GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001347

Efficient allocation of disaggregated RAN functions and Multi-access Edge Computing services

Luciano de S. Fraga*, Gabriel Matheus Almeida*, Sand Correa*, Cristiano Both†, Leizer Pinto*, Kleber Cardoso*
*Universidade Federal de Goiás, Brazil, †University of Vale do Rio dos Sinos, Brazil.
E-mail: {lucianofraga, gabrielmatheus, sand, leizer, kleber}@inf.ufg.br*, cbboth@unisinos.br†

***Abstract*—Openness, virtualization and disaggregation of functions represent the state-of-the-art (SOTA) for optimal management and orchestration of Radio Access Network (RAN) resources. However, in 5G and beyond networks, virtualized RAN (vRAN) functions may commonly share computing resources with Multi-access Edge Computing (MEC) services. This paper introduces a new problem formulation that jointly optimizes vRAN functions and MEC services respecting the maximum acceptable delay of the applications. We show that our model achieves better solutions than a SOTA approach and it is also more flexible. We also present a heuristic solution that is able to achieve near optimal results for real-world networks.**

***Index Terms*—Virtualized RAN (vRAN), Multi-access Edge Computing (MEC) services, centralization level, functional splits.**

I. Introduction

Mobile operators are adopting Radio Access Network virtualization (vRAN) to accommodate the high demand for mobile services at an affordable cost [1]. In vRAN, the base station (BS) functions are disaggregated, virtualized, and executed in general-purpose computing resources (CRs) deployed across the network. These CRs play the role of three types of vRAN nodes: virtual Centralized Unit (vCU), virtual Distributed Unit (vDU), and Radio Unit (RU). The BS function disaggregation follows a certain *functional split* [2], which is usually chosen based on resource availability and the network load. Table I shows the specification of the functional split options for an RU with the following configuration: 100 MHz spectrum bandwidth, 32 antenna ports, 8 MIMO layers, and 256 QAM modulation [3]. As shown in the table, each split option imposes a (maximum) latency and a (minimum) bitrate value that must be satisfied in the communication among vRAN nodes (i.e., vCU, vDU, and RU) even if the nodes are running on different CRs.

Split Option	Functional Split	One-way Latency	Bitrate (Gbps) DL	UL
O1	RRC – PDCP	10 ms	4	3
O2	PDCP – High RLC	10 ms	4	3
O3	High RLC – Low RLC	10 ms	4	3
O4	Low RLC – High MAC	1 ms	4	3
O5	High MAC – Low MAC	< 1 ms	4	3
O6	Low MAC – High PHY	250 μs	4.13	5.64
O7	High PHY – Low PHY	250 μs	86.1	86.1
O8	Low PHY – RF	250 μs	157.3	157.3

TABLE I: Split options and their associated latency and bitrate.

978-1-6654-3540-6/22 © 2022 IEEE

The vRAN technology brings advantages such as resource pooling, simpler update roll-ups, and cheaper management and control, resulting in cost-efficient and high-performance RAN operation [4]. However, the design and specification of the placement of the vRAN nodes and the BS functions they will execute is a non-trivial decision. As illustrated in Figure 1, the combination of eight split options and three vRAN nodes results in nine Viable Next-generation RAN Configurations (VNCs) for each BS [5]. Each VNC generates a different demand over the crosshaul and the CRs (e.g., processing, memory, and storage).

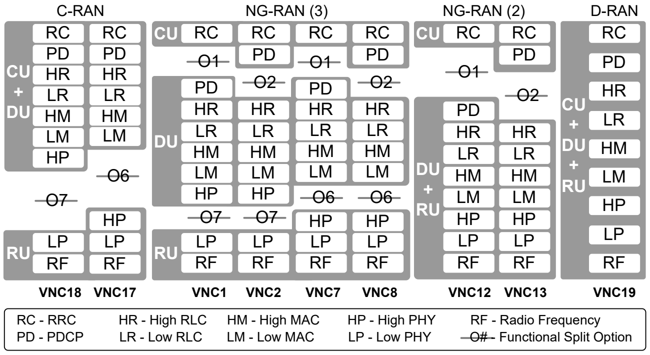


Fig. 1: Nine viable Next-generation RAN Configurations.

The complexity of designing vRANs is further exacerbated when the vRAN nodes need to accommodate Multi-access Edge Computing (MEC) traffic. MEC is a key enabler of 5G/B5G networks, considered instrumental for such networks to support low-latency (e.g., mixed reality, tactile Internet, and autonomous driving) or high-bandwidth (e.g., video streaming for surveillance) applications [6]. To illustrate, traditional topologies such as ring-based ones impose a significant delay between several RUs and computing resources. Therefore, it is necessary to deploy the MEC service at RU to meet the needs of ultra-low-latency applications, constraining the eligible VNC to D-RAN (VNC 19). Conversely, C-RAN (VNC 18 and VNC 17) are preferable to ensure cost gains in resource pooling and centralization of vRAN functions. However, such VNCs may be prohibitive to ultra-low-latency or throughput-hungry MEC services. Therefore, there is an intricate coupling between the design of vRANs and the deployment of MEC services. In this context, the problem for deciding (i) the best VNC for each BS, (ii) where to place the various MEC services, and (iii) how to route the legacy (non-MEC) and MEC traffic between RUs, vDUs, and vCUs, is challenging since MEC services can be diverse, i.e., some create massive data, while others have ultra-low-latency needs.

Related Work: Despite the coupling between the vRAN design and the MEC service deployment, these problems are usually treated separately in the literature. In the last few years, vRAN design has been the subject of several works in the literature [4], [5], [7]–[9]. In such works, centralization of vRAN functions, i.e., the amount of virtual network functions (VNFs) from different BSs running in the same CR, is the main goal as it improves the performance of RAN operations through, for example, carrier aggregation and coordinated multipoint [2], [10]. The works in [7], [8] consider only one functional split option per BS. In [4], [5], the authors investigate the possibility of using two or more functional splits. In [5], we introduce PlaceRAN, a flexible framework supporting any number of functional split options. PlaceRAN allows an RU to be served by any CR playing as vDU or vCU. This feature makes the allocation more flexible because the role of CRs is not fixed, enabling them to act as vDU or vCU as long as the latency and throughput requirements are satisfied. In [9], we extend PlaceRAN to allow flow splitting in a pair of vRAN functions. This extension allows the traffic between any pair of vRAN nodes to be split and routed between multiple paths. Flow splitting benefits the minimization of used CRs and improves the centralization of VNFs, mainly when the network links present bottlenecks. MEC server deployment is treated in [11]–[13]. Latency and deployment cost are usually considered the optimization criterion in such works. In [11], the authors propose a placement scheme of deploying K MEC servers at BSs to minimize the deployment cost and latency. The work in [12] presents a mathematical model to minimize the deployment cost, and latency of MEC servers deployed hierarchically at BS and the metro levels to compensate for the relatively small capacity at BSs. In [13], the authors propose a hierarchical deployment scheme based on C-RAN to exploit the tradeoff between deployment cost and average latency by jointly optimizing MEC server deployment, requests allocation, and routing. However, none of these works address the placement of vRAN nodes and the functional split options. Indeed, we are aware of only one work [10] that deals with the joint problem of designing vRANs and deploying MEC services. In such work, the authors propose an optimization model that minimizes vRAN costs and maximizes MEC performance by considering function splits, the fronthaul routing paths, and the placement of MEC functions. However, in [10] only four split options are considered per BS. Moreover, there is a single CU located at a fixed point. Therefore, to the best of our knowledge, there is no work yet in the literature that proposes a general formulation to the problem of jointly addressing the vRAN design and the deployment of MEC services on realistic operational networks.

Contribution: This work introduces a new problem formulation that jointly optimizes the placement of vRAN functions and MEC services. The main objective is to maximize the centralization level of vRAN functions and minimize the number of computing resources necessary for running all the VNFs (vRAN and MEC services) while respecting the maximum acceptable delay of the MEC applications. As in our previous work [9], our formulation is general, and (i) supports any number of functional split options, (ii) allows an RU to be served by any CR playing as vDU or vCU, and (iii) supports flow splitting. To this end, we formulate our framework as two single objectives Mixed Integer Programming (MIP) problems. In the first, we aim to minimize the latency in each BS, and in the second, we goal to maximize the centralization level by accepting a controlled increase in the latency. This approach provides optimal placement of VNFs from vRAN and MEC services, supporting different MEC latency constraints, no matter how limiting they are. We also propose an approximate algorithm to solve large-scale topologies since this problem is NP-hard. Finally, we show the advantage of our solutions (both optimal and heuristic) compared to a state-of-the-art approach.

This work is organized as follows. In Section II, we present the system model and problem formulation. Section III describes the heuristic approach. Section IV presents the performance evaluation. Final considerations and future work are in Section V.

II. System Model and Problem Formulation

We consider a NG-RAN composed by a set $\mathcal{B} = \{b_1, b_2, ..., b_{|\mathcal{B}|}\}$ of RUs, a set $\mathcal{T} = \{t_1, t_2, ..., t_{|\mathcal{T}|}\}$ of forwarding nodes, and a set $\mathcal{C} = \{c_1, c_2, ...c_{|\mathcal{C}|}\}$ of CRs. Each forwarding node $t_k \in \mathcal{T}$ may be connected to RUs, CRs, the core network, or other forwarding nodes. A processing capacity c_m^{Proc} characterizes each CR $c_m \in \mathcal{C}$. We also define the graph $G = (\mathcal{V}, \mathcal{E})$ to represent this NG-RAN and the core network v_0, with $\mathcal{V} = \{v_0\} \cup \mathcal{B} \cup \mathcal{T} \cup \mathcal{C}$ being the set of nodes and $\mathcal{E} = \{e_{ij}; v_i, v_j \in \mathcal{V}\}$ the set of links. Each link $e_{ij} \in \mathcal{E}$ has a transmitting capacity e_{ij}^{Cap} and a latency e_{ij}^{Lat}.

We define $\mathcal{P}_l$ as the set of tree structures that guarantee a path to each RU $b_l \in \mathcal{B}$ (root node) from both: the core network (leaf node) and the MEC Server (root, leaf or internal node). Each tree $p \in \mathcal{P}$ is composed of four subpaths: p_{Bh} (backhaul, i.e., core ↔ CU, or core ↔ CU+DU, or core ↔ CU+DU+RU), p_{Bhm} (backhaul-MEC, i.e., MEC server ↔ CU, or MEC server ↔ CU+DU, or MEC server ↔ CU+DU+RU), p_{Mh} (midhaul, i.e., CU ↔ DU, or CU ↔ DU+RU), and p_{Fh} (fronthaul, i.e., RU ↔ DU, or RU ↔ CU+DU), in which at least the subpath p_{Bh} is not empty.

We consider that a VNF runs part of the RAN protocol stack (except RF and Low PHY protocol, that aways run in the RU). VNFs are labeled in increasing order, starting from PHY Low with f_1 and ending at RRC with f_7. We define $\mathcal{F} = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$ as the set of disaggregated RAN VNFs, where the distribution must follow one of the VNCs of the set $\mathcal{D} = \{D_1, D_2, ..., D_{|\mathcal{D}|}\}$.

Similarly we denote by f_{mec} the MEC service to be placed in the MEC server, which, in turn, could be allocated in any CR as long as it obeys the stack of functions $\mathcal{F}$, being deployed after the VNF f_7. Each BS $b_l \in \mathcal{B}$ has a demand of MEC requests upper bounded by μ_l (requests/s), that generate the bitrate λ_l (Mb/s) to be routed from the MEC server. Also, the parameter γ^{mec} represents the computing demand for each Mb/s generated by the VNF f_{mec}.

Our problem has the following goals: (i) minimize the delay of a service allocated in a MEC server; (ii) maximize the centralization level while minimize the number of CRs that are used to run the VNFs. The mathematical formulation that we have developed for this problem is presented bellow.

Let us denote the binary decision variables $x_l^{p,r}$ and w_l^m that represent respectively, which pair of tree $p \in \mathcal{P}$ and VNC $D_r \in \mathcal{D}$ is selected to serve a RU $b_l \in \mathcal{B}$, and if a CR $c_m \in \mathcal{C}$ is selected to execute a MEC service that serve a RU $b_l \in \mathcal{B}$.

Goal (i). To minimize the routing and processing delay for all the base stations that are served by a MEC service, we propose the following objective function:

$$\text{minimize} \sum_{b_l \in \mathcal{B}} \mu_l D_l(x_l^{p,r}, w_l^m). \tag{1}$$

The delay function $D_l(x_l^{p,r}, w_l^m)$ is defined as:

$$\begin{aligned} D_l(x_l^{p,r}, w_l^m) = \sum_{D_r \in \mathcal{D}} \max_{P \in \mathcal{P}_l} \\ \Big\{ \sum_{e_{ij} \in \mathcal{E}} \Big((s_l x_l^{p,r} e_{ij}^{Lat})(y_{eij}^{p_{Bhm}} + y_{eij}^{p_{Mh}} + y_{eij}^{p_{Fh}}) \Big) \Big\} \\ + \sum_{c_m \in C} \delta_1 \Big(\lambda_l w_l^m (\frac{\gamma^{mec}}{c_m^{Proc}}) \Big) + \delta_2 \Big(\lambda_l w_l^m (\frac{\gamma^{mec}}{c_m^{Proc}}) \Big)^2, \end{aligned} \tag{2}$$

where the input data $s_l \in \{0,1\}$ indicates if a RU has demand of a MEC service, and $y_{e_{ij}}^k$ with $k \in \{p_{Bh}, p_{Mh}, p_{Fh}, p_{Bhm}\}$ represents if the link e_{ij} is part of backhaul, midhaul, fronthaul or backhaul-MEC. The two parameters δ_1 and δ_2 represent the relationship between load and delay [10].

Constraints. To ensure that the sub-path p_{Bhm} has a connection with the MEC server selected for a RU, we formulate the constraint (3), where the mapping function $V(p_{Bhm}, c_m) \in \{0,1\}$ (from the input) defines if the sub-path p_{Bhm} has the CR c_m as it source (downlink). Constraint (4) enables only 1 CR to serve as a MEC server for the RU $b_l \in \mathcal{B}$ with MEC demand:

$$\begin{aligned} s_l x_l^{p,r} V(p_{Bhm}, c_m) = s_l x_l^{p,r} w_l^m, \\ \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D}, c_m \in \mathcal{C}, \end{aligned} \tag{3}$$

$$\sum_{c_m \in \mathcal{C}} w_l^m = s^l, \forall b_l \in \mathcal{B}. \tag{4}$$

The following constraints (5)-(10) enable the flow split among several tree structures. Let the decision variable $y_l^{p,r} \in [0,1]$ represents the amount of traffic flow of RU $b_l \in \mathcal{B}$, using VNC $D_r \in \mathcal{D}$, that must cross the tree $p \in \mathcal{P}$. The constraint (5) assures that all RUs of $\mathcal{B}$ are served, employing potentially more than one tree of $\mathcal{P}_l$. Constraint (6) ensures that the amount of traffic flow is never negative. The delivery of the whole traffic flow is assured by constraint (7). The number of CRs that serve a RU as one of the three RAN Units must be equal to the number of CRs of the chosen VNC (returned by $F_{CR}(D_r)$) as defined in constraint (8), the mapping function $H(p_{Bhm})$ returns 1 if p_{Bhm} is not empty and 0 otherwise. Constraints (9) and (10) assure that if a tree $p \in \mathcal{P}$ is chosen to serve a RU $b_l \in \mathcal{B}$ (i.e., $x_l^{p,r} = 1$), then at least some traffic flow must be routed across that tree (i.e., $y_l^{p,r} > 0$). Otherwise if $x_l^{p,r} = 0$ then no traffic will be routed through that tree from that RU. Moreover, the parameter $0 \leq \varepsilon \leq 1$ allows for controlling the granularity of the flow split.

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

$$y_l^{p,r} \geq 0, \qquad \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D}, \tag{6}$$

$$\sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} y_l^{p,r} = 1, \qquad \forall b_l \in \mathcal{B}, \tag{7}$$

$$\begin{aligned} \sum_{c_m \in \mathcal{C}} \left\lceil \frac{\sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} (x_l^{p,r} u_m^p (1 - w_l^m H(p_{Bhm})))}{|\mathcal{P}_l|} \right\rceil \\ = F_{CR}(D_r), \forall b_l \in \mathcal{B}, \end{aligned} \tag{8}$$

$$y_l^{p,r} \leq x_l^{p,r}, \qquad \forall p \in \mathcal{P}_l, b_l \in \mathcal{B}, D_r \in \mathcal{D}, \tag{9}$$

$$x_l^{p,r} - y_l^{p,r} \leq 1 - \varepsilon, \qquad \forall p \in \mathcal{P}_l, b_l \in \mathcal{B}, D_r \in \mathcal{D}. \tag{10}$$

We denote as d^{th} the minimum value (in ms), for each RU, where any delay below that does not guarantee real benefit to users and, the follow constraint define this limit:

$$D_l(x_l^{p,r}, w_l^m) \geq d^{th}, \forall b_l \in \mathcal{B}. \tag{11}$$

Each RU can only enable one VNC , as define in equation (12), where the mapping function $W(D_r, b_l)$(from the input) defines if the RU $b_l \in \mathcal{B}$ is related to the VNC $D_r \in \mathcal{D}$:

$$\sum_{D_r \in \mathcal{D}} W(D_r, b_l) = 1, \forall b_l \in \mathcal{B}. \tag{12}$$

The transmitting capacity e_{ij}^{Cap} of every link e_{ij} must not be exceeded, as described by the following constraint:

$$\begin{aligned} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \Big[y_l^{p,r} \Big(y_{e_{ij}}^{p_{Bh}} \alpha_{Bh}^r + y_{e_{ij}}^{p_{Mh}} \alpha_{Mh}^r + y_{e_{ij}}^{p_{Fh}} \alpha_{Fh}^r \\ + y_{e_{ij}}^{pBhm} y_{e_{ij}}^{pMh} y_{e_{ij}}^{pFh} \Big(s_l \lambda_l \Big) \Big) \Big] \leq e_{ij}^{Cap}, \qquad \forall e_{ij} \in \mathcal{E}, \end{aligned} \tag{13}$$

where α_{Bh}^r, α_{Mh}^r, and α_{Fh}^r represent the demands for bitrate in the backhaul, midhaul, and fronthaul of a VNC $D_r \in \mathcal{D}$.

Each $D_r \in \mathcal{D}$ tolerates a maximum latency in each sub-path (backhaul, midhaul, and fronthaul) of the tree $p \in P_l$, which is described by the following constraints:

$$\sum_{eij \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Bh}} e_{ij}^{Lat} \leq \beta_{Bh}^r, \ \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D}, \tag{14}$$

$$\sum_{eij \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Mh}} e_{ij}^{Lat} \leq \beta_{Mh}^r, \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D}, \tag{15}$$

$$\sum_{eij \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Fh}} e_{ij}^{Lat} \leq \beta_{Fh}^r, \ \forall b_l \in \mathcal{B}, p \in \mathcal{P}_l, D_r \in \mathcal{D}, \tag{16}$$

where β_{Bh}^r, β_{Mh}^r, and β_{Fh}^r represent the maximum latency tolerated in the backhaul, midhaul, and fronthaul, respectively, of a tree $p \in \mathcal{P}_l$ that transports a specific VNC $D_r \in D$.

Finally, the VNFs selected to run in a CR $c_m \in \mathcal{C}$ must not exceed its processing capacity c_m^{Proc} , as represented by the following constraint:

$$\sum_{b_l \in \mathcal{B}} w_l^m \lambda_l \gamma^{mec} + \sum_{f_s \in \mathcal{F}} \left[\sum_{b_l \in \mathcal{B}} \left\lceil \frac{\sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} (x_l^{p,r} u_m^p)}{|\mathcal{P}_l|} \right\rceil \times \sum_{D_r \in \mathcal{D}} M(c_m, f_s, b_l, D_r) \gamma_m^s \right] \leq c_m^{Proc}, \forall c_m \in \mathcal{C}, \tag{17}$$

where γ_m^s and γ^{mec} is the computing demand of the VNFs $f_s \in \mathcal{F}$ and f_{mec} respectively. The mapping function $M(c_m, f_s, b_l, D_r) \in \{0,1\}$ (from the input), indicates if the CR $c_m \in \mathcal{C}$ runs the VNF $f_s \in \mathcal{F}$ from the RU $b_l \in \mathcal{B}$, according to the VNC $D_r \in \mathcal{D}$.

Goal (ii). To jointly minimize the number of CRs select to run the VNFs and maximize the centralization level i.e., the number of grouped RAN VNFs that is deployed in a single CR, we define the following objective function:

$$\text{minimize} \sum_{c_m \in \mathcal{C}} \left\lceil \frac{\sum_{b_l \in \mathcal{B}} \sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} (x_l^{p,r} u_m^p)}{|\mathcal{P}_l|} \right\rceil - \sum_{c_m \in \mathcal{C}} \sum_{f_s \in \mathcal{F}} \left(\sum_{b_l \in \mathcal{B}} \left\lceil \frac{\sum\limits_{p \in \mathcal{P}_l} \sum\limits_{D_r \in \mathcal{D}} (x_l^{p,r} u_m^p)}{|\mathcal{P}_l|} \right\rceil M(c_m, f_s, b_l, D_r) - \left\lceil \sum_{b_l \in \mathcal{B}} \left\lceil \frac{\sum\limits_{p \in \mathcal{P}_l} \sum\limits_{D_r \in \mathcal{D}} (x_l^{p,r} u_m^p)}{|\mathcal{P}_l|} \right\rceil M(c_m, f_s, b_l, D_r) \Big/ |\mathcal{B}| \right\rceil \right), \tag{18}$$

where the first (positive) part represents the minimization of the number of CRs and the second (negative) part represents the maximization of the centralization level.

Now let (P1) be the mono-objective problem composed by the objective function (1) and constraints (3)–(17), and let $(x_l^{p,r^*}, w_l^{m^*})$ be the optimal solution to (P1). To achieve simultaneously the goals (i) e (ii) we define a second mono-objective problem (P2) composed by the objective function (18) and constraints (3)–(17), and (19):

$$D_l(x_l^{p,r}, w_l^m) = D_l(x_l^{p,r^*}, w_l^{m^*}), \forall b_l \in \mathcal{B}. \tag{19}$$

Constraint (19) defines that the latency achieved in (P2) must be the same as the optimal value, obtained by solving (P1). Due to this limit, we have no flexibility when optimizing the centralization in (P2). To solve this problem we define the constant Δ as a value (in ms) to be added in the latency found in (P1). Then in constraint (20), we allow the latency in (P2) to be greater than the value found in (P1), in exchange for an improvement of the centralization in (P2). Lastly, (P2) now is composed by the objective function (18) and constraints (3)–(17), and (20):

$$D_l(x_l^{p,r}, w_l^m) \leq \Delta + D_l(x_l^{p,r^*}, w_l^{m^*}), \forall b_l \in \mathcal{B}. \tag{20}$$

III. Optimal and Heuristic Solution

Our optimization problem presents the features of a *location-allocation* (LA) problem as described in [14], where a set of resources (CRs) are chosen to play different roles (RAN Units, MEC server) in order to attend the demand of the clients (users associated to a BS). The *location-allocation* class problems are NP-Hard and, in order to scale our solution to larger topologies we propose a iterative non-deterministic heuristic approach [15]. On the optimal approach we jointly reach the goals (i) and (ii) by solving the problems P1 and P2. On the heuristic approach we solve the same problems, but limiting the number of centralization points, that are randomly chosen.

Algorithm 1 summarizes our heuristic solution. The inputs are the same as the optimal model and the value $heuristicTL$. The outputs are the solution of P2, and the decision variables defined in section II. The heuristic begins initiating k with a number of CRs closest to the core $(1 < k < |C|)$ randomly chosen ($l.$ 2). The set of k shortest CRs to the core is assigned to ksc ($l.$ 3), then we perform the solution of P1 and P2 as defined in section II, but with the centralization happening only in the CRs assigned to ksc ($l.$ 4-5). The solution of the problem is constrained by a time limit ($heuristicTL$), the heuristic is interrupted, and the current solution is presented when the limit is achieved ($l.$ 1).

Algorithm 1 Heuristic approach.

```
   Input: All inputs from the optimal model, $heuristicTL$
   Output: $p2\_solution$, $x_l^{p,r}$, $y_l^{p,r}$, $w_l^m$
1: while $time \leq heuristicTL$ do
2:     $k \leftarrow$ RANDOM($|\mathcal{C}|$)
3:     $ksc \leftarrow$ K_SHORTEST_CRS($k$)
4:     $p1\_solution \leftarrow$ P1($ksc$)
5:     $p2\_solution \leftarrow$ P2($ksc, p1\_solution$)
6: Update $x_l^{p,r}$, $y_l^{p,r}$, $w_l^m$ accordingly to $p2\_solution$
```

IV. Evaluation

We implement our optimal and heuristic models using Python 3.6.9, docplex 2.20.204 and the IBM optimizer CPLEX 20.1.0. The experiments were performed in a Virtual Machine (VM) with Ubuntu 18.04, 16 vCPUs, 256 GB RAM, and 40 GB of the virtual disk. The VM is hosted in a server DELL PowerEdge M620 with two Intel Xeon E5-2650 @ 2 GHz. We evaluate our model using two RAN topologies: (i) a traditional ring-based network (T1) used in the 5G-crosshaul project [16]; and (ii) a next-generation hierarchical RAN (T2) inspired by the PASSION project [17]. T1 has 30 nodes where 28 are RUs and T2 has 32 nodes where 30 are RUs.

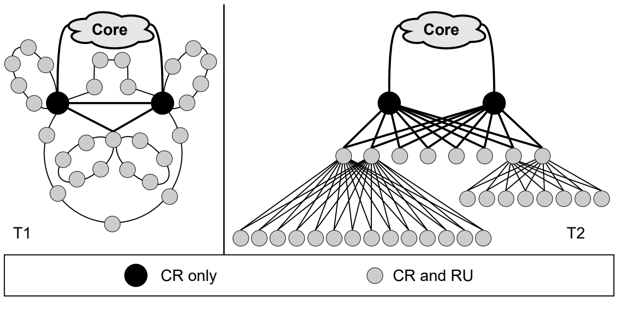


Fig. 2: RAN topologies.

Table II summarizes the parameters employed in this evaluation. The processing requirements (γ^{mec}), the network load requeriments (λ_l), and the relationship between load and delay (δ_1/δ_2) of a typical MEC service is obtained from [10], based on the profiling of a face recognition application. The processing load of each RAN protocol employed in this work is based on [9] and is illustrated in Table III.

We first compare the results achieved by our optimal solution with the ones produced by a SOTA solution presented in [10]. We call the SOTA proposal as Restricted VNCs and Fixed CU (RVFC) since it supports only four split options (VNCs 18, 17, 13 and 19) and can only centralize functions in one CR. We define the closest CR to the core (i.e., one of the *CR only* nodes in Fig. 2) as the centralization point in RVFC.

TABLE II: Parameters employed in the evaluation.

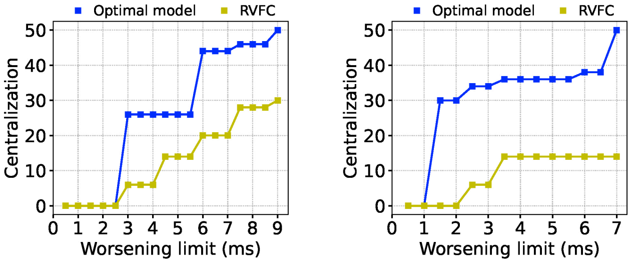
Parameter	T1	T2
$\|\mathcal{B}\|$	28	30
$\|\mathcal{C}\|$	30	32
$\|\mathcal{D}\|$	9	9
c_m^{Proc} (cores)	{8, 16, 32}	{8, 16}
e_{ij}^{Cap} (Gbps)	{5, 10, 50}	{50, 100, 800}
e_{ij}^{Lat} (ms)	{1.38, 2.86}	{1.02, 2.07, 3.22, 5.79, 6.65}
γ^{mec} (RCs)	0.1/Mbps	0.1/Mbps
λ_l (Mb/s)	10	10
δ_1/δ_2 (ms)	0.25	0.25
d^{th} (ms)	0	0

TABLE III: CPU load for the RAN protocol stack.

RAN Protocol	CPU Load (RCs)
RRC (f_7)	0.49 (γ_m^7)
PDCP (f_6)	0.49 (γ_m^6)
High RLC (f_5)	0.0245 (γ_m^5)
Low RLC (f_4)	0.0245 (γ_m^4)
High MAC (f_3)	0.343 (γ_m^3)
Low MAC (f_2)	0.343 (γ_m^2)
High PHY (f_1)	0.833 (γ_m^1)
Low PHY	2.352
Total	4.9

Fig. 3 presents the number of centralized functions achieved by both models considering different worsening limits (WL). These results represent the optimal solution from (P1) degraded by Δ. Fig. 3a and Fig. 3b show the maximum centralization achieved by both models in topologies T1 and T2, respectively. As expected, as WL increases, centralization also increases. In T1, the maximum centralization is obtained when WL is 9ms, while in T2, the maximum centralization is obtained when WL is 7ms. Our optimal model obtains the same performance or outperforms RVFC in all configurations in both topologies. Moreover, our optimal model achieves centralization levels around 1.6 times higher than RVFC in T1, while in T2, this gain is 2.5 times. These results evidence the benefits of enabling VNF centralization in any CR and employing a broader set of split options.

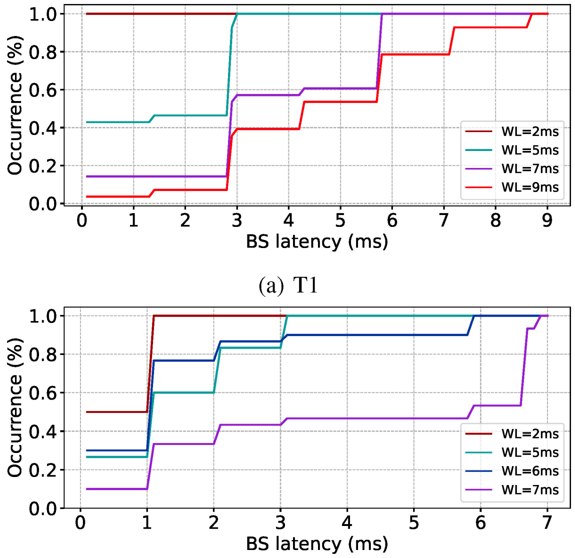
To better understand the results achieved by our optimal model, Fig. 4 shows the Cumulative Distribution Function (CDF) of BSs whose MEC service latency is bellow or equal the worsening limit value in both topologies (T1 - Fig. 4a and T2 - Fig. 4b). These results illustrate the percentage of BSs whose latency are not affected by the flexibility imposed by WL. For example, in both topologies, the MEC service delay is bellow 5 ms (T1) and 6 ms (T2) in at least more than half of the BS when WL is set to 9 ms (T1) and 6ms (T2). Our model can guarantee all the BSs with latency below 1ms in T1 and 50% in T2 when WL is set to 2ms.



(a) T1

(b) T2

Fig. 3: Centralization as a latency function worsening limit.

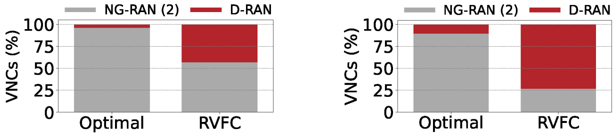


(a) T1

(b) T2

Fig. 4: Percentage of BSs per latency values.

Fig. 5 shows the distribution of chosen VNCs for our optimal model and RVFC when maximum centralization is allowed. In both topologies, our optimal model is able to take advantage of the network and computing resources and choose for a VNC with two vRAN nodes instead of D-RAN. Such decision enables greater centralization than RVFC. Indeed, due to the topologies characteristics used in this work (all links have delay greater than 0.25ms), the VNCs with three vRAN nodes and C-RAN could not be chosen. Finally, we compare the performance of our optimal model, RVFC, and our proposed heuristic in terms of achieved centralization and solution time for topology T2, varying the number of nodes in the topology and seting WL to be greater enough to allow maximum centralization. The top of Fig. 6 shows the results for the achieved centralization. With 32, 48 and 64 nodes, our optimal model and the heuristic achieve the same performance.



(a) T1

(b) T2

Fig. 5: Percentage of VNCs chosen fo each model.

The optimal solution could no be achieved with more than 64 nodes due to RAM exhaustion. RVFC, on the other hand, is able to scale, but it is outperformed by our heuristic in all settings regarding centralization. The bottom of Fig. 6 shows the results for the solution time. Since RVFC is based on a simpler formulation than our optimal solution and our proposed heuristic, it performs better than our models in terms of execution time. Therefore, we conclude that a more general model such as our optimal model achieves better centralization results than RVFC but does not scale for bigger topologies. On the other hand, our heuristic can achieve better centralization results than RVFC in a feasible time.

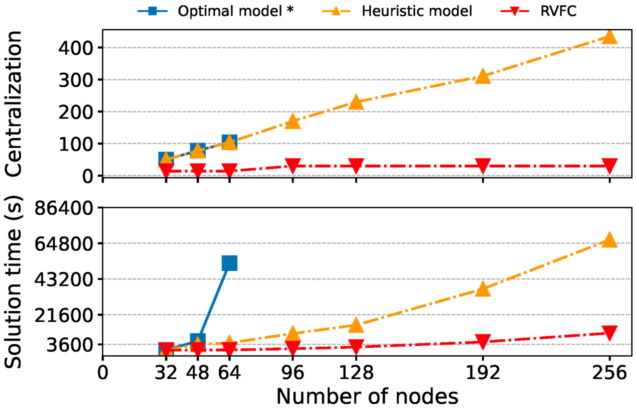


Fig. 6: Maximum centralization (top) and solution time (bottom) as a function of the number of nodes.

V. Conclusion and Future Work

In this paper we introduce a solution to vRAN/MEC placement problem. We address the trade-off between minimization of MEC services latency and the maximization of vRAN functions centralization employing two single objective optimal problems. Our model allows the choice of up to two functional splits and the routing of the data flow among several tree structures. To achieve scalability we also present an iterative non-deterministic heuristic. Finally, we show the advantages of our two solutions in comparison with the SOTA vRAN/MEC placement. In future work, we intend to investigate the dynamic VNF placement due to time varying demand considering both vRAN and MEC services.

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] J. A. Ayala-Romero *et al.*, "VrAIn: A Deep Learning Approach Tailoring Computing and Radio Resources in Virtualized RANs," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019.

[2] L. M. P. Larsen *et al.*, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 146–172, 2019.

[3] 3rd Generation Partnership (3GPP), "Study on New Radio Access Technology; Radio Access Architecture and Interfaces (Release 14)," Tech. Rep., 2017.

[4] F. W. Murti *et al.*, "An Optimal Deployment Framework for Multi-Cloud Virtualized Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2251–2265, 2021.

[5] F. Z. Morais *et al.*, "PlaceRAN: optimal placement of virtualized network functions in Beyond 5G radio access networks," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2022.

[6] P. Mach *et al.*, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[7] A. Garcia-Saavedra *et al.*, "FluidRAN: Optimized vRAN/MEC Orchestration," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 2366–2374.

[8] 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.

[9] G. M. Almeida *et al.*, "Optimal joint functional split and network function placement in virtualized ran with splittable flows," *IEEE Wireless Communications Letters*, pp. 1–1, 2022.

[10] A. Garcia-Saavedra *et al.*, "Joint optimization of edge computing architectures and radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2433–2443, 2018.

[11] Q. Fan *et al.*, "Cost Aware cloudlet Placement for big data processing at the edge," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[12] Z. Liu *et al.*, "Cost Aware Mobile Edge Computing Hierarchical Deployment in Optical Interconnection Network," in *Asia Communications and Photonics Conference (ACP)*, 2018, pp. 1–3.

[13] Z. Liu *et al.*, "Hierarchical MEC servers deployment and user-MEC server association in C-RANs over WDM ring networks," *Sensors*, vol. 20, no. 5, p. 1282, 2020.

[14] Z. Azarmand *et al.*, "Location allocation problem," in *Facility location*, Springer, 2009, pp. 93–109.

[15] S. A. Khan *et al.*, "Iterative non-deterministic algorithms in on-shore wind farm design: A brief survey," *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 370–384, 2013.

[16] 5G-crosshaul Project, *5G-crosshaul, D1.2: final 5G-crosshaul system design and economic analysis*, Available at http://163.117.166.92/wp-content/uploads/2018/01/5G-CROSSHAUL_D1.2.pdf. Accessed 24/10/2021, 2017.

[17] PASSION Project, Available at http://www.passion-project.eu/project/. Accessed 24/10/2021, 2020.