A Coverage-Aware VNF Placement and Resource Allocation Approach for Disaggregated vRANs

Victor Hugo L. Lopes*‡, Gabriel Matheus Almeida*, Aldebaro Klautau†, Kleber Cardoso*
*Universidade Federal de Goiás, Brazil, †Universidade Federal do Pará, Brazil.
‡Instituto Federal de Goiás, Brazil.
E-mail: {victor.lopes, gabrielmatheus, kleber}@inf.ufg.br*, aldebaro@ufpa.br†

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10000776

***Abstract*—Disaggregated and virtualized RANs (vRANs) offer the opportunity for flexible and efficient use of computing resources through the proper placement of the RAN Virtualized Network Functions (VNFs). However, many works neglect the necessary coordination between VNF placement and the processing of the RAN tasks inside these VNFs. This can negatively impact important tasks such as resource scheduling and interference control. In this work, we introduce a new approach for VNF placement that is aware of the wireless coverage and its associated tasks. Our solution was designed in the context of O-RAN architecture, exploring functionalities of monitoring and closed-loop decision making. Simulation results illustrate the benefits of our solution, mainly related to improvements for edge users who are exposed to the worse conditions of spectral efficiency and throughput.**

***Index Terms*—Virtualized RAN, VNF placement, functional splits, interference control, resource allocation.**

I. Introduction

Several research and standardization efforts have been carried out in order to overcome the limitations of the monolithic Radio Access Network (RAN) deployments. One of these efforts resulted in the proposal of the Cloud RAN architecture (C-RAN), in which the most part of the baseband processing functions is centralized in virtualized cloud data centers so that they can be deployed far from the radio units and integrated via high-speed transport networks.

As in the most diverse areas, wireless communication systems are increasingly impacted by the virtualization. The Network Function Virtualization (NFV) paradigm enabled the use of VNFs in order to build virtualized RANs (vRANs). Thus, it becomes attractive to implement access networks with architectures that provide a more robust means for resource management, in which the aggregation of VNFs allows important improvements, such as in signal processing [1] and interference mitigation. Such evolution contributed to the emergence of initiatives towards the use of open and standardized interfaces, as seen in the O-RAN architecture [1], [2].

The set of decisions that need to be taken in order to obtain the proper positioning of the VNFs of these new disaggregated RANs (i.e., making efficient use of the resources) and also to maximize the spectral efficiency become challenging. While this issue has already been tackled in literature, the state-of-the-art [3] still has some relevant limitations. First, a simplified version of the vRAN is adopted in terms of routing and availability of computing resources. Second, O-RAN architecture is not taken into consideration, neglecting its components, the monitoring aspects, and machine learning closed loops. In this context, we introduce a new solution for VNF placement and resource allocation based on a sophisticated framework [4] which is modified to take into account the channel quality of the UEs. Additionally, our proposal is designed to be fully adherent to O-RAN specifications.

The rest of this paper is organized as follows. Section II presents a brief overview of the most relevant concepts and the related work. Section III introduces the system model and the problem formulation. Section IV describes the proposed solution. In Section V, we present and discuss the results of the performance evaluation. Section VI presents final remarks and future work.

II. Background and Related Work

Coordinated Multi-Point transmission: As a way to improve the performance of 4G networks, a series of new technologies were introduced in LTE-A, mainly with a focus on optimizing the performance in both uplink (UL) and Downlink (DL), such as heterogeneous networks, carrier aggregation, and Multiple Input Multiple Output (MIMO). Such technologies, among others, enabled new deployment scenarios. However, due to the emergence of new challenges involving interference mitigation in these scenarios, another technology was also included, the Coordinated Multi-Point transmission/reception (CoMP). CoMP is a technique applied both in UL and DL that aims to provide better performance of the network capacity as a whole and in particular to provide greater performance to users located at the edges of the cells, suffering impacts from inter-cells interference.

From an architectural perspective, two CoMP approaches are considered, one in which the control occurs autonomously and decentralized in each BS, and another in which it employs centralized control mechanisms. In the first approach, CoMP employs a signaling channel between BSs and is affected by both signaling delay and generated overhead. In the second approach, a set of BSs is selected to compose a cluster (or CoMP set), associated with an anchor BS, responsible for the centralized management of all radio resources, baseband data transmission, and control decisions via transport network links. In this type of CoMP, the delay and overhead can be reduced,

and the management of intra-cell radio resources is facilitated. However, the size of the cluster is limited by the capabilities of the anchor BS [3] and the transport network links.

When CoMP is employed in DL, two different approaches can be adopted: i) Coordinated Scheduling (CS) / Coordinated Beamforming (CB); and ii) Joint Transmission. In both cases, transmissions for each UE continue to be carried out individually by the associated BS, as done when CoMP is not employed, but the scheduling decisions for transmissions are dynamically coordinated by the cluster anchor. Joint transmission is not considered in this work.

VNF placement: In disaggregated vRANs, it is possible to divide the RAN protocol stack into up to three parts, which run on three RAN nodes: central unit (CU), distributed unit (DU), and radio unit (RU). Such divisions are called functional splits and enable a set of features that improve network management [5]. Functional splints allow virtualized network functions to run on different RAN nodes spread across the network. In addition, thanks to the virtualization of RAN functionalities, it is possible to run virtualized functions on general-purpose hardware (denoted as computing resources - CRs), bringing several benefits to mobile operators in terms of operational cost and deployment.

The centralization of VNFs is widely investigated in the literature, being a regular objective of several RAN efficiency works. In this way, the vRAN placement problem consists of positioning the stack of virtualized network functions from all RUs on the network, considering the requirements of functional splits and the capacity constraints of the network nodes, e.g., link capacities, latency, and processing capacity. Each functional split has throughput and latency requirements based on the proximity of the radio functions, i.e., splits closer to the radio function (RF) demand more resources, while splits closer to the top of the stack demand less resources.

In this way, the vRAN placement problem can be defined as an optimization problem that pursues an objective (e.g. centralization and energy efficiency) subject to the transport network shared resources, the functional splits requirements, and the CRs processing capacities.

Related work: Methods for VNF placement taking into account aspects of the cell coverage area have recently attracted attention [3], [4], [6]–[8]. However, there are several aspects that need to be jointly considered in order to make the best use of the benefits arising from these architectures. Thus, it is observed that most recent studies in the literature deal with these aspects separately, as summarized in Table I. Thus, such studies may not be applicable to flexible deployment scenarios, as considered in our proposal, in which they depend on predetermined CU-DUs sets, based on dedicated transport networks, or deal only with fixed functional splits.

The most related work is [3], where five design factors are jointly considered. Although they also deal with clustering, in order to optimize network efficiency by performing CoMP, they must impose changes in the form of the RAN operation, as in defining the policy for user association, which makes its adoption in architecture such as O-RAN difficult. Our approach is fully flexible, as we consider CRs that can perform as any RAN node, even concurrently when serving RUs in disjoint clusters. Additionally, because we consider cross-haul, we perform the end-to-end routing, enabling functional splits not just for the fronthaul.

TABLE I: Related work overview.

Aspects	[4]	[6]	[7]	[8]	[3]	This Work
CoMP			X		X	X
Dynamic clustering		X	X	X	X	X
Channel measurement		X	X		X	X
O-RAN compliant	X					X
Flexible VNF Placement	X					X

III. System Model and Problem Formulation

The considered system model is presented in Fig. 1, and follows the 3GPP standardizations [9], and the O-RAN specifications [2]. From the cellular network perspective, it comprises a set of three-sectors cells sites, composed of disaggregated BSs, operating in a frequency reuse factor one. Thus, although this adopted reuse factor tends to allow spectral efficiency improvements, cell-edge users tend to suffer greater degradation in signal quality, especially due to inter-cell interference.

Regarding the vRAN domain, the considered topology is composed of: i) a set of $\mathcal{B} = \{b_1, b_2, ..., b_{|\mathcal{B}|}\}$ RUs that host the Low PHY sublayer and the RF functionalities required by the lower layer functional split; ii) a set of $\mathcal{C} = \{c_1, c_2, ..., c_{|\mathcal{C}|}\}$ CRs that can process the allocated VNFs, according to the c_m^{Proc} computing processing capacity available in each CR; and iii) a set of $\mathcal{T} = \{t_1, t_2, ..., t_{|\mathcal{T}|}\}$ transport nodes, which may connect to the network core, to any RUs and/or CRs, or each other. It is important to note that although each CR can also be described through its RAM and storage capacities, the processing capacity tends to be exhausted before the others, which can generate impacts on the network delay.

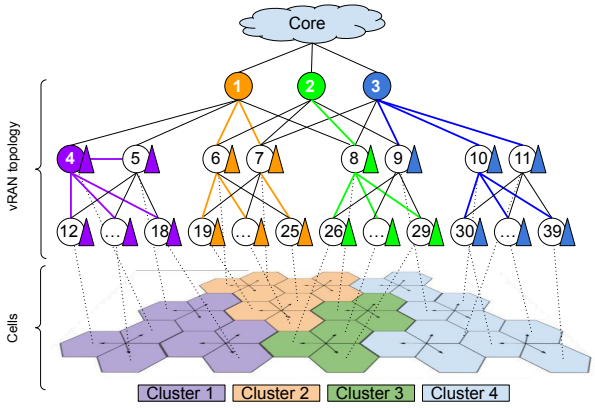

Fig. 1: Overview of the coverage-aware disaggregated vRAN.

The network flow has as source the core of the network and as destination the RU, when operating in DL, or the reverse flow for UL. Without loss of generality, although this work only considers the DL operation, it can also be applied in the UL, when considering the necessary changes in the other components of the system. Thus, the network

flow must follow one path p from the set $\mathcal{P}_l$ of the k shortest paths from the network core to each destination RU $b_l \in \mathcal{B}$. Each path $p \in \mathcal{P}_l$ is formed by three subpaths that form the transport network, namely, backhaul (p_{BH}), midhaul (p_{MH}), and fronthaul (p_{FH}), where at least one of the subpaths is non-empty.

We consider the RAN protocol stack as VNFs (except the RF protocol, as described previously), labeled in increasing order, starting from PHY Low as f_1 and ending at RRC as f_8, where we can define $\mathcal{F} = \{f_1, ..., f_8\}$ as the set of disaggregated RAN VNFs that must be allocated correctly. In order to do so, we define a set $\mathcal{D} = \{D_1, D_2, ..., D_{|\mathcal{D}|}\}$ of industry disaggregated RAN combinations (DRCs), in line with [4], as depicted in Fig. 2.

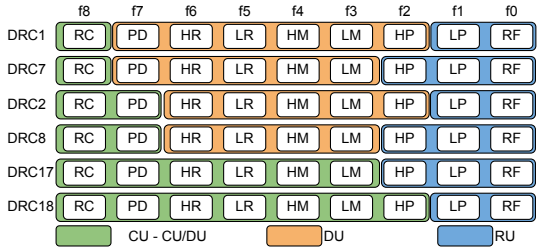

Fig. 2: DRCs that enable CoMP operations [4].

A. Computational resource consumption model

From [10] we can calculate the computational resource consumption (i.e., the computational resource cost), in terms of the amount of processing Giga Operations per Second (GOPS) for all VNFs in each functional split $s \in |S|$ allocated at the CR $c_m \in |C|$ in the TTI t, using the equation

$$\mathcal{Q}_{s,t}^{c_m} = \sum_{f=1}^{F} \sum_{u=1}^{U} \alpha_{s,f} \left(3a + a^2 + \frac{o_u r_u l}{3}\right) \frac{\Gamma_u}{5}, \tag{1}$$

where a defines the number of RU antennas, o_u and r_u are the maximum modulation order and the coding rate for the user u, respectively, l is the number of RU MIMO layers, and $\Gamma_u \in \mathbb{Z}^+$ is the amount of resource blocks (RBs) allocated for the user u, and $\alpha_{s,f} \in \mathbb{R}^+$ is the scaling factor which maps the computational complexity of the VNF f running on each node of the vRAN in relation to the total computational resources required by the functional split s. According to Eq. (1) the selected modulation and coding scheme (MCS) (from the reported channel quality indicator - CQI) for each UE is directly proportional to the computational consumption of the allocated functional split, and should not be neglected when planning and managing the disaggregated vRAN.

B. Problem formulation

To formulate the problem, we define the decision variable $x_l^{p,r} \in \{0, 1\}$, which represents the allocation configuration to serve the RU $b_l \in \mathcal{B}$ with the DRC $D_r \in \mathcal{D}$ through the path $p \in \mathcal{P}_l$. Our objective is to allocate all BSs functions while minimizing the number of clusters in the vRAN, subject to the functional splits requirements (latency and throughput), the CRs processing capacity and links capacity. Table II summarizes the parameters and variables used in the system model, problem formulation, and solution. Therefore, to represent the objective, we define the following equation:

$$\Phi = \sum_{c_m \in \mathcal{C}} \left\lceil \frac{\sum_{b_l \in \mathcal{B}} \sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} (x_l^{p,r} \zeta(b_l, c_m))}{|\mathcal{B}|} \right\rceil \tag{2}$$

Equation (2) represents the number of clusters generated by the solution. $\zeta = \{0, 1\}$ represents whether RU b_l is part of the cluster anchored by CR c_m. Note that the numerator part represents the c_m cluster size. However, the information we are looking for is whether or not there is a cluster in c_m. In this case, we use the ceiling function to get such information using $|\mathcal{B}|$ as the denominator, i.e., the maximum possible size of a cluster, which is the largest possible number in the numerator.

As previously described, in this work we take into account the RU coverage area, aiming to guarantee better results in terms of interference control through CoMP. In this way, we define a subset of RUs $\mathcal{B}_l \subseteq \mathcal{B}$ that contains all the neighboring RUs of RU b_l. The neighborhood relationship is performed through a graph $G^* = \{V^*, E^*\}$ passed to the model as input, where its vertices are the RUs and the edges represent the interference between the RUs, i.e., if a pair of RUs interfere with each other, then there is an edge between their respective vertices.

We use this information to define cohesively clusters, considering their RUs coverage area. We define that the subset of RUs of a cluster must generate a connected subgraph in the graph G^*, i.e, the coverage area of the RUs of a cluster must be continuous. This is represented by the following equations:

$$x_l^{p,r} \zeta(b_l, c_m) \leq \sum_{b_i \in \mathcal{B}_l} x_i^{p,r} \zeta(b_i, c_m) \quad \forall c_m \in \mathcal{C}, b_l \in \mathcal{B}, p \in \mathcal{P}_l, r \in \mathcal{D} \tag{3}$$

$$\zeta(b_i, c_m)\zeta(b_l, c_m) \leq P(b_i, b_j, c_m) \quad \forall b_i, b_l \in \mathcal{B}, c_m \in \mathcal{C} \tag{4}$$

Equation (3) defines that a RU must belong to a cluster that has at least one of its neighboring RUs (except in cases of D-RAN). Equation (4) defines that every pair of RUs in a cluster must form a connected subgraph in G^*.

To define the constraints on network resources and functional splits requirements, we use our formulation presented in [4]. The equations considered are presented below:

$$\sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} x_l^{p,r} = 1, \quad \forall b_l \in \mathcal{B} \tag{5}$$

$$\sum_{D_r \in \mathcal{D}} \sum_{b_l \in B} \sum_{p \in P_l} \left[x_l^{p,r} \left(y_{e_{ij}}^{p_{Bh}} \alpha_{Bh}^r + y_{e_{ij}}^{p_{Mh}} \alpha_{Mh}^r + y_{e_{ij}}^{p_{Fh}} \alpha_{Fh}^r \right) \right] \leq e_{ij}^{Cap}, \quad \forall e_{ij} \in \mathcal{E} \tag{6}$$

$$\sum_{e_{ij}\in\mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Bh}} e_{ij}^{Lat} \leq \beta_{Bh}^r, \quad \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D} \quad (7)$$

$$\sum_{e_{ij}\in\mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Mh}} e_{ij}^{Lat} \leq \beta_{Mh}^r, \quad \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D} \quad (8)$$

$$\sum_{e_{ij}\in\mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Fh}} e_{ij}^{Lat} \leq \beta_{Fh}^r, \quad \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D} \quad (9)$$

$$\sum_{f_s\in\mathcal{F}} \sum_{b_l\in\mathcal{B}} \sum_{D_r\in\mathcal{D}} \sum_{p\in\mathcal{P}_l} x_l^{p,r} u_m^p M(c_m, f_s, b_l) \gamma_m^s \leq c_m^{Proc}, \quad \forall c_m \in \mathcal{C} \quad (10)$$

Equation (5) represents that only one path and DRC configuration can be chosen for an RU. Equation (6) represents that the transmitting capacity of every link must not be exceeded. Equations (7) - (9) represent the requirement of functional splits regarding latency in backhaul, midhaul and fronthaul, respectively. Finally, Equation (10) represents that the CRs processing capacity must not be exceeded.

TABLE II: Parameters and variables.

Parameter	Definition
$\mathcal{B}$	Set of BSs (RUs)
$\mathcal{B}_l$	Subset of RUs neighboring RU b_l
$\mathcal{C}$	Set of CRs
$\mathcal{T}$	Set of transport nodes
$\mathcal{E}$	Set of links
e_{ij}^{Cap}	Transmission capacity of link e_{ij}
e_{ij}^{Lat}	Estimated latency e_{ij}
$\mathcal{P}_l$	Set of k-shortest paths
$\mathcal{F}$	Set of disaggregated RAN VNFs
$\mathcal{D}$	Set of VNCs
u_p^m	Indicates if $c_m \in \mathcal{C}$ is part of $p \in \mathcal{P}_l$
$M(c_m, f_s, b_l)$	Indicates if $c_m \in \mathcal{C}$ runs $f_s \in \mathcal{F}$ from $b_l \in \mathcal{B}$
$y_{e_{ij}}^{p_{BH}}$	Indicates if e_{ij} is part of the backhaul
$y_{e_{ij}}^{p_{MH}}$	Indicates if e_{ij} is part of the midhaul
$y_{e_{ij}}^{p_{FH}}$	Indicates if e_{ij} is part of the fronthaul
α_{BH}^r	The associated demands for bitrate in the backhaul for $D_r \in \mathcal{D}$
α_{MH}^r	The associated demands for bitrate in the midhaul for $D_r \in \mathcal{D}$
α_{FH}^r	The associated demands for bitrate in the fronthaul for $D_r \in \mathcal{D}$
β_{BH}^r	The maximum latency tolerated in the backhaul for $D_r \in \mathcal{D}$
β_{MH}^r	The maximum latency tolerated in the midhaul for $D_r \in \mathcal{D}$
β_{FH}^r	The maximum latency tolerated in the fronthaul for $D_r \in \mathcal{D}$
γ_m^s	The computing demand of $f_s \in \mathcal{F}$
c_m^{Proc}	Processing capacity of $c_m \in \mathcal{C}$
$\zeta(b_l, c_m)$	Indicates if RU b_l is is in the c_m cluster
$P(b_i, b_j, c_m)$	Indicates if there is a path from b_i to b_j between the c_m cluster vertices.
Variable	**Definition**
$x_l^{p,r} = \{0, 1\}$	Represents which pair of $p \in \mathcal{P}_l$ and $D_r \in \mathcal{D}$ is selected to serve $b_l \in \mathcal{B}$

Our optimization model can be represented as follows:

minimize Φ
subject to :
Constraints (3) – (10)

IV. The Proposed Solution

The O-RAN architecture has two functional components, the non-real-time (non-RT) radio intelligent controller (RIC) and the near-RT RIC [11]. While the former supports tasks with longer timescales (seconds or minutes), the latter deal with shorter timescales tasks (between 10 ms to 1 s). The non-RT RIC comprises the non-RT RIC framework, that exposes the services of the non-RT RIC applications (rApps), and the near-RT RIC supports multiple third-part applications (xApps). The rApps and the xApps are modular applications that can provide specialized services, supporting both the management tasks, as well the necessary adaptations for the vRAN enhancements. In this sense, the proposed solution involves the inclusion of two new components to the network architecture. Such components are presented in a high-level view in Fig. 3: i) the VNF Controller (VNF-C), that can be implemented as a xApp; and ii) the VNF Placement (VNF-P), working as a rApp.

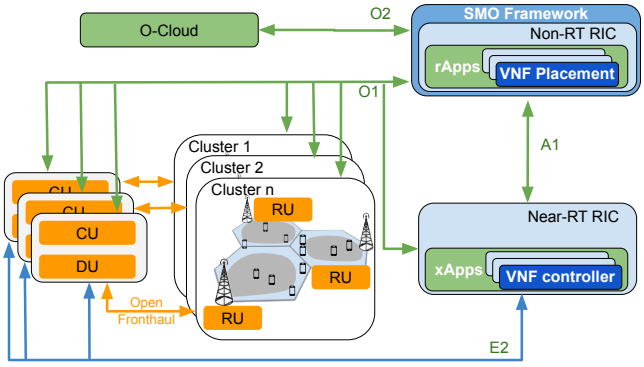

Fig. 3: The proposed solution overview, with the inclusion of the rApp and the xApp.

The adopted architecture enables the three independent control loops required by the O-RAN, which, in addition to the non-RT and near-RT loops, require real-time control operations responsible for legacy radio operations and radio resource management, including user/resource scheduling, HARQ operations, for example, and all tasks related to the joint operation of the CU, DU, and RU nodes.

The VNF-C is responsible for monitoring the topology in terms of the computational capacity usage of the CRs, as well as the average CQIs reported by each associated UE. The monitoring task can generate both periodic and aperiodic reports.

An aperiodic report is generated to inform about clusters where the CR running DU/CU-DU VNFs has exhausted its processing capacity. When this occurs, in order to not generate delays in the VNFs processing time (computational outage [12]) and, consequently, degradation of the system metrics, the CQI downgrade procedure must be employed. In this procedure, the maximum CQI index assigned for each UE within the cluster is limited to CQI_{max}. According to the Eq. (1), the CQI_{max} value needs to be iteratively reduced until the required computational consumption fits the available computational capacity. Therefore, the CQI downgrade seeks

to avoid affecting edge UEs, as it starts affecting those with higher CQIs.

Concerning the periodic reports, the VNF-C performs the measurement of the topology parameters for every Δt elapsed TTIs. In this way, the focus is to inform about situations of a decrease in the computational demand of the CRs, at levels where a new DRC can be added to the cluster. For this, the average CQI reported by the UEs in the last Δt measurement is considered. Again using the Eq. (1), we can get the drop in average CQI needed to add a new DRC to the cluster.

In both cases, once a new VNF placement solution must be established, the VNF-C is responsible for making an asynchronous call to the VNF-P. When the VNC-P is called, it obtains the inputs to the problem, as described in the optimization model formulated in Sec. III, and starts the search for the optimal VNF placement solution. As soon as a new solution arrives, it is immediately deployed if, and only if, it is more suitable for the current situation, i.e., if it has a smaller number of clusters (for the periodic event) or with adequate computational capacity (for the aperiodic event). It is important to note that the VNF-C is placed into the near-RT RIC as the information under monitoring may incur delays, especially with regard to the tasks involved in measuring and reporting CQIs.

V. Evaluation

To compare the performance of the proposal, simulations were performed considering a very detailed environment, composed of a communication system, responsible for the network simulation, including the entire wireless propagation environment, and the optimization system. The communication system was implemented using the Vienna LTE-A downlink system level simulator [13]. The VNF-C xApp was implemented in Matlab®, which can read and change the CQI report tables of the communication system simulator, in order to evaluate the performance of the CQI downgrade, as well as to calculate the computational demand required by each cluster. The VNF-P rApp was implemented em python, integrated into the CPLEX® solver running the formulated problem described in Sec. III.

In order to check the network performance against VNF placement solutions with different cluster sizes, Fig. 4 presents the spectral efficiency per channel use (bits/cu), observed for both edge and non-edge UEs. The *No-CoMP* solution represents the case where each BS is defined as a D-RAN, so that CoMP cannot be performed. The other solutions vary the size of the clusters. When using a greater number of clusters (e.g. c=12 clusters), the smaller the size of each cluster must be, or the opposite in the case where the number of clusters is smaller (e.g. c=2 clusters). When there is only 1 cluster (c=1), the CoMP is global, in which all BSs in the network cooperate. The considered topology has 25 CRs, with 21 RUs, each one representing a BS, in which the restrictions applied to links and CRs (e.g. capacities and delays) are not considered, allowing any of the implemented solutions, without loss of generality. The other parameters are described in the Table III. Several simulations were performed, in which each one considers randomly positioned UEs, according to a uniform distribution, with approximately 20% of the UEs at the edge of the cells. Thus, the averages computed for each solution are presented.

TABLE III: Simulation parameters.

Parameter/configuration	Value
Antennas (Tx/Rx)	4 / 2
Channel model	ITU-R Pedestrian-A
Inter BS distance	600 m
Number of cells	21
Constant UEs per cell	4
UE speed	8 Km/h
Frequency	2.14 GHz
Bandwidth	20 MHz
MIMO tx mode	Closed Loop Spatial Multiplexing
Traffic model	Full buffer
MIMO layers	2
Scheduler	Round Robin - CoMP CB

As can be seen in Fig. 4, any of the solutions that perform the clustering of BSs and the correct positioning of the VNFs in order to enable the use of CoMP, tend to improve the spectral efficiency of the entire network, directly proportional to the size of the cluster. Although the implementation of solutions that generate very large clusters (e.g. c=1) may be unfeasible in real topologies, as they must require links and CRs with very high capacities, the feasible solutions prove to be capable of improving the network performance. Compared with the No-CoMP solution, the solution with 12 clusters (c=12) presents an average increase of 36.3% in the spectral efficiency of the edge UEs and 17.4% for the non-edge users. When $c = 2$, edge UEs can have 110% better performance, and 41.5% better performance for the non-edge UEs.

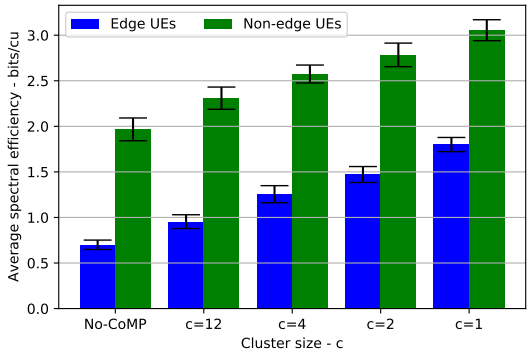

Fig. 4: Spectrum efficiency as a function of the cluster size.

In a second experiment, we seek to evaluate the performance of the proposal in the running system, in which the average demand that each RU is receiving at a given moment is observed, in order to serve as input to the optimization model to obtain the best solution. In this solution, it is expected that we will have the smallest possible number of clusters and, consequently, clusters with the largest possible size, as previously highlighted. It is important to emphasize that the solution found should only consider clusters that support the recent demand. Otherwise, the performance of the network will be impacted by the CQI downgrade procedure.

The same topology used in the former experiment was considered (Tab. III). However, the topology constraints are now fully observed as defined in the problem formulation (Sec. III-B). The network starts with no ability to perform CoMP, since all RUs implement only D-RANs. Due to periodic monitoring ($\Delta t = 100$), or at any time when triggered by aperiodic monitoring, VNF-C calculates network metrics and requests a new solution whenever necessary, as described in Sec. IV. For simplicity, it was assumed that the solution is delivered by VNF-P before the next periodic event. In fact, although the time required to obtain the optimal solution cannot be synchronized, this approach only seeks to simplify the experiment, without loss of generality.

The results are presented in Fig. 5, which demonstrates the average throughput obtained by the UEs in a set of 10 simulations. In each simulation, 4 UEs are randomly placed in each BS, according to a uniform distribution. The UEs move at a constant speed, in a straight line, assuming a random direction. If a UE enters the coverage area of another BS, the handover is made immediately. For simplicity, a constant traffic model (full buffer) is assumed. Again, *No-CoMP* indicates when RUs only implement D-RANs and do not perform CoMP, and *Dynamic* indicates the result of the dynamic cluster size adjustment made by our proposal. Solid lines are used to represent the edge UEs and dashed lines for the non-edge UEs.

It can be seen that the proposal is capable of improving the performance of the network as a whole. As verified in the previous experiment (Fig. 4), CoMP optimizes the spectral efficiency of the network, without compromising the non-edge UEs. By making decisions about the size of the clusters in parallel to the running system, and respecting the quality of the channels of the UEs over time, the system can guarantee the appropriate size of the cluster in order to serve the maximum number of UEs, reducing the impact of interference. As can be observed, the proposed method allows the edge UEs to present an average performance of 73% higher, and 26% higher for the

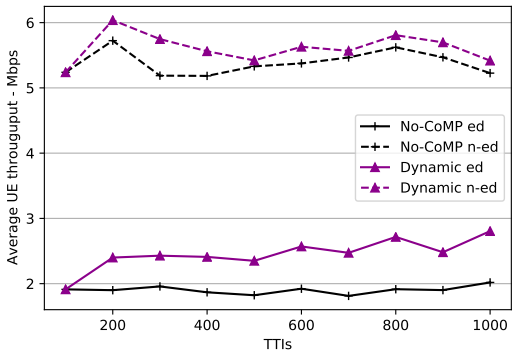

Fig. 5: Average UE throughput along the time while the clusters are dynamically reconfigured.

VI. Conclusion

In this work, we presented a solution for VNF placement and resource allocation based in advanced framework that was modified to take into account the channel quality of the UEs. Our proposal was designed to be integrated in O-RAN architecture, basically, working as a new xApp and a new rApp. Our performance evaluation illustrates the potential benefits of our solution, mainly the proper integration between multiple closed loops involving different, but correlated decision-making processes. In future work, we plan to explore machine learning techniques to improve our solution, for example, through traffic forecasting and user mobility in order to anticipate aperiodic reports concerning exhaustion of the allocated computing resources.

Acknowledgements

This work was supported in part by MCTIC/CGI.br/São Paulo Research Foundation (FAPESP) through the Project Smart 5G Core And MUltiRAn Integration (SAMURAI) under Grant 2020/05127-2, by CNPq through the Project Universal under Grant 405111/2021-5, and by RNP/MCTIC, Grant No. 01245.010604/2020-14, under the 6G Mobile Communications Systems project.

References

[1] M. Polese *et al.*, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," *arXiv preprint arXiv:2202.01032*, 2022.

[2] O. Alliance, "Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN," *Technical Specification, Apr*, 2020.

[3] C.-Y. Chang *et al.*, "FlexDRAN: Flexible Centralization in Disaggregated Radio Access Networks," *IEEE Access*, vol. 10, pp. 11 789–11 808, 2022.

[4] F. Z. Morais *et al.*, "PlaceRAN: optimal placement of virtualized network functions in Beyond 5G radio access networks," *IEEE Transactions on Mobile Computing*, 2022.

[5] A. Garcia-Saavedra *et al.*, "WizHaul: On the Centralization Degree of Cloud RAN Next Generation Fronthaul," *IEEE Transactions on Mobile Computing*, vol. 17, no. 10, pp. 2452–2466, 2018.

[6] K. Sundaresan *et al.*, "FluidNet: a flexible cloud-based radio access network for small cells," *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 2, pp. 915–928, 2016.

[7] A. A. A. Ari *et al.*, "Resource allocation scheme for 5G C-RAN: A Swarm Intelligence based approach," *Computer Networks*, vol. 165, p. 106 957, 2019.

[8] I. Koutsopoulos, "The Impact of Baseband Functional Splits on Resource Allocation in 5G Radio Access Networks," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, IEEE, 2021, pp. 1–10.

[9] 3GPP-TR21.916, "System Architecture for the 5G (Release 16)," Technical Recommendation 21.916, 2020.

[10] E. Sarikaya *et al.*, "Placement of 5G RAN Slices in Multi-tier O-RAN 5G Networks with Flexible Functional Splits," in *2021 17th International Conference on Network and Service Management (CNSM)*, IEEE, 2021, pp. 274–282.

[11] A. Garcia-Saavedra *et al.*, "O-RAN: Disrupting the virtualized RAN ecosystem," *IEEE Communications Standards Magazine*, 2021.

[12] D. Bega *et al.*, "CARES: Computation-aware scheduling in virtualized radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 7993–8006, 2018.

[13] M. Rupp *et al.*, *The Vienna LTE-Advanced Simulators: Up and Downlink, Link and System Level Simulation*, 1st ed., ser. Signals and Communication Technology. Springer Singapore, 2016.