

$$\mathrm{s.t.:} \quad \sum_{m=1}^{M} t_{im} \leq 1, \forall i \in N \quad (18)$$

$$\sum_{b=1}^{B} w_{imb} \geq t_{im}, \forall m \in M, \forall i \in N \quad (19)$$

$$M_2 t_{im} \geq \sum_{b=1}^{B} w_{imb}, \forall i \in N, \forall m \in M \quad (20)$$

$$\sum_{m=1}^{M} \widetilde{r}_{im} \leq \lambda_i, \forall i \in N \quad (21)$$

$$\sum_{i=1}^{N} w_{imb} \leq 1, \forall m \in M, \forall b \in B \quad (22)$$

$$\sum_{m' \neq m} \sum_{i' \neq i} \overline{p_{RB}} h_{i'm'} w_{i'm'b} \leq I^{MAX} + M_3(1 - w_{imb}), \forall i \in N, \forall m \in M, \forall b \in B \quad (23)$$

$$t_{im}, w_{imb} \in \{0,1\}, \forall i \in N, \forall m \in M, \forall b \in B \quad (24)$$

Constraint (18) expresses that each UE should be attached at most one *gNB*. Constraint (19) specifies that UE_i can get more than one RB when it is attached to gNB_m. In (20), the total amount of allocated RBs to UE_i in gNB_m is constrained by the upper bound limit $M_2 = B$. In (21), the final served throughput for UE_i should be less than what is required with λ_i. In (22), each RB is assigned to only one *UE*. Finally, (23) expresses the interference constraint for each allocated RB, where I^{MAX} refers to the interference threshold and M_3 is a Big-M constant to tolerate interference on unallocated RBs.

D. FSRRA Problem Formulation

Considering the two aforementioned sub-problems, our joint Functional Split and Radio Resource Allocation (FSRRA) problem can be thus formulated as an Integer Linear Program (ILP), as follows:

$$\mathcal{LP}_3: \ \mathrm{Max}\, \theta \mathcal{LP}_2 - \mu \mathcal{LP}_1$$
$$\mathrm{s.t.:} \quad (1)-(24)$$

Our objective is to find the trade-off between the total served user throughput expressed in $\mathcal{LP}_2$ weighted by θ and the total RAN deployment cost expressed in $\mathcal{LP}_1$ weighted by μ. Note that, in $\mathcal{LP}_3$, all user requirements in terms of latency and throughput, as well as computational, link and latency requirements of each split, have been taken into account.

V. Proposal: RAN-USA

In this section, we resolve the above FSRRA problem formulated in $\mathcal{LP}_3$. This problem is classified as a non-deterministic polynomial hard problem [46], which requires exhaustive search in the solution space in order to converge to optimal solutions. Hence, general-purpose linear solver [46] struggles to converge in case of high-scale of *UE* numbers. For this reason, there is a need for designing a heuristic to solve the formulated FSRRA problem in a reasonable time with a near-optimal solution.

In this context, we propose an adaptive approach of the Particle Swarm Optimization (PSO) Algorithm [8], called RAN User-centric Slice Allocation (RAN-USA), to solve our problem expressed in $\mathcal{LP}_3$. Our heuristic proceeds as follows: during the **initialization** stage, an initial set of feasible

solutions is generated by affecting for each *UE*: i) attached *gNB*, proportions of RBs and ii) split selection. Then, to solve the problem, our proposal proceeds iteratively on two folds. First, a better **`radio allocation`** ((i.e., $UE-gNB$ attachment and radio resource allocation) is explored. Second, a user functional split selection based on the shortest path algorithm is performed to find the optimal **`split selection`** for the already generated radio configuration in the first phase of current iteration. In the following, we detail these steps.

A. Initialization Stage

Each particle p, $p \in \{1, .., P\}$, is characterized by a position $\mathcal{S}^p$, a velocity $\mathcal{V}^p$ and a best local position $\mathcal{S}^{L,p}$. The first attribute (i.e., position) corresponds to a candidate slice allocation solution. The second attribute (i.e., velocity) expresses the change vector that allows the particle to evolve to a next position. The third attribute (i.e., best local position) memorizes the best achieved local solution. Candidate solutions are evaluated through the utility function $\mathcal{U}_{\mathcal{F}}$ expressed in $\mathcal{LP}_3$. We also denote by $\mathcal{S}^G$, the best achieved solution among all best local solutions, $\mathcal{S}^{L,p}$, $\forall\ p \in \{1, .., P\}$.

In what follows, we design position $\mathcal{S}^p$ of particle p as a 3-D matrix of $[N \times M \times (B+1)]$. The entry $\mathcal{S}^p_{imb}$ is a binary variable that takes the value 1 if UE_i is allocated the RB_b in gNB_m. Furthermore, we affect split k to UE_i in gNB_m. Formally, $\mathcal{S}^p_{im,(B+1)} = k$. The velocity component V^p of particle p is expressed as a 2-D matrix of $[N \times M]$. The entry $\mathcal{V}^p_{im}$ expresses the number of RBs to be added or removed in the next iteration for UE_i in gNB_m. Formally, $\mathcal{V}^p_{im}$ is in $[-B^{MAX}_{im}, +B^{MAX}_{im}]$, where B^{MAX}_{im} is the upper bound limit for UE_i allocation in gNB_m to satisfy his throughput λ_i.

Initially, each UE_i, $i \in \{1, .., N\}$ is attached to a random gNB_m, $m \in \{1, .., M\}$, with a random number n_{RB} of RBs. Meanwhile, we ensure that constraints (18-23) are satisfied. Constraint (18) leads to $\mathcal{S}^p_{im'b} = 0,\ \forall\ m' \neq m$. Constraint (19), (20) and (21) express that, $\mathcal{S}^p_{imb}$ can be positive n_{RB} times, where n_{RB} is in $[-B^{MAX}_{im}, +B^{MAX}_{im}]$. We privilege RBs suffering less interference level. Constraint (22) implies that $\mathcal{S}^p_{i'mb} = 0,\ \forall\ i' \neq i$. Finally UE_i is assigned a random split k and a random velocity $\mathcal{V}^p_{im}$ in $[-B^{MAX}_{im}, +B^{MAX}_{im}]$.

B. Slice Allocation Based on Particle Swarm Optimization

Iteratively, each particle p evolves towards a new position $\mathcal{S}_p$ after updating its velocity $\mathcal{V}^p$ as stated in equation $(E21)$ below. The velocity update process is formulated in equation $(E22)$, where the new velocity is constructed based on the current velocity $\mathcal{V}^p$, current position $\mathcal{S}^p$, $\mathcal{S}^{L,p}$ and $\mathcal{S}^G$. Wherein, we integrate the coefficient ε to improve the random nature of the evolution process. More specifically, we define the first action, i.e., following the best local particle $\mathcal{S}^{L,p}$ with probability ε and a second action, i.e., following the best global particle $\mathcal{S}^G$ with probability $1-\varepsilon$.

$$\mathcal{S}^p = \mathcal{S}^p \oplus \mathcal{V}^p \tag{E21}$$

$$\mathcal{V}^p = \mathcal{V}^p \cap \left[\varepsilon \otimes \left(\mathcal{S}^{L,p} \ominus \mathcal{S}^p\right) + (1-\varepsilon) \otimes \left(\mathcal{S}^G \ominus \mathcal{S}^p\right)\right] \tag{E22}$$

Algorithm 1: `RAN-USA`

Inputs: I^{MAX}, C^D_{MAX}, C^C_{MAX}, R_{MAX}, P, E_{MAX}, ε
$\lambda_i, \nu_i, \forall\ i \in \{1, .., N\}, \forall\ m \in \{1, .., M\}$
$\alpha, \beta, \gamma, \theta, \mu, \nu_k, g_k, f_k, \forall\ k \in \{0, .., K\}$
Output: $\mathcal{S}^G$ with the best utility function from $\mathcal{LP}_3$
Begin

```
for p = 1 to P do
    for i = 1 to N do
        m ← random (M)
        n_RB ← random ([−B^MAX_im, +B^MAX_im])
        S^p_imb ← 1; n_RB times; {priory to RBs with less interference level}
        S^p_im,B+1 ← random (K)
        V^p_im ← random([−B^MAX_im, +B^MAX_im])
    end for
end for
while iter <E_MAX do
    for p = 1 to P do
        if U_F(S^p) >U_F(S^{L,p}) then
            S^{L,p} ← S^p
        end if
        if U_F(S^p) >U_F(S^G) then
            S^G ← U_F(S^p)
        end if
    end for
    for p = 1 to P do
        r ← random ([0, 1])
        for i = 1 to N do
            for m = 1 to M do
                if r < ε then
                    V̂^p_im ← Σ_{b=1}^{B} S^{L,p}_imb - Σ_{b=1}^{B} S^p_imb
                else
                    V̂^p_im ← Σ_{b=1}^{B} S^G_imb - Σ_{b=1}^{B} S^p_imb
                end if
                if r̃_im(V̂^p_im) > r̃_im(V^p_im) then
                    V^p_im ← V̂^p_im
                end if
            end for
        end for
    end for
    for p = 1 to P do
        for i = 1 to N do
            for m = 1 to M do
                n_RB ← Σ_{b=1}^{B} S^p_imb + V^p_im
                S^p_imb ← 1; n_RB times; {priory to RBs with less interference level}
            end for
        end for
        Run UFSS as described in Section V-B2
    end for
```

More specifically, we adopt the ε-greedy method to alternate between following i) the best local particle $\mathcal{S}^{L,p}$ with probability ε and ii) the best global particle $\mathcal{S}^G$ with probability $1-\varepsilon$. In doing so, the challenge is to find the balance between using local knowledge (exploitation) and

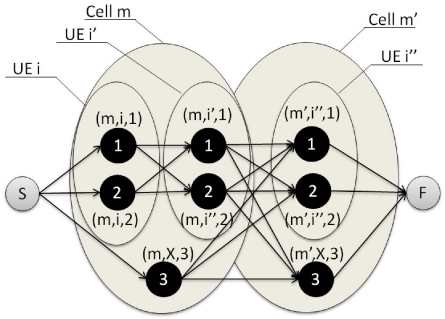


Fig. 3. *Graph G for* UFSS *algorithm.*

investigating other options by following the global knowledge (exploration). We believe that this is essential to ensure the investigation of the entire search space before converging to the near-optimal solution. RAN-USA is described in Algorithm 1.

1) Radio Resource Optimization: In this phase, we evaluate the utility function $\mathcal{U}_{\mathcal{F}}(\mathcal{S}^p)$ of each particle p and update $S^{L,p}$ and $\mathcal{S}^G$ accordingly. Afterwards, each particle p updates its velocity $\mathcal{V}^p$ and its new position $\mathcal{S}^p$. Then, our algorithm User Functional Split Selection, denoted by UFSS is performed to calculate the optimal split selection for each *UE* in particle p based on the newly generated radio configuration.

2) User Functional Split Selection Based on Shortest Path Algorithm (UFSS): In each iteration, once the $UE-gNB$ attachment and RB allocation is updated for particle p, our UFSS algorithm is executed to define the optimal split configuration. Then, we formulate the functional split selection optimal strategy as a shortest path problem. More specifically, we model all split possibilities as a Directed Acyclic Graph (DAG), G, with almost $N \times K$ nodes. Each node (m, i, k) is either a user split k for UE_i in gNB_m, or a cell split k for all *UE*s attached to gNB_m. Then, we consider only split node (m, i, k) which satisfy *UE* latency requirements, which receives links from other nodes with weights expressing the deployment cost from selecting the node (m, i, k). The weight of each ongoing link to node (m, i, k) is defined as:

$$\alpha \frac{\mathrm{C}^{\mathrm{D}}_{\mathrm{imk}}}{\mathrm{C}^{\mathrm{D}}_{\mathrm{MAX}}} + \beta \frac{\mathrm{C}^{\mathrm{C}}_{\mathrm{imk}}}{\mathrm{C}^{\mathrm{E}}_{\mathrm{MAX}}} + \gamma \frac{\mathrm{R}_{\mathrm{imk}}}{\mathrm{R}_{\mathrm{MAX}}} \qquad (E23)$$

The Graph G without link weights is depicted in Fig. 3, for $M = 2$ *gNBs*, $N = 3$ *UEs* and $K = 3$ splits. Note that there are two extra nodes: s and f. S is a starting point, that is connected to the all split possibilities of first UE_i which are nodes (m, i, k); $k \in \{1, .., 3\}$. And all split possibilities of last $UE_{i''}$ are connected to node f, where f is a finish point for the directed graph G with all the ongoing link weighted by zero. In Fig. 3, both UE_i and $UE_{i'}$ are attached to gNB_m and $UE_{i''}$ is attached to $gNB_{m'}$. Each *UE* can be assigned to only two user splits, i.e., 1 and 2, and one cell split, i.e., 3. Herein, node (m, i, k) denotes the selection of user split k for UE_i in gNB_m, while node (m, X, k) expresses the selection of cell split k to all *UEs* in gNB_m.

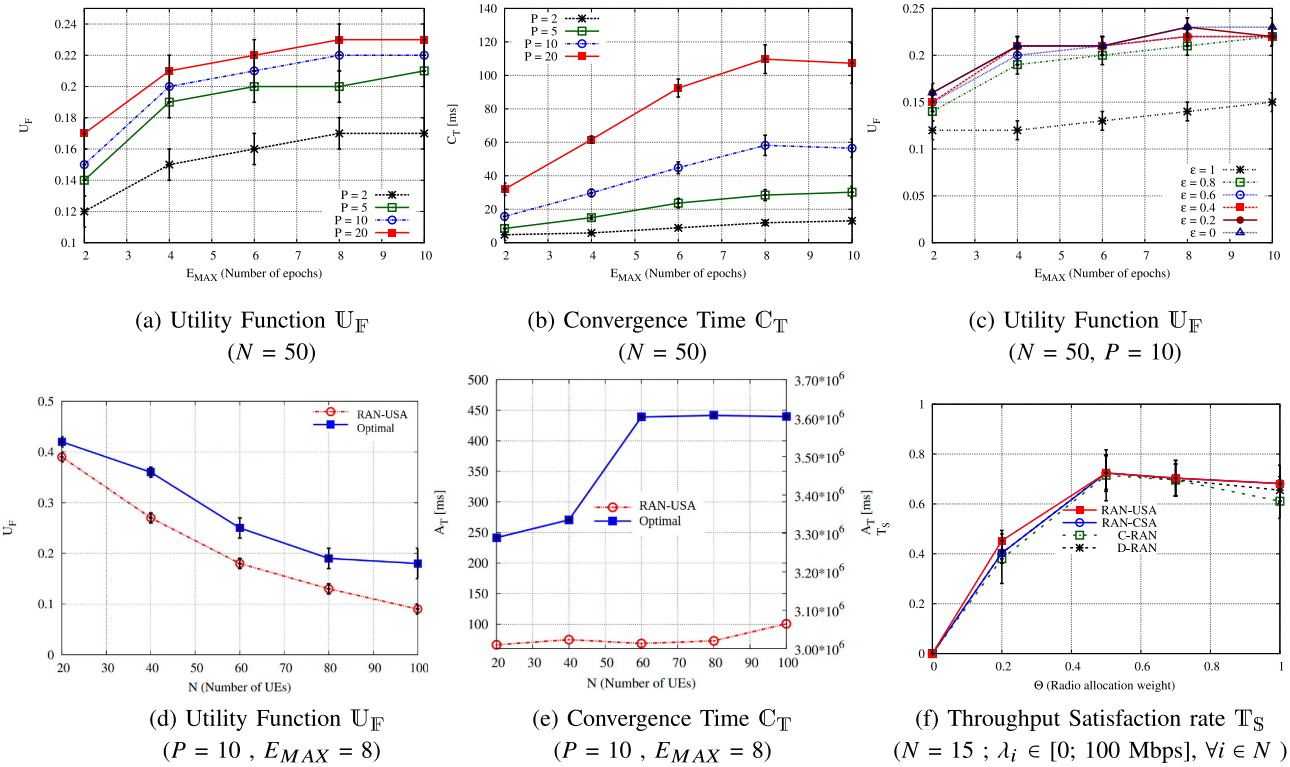
In doing so, path P from s to f in graph G, corresponds to a selection strategy of functional splits for *UE*s that are already attached to different *gNB*s with a given radio load. It is worth noting that, the sum of links' costs traversed by path P is equal to the deployment cost expressed in $\mathcal{LP}_1$. A Path $P^\star$ of minimum cost corresponds to the optimal functional split decision that minimizes the overall deployment cost. The problem of calculating the optimal functional split selection is equivalent to finding a min-cost path in a DAG. The latter is resolved through the Dijkstra algorithm in $O(|E| + |V|log|V|)$ time, where $|E|$ and $|V|$ are the number of edges and vertices. In our graph G, there exist $O(NK)$ nodes and $O(NK^2)$ links. So, finding the min-cost path takes $O(NK^2 + NKlog(NK))$. At the end, the entire RAN-USA algorithm with the UFSS approach runs with a complexity of $O(E_{MAX}\, PMB(NK)^2 logNK))$.

VI. Performance Evaluation

In this section, we gauge the performance of our proposed RAN-USA based on extensive simulations using our JAVA-based simulator. First, we describe the simulation environment setup and detail the various performance metrics. Then, we analyze the obtained results and discuss the effectiveness of our proposal compared with: i) commercial standard solvers such as IBM's ILOG CPLEX solver, ii) full Centralized deployment approach (i.e., C-RAN), iii) full Decentralized deployment approach (i.e., D-RAN), and iv) Cell-centric Split Allocation approach denoted by RAN-CSA. Note that the interference mitigation is inherently implemented in the C-RAN approach, while we assume that this mechanism is adopted for the D-RAN case. We set the number of split options K to 9, where $Split_6$, $Split_7$ and $Split_8$ are cell-centric, while $Split_0$, $Split_1$, $Split_2$, $Split_3$, $Split_4$ and $Split_5$ are user-centric. To the best of our knowledge, there is no simulator for RAN slice orchestration with user functional split selection deployment so far.

A. Simulation Setup

We simulate our Cloud-RAN infrastructure with respect to our model described in Section IV. We consider N *UE*s uniformly distributed in an OFDMA based cellular network. Table I reports the simulation parameters that have been used for our simulations [47]. Based on the radio parameters in Table I, and according to the work in [43], each cell has a downlink maximum throughput of 75 Mbps. Hence, for a network of 100 *UEs*, we assume that each *UE* requires a data rate in the range of [0, 1] Mbps. Note that, for a network of 7 *gNBs*, our simulations show that the system straggles to respond to all *UEs* demands in terms of throughput for N in [80, 100]. That is why we have fixed N to 100 users. In addition, according to [1], we assume that eMBB *UE*s require a latency in {1, 2, 3, 4} *ms*, while uRLLC *UE*s require a latency in {0.1, 0.2, 0.3, 0.4, 0.5} *ms*. Besides, *UEs* positions are randomly generated for each execution. Then, according to their positions, we calculate *CQI*, *MCS* and *TBSI* between each *UE* and *gNB* in order to approximate the linear function of generated TBS ($\widetilde{TBS}$) between each *UE* and *gNB*. It is worth noting that our obtained results correspond to the average of 30 simulations with a confidence interval set to 95% to approximate the average of calculated metrics according to their standard deviation.



(a) Utility Function $\mathbb{U}_{\mathbb{F}}$ ($N = 50$)

(b) Convergence Time $\mathbb{C}_{\mathbb{T}}$ ($N = 50$)

(c) Utility Function $\mathbb{U}_{\mathbb{F}}$ ($N = 50$, $P = 10$)

(d) Utility Function $\mathbb{U}_{\mathbb{F}}$ ($P = 10$, $E_{MAX} = 8$)

(e) Convergence Time $\mathbb{C}_{\mathbb{T}}$ ($P = 10$, $E_{MAX} = 8$)

(f) Throughput Satisfaction rate $\mathbb{T}_{\mathbb{S}}$ ($N = 15$; $\lambda_i \in [0; 100$ Mbps$], \forall i \in N$)

Fig. 4. Convergence evaluation.

TABLE I
SIMULATION PARAMETERS

Number of $gNBs$	M	7
Number of UEs	N	100
Inter-cell distance		50 m
Number of RBs	B	100
Spectrum Bandwidth	W	20 MHz
Antenna mode	A	1, SISO
Average RB power	$\overline{P_{RB}}$	10 mW
Average cell power	$\overline{P_{gNB}}$	1 $Watt$
Transmit power gain	Gtx	8 dBi
Shadowing coefficient	Ω	5 dB
Thermal Noise		-174 dBm/Hz
SINR threshold	$SINR^{MAX}$	10 dB
Path loss model D in Km	PL	$148.1 + 37.6\ log(D)$
Fading coefficient	ρ	$U(0, 1)$
Channel gain	h	$10^{-PL/20} \cdot \sqrt{Gtx \cdot \Omega} \cdot \rho$
conversion coefficient	c_{14}	10^{-3}
Interference threshold	I^{MAX}	$\frac{h * \overline{P_{RB}}}{SINR^{MAX}} - \sigma^2$
Midhaul capacity	R_{MAX}	3686, 4 Mbps [43]
Computational capacity	C^D_{MAX} C^C_{MAX}	960 GOPS [41]

B. Performance Metrics

We rely on the following metrics to gauge the performance of our proposal RAN-USA compared with baseline strategies.

- $\mathbb{U}_{\mathbb{F}}$ is the Utility Function in LP_3 expressing the trade-off between the served throughput and the deployment cost.
- $\mathbb{C}_{\mathbb{T}}$ is the average Convergence Time for user-centric Allocation in *ms*.
- $\mathbb{T}_{\mathbb{S}}$ is the Throughput Satisfaction rate expressing the ratio between the overall served and requested throughputs.
- $\mathbb{C}_{\mathbb{D}}$ is the Cost of Deployment expressing the computational and link resource usage as defined in LP_1, which is expressed as the weighted sum of the resource usage in i) the DU sites weighted by α, ii) the CU sites weighted by β, and iii) the midhaul link weighted by γ.
- $\mathbb{L}_{\mathbb{T}}$ is the Latency penalty of Total UEs expressed as $\sum_i \frac{\nu_k - \nu_i}{\nu_i}$, $\forall k \in K, \forall i \in N$ where, ν_k is the latency of split k in the midhaul and ν_i is the required latency from UE_i.
- $\mathbb{S}$ corresponds to the percentage of Splits.

C. Simulation Results

1) Convergence Analysis: First, we evaluate the impact of the number of particles P and the number of epochs E_{MAX} on the solution quality (i.e., utility function $\mathbb{U}_{\mathbb{F}}$ and the convergence time $\mathbb{C}_{\mathbb{T}}$). Fig. 4(a) assesses the performance of RAN-USA with different swarm population size P, while varying the number of epochs E_{MAX}. Indeed, for a fixed number of UEs (i.e., $N = 50$), we can observe that the utility function $\mathbb{U}_{\mathbb{F}}$ of each swarm population is increasing when E_{MAX} grows up. Besides, it is straightforward to see that the size of P impacts the quality of solution. In particular, the curves corresponding to $P = 10$ and $P = 20$ have close values that outperform both $P = 5$ and $P = 2$. Then, it is interesting to see that $\mathbb{U}_{\mathbb{F}}$ keeps stable starting from $E_{MAX} = 8$.

In Fig. 4(b), we study the impact of the swarm population size P on the convergence time $\mathbb{C}_{\mathbb{T}}$. It is clear to see that

when the number of particles P increases, the convergence time $\mathbb{C}_{\mathbb{T}}$ increases as well. Such a behavior is predictable, as the solution quality is enhanced as soon as P is increased, which in turn, requires more computation time to solve the problem. In particular, the curve corresponding to $P = 20$ costs much more computational time than the curves corresponding to $P = 10$, $P = 5$ and $P = 2$. We fix P to 10 and E_{MAX} to 8.

Fig. 4(c) assesses the convergence behavior of RAN-USA with different values of ε, while varying $ITER_{MAX}$. Indeed, for a fixed number of UEs (i.e., $N = 50$) and a fixed number of particles (i.e., $P = 10$), we can observe that $\mathbb{U}_{\mathbb{F}}$ increases when $ITER_{MAX}$ increases. We recall that ε is the probability of a particle to follow the local best position according to the equation (E22). As depicted in Fig. 4(c), when $\varepsilon = 1$, i.e., particles only follow their best local positions, the algorithm struggles to find an optimal solution. Meanwhile, the solution quality is enhanced when ε is less than 0.8. This proves that particles need to collaborate with each other to fasten the convergence process. It is interesting to see that, when $ITER_{MAX}$ is lower than 8 epochs, the curves corresponding to $\varepsilon = 0.2$ outperforms the one corresponding to $\varepsilon = 0$. This can be explained that RAN-USA rather favors a trade-off between exploitation (ε) and exploration $(1 - \varepsilon)$ to achieve better results. Hence, in our subsequent experiments, we fix the balance point of exploitation-exploration, ε to 0.2.

In the following, we also vary the number of users N in $[20, 100]$ with a rate of uRLLC UEs equal to 40% in each iteration. We set P and E_{MAX} to 10 and 8, respectively. We aim to evaluate the performance of RAN-USA in case of high density of UEs. In Fig. 4(d), we compare RAN-USA approach with the optimized solution generated by the solver CPLEX. It is straightforward to see that our solution approach generates near optimal solutions when the number of UEs N is equal to 20. Whereas, when N is higher than 20, our proposed approach achieves a lower utility function with a gap of 28%.

With regards to scalability, Fig. 4(e) illustrates the average resolution time $\mathbb{A}_{\mathbb{T}}$ of the different strategies versus the number of UEs N. Note that the Transmission Time Interval (TTI) in C-RAN is equal to 1 millisecond according to [48]. It is straightforward to see that the non-scalable optimal solution takes a significantly longer time than RAN-USA to solve one instance of the optimization problem. Indeed, the optimal solution struggles to scale, as it takes several minutes to solve instances of N. In contrast, RAN-USA can easily solve any size of instance (i.e., N in [20, 100]) in the range of [66; 100] milliseconds. Eventually, RAN-USA is able to take an up-to-date decision and execute it after 100 TTI period. Unfortunately, Optimal-Split is not able to do so since its decision, once taken, will be already obsolete and hence not applicable.

Fig. 4(f) illustrates $\mathcal{T}_S$, with respect to the radio allocation weight (θ). Wherein, the radio configuration is scaled up to $M = 7$ gNBs with 8x8 MIMO mode. Then, we fixed the number of UEs N to 15, while they requesting a service throughput in the range of [0; 100] Mbps. Additionally, we assume that the radio allocation weight θ is increasing in the range of [0; 1] while the split allocation weight μ is decreasing in the range of [0; 1]. As depicted, the throughput satisfaction is almost enhanced while θ is increasing. Furthermore, $\mathcal{T}_S$ reaches its maximum value at 0.72. This is explained by the fact that, when the throughput demand is high, radio resources become scarce which makes the selection of the appropriate set of resource blocks extremely challenging. Note that RAN-USA outperforms the other baseline scenarios with 3.8%. This is explained by the fact that RAN-USA aims to optimize the computational and the link resource consumption, while lightly enhancing users throughput satisfaction rate.

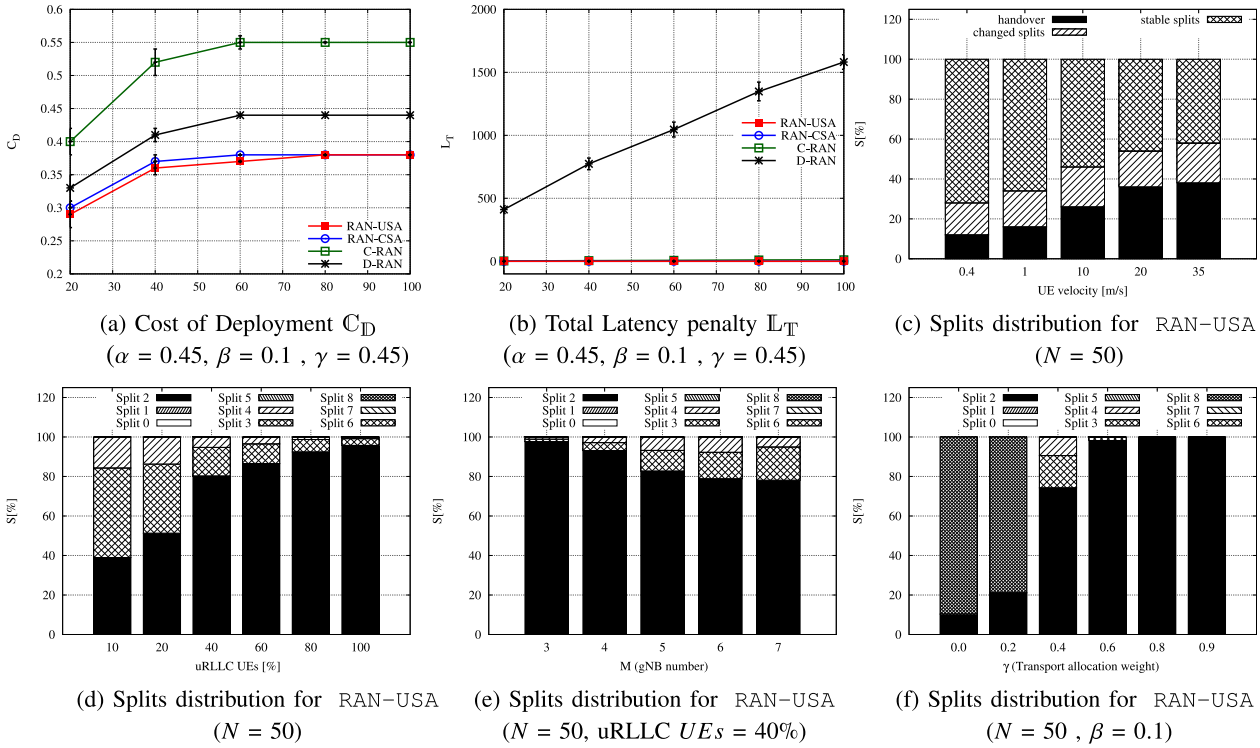
2) Performance Analysis: Hereafter, we fix θ and μ to 0.5 each to fairly weight both radio and virtual infrastructure allocation schemes, and we propose here to study the trade-off between the DU computational and the link resource allocation costs. We also fix the RCC computational consumption weight β to 0.1 as cloud data centers are natively efficient in power consumption. We assume that α and γ are both equal to 0.45 to emphasize the trade-off issue between minimizing DU computational cost weighted by α and optimizing the link resource usage weighted by γ.

In Fig. 5(a), we illustrate C_D with respect to the number of *UE*s. It is straightforward to see that, our approach RAN-USA further optimizes the computational and link resource usage cost comparing to baseline approaches. Indeed, our proposal is user-centric, hence, it adopts a fine grained approach to optimize the resource allocation. It is worth pointing out that C-RAN and D-RAN achieve higher cost of deployment. As a matter of fact, the C-RAN approach allocates constantly the full transport link bandwidth, while, the D-RAN approach utilizes all the computational resources in DU sites.

Fig. 5(b) illustrates the penalty $\mathcal{L}_T$ as a function of the *UE*s' number. As we can see, both RAN-USA and RAN-CSA approaches keep a zero penalty, which means that all *UE*s are constantly served with splits satisfying their latency requirement. However, the D-RAN approach causes high penalty because all *UE*s are served with $Split_0$ that implies a latency in the order of 10 *ms*, which obviously violates the latency requirements of both eMBB and uRLLC *UE*s. C-RAN also implies a latency penalty in the range of [1; 5] for some uRLLC *UEs* requiring a latency less than 0.2 ms.

Fig. 5(c) depicts the splits distribution for RAN-USA, while increasing the *UE*s' velocity in the range of [0; 35 m/s] for a fixed number of *UE*s (i.e., $N = 50$). Note that the velocity increases to simulate a pedestrian user (0.4 m/s), up to a high mobility use case (35 m/s). We can observe that the number of handovers increases obviously with the increase of the *UE*s' velocity. At the same time, the number of stable splits decreases since new split configuration options become possible to re-establish the balance point between radio allocation and energy consumption according to new users' position and resource availability. However, triggering such a solution at each user event (arrival/departure/mobility) is clearly not practical since the allocation will impact all existing users, along with the required service performance. Therefore, we propose to perform our optimization process in proactive manner and to trigger it periodically after a predefined time period $T = 100$ ms according to Fig. 4(e).

Fig. 5(d) assesses the split selection strategy of RAN-USA with different percentage of uRLLC *UEs*. Indeed, for a fixed number of UEs (i.e., $N = 50$) and a fixed number of *gNBs* (i.e., $M = 7$), it is straightforward to see that, our approach



(a) Cost of Deployment $\mathbb{C}_{\mathbb{D}}$ (α = 0.45, β = 0.1 , γ = 0.45)

(b) Total Latency penalty $\mathbb{L}_{\mathbb{T}}$ (α = 0.45, β = 0.1 , γ = 0.45)

(c) Splits distribution for RAN−USA (N = 50)

(d) Splits distribution for RAN−USA (N = 50)

(e) Splits distribution for RAN−USA (N = 50, uRLLC UEs = 40%)

(f) Splits distribution for RAN−USA (N = 50 , β = 0.1)

Fig. 5. Performance evaluation.

favors $Split_2$, $Split_3$ and $Split_4$. $Split_0$ and $Split_1$ are excluded since they induce a latency of 10 *ms*, which does not satisfy neither eMBB nor uRLLC flows. Furthermore, $Split_5$, $Split_6$, $Split_7$ and $Split_8$ are not selected because they generate a high traffic in the midhaul, which impacts the deployment cost. Instead, our approach achieves a trade-off between DU computational usage and link resource usage by adopting a partial centralization scheme. Specifically, $Split_2$ increases proportionally with the uRLLC *UE*s density. This emphases the fact that *UEs* with stringent latency requirement less that 0.2 ms restraint *gNBs* to deploy only $Split_2$, excluding necessary other split options even for other attached *UEs*. With reference to Section III-A, $Split_2$ leads to a high computational deployment cost comparing to other feasible splits. To counteract this side-effect, RAN−USA selects $Split_3$ and $Split_4$ in other *gNBs* to centralize more functions in the cloud.

Fig. 5(e) illustrates the impact of the *gNB* number (i.e., *M*) on the split selection strategy $\mathbb{S}$. For a fixed number of UEs (i.e., $N = 50$), uRLLC *UEs* percentage is fixed to 40% and *M* in the range of [3; 7], we can see that the deployment of $Split_2$ decreases, while the adoption for $Split_3$ and $Split_4$ increases. The reason behind this is that, RAN−USA is not anymore constrained to deploy $Split_2$ in some *gNBs*. Instead, RAN−USA finds a greater flexibility to deploy other splits to achieve the trade-off between DU computational usage and link resource usage.

Finally, in Fig. 5(f), we study the trade-off between the DU computational cost, which is weighted by α and the link resource usage, which is weighted by γ. Therefore, we assume that γ is increasing in the range of [0; 1], while α is decreasing in the range of [0; 1]. As depicted in Fig. 5(f), our solution adopts $split_2$ and $split_8$ when γ is lower than 0.4 (i.e., α is higher than 0.6). Then, when γ is equal to 0.4, the algorithm adopts mainly $split_2$, $split_3$ and $split_4$ until γ reaches 0.6. Afterwards, $split_2$ is constantly deployed. The reason behind this behavior is that RAN−USA adopts splits with minimum DU computational cost when α is high (namely $split_8$), while $split_2$ is served for some uRLLC *UEs*. When γ is high, RAN−USA favors splits with minimum traffic flow in the midhaul (namely $split_2$). It is interesting to see that when γ is equal to 0.4 and α is fixed to 0.6, the trade-off is achieved by deploying simultaneously $split_2$, $split_3$ and $split_4$.

VII. Conclusion

5G-RAN stakeholders aim to build a RANaaS concept with innovative RAN infrastructure to address the new 5G applications requirements. In this context, the slice concept is introduced in order to handle the heterogeneity of new use-cases. Despite the great advances achieved by RAN functional split standardization, there is still a coarse grained approach in the deployment process. In this paper, we propose a RAN User-centric Slice Allocation approach RAN−USA. Wherein, each user is assigned a proportion of radio and a split option. At the end, multiple user slices are created and managed on top of the physical infrastructure tailored to users' requirements. Our contribution is twofold. First, we put forward an oriented service Framework for user slice allocation. Second, we propose a heuristic based on Particle Swarm Optimization that jointly optimizes radio, link and computational resource allocation. Based on Particle Swarm Optimization, RAN−USA is scalable and achieves optimized user-centric slice allocation solution in a satisfactory time. Based on extensive simulations, we have shown that RAN−USA achieves good performances in terms of total throughput satisfaction and deployment cost.

REFERENCES

[1] *Study on Scenarios and Requirements for Next Generation Access Technologies*, 3GPP Standard TS 38.913 V.14.3.0, Jul. 2017.

[2] *5G End-to-End Architecture Framework*, Next Gener. Mobile Netw. (NGMN) Alliance, Frankfurt, Germany, Sep. 2019.

[3] *5G; Management and Orchestration; Architecture Framework*, 3GPP Standard TS 28.533 V.15.1.0 Release 15, Apr. 2019.

[4] *NG-RAN; Architecture Description (Release 15)*, 3GPP Standard TS 38.401 V.15.5.0, May 2019.

[5] "C-RAN the road towards green RAN white paper," China Mobile Res. Inst., Beijing, China, White Paper, Oct. 2011.

[6] *Network Sharing; Architecture and Functional Description (Release 11)*, 3GPP Standard TS 23.251, Jan. 2009.

[7] S. Khatibi, L. Caeiro, L. S. Ferreira, L. M. Correia, and N. Nikaein, "Modelling and implementation of virtual radio resources management for 5G cloud RAN," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, p. 128, Jul. 2017.

[8] M. Mira, T. Chahed, L. Chen, J. Elias, and F. Martignon, "A two-level auction for resource allocation in multi-tenant C-RAN," *Comput. Netw.*, vol. 135, pp. 240–252, Apr. 2018.

[9] T. LeAnh, N. H. Tran, D. T. Ngo, and C. S. Hong, "Resource allocation for virtualized wireless networks with backhaul constraints," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 148–151, Jan. 2017.

[10] C. Gao, G. Ozcan, J. Tang, M. C. Gursoy, and W. Zhang, "R-cloud: A cloud framework for enabling radio-as-a-service over a wireless substrate," in *Proc. IEEE Int. Conf. Netw. Protocols (ICNP)*, Singapore, 2016, pp. 1–10.

[11] Y. Zhu, H. Yu, R. A. Berry, and C. Liu, "Cross-network prioritized sharing: An added value MVNO's perspective," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Paris, France, 2019, pp. 1549–1557.

[12] I. Koutsopoulos, "Optimal functional split selection and scheduling policies in 5G Radio Access Networks," in *Proc. IEEE ICC Workshops*, 2017, pp. 993–998.

[13] H. Hirayama, Y. Tsukamoto, S. Nanba, and K. Nishimura, "RAN slicing in multi-CU/DU architecture for 5G services," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, Honolulu, HI, USA, 2019, pp. 1–5.

[14] C. Song et al., "Hierarchical edge cloud enabling network slicing for 5G optical fronthaul," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 11, no. 4, pp. B60–B70, Apr. 2019.

[15] C.-Y. Chang and N. Nikaein, "RAN runtime slicing system for flexible and dynamic service execution environment," *IEEE Access*, vol. 6, pp. 34018–34042, 2018.

[16] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[17] H. Davit and R. Riggio, "Flexible functional split in 5G networks," in *Proc. IEEE Int. Conf. Netw. Serv. Manag.*, 2017, pp. 1–9.

[18] G. Tseliou, F. Adelantado, and C. Verikoukis, "NetSliC: Base station agnostic framework for network SliCing," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3820–3832, Apr. 2019.

[19] K. Park, R. Rothfeder, S. Petheram, F. Buaku, R. Ewing, and W. H. Greene, "Linear regression," in *Basic Quantitative Research Methods for Urban Planners*. New York, NY, USA: Routledge, 2020, ch. 12.

[20] X. Yu et al., "Set-based discrete particle swarm optimization based on decomposition for permutation-based multiobjective combinatorial optimization problems," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2139–2153, Jul. 2018.

[21] Y. L. Lee, J. Loo, and T. C. Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *Proc. Int. Conf. Adv. Electr. Electron. Syst. Eng.*, 2016, pp. 414–420.

[22] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, "Network slicing games: Enabling customization in multi-tenant mobile networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 662–675, Apr. 2019.

[23] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "OpenRAN: A software-defined RAN architecture via virtualization," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 549–550, 2013.

[24] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *Proc. ACM SIGCOMM Workshop Hot Topics Softw. Defined Netw.*, 2013, pp. 25–30.

[25] T. Chen, H. Zhang, X. Chen, and O. Tirkkonen, "SoftMobile: Control evolution for future heterogeneous mobile networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 70–78, Dec. 2014.

[26] M. Bansal, J. Mehlman, S. Katti, and P. Levis, "OpenRadio: A programmable wireless dataplane," in *Proc. 1st Workshop Hot Topics Softw. Defined Netw.*, 2012, pp. 109–114.

[27] W. Wu, L. E. Li, A. Panda, and S. Shenker, "PRAN: Programmable radio access networks," in *Proc. ACM Workshop Hot Topics Netw.*, 2014, pp. 1–7.

[28] 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 *Proc. Int. Conf. Emerg. Netw. EXp. Technol.*, New York, NY, USA, 2016, pp. 427–441.

[29] E. Coronado, S. N. Khan, and R. Riggio, "$5G - EmPOWER$: A software-defined networking platform for 5G radio access networks," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 2, pp. 715–728, Jun. 2019.

[30] "Network function virtualisation (NFV); reliability; report on the resilience of NFV-MANO critical capabilities, V.1.1.2," ETSI, Sophia Antipolis, France, ETSI Rep. ETSI GR NFV-REL 007, Sep. 2017.

[31] J. Bartelt et al., "5G transport network requirements for the next generation fronthaul interface," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, p. 89, May 2017.

[32] A. Rostami et al., "Orchestration of RAN and transport networks for 5G: An SDN approach," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 64–70, Apr. 2017.

[33] R. B. Williams, C.-T. Lee, and L.-P. Yuan, "Adding multi-tenant awareness to a network packet processing device on a software defined network (SDN)," U.S. Patent App. 16 417 684, 2015.

[34] B. Ojaghi, F. Adelantado, E. Kartsakli, A. Antonopoulos, and C. Verikoukis, "Sliced-RAN: Joint slicing and functional split in future 5G radio access networks," in *Proc. IEEE ICC*, 2019, pp. 1–6.

[35] L.-H. Shen, Y.-T. Huang, and K.-T. Feng, "CoMP-enhanced flexible functional split for mixed services in beyond 5G wireless networks," 2021. *arXiv:2112.04079*.

[36] S. Matoussi, I. Fajjari, N. Aitsaadi, and R. Langar, "User slicing scheme with functional split selection in 5G cloud-RAN," in *Proc. IEEE WCNC*, Seoul, South Korea, 2020, pp. 1–8.

[37] Y. Virginia, and A. Amandi, "Project scheduling: A multi-objective evolutionary algorithm that optimizes the effectiveness of human resources and the project makespan," *Eng. Optim.*, vol. 45, no. 1, pp. 45–64, 2013.

[38] J. Zhang, W. Wang, Y. Zhao, and C. Cattani, "Multiobjective quantum evolutionary algorithm for the vehicle routing problem with customer satisfaction," *Math. Problems Eng.*, vol. 2012, Dec. 2012, Art. no. 879614.

[39] "Technical specification group radio access network; study on new radio access technology: Radio access architecture and interfaces," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 38.801, Mar. 2017.

[40] S. Matoussi, I. Fajjari, S. Costanzo, N. Aitsaadi, and R. Langar, "5G RAN: Functional split orchestration optimization," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 7, pp. 1448–1463, Jul. 2020.

[41] U. Dötsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Tech. J.*, vol. 18, no. 1, pp. 105–128, Jun. 2013.

[42] D. Szczesny, A. Showk, S. Hessel, A. Bilgic, U. Hildebrand, and V. Frascolla, "Performance analysis of LTE protocol processing on an ARM based mobile platform," in *Proc. IEEE Int. Symp. System-on-Chip*, Tampere, Finland, 2009, pp. 56–63.

[43] *Small Cell Virtualization: Functional Splits and Use Cases Rel. 6.0*, Small Cell Forum, Dursley, U.K., Jan. 2016.

[44] J. Lee and S. Leffer, *Mixed Integer Nonlinear Programming*, vol. 154. New York, NY, USA: Springer, 2011.

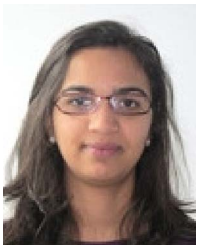
[45] *Physical Layer Procedures, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA)*, 3GPP Standard TS 36.213, May 2016.

[46] A. Mignotte and O. Peyran, "Reducing the complexity of ILP formulations for synthesis," in *Proc. Int. Symp. Syst. Synth.*, 1997, pp. 58–64.

[47] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "A dynamic resource allocation framework in LTE downlink for cloud-radio access network," *Comput. Netw.*, vol. 140, pp. 101–111, Jul. 2018.

[48] "Study on new radio access technology; radio interface protocol aspects (release 14)," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 38.804 V1.0.0, 2017.

Salma Matoussi received an Engineer Diploma degree (Hons.) and the M.S. degree in computer science and networking from the National School of Computer Science, Tunisia, in 2012 and 2013, respectively, and the Ph.D. degree in computer science from Sorbonne University, France, in 2021. She is a Research Data Engineer with Gandi SAS. Before joining Sorbonne University, she worked as a Software Engineer with Iliade Consulting and Focus International. Her research interests include data science and resource allocation optimization in a cloud environment.



Ilhem Fajjari received the Ph.D. degree (Hons.) in computer sciences from Pierre and Marie Curie University (Paris 6), France, in 2012. From 2012 to 2014, she worked as a Research Project Leader of Network Virtualization with VirtuOR Startup. She is a Research Project Leader of Cloud-Native Network Function Orchestration with Orange Labs. Her main research interests include cloud, network function virtualization, orchestration, and optimization of communication networks. She is the TPC Co-Chair of IEEE/IFIP CIoT'18. She served as a Co-Guest Editor of the Special Issue "Cloud Edge Computing in the IoT", *Annals of Telecommunications* (Springer). She is also an Active TPC Member in several international conferences, including IEEE ICC, IEEE GLOBECOM, and IEEE LCN.

Nadjib Aitsaadi (Member, IEEE) is a Full Professor of Computer Science with UVSQ Paris-Saclay University. He is the coauthor of many IEEE/IFIP major journals and conferences. Also, he is involved in many European projects, such as SARWS, TILAS, and GOLDFISH. His main research interest is the optimization on of QoS in cellular and wired networks. He is very active in the IEEE ComSoc Information Infrastructure and Networking Technical Committee and he is running now the chair position.

Rami Langar (Member, IEEE) received the M.Sc. degree in network and computer science from University Pierre and Marie Curie (currently, Sorbonne University) in 2002, and the Ph.D. degree in network and computer science from Telecom ParisTech, Paris, France, in 2006. He was a Postdoctoral Research Fellow with the School of Computer Science, University of Waterloo, Waterloo, ON, Canada, from 2006 to 2008, and an Associate Professor with LIP6, University Pierre and Marie Curie from 2008 to 2016. He has been currently a Full Professor affiliated with University Gustave Eiffel, France, and École de Technologie Supérieure, Montreal, Canada, since 2016 and 2021, respectively. His research interests include resource management in future wireless systems, network slicing in 5G/5G+/6G, software-defined wireless networks, mobile cloud offloading, and green networking. He was the Chair of the IEEE ComSoc Technical Committee on Information Infrastructure and Networking from January 2018 to December 2019.