IEEE ComSoc IEEE Open Journal of the Communications Society

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version 11 January, 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.011100

Optimizing Energy Consumption for vRAN Placement in O-RAN Systems with Flexible Transport Networks

WILLIAM T. PIRES-JR[1], GABRIEL M. ALMEIDA [1], SAND L. CORRÊA[1], CRISTIANO B. BOTH[2], LEIZER L. PINTO[1], AND KLEBER V. CARDOSO.[1]

[1]Universidade Federal de Goiás, Brazil
[2]University of Vale do Rio dos Sinos, Brazil

CORRESPONDING AUTHOR: William T. Pires-Jr (e-mail: williamtpjunior@discente.ufg.br).

ABSTRACT Virtualized RAN (vRAN) matches O-RAN Alliance specifications while transitioning towards virtualized functions on general-purpose computing platforms. However, the energy consumption of these systems remains a major concern. Although this issue has been addressed in the literature, previous works oversimplify routing decisions, overlook the benefits of flexible split choices, or neglect the energy consumption of the transport network. Additionally, most studies employing optimal solutions exhibit very limited scalability due to their high computational time. In this work, we present a comprehensive and efficient Mixed Integer Linear Programming model to minimize the energy consumption of O-RAN systems, addressing the limitations of current approaches. We also design and implement a synthetic data generator to evaluate our model across various network usage profiles, topologies, and reasonable-size networks. We achieved valuable insights and promising results in our evaluation. For example, our results show that when devices require high throughput, the transport network incurs significant energy costs and reduces the centralization rate. We also observed that hierarchical RAN topologies can achieve greater energy efficiency than ring topologies, with our approach enabling up to 15% more centralization while saving around 28% of energy and consuming at least one order of magnitude less time than other strategies.

INDEX TERMS End-to-end resource allocation, energy efficiency, O-RAN functional splits, virtualized RAN.

I. Introduction

THE Open Radio Access Network (O-RAN) Alliance is an attempt in the telecom industry to redesign the Radio Access Network (RAN) technologies [1]. The main principles supported by O-RAN are open interfaces for hardware and software interoperability between different manufacturers, the deployment of RAN intelligent controllers beyond data/control plane separation as stated by Software Defines Network (SDN) concept, and the virtualization of the RAN [2]. The introduction of the Network Function Virtualization (NFV) paradigm allows the RAN protocol stack associated with a given Base Station (BS) to be converted into Virtualized Network Functions (VNFs) to be processed by General Purpose Processors (GPPs). This transformation decouples radio software components from the underlying hardware, leading to improved management flexibility and cost reductions in RAN deployments by enabling infrastructure sharing [3].

The pooling of Baseband Units (BBUs) from multiple BSs in a single location is the main proposal of Centralized RAN (C-RAN). Virtualized RAN (vRAN) leverages C-RAN architecture by introducing virtualization of VNFs. However, this complete centralization requires a high capacity and a low-latency transport network, which can hinder practical deployment. To address these challenges, the O-RAN architecture specification proposes the disaggregation of the RAN protocol stack to realize such decoupling and take advantage of VNFs centralization, while presenting more flexible requirements. This centralization brings several benefits to the network, such as reducing the signaling required for mobile devices, allowing the most efficient utilization of hardware, and enabling Coordinated Multi-Point (CoMP) techniques. Disaggregation in O-RAN is realized through the split of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

the RAN protocol stack according to eight options and the distribution of VNFs between up to three logical units: the Radio Unit (RU), the virtual Distributed Unit (vDU), and the virtual Centralized Unit (vCU) [4].

Figure 1 shows an example of a RAN topology according to O-RAN specification. RU processes the parts related to Radio Frequency (RF) and Low-PHY, while the upper layers are processed in DU and CU. A Distributed Unit (DU) may aggregate the flow from a set of nearby RUs, while a Centralized Unit (CU) may further centralize the processing of the flow coming from multiple DUs. CUs and DUs can be virtualized, i.e., vCU and vDU, and processed on GPPs located in Computing Resources (CRs) spread across different network points, as illustrated in the figure. The disaggregated RAN units are connected to the core network through a crosshaul transport network [5], where the routing from RAN to core can be divided into up to three components: (i) fronthaul: path from RU to vDU; (ii) midhaul: path from vDU to vCU; and (iii) backhaul: path from vCU to core.

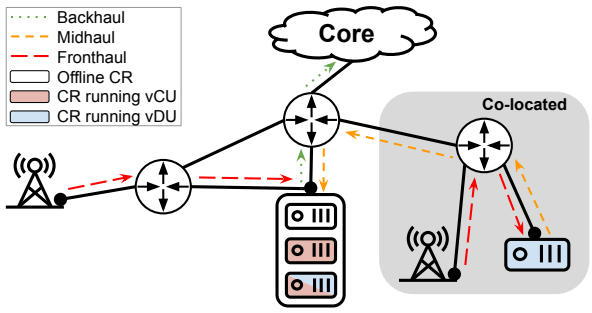


FIGURE 1. Example of vRAN topology and transport network route.

The problem of determining in which CR the VNFs of each BS should be processed (i.e., RAN VNF placement) has been widely studied in the literature [2], [3], [6], [7]. However, the energy consumption of vRAN systems remains a major concern because to offer high throughput links to users, 5G and Beyond networks require higher RAN densification [8]. In addition, the CRs spread along the network will foster computationally demanding use cases, further intensifying energy usage [9]. Indeed, operators are currently under pressure to reduce their energy consumption as telecoms account for 2-3% of total global energy demand, and energy cost represents 20-40% of mobile network operators' expenses [10].

Four challenges emerge to design an effective model to minimize the energy consumption of advanced vRAN systems based on O-RAN:

1) *System comprehensiveness*: modeling the energy consumption of all the system components from the RU to the core is particularly challenging because it depends on an accurate representation of the digital operations at the radio site, the computing and optical transport resources demanded by such operations, and the routing possibilities across the crosshaul toward the core.
2) *Network flexibility*: representing the system components with flexibility is also challenging as the model must consider all possible disaggregated RAN units (RU, vDU, and vCU), the segments between those units (fronthaul, midhaul, and backhaul), and all possible functional split options.
3) *CR and transport adaptability*: given the dynamic characteristic of mobile network usage, the load on RUs may vary. Load variation impacts CR requirements and transport network usage, raising the need to calculate new solutions periodically. However, migrating a VNF processing from one CR node to another can result in additional energy costs on the transport network. Therefore, energy cost overhead due to migration must be accounted for when providing a new solution.
4) *Fast response time*: wireless conditions change very fast in mobile networks due to the high mobility of users [11]. To enable control and optimization of RAN resources, the O-RAN architecture defines a non-real time intelligent controller, where rApps operates within control loops of 1 second or more [12]. Therefore, for practical evaluation, the model must provide a solution at the same time granularity.

Although previous works [13]–[19] have addressed the problem of minimizing the energy consumption of vRAN systems, they either oversimplify routing decisions, overlook the benefits of flexible split choices, or neglect the energy consumption of the transport network. In addition, most of the work employing optimal solutions exhibits very limited scalability due to their high computational time. To fill this gap, in this work, we propose an Mixed Integer Linear Programming (MILP) model that addresses the four challenges highlighted in the literature. Our optimization model (i) considers relevant components related to energy consumption in the RAN and transport network, (ii) allows each BS to implement a different functional split option, (iii) considers the routing decision of each BS until the core, enabling a given CR to work both as vCU and vDU, (iv) considers the energy cost of migrating to a new solution, and (v) provides an efficient linear formulation capable of solving instances of considerable size while ensuring a lower bound centralization of VNFs.

To evaluate our optimization model, we have implemented a synthetic load generator based on Markov Chains to simulate the demand variation of RUs. In addition to our synthetic data generator, we have developed a network topology generator based on a real-world RAN topology database with latitude and longitude data. The topology generator accounts for specific characteristics of vRAN networks and produces a topology graph accordingly. Using the load and the topology generators, we assessed the performance of our optimization model across various network usage profiles, topologies,

and reasonable-size networks. We observed that although considering more features and challenges than previous optimal energy-efficient models, our model finds the optimal solution in at least one order of magnitude less time than other strategies proposed in the literature. We also verified valuable insights that dynamic orchestration algorithms can further explore. For instance, our results show that when user devices require high throughput, the transport network incurs significant energy costs and reduces the centralization rate. We also observed that hierarchical RAN topologies can achieve greater energy efficiency than ring topologies, with our model enabling up to 15% more centralization while saving around 28% of energy. To summarize, the main contributions of this work are:

- We formulate a flexible and comprehensive Integer Linear Program (ILP) model to minimize the energy consumption of vRAN systems.
- Given the efficiency of our linear model, we can obtain the optimal solution for instances of considerable size in extremely fast time (milliseconds).
- Since the ILP model represents a NP-Hard problem, we also propose a heuristic algorithm with polynomial complexity to provide a faster solution for instances where an optimal solution cannot be found within an acceptable time frame.
- We provide valuable insights on how service requirements and network topologies impact both energy consumption and the centralization rate of the deployment.
- We make the source code and data related to the optimization model, along with the heuristic algorithm, publicly available in a GitHub repository[3], as well as the load and topology generators.

This article is organized as follows. Section II describes the related work. Section III shows the system model and the problem formulation. Section IV presents the load and topology generators. Section V discusses the results. Final considerations are made in Section VI.

II. Related Work

The VNF placement problem in vRAN is usually studied in the literature aiming to maximize VNF centralization (i.e. the number of VNFs from different RUs being processed on the same CR), while constrained by transport network and processing capacity [2], [3], [6], [7]. Nevertheless, the work in [21] shows that centralizing VNFs is not guaranteed to bring energy-related benefits. It formulates the energy efficiency in vRAN as a bi-objective function and, therefore, analyzes a tradeoff between centralization and energy consumption, but not a decisive solution deployed periodically.

Some works have also investigated the VNF placement problem in vRAN, seeking to minimize the network's energy consumption. The authors in [13] propose a heuristic to solve the association problem between CU, DU, and User Equipment (UE), aiming to minimize network energy consumption while reducing mobile device handovers. However, they consider DU and RU with fixed association, disregarding the impact of dynamic choices of RAN split options, which may lead to under-utilization of resources during off-peak moments when a DU can handle a higher number of RUs. GreenRAN [14] formulates the problem of VNF placement in vRAN to minimize energy consumption as a quadratic integer program. Given its complexity, the authors propose a metaheuristic to solve a relaxed version of the problem. VNF migration cost and the split of RAN into up three units are considered. However, the excessive simplification of network routing and the omission of the energy consumption from the transport network can result in sub-optimal decisions in scenarios where the transport network's energy cost dominates the problem.

A reinforcement learning algorithm is proposed in [15] to decide the placement of VNFs among a set of DUs with different capacities and the scheduling of radio resource blocks for each UE demand, aiming to reduce energy consumption by turning off idle DUs while ensuring heterogeneous latency requirements for devices. However, the energy consumption of the transport network is not represented and only a single split option, C-RAN, is considered, disregarding possible energy benefits from more flexible split options. In addition, C-RAN is the most challenging split option to implement in practice due to its strict requirements on the transport network [22], which limits the practicality of the solution.

In [16], the authors formulate the problem of deciding the RAN function split performed per network slice to reduce the energy consumption of CRs and transport network. A heuristic is proposed to present a good solution in a reasonable time. However, the routing is oversimplified, limiting the association between RU and DU. This may prevent associations with better energy efficiency. The problem of VNFs split assignment for each BS in a set of time slots is solved using a deep reinforcement learning approach in the article [17]. Both in [16] and [17], only the energy consumption of the midhaul is considered, whereas the energy consumption of the backhaul and midhaul may be significant enough to impact the quality of the solution. Zorello et al. [18] studied the CU and DU placement problem while considering different 5G service requirements for bandwidth and latency. A heuristic algorithm is proposed and compared with the optimal solution. They showed that routing choices have a significant impact on power consumption. In [19], the authors formulate the problem of minimizing energy consumption in a C-RAN scenario. However, C-RAN represents only one of the possible functional splits of the RAN. In [20], joint allocation of cloud, network, and radio resources per UE is performed in a cell-free mobile network to reduce power consumption. The authors consider a single cloud connected to all RUs, which simplifies the routing and may

[3]https://github.com/LABORA-INF-UFG/paper-WGSCLK-2025

TABLE I. Summary of related work in vRAN energy consumption

Article	Comprehensiveness			Flexibility		Solution	Instance Size*	Response Time
	RAN	TNet	Mig	Splitting	Routing			
Apt-RAN [13]	●	○	●	●	○	Optimal†	50 BSs	10^4s
GreenRAN [14]	●	○	●	●	○	Metaheuristic	25 BSs	10^2s
Mollahasani et al. [15]	●	○	○	○	○	Reinforcement Learning	–	–
Sen et al. [16]	●	◐	○	●	○	Optimal†	30 Slices	10^1s
Amiri et al. [17]	●	◐	○	●	○	Deep Reinforcement Learning	–	–
Zorello et al. [18]	●	●	○	○	●	Optimal†	10 BSs	10^3s
OptiLoop [19]	●	●	○	○	●	Heuristic	–	–
Demir et al. [20]	●	●	○	●	◐	Heuristic	–	–
Our Proposal	●	●	●	●	●	Optimal	50 BSs	10^{-1}s

Symbols: ● Considered - ◐ Partially considered - ○ Not considered.

RAN and TNet represent the energy consumption of the radio access network and the transport network, respectively.

Mig represents the energy cost of migration.

* Instance size is related to the response time column. We omit this value when solution response time is not available.

† This work also presents a heuristic solution; however, we consider the optimal solution for comparability.

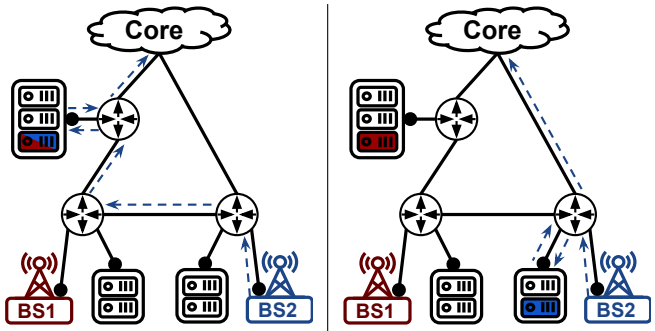
not represent all possible topologies. Moreover, none of these works ([15]–[20]) considers the energy consumption of VNF migration, hindering the use of the proposed solutions with varying network load over time, as small changes in load may trigger a migration of VNFs where the energy savings from the new solution do not outweigh the migration overhead.

Table I summarizes the challenges addressed in the related literature and the instance size and response time achieved by the existing solutions. To the best of our knowledge, our work is the first to address all the considered challenges, as illustrated in the table. However, it is important to highlight that given the NP-hard nature of the addressed problem, using off-the-shelf solvers to obtain an optimal solution still involves exponential complexity, making it impractical for solving larger topologies. As shown in Table I, our efficient formulation can rapidly achieve an optimal solution for instances of considerable size, providing a robust baseline to evaluate non-optimal solutions in future works.

III. System Model and problem statement

We consider a RAN topology composed of a set $\mathcal{B} = \{b_1, b_2, \ldots, b_{|\mathcal{B}|}\}$ of RUs and a set of general-purpose servers $\mathcal{H} = \{h_1, h_2, \ldots, h_{|\mathcal{H}|}\}$ capable of processing the RAN VNFs. RUs and servers are connected through a set of transport network nodes $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$. We define the set of CRs as $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ where $c_m \subseteq \mathcal{T} \cup \mathcal{H}$ represents a group of co-located switches and servers. We represent the RAN topology as a graph $G = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V} = \{v_0\} \cup \mathcal{B} \cup \mathcal{T} \cup \mathcal{H}$ denotes the vertices, where v_0 represents the core network, and $\mathcal{E} = \{e_{ij} \mid v_i, v_j \in \mathcal{V}\}$ represents the set of edges corresponding to the network links connecting the nodes. This graph structure is flexible enough to enable the representation of any possible RAN topology design, from a traditional distributed RAN topology to infrastructures featuring cloud and edge computing centers.

Furthermore, in practical network deployment, the graph-based structure of our system model supports the implementation of our solution as a rApp in the Non-RT RIC of the O-RAN architecture [23]. This is particularly justified as the rApp manages the placement of virtualized radio functions across multiple base stations, leveraging a broader network perspective [24], [25].



(a) Long route. (b) Short route.

FIGURE 2. Example of different routes in the same vRAN topology.

Routing – All data traffic originates (downlink) or terminates (uplink) at the core network v_0. The best route for a BS depends highly on the power efficiency of the transport network and processing equipment, as well as the data load generated at the BS. In a scenario where the cost of turning on a server has the greatest impact on energy consumption, a longer route, as illustrated in Figure 2(a), can lead to lower energy costs. Conversely, if the energy cost of the transport network is the dominant factor, a shorter route, as shown in Figure 2(b), can be the most beneficial choice. As the size of the topology increases, this trade-off becomes more complex and must be considered in the formulation. Without loss of generality, we consider only the downlink flow. We define $\mathcal{P}_l$ as the set of paths

from each RU $b_l \in \mathcal{B}$ to the core network v_0, where each path $p \in \mathcal{P}_l$ may be split into up to three sub-paths: p_{Bh} (backhaul), p_{Mh} (midhaul), and p_{Fh} (fronthaul). Moreover, we consider a crosshaul transport network, i.e., a link $e_{ij} \in \mathcal{E}$ can serve any sub-path combination.

Virtual Network Functions – For each RU $b_l \in \mathcal{B}$, our objective is to determine the best CR to deploy the RU set of RAN VNFs, denoted by $\mathcal{F} = \{f_1, f_2, f_3, f_4, f_5\}$ representing High-PHY, MAC, RLC, PDCP, and RRC functions respectively. The VNF stack may be partitioned up to two times according to the functional splits $\mathcal{D} = \{O_0, O_1, O_2, O_4, O_6, O_9\}$ defined in Table II (according to [26]), which enhances the flexibility of the solution by allowing parts of the VNF stack to be deployed in different CRs. The Low-PHY function is not virtualized and is assumed to always be deployed at the RU. For our formulation, option O_0 denotes the Disaggregated RAN (D-RAN), where no split is performed, and all the VNFs run on a server co-located with the RU.

TABLE II. Functional split options for the RAN

SPLIT	L-PHY	H-PHY	MAC	RLC	PDCP	RRC
1	RU					CU
2	RU				CU	
4	RU			CU		
6	RU		CU			
7.2	RU	CU				

A. VNF Processing

Let the binary decision variable $x_l^{p,r}$ denote if the path $p \in \mathcal{P}_l$ was chosen for the $b_l \in \mathcal{B}$ using the functional split $D_r \in \mathcal{D}$. However, during the decision, we must ensure that the assigned servers selected by the path $p \in \mathcal{P}_l$ have the required processing capacity. To achieve this, we formulate the computing resource cost for each VNF in terms of Giga Operations Per Second (GOPS) as described in the following.

The High-PHY layer runs pre-coding (relating to channel prediction and reciprocity calibration), modulation, and resource element mapping. Similar to [20], [27], we formulate the processing of those operations as follows:

$$C_{precoding} = \frac{N_{used}}{T_s \tau_c 10^9}\big(8N_l\tau_p^2 + 8N_l^2(\tau_p + Load(b_l))\big) + \frac{N_{used}\tau_d}{T_s\tau_c 10^9}\big(8N_l Load(b_l)\big) + \frac{N_{used}}{T_s\tau_c 10^9}\big(8N_l Load(b_l)\big) + \frac{N_{used}}{T_s\tau_c 10^9}\Big((4N_l^2 + 4N_l)\tau_p + 8N_l^2 Load(b_l) + \frac{8(N_l^3 - N_l)}{3}\Big), \quad (1)$$

$$C_{modulation} = 1.3N_l\left(\frac{N_{bits}}{16}\right)^{1.2}, \quad (2)$$

$$C_{mapping} = 1.3Load(b_l)\left(\frac{N_{bits}}{16}\right)^{1.2}\left(\frac{SE_0}{6}\right)^{1.5}, \quad (3)$$

where N_{used} denotes the number of sub-carriers used. T_s is related to the duration of the Orthogonal Frequency Division Multiplexing (OFDM) symbol. τ_c, τ_p, and $\tau_d = \tau_c - \tau_p$ represent the number of samples per coherence block, the number of received samples during the training phase, and the number of samples during downlink data transmission, respectively. N_l stands for the number of antennas in RU $b_l \in \mathcal{B}$. $Load(b_l)$ is a function over the input data that returns the current number of devices associated to RU $b_l \in \mathcal{B}$. N_{bits} represents the number of bits used for data quantization and SE_0 denotes the spectral efficiency of the channel. Therefore, the number of GOPS required for High-PHY layer processing is given by:

$$C_{HighPHY} = C_{precoding} + C_{modulation} + C_{mapping}. \quad (4)$$

Channel coding, system control, and data redirection to the core network operations are implemented in the higher (or superior) layers of the RAN VNF stack. Similar to [27], the number of GOPS required by all these superior layers is calculated as follows:

$$C_{SupLayers} = 1.3Load(b_l)\left(\frac{N_{bits}}{16}\right)^{1.2}\left(\frac{SE_0}{6}\right) + 2.7\sqrt{N_l}\left(\frac{N_{bits}}{16}\right)^{0.2} + 8Load(b_l)\left(\frac{SE_0}{6}\right). \quad (5)$$

To determine the processing required by each VNF individually, given the aggregate load of the superior layers $C_{SupLayers}$ from (5), we utilize proportions based on the CPU utilization profiles observed in the OpenAirInterface (OAI) implementation [28]. Finally, the total processing load γ_w for a given server $h_w \in \mathcal{H}$ is calculated as follows:

$$\gamma_w = \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} \Big[F_1(h_w, D_r, p) C_{HighPHY} + F_2(h_w, D_r, p)\Big(0.4\ C_{SupLayers}\Big) + F_3(h_w, D_r, p)\Big(0.028\ C_{SupLayers}\Big) + F_4(h_w, D_r, p)\Big(0.286\ C_{SupLayers}\Big) + F_5(h_w, D_r, p)\Big(0.286\ C_{SupLayers}\Big)\Big], \quad (6)$$

where $F_n(h_w, D_r, p)$, based on the input data, returns 1 when server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ according to functional split $D_r \in \mathcal{D}$ and route $p \in \mathcal{P}_l$; otherwise, it returns 0.

B. Optimization Model

We formulate the problem of RAN VNF placement to minimize the energy consumption as a MILP model. Moreover, we decompose the energy consumption into three main components, which are detailed below.

vRAN energy consumption – The energy consumed by the general purpose servers processing the VNFs of all RUs

$b_l \in \mathcal{B}$ characterizes the vRAN energy consumption. We use a traditional energy model [29] to estimate this energy consumption, in which the total energy consumed by a server $h_w \in \mathcal{H}$ over the period T includes a static power component P_w^{idle} consumed whether the server is turned on, and a dynamic load-dependent power consumption $P_w^{busy} - P_w^{idle}$, defined as:

$$E_{vRAN} = \sum_{h_w \in \mathcal{H}} T\Big[\psi_w^{on} P_w^{idle} + \big(\gamma_w / C_w^{cap}\big)\big(P_w^{busy} - P_w^{idle}\big)\Big], \tag{7}$$

where γ_w is the total load assigned to server $h_w \in \mathcal{H}$ as calculated in (6), C_w^{cap} represents the number of GOPS the server $h_w \in \mathcal{H}$ can perform. ψ_w^{on} is a ceiling function that ensures that a server $h_w \in \mathcal{H}$ is counted as active if, and only if, at least one VNF of any RU is assigned to it. This ceiling function is defined as follows:

$$\psi_w^{on} = \left\lceil \sum_{f_n \in \mathcal{F}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} u_w^p M(h_w, f_n, b_l, D_r)}{|\mathcal{F}|\ |\mathcal{B}|} \right\rceil, \tag{8}$$

where $u_w^p \in \{0,1\}$ is based on the input data, indicating whether a server $h_w \in \mathcal{H}$ is part or not of the route $p \in \mathcal{P}_l$. The mapping function $M(h_w, f_n, b_l, D_r)$ returns 1 when the server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ from RU $b_l \in \mathcal{B}$ according to functional split $D_r \in \mathcal{D}$.

Transport network energy consumption – We consider an optical transport network dedicated to the RAN infrastructure where the energy consumption comes from (i) Ethernet switches, allowing higher routing flexibility by enabling packet aggregation, and (ii) Dense Wavelength Division Multiplexing (DWDM) pluggable transceivers that can be directly installed in radio devices and switches [30], [31]. Each link $e_{ij} \in \mathcal{E}$ is characterized by its data transmission capacity $R_{e_{ij}}^{tr}$ and power consumption $P_{e_{ij}}^{tr}$ of the transceivers at each end of the link. For each topology node $v_k \in \mathcal{V}$, the function $S(v_k) \in \{0,1\}$ indicates whether v_k is a packet switch, while $P_{v_k}^s$ represents the power consumed by each switch port. Finally, The transport network energy consumption is defined as follows:

$$E_{TNet} = \sum_{e_{ij} \in \mathcal{E}} \left[T \frac{\gamma_{e_{ij}}}{R_{e_{ij}}^{tr}} \Big(2P_{e_{ij}}^{tr} + S(v_j) P_{v_j}^s + S(v_i) P_{v_i}^s \Big) \right]. \tag{9}$$

$\gamma_{e_{ij}}$ in (9) represents the total throughput over link e_{ij}, which is defined as:

$$\gamma_{e_{ij}} = \sum_{D_r \in \mathcal{D}} \sum_{b_l \in B} \sum_{p \in \mathcal{P}_l} x_l^{p,r} R^l \left(y_{e_{ij}}^{p_{Bh}} \alpha_{Bh}^{r,l} + y_{e_{ij}}^{p_{Mh}} \alpha_{Mh}^{r,l} + y_{e_{ij}}^{p_{Fh}} \alpha_{Fh}^{r,l} \right), \tag{10}$$

where the data throughput $R^l = Load(b_l) R^{\text{dev}}$ generated by RU $b_l \in \mathcal{B}$ is a dynamic parameter that fluctuates over time, and can be estimated by the number of devices connected to the RU at a given instant and the mean throughput generated by these devices. $y_{e_{ij}}^{p_{Bh}}$, $y_{e_{ij}}^{p_{Mh}}$, and $y_{e_{ij}}^{p_{Fh}}$ indicate if the link e_{ij} is part of the backhaul, midhaul, or fronthaul, respectively, for the path $p \in \mathcal{P}_l$. For each RU $b_l \in \mathcal{B}$, the chosen functional split $D_r \in \mathcal{D}$ increases the data rate required at the backhaul, midhaul, and fronthaul by the factors $\alpha_{Bh}^{r,l}$, $\alpha_{Mh}^{r,l}$, and $\alpha_{Fh}^{r,l}$, respectively. As the solution is expected to be effective during the time period T, we consider the estimated number of active links $\gamma_{e_{ij}} / R_{e_{ij}}^{tr}$ as a possible non-integral real value to account for the deactivation of the link during periods of inactivity.

VNF migration energy consumption – To ensure that the energy cost overhead due to migration does not outweigh the energy gains in the new solution, we use a linear approximation model based on empirical data, similar to [13], [14]. As shown in Section V, VNF migration is the component with the lowest impact on total energy consumption. Therefore, despite the possible inaccuracies inherent in linear approximations, this approach has a limited impact on the quality of the solution and complies with our objective of preserving the linearity of the formulation. Considering that each VNF is processed individually in its own Virtual Machine (VM), allowing for flexible function placement, we estimate the energy cost of migrating a specific VNF $f_n \in \mathcal{F}$ as $E_{f_n} = aV_{fn} + b$, where V_{fn} is the memory volume of the VM hosting the VNF $f_n \in \mathcal{F}$, and coefficients a and b are derived from experimental observations [32], which maps the data traffic of VM migration to energy consumption. Lastly, we define the total VNF migration energy consumption by the following equation:

$$E_{Mig} = \sum_{h_w \in \mathcal{H}} \sum_{f_n \in \mathcal{F}} \sum_{D_r \in \mathcal{D}} \sum_{b_l \in \mathcal{B}} \sum_{p \in P_l} \big[1 - N(h_w, f_n, b_l)\big] x_l^{p,r} u_w^p M(h_w, f_n, b_l, D_r) E_{f_n}, \tag{11}$$

where $N(h_w, f_n, b_l)$ is defined over the input data, resulting in 1 when server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ from RU $b_l \in \mathcal{B}$ for the previous VNF deployment. Otherwise, it returns 0.

The objective is to minimize the total energy consumption needed to process the vRAN VNFs. This total energy consumption is impacted by the hardware processing the VNFs, the transport network connecting the CRs, and the migration of VNFs given a previous deployment. Therefore, we define the objective function as follows:

$$\underset{x_l^{p,r}, \forall b_l, \forall D_r, \forall p}{\text{minimize}} \quad E_{vRAN} + E_{TNet} + E_{Mig} \tag{12a}$$

subject to

$$\sum_{c_m \in \mathcal{C}} \sum_{f_n \in \mathcal{F}} \Bigg(\sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} v_{m,w}^{p} M(h_w, f_n, b_l, D_r) - \psi_{m,n}^{single} \Bigg) \geq \rho^c, \tag{12b}$$

$$\sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} = 1, \quad \forall b_l \in \mathcal{B}, \tag{12c}$$

$$\gamma_{e_{ij}} \leq e_{ij}^{Cap}, \quad \forall e_{ij} \in \mathcal{E}, \tag{12d}$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Bh}} e_{ij}^{L} \leq \beta_{Bh}^{r}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{12e}$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Mh}} e_{ij}^{L} \leq \beta_{Mh}^{r}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{12f}$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Fh}} e_{ij}^{L} \leq \beta_{Fh}^{r}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{12g}$$

$$\gamma_w \leq C_w^{cap}, \quad \forall h_w \in \mathcal{H}, \tag{12h}$$

$$x_l^{p,r} \in \{0,1\}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l. \tag{12i}$$

We define centralization as the number of VNFs from different RUs being processed in the same CR, i.e., in servers at the same geographical location. The constraint in (12b) ensures a lower bound centralization ρ^c of VNFs, where $v_{m,w}^{p}$, defined over the input data, indicates whether a server $h_w \in \mathcal{H}$ is associated with CR $c_m \in \mathcal{C}$ and is part of route $p \in \mathcal{P}_l$. The term ψ_m^{single} is an expression and assures that at least two RUs $b_l \in \mathcal{B}$ must have the same type of VNF $f_n \in \mathcal{F}$ allocated to the same CR $c_m \in \mathcal{C}$ to count as centralization, which is formulated as:

$$\psi_{m,n}^{single} = \left\lceil \sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} v_{m,w}^{p} M(h_w, f_n, b_l, D_r)}{|\mathcal{F}| \ |\mathcal{B}|} \right\rceil. \tag{13}$$

For each RU $b_l \in \mathcal{B}$, only one combination of route $p \in \mathcal{P}_l$ and functional split $D_r \in \mathcal{D}$ must be assigned, as represented in the constraint (12c).

Each link $e_{ij} \in \mathcal{E}$ has a maximum data rate capacity e_{ij}^{Cap} defined by the number of fibers and data rate of transceivers composing it. This maximum capacity must not be exceeded, as represented in the constraint (12d). Furthermore, scenarios where the transport network is shared with other services can be considered by extending constraint (12d).

Depending on the functional split $D_r \in \mathcal{D}$ chosen, different latencies must be granted at *fronthaul* (β_{Fh}^{r}), *midhaul* (β_{Mh}^{r}), and *backhaul* (β_{Bh}^{r}) of path $p \in \mathcal{P}_l$. Since each link $e_{ij} \in \mathcal{E}$ incurs in delay e_{ij}^{L} according to its capacity, distance between nodes, number of hops, and packet queue size, the chosen path $p \in \mathcal{P}_l$ must ensure the latency required by functional split $D_r \in \mathcal{D}$, as defined in the constraints (12e) – (12g).

The VNFs assigned to a given server $h_w \in \mathcal{H}$ must not exceed its maximum processing capacity C_w^{cap}, as defined in the constraint (12h). Finally, the constraint in (12i) specifies that the decision variable is binary.

C. Linearization

The problem stated in (12) represents an integer program formulation, which is known to be an NP-hard problem, as demonstrated by the authors of [7]. However, the ceiling function in (8) and (13) renders those constraints discontinuous. This approach makes the model nonlinear, i.e., unsupported by MILP solvers and dependent on inefficient solutions. Therefore, we linearize these constraints by introducing two new integer decision variables, y_w and y_m, for each server, $h_w \in \mathcal{H}$ and CR $c_m \in \mathcal{C}$. Regarding the former and considering the following relaxation of ψ_w^{on}:

$$\psi_w^{on'} = \sum_{f_n \in \mathcal{F}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} u_w^{p} M(h_w, f_n, b_l, D_r)}{|\mathcal{F}| \ |\mathcal{B}|}, \tag{14}$$

we can mimic the behavior of the ceiling function by adding the following constraints:

$$y_w \geq \psi_w^{on'}, \tag{15}$$

$$y_w \leq \psi_w^{on'} + 1 - \epsilon. \tag{16}$$

Since the lowest value $\psi_w^{on'}$ can assume is $1/(|\mathcal{F}| \ |\mathcal{B}|)$, the integrity parameter ϵ must be chosen as $\epsilon < 1/(|\mathcal{F}| \ |\mathcal{B}|)$. This assumption ensures that y_w always remains equivalent to ψ_w^{on}, even when $\psi_w^{on'}$ approaches an integer value, thereby preventing any inconsistencies in the formulation. Similarly, for y_m, we define $\psi_m^{single'}$ as the relaxation of ψ_m^{single}:

$$\psi_m^{single'} = \sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} v_{m,w}^{p} M(h_w, f_n, b_l, D_r)}{|\mathcal{F}| \times |\mathcal{B}|}, \tag{17}$$

and formulate the following constraints:

$$y_{m,n} \geq \psi_{m,n}^{single'}, \tag{18}$$

$$y_{m,n} \leq \psi_{m,n}^{single'} + 1 - \epsilon. \tag{19}$$

To conclude, after replacing ψ_w^{on} with y_w in (7) and ψ_m^{single} with y_m in (12b), we end up with:

$$\underset{\substack{x_l^{p,r}, \forall b_l, \forall D_r, \forall p \\ y_{m,n}, \forall c_m, \forall f_n \\ y_w \forall h_w}}{\text{minimize}} \quad E_{vRAN} + E_{TNet} + E_{Mig}.$$

subject to

$$\sum_{c_m \in \mathcal{C}} \sum_{f_n \in \mathcal{F}} \left(\sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} v_{m,w}^{p} M(h_w, f_n, b_l, D_r) - y_m \right) \geq \rho^c,$$

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

$$\gamma_{e_{ij}} \leq e_{ij}^{Cap}, \quad \forall e_{ij} \in \mathcal{E},$$

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

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

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

$$\gamma_w \leq C_w^{cap}, \quad \forall h_w \in \mathcal{H},$$

$$y_w \geq \psi_w^{on'}, \quad \forall h_w \in \mathcal{H},$$

$$y_w \leq \psi_w^{on'} + 1 - \epsilon, \quad \forall h_w \in \mathcal{H},$$

$$y_{m,n} \geq \psi_{m,n}^{single'}, \quad \forall c_m \in \mathcal{C}, f_n \in \mathcal{F},$$

$$y_{m,n} \leq \psi_{m,n}^{single'} + 1 - \epsilon, \quad \forall c_m \in \mathcal{C}, f_n \in \mathcal{F},$$

$$x_l^{p,r} \in \{0,1\}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l,$$

$$y_w, y_{m,n} \in \mathbb{Z}, \quad \forall c_m \in \mathcal{C}, h_w \in \mathcal{H}, f_n \in \mathcal{F},$$

which represents a more efficient MILP formulation, as illustrated by the results presented in Section V.

Table III summarizes all the decision variables, sets, and data parameters used throughout the formulation.

D. Heuristic

Problems formulated as MILP are known to be NP-Hard, presenting non-polynomial complexity to solve. To obtain a faster solution for instances that cannot be solved within an acceptable time using the MILP model, we propose a heuristic solution.

Algorithm 1: VNF Placement For The First Level

Input : $L_n, \mathcal{G}, \mathcal{D}^*, \mathcal{P}, \mathcal{S}^{-1}$.
Output: Partial set of associations between BS, route and functional split $\mathcal{S}$.

$\mathcal{S} \leftarrow \varnothing$
Sort set L_n, for each n, by BS load
for $v_i \in L_1$ **do**
  if $v_i \notin \mathcal{B}$ **then**
    continue
  $p \leftarrow$ No-split route for v_i with the most power-efficient available server
  if *Migrating to* (v_i, p, O_0) *is worthwhile* **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, O_0)$

Algorithm 2: VNF Placement For Level 2 and Beyond

Input : $L_n, \mathcal{G}, \mathcal{D}^*, \mathcal{P}, \mathcal{S}^{-1}, \mathcal{S}$.
Output: Complete set of associations between BS, route and functional split $\mathcal{S}$.

Sort set L_n, for each n, by BS load
for $n > 1$ **do**
  for $v_i \in L_n$ **do**
    if $v_i \notin \mathcal{B}$ **then**
      continue
    $\mathcal{C}_{cands.} \leftarrow$ CRs with turned on server
    Reverse sort $\mathcal{C}_{cands.}$ by number of associated BSs
    $p_0 \leftarrow$ No-split route for v_i with the most power-efficient available server
    candAssociation $\leftarrow (v_i, p_o, O_0)$
    feasible $\leftarrow$ False
    for $c_m \in \mathcal{C}_{cands.}$ **do**
      for *route* p *that contains candidate CR* c_m **do**
        for $O_j \in \mathcal{D}^*$ **do**
          if (v_i, p, O_j) *is unfeasible* **then**
            continue
          feasible $\leftarrow$ True
          if (v_i, p, O_j) *consumes less energy than candAssociation* **then**
            candAssociation $\leftarrow (v_i, p, O_j)$
    if *feasible* **then**
      if *Migrating to* (v_i, p, O_j) *is worthwhile* **then**
        $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, O_j)$
    else
      $c_m \leftarrow$ a CR with available server to turn on, prioritizing already active CRs
      `Assuming that the BS load in` v_i `is equivalent to the load of all` v_x `not yet associated in` L_n
      if $\exists p \in \mathcal{P}$ *and* $\exists O_r \in \mathcal{D}^*$ *such that* (v_i, p, O_r) *is feasible and consumes less energy than performing no-split* **then**
        if *Migrating to* (v_i, p, O_r) *is worthwhile* **then**
          $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, O_r)$
      else
        $p \leftarrow$ No-split route for v_i with the most power-efficient available server
        if *Migrating to* (v_i, p, O_0) *is worthwhile* **then**
          $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, O_0)$
        Try to apply no-split option for the remaining nodes in distance n
        Iterate loop in line 2

TABLE III. Sets, Input Data, Decision Variables, and Expressions

	Notation	Description
Sets	$\mathcal{B}$	Set of RUs
	$\mathcal{H}$	Set of general-purpose servers
	$\mathcal{T}$	Set of transport network nodes
	$\mathcal{C}$	Set of CRs
	$\mathcal{G}$	Topology graph where $G = (\mathcal{V}, \mathcal{E})$
	$\mathcal{V}$	Set of topology nodes where $\mathcal{V} = \{v_0\} \cup \mathcal{B} \cup \mathcal{T} \cup \mathcal{H}$
	$\mathcal{E}$	Set of topology links where $\mathcal{E} = \{e_{ij} \mid v_i, v_j \in \mathcal{V}\}$
	$\mathcal{P}_l$	Set of k-shortest paths from each RU $b_l \in \mathcal{B}$ to the core network v_0
	$\mathcal{F}$	Set of VNFs
	$\mathcal{D}$	Set of functional splits
Input Data	ρ^c	VNF centralization lower bound
	C_w^{cap}	Processing capacity of server $h_w \in \mathcal{H}$
	P_w^{busy}	Load-dependent power consumption of server $h_w \in \mathcal{H}$
	P_w^{idle}	Static power consumption of server $h_w \in \mathcal{H}$
	E_{f_n}	Memory volume of VM hosting the VNF $f_n \in \mathcal{F}$
	e_{ij}^{Cap}	Maximum link data rate
	e_{ij}^{L}	End-to-end link latency
	$R_{e_{ij}}^{tr}$	Transceiver transmission capacity
	$P_{e_{ij}}^{tr}$	Transceiver power consumption
	$P_{v_i}^{s}$	Ethernet switch port power consumption
	T	Expected solution deployment duration
	T_s	OFDM symbol duration
	N_{used}	Number of sub-carriers used
	N_{bits}	Number of bits used for data quantization
	N_l	Number of antennas in RU $b_l \in \mathcal{B}$
	u_w^p	Indicates whether a server $h_w \in \mathcal{H}$ is part of the path $p \in \mathcal{P}_l$
	$v_{m,w}^p$	Indicates whether a server $h_w \in \mathcal{H}$, part of the path $p \in \mathcal{P}_l$, is associated with CR $c_m \in \mathcal{C}$
	SE_0	Channel spectral efficiency
	τ_c	Number of samples per coherence block
	τ_p	Number of received samples during training phase
	τ_d	Number of received samples during downlink data transmission
Decision Variables and Expressions	$x_l^{p,r}$	Binary decision variable indicating that path $p \in \mathcal{P}_l$ was chosen for RU $b_l \in \mathcal{B}$ using functional split $D_r \in \mathcal{D}$
	y_w, y_m	Integer decision variables that mimic ψ_w^{on} and $\psi_{m,n}^{single}$
	γ_w	Processing load for server $h_w \in \mathcal{H}$
	$\gamma_{e_{i,j}}$	Total throughput over link $e_{i,j}$
	ψ_w^{on}	Indicate if server $h_w \in \mathcal{H}$ is assigned to the processing of any VNF and needs to be activated
	$\psi_{m,n}^{single}$	Indicate if at least two different RUs deploy the same VNF $f_n \in \mathcal{F}$ in CR $c_m \in \mathcal{C}$

Algorithms 1 and 2 present the heuristic to assign a route between each BS in the topology and the core network, as well as a functional split for the association. As input, we consider the sets L_n of nodes with distance n (in number of hops) from the core, the RAN topology graph $\mathcal{G}$, the set of functional splits $\mathcal{D}^*$ excluding the no-split option O_0, the set of routes $\mathcal{P}$ from every BS until the core, and the associations $\mathcal{S}^{-1}$ currently deployed in the network. Additionally, Algorithm 2 receives the partial solution from Algorithm 1.

Algorithm 1 creates a partial solution by assigning a route and functional split to every BS in the first hop from the core, if any, that cannot centralize their VNFs. In line 2 we sort the nodes in each level n (i.e., nodes with the same distance from the core) by BS load, so we associate the nodes with the least load first. In lines 3 to 8 a route with a no-split option is associated for BSs at level 1. To address the energy cost of VNF migration, in line 7 we evaluate whether the energy savings from the new association compensate for the energy overhead incurred by migrating from the previous solution.

For levels 2 and beyond, in Algorithm 2, we evaluate whether performing a split and using an active CR in the upper levels is more efficient than turning on a local server and opting for a no-split option, as detailed in lines 6 to 21. If none of the evaluated associations are feasible, we determine whether a new server should be activated in the upper levels or if the no-split option should be assigned to the remaining BSs in the current level, as addressed in lines 22 to 33. This decision is based on the assumption that the total load in the remaining BSs at this level is generated by the current BS being evaluated. This assumption implies that the greater the difference in efficiency between CRs and transport network links, the farther from optimal the heuristic solution is expected to be.

Complexity analysis – The proposed heuristic algorithm can find a satisfactory solution in polynomial time, as shown in Section V. In algorithm 1, line 2 will perform at worst $O(|\mathcal{V}|)$ sorting operations. Using the quick sort algorithm [33] for reference, this would be limited to a total complexity of $o(|\mathcal{V}|^3)$, since, by the definition of L_n, it is not possible to have $|\mathcal{V}|$ levels with $|\mathcal{V}|$ CRs. Lines 3 to 8 execute at most $O(|\mathcal{V}|)$ times. For algorithm 2, line 1 performs the same sort as already described. Lines 2 and 3 iterate over each CRs once, while line 11, in the worst case, handles all CRs again. Line 12 iterates through all routes for a given CR, and considering the use of k-shortest path algorithm, we have k paths where each is further divided into 1 route without splits, $|hops|-1$ routes with a single split, and $\sum_{j=1}^{|hops|-2} j = O(|hops|^2)$ routes with two splits. In the worst case, all paths have $|hops| = |\mathcal{V}| - 1$, limiting the number of iterations in line 12 to $O(k \cdot |\mathcal{V}|^3)$. Finally, lines 2 to 18 have complexity $O(k \cdot |\mathcal{V}|^5)$. The evaluation in line 25 is done similarly, having the same complexity.

IV. Synthetic Load Generator

Academia and industry have made significant strides in developing synthetic data generators to overcome the scarcity of publicly available real-world data [34]–[36]. Following the methodology proposed in [34], this work formulates a

synthetic data generator based on Markov Chains to simulate the demand variation of RUs. In our load generator, each RU is characterized by its region type (e.g., urban, rural, center, suburb), the time of trace capture (e.g., hour), the day type of capture (e.g., weekday (WD) or weekend (WE)), and the number of associated users. While the load generator presented in [34] considers a transition function for each day of the week, resulting in higher computational complexity in terms of time and space to compute the Markov Chain transition probability, our synthetic load generator minimizes the complexity of the Markov Chain model by representing the days of the week according to their type, e.g., weekday and weekend. This formulation strategy leverages the similarity of the mobility pattern between weekdays and weekends [37], minimizing the complexity of the Markov Chain model while maintaining its generality and accuracy in capturing demand variations.

We consider a set of states $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$, where each state $s \in \mathcal{S}$ is defined by the following tuple:

$$s = (A(b_l), DT, q, \delta), \tag{21}$$

where $A(b_l)$ represents the region type of RU $b_l \in \mathcal{B}$, $DT \in \{WD, WE\}$ denotes the day type, $q \in [0, 24]$ defines the hour of the day, and $\delta \in \mathbb{Z}_{\geq 0}$ represents the number of users associated to RU $b_l \in \mathcal{B}$ at a time q. For example, the tuple $s = (\text{Suburb}, \text{WE}, 9, 45)$ specifies that RU $b_l \in \mathcal{B}$ is in a suburb region on the weekend, and are 45 users associated to this RU at 9:00 AM.

The transition function calculates the probability of a RU in a region type $A(b_l)$ moving from demand δ_i at time q_i to demand δ_j at time q_j, with $i < j$. This function is defined as:

$$\rho = (A(b_l), DT, q_i, \delta_i, q_j, \delta_j). \tag{22}$$

Algorithm 3 describes how we calculate the transition probability of our model, defining the set of states $\mathcal{S}$ and the probabilistic transition function $\rho(s_i, s_j)$ for all pairs of states. As input, we consider the set of RUs $\mathcal{B}$, the set of region types A, the set of days D, the set of times Q, and the number of users associated with each RU, i.e., Δ. Lines 2 to 18 calculate the set of states based on the combination of region type $A(b_l)$, day d, time q, and the number of users associated with the RU, $\Delta(b_l, d, q)$. This loop also computes $p(s_i, s_j)$, representing the frequency of transitions from state s_i to state s_j in the data. This information about the occurrence of transition is used in the second loop (lines 19 and 20) to calculate the probabilistic transition function between states s_i and s_j.

In addition to our synthetic data generator, we have designed a network topology generator that accounts for the various region types within the topology. The implementation incorporates maximum fiber distance between hops, the number of hops, redundant routing paths, RU coverage range, and regional RU density. The output of this generator includes the position of RUs, link lengths, and path latencies, calculated using Manhattan distance. This comprehensive approach ensures an accurate and realistic representation of vRAN topologies. We use the number of associated users each time as the primary information to represent the demand variation of RUs, which differs from the authors of [34], that used as primary information traffic load. We opt for this strategy because this type of information is more readily available. Our formulation is dynamic since we can consider different user applications that impact the traffic on RUs.

Algorithm 3: Estimation of Markov Chain transition probabilities

Input : $\mathcal{B}, A, D, Q, \Delta$.
Output: Set of states $\mathcal{S}$ and the transition function ρ.

1 $\mathcal{S} \leftarrow \varnothing$, $P \leftarrow \varnothing$, $\rho \leftarrow \varnothing$
2 **for** $b_l \in \mathcal{B}$ **do**
3 **for** $d \in D$ **do**
4 **if** $d \in \{$*Mon., Tue., Wed., Thu., Fry.*$\}$ **then**
5 $DT \leftarrow$ WD
6 **else**
7 $DT \leftarrow$ WE
8 **for** $q \in Q$ **do**
9 $s_i \leftarrow (A(b_l), DT, q, \Delta(b_l, d, q))$
10 $\mathcal{S} \leftarrow s_i$
11 **if** $q + 1 \in Q$ **then**
12 $s_j \leftarrow (A(b_l), DT, q + 1, \Delta(b_l, d, q + 1))$
13 $\mathcal{S} \leftarrow s_j$
14 **if** $p(s_i, s_j) \notin P$ **then**
15 $p(s_i, s_j) \leftarrow 1$
16 $P \leftarrow p(s_i, s_j)$
17 **else**
18 $p(s_i, s_j) \leftarrow p(s_i, s_j) + 1$
19 **for** $s_i, s_j \in \mathcal{S}$ **do**
20 $\rho(s_i, s_j) \leftarrow \frac{p(s_i, s_j)}{\sum_{s_k \in \mathcal{S}} p(s_i, s_k)}$

We compare one random instance generated by synthetic data and the dataset's real-world data to illustrate the behavior of our synthetic load generator. Figure 3 presents this comparison in which we considered 170 hours and included data from 50 RUs. Regarding the distribution of RUs by region, we selected 15% of center RUs and 85% of suburban RUs, mirroring the distribution in the real-world dataset. Despite the lower number of center RUs, the synthetic data preserves the behavior of the real-world data, as illustrated in the figure. Due to the higher proportion of suburban RUs, the overall peaks in the synthetic data are lower than those presented by the center RUs, properly reflecting the real-world demand variation patterns.

To ensure a comprehensive analysis, we generated 30 distinct instances of synthetic data, accounting for its non-deterministic nature. Table IV presents a statistical comparison between the real-world dataset and the synthetic data produced by our model. The results indicate a strong

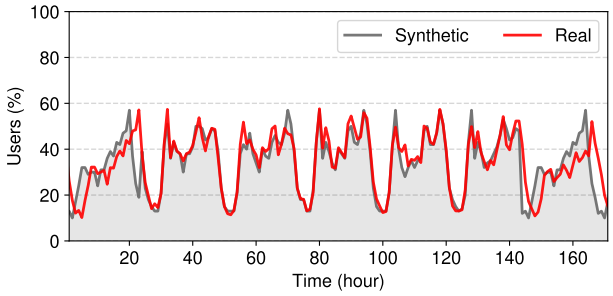


FIGURE 3. Comparison of demand variation patterns of synthetic and real-world data.

alignment between the two, demonstrating the accuracy of our generator in replicating key characteristics of the real-world dataset.

The close mean values suggest that the synthetic data effectively captures the overall trend of the original data. Additionally, the similarities in medians and quartiles confirm that the distribution of the synthetic data closely mirrors that of the real-world dataset, preserving its overall shape. The standard deviation and variance values further indicate that the synthetic data maintains a comparable level of variability. Lastly, the consistent minimum and maximum values show that the range of the synthetic data aligns well with the real-world dataset, reinforcing the reliability of our generation process.

TABLE IV. Synthetic data and real-world data statistics

	Real	Synthetic
Mean	28.4	21.9
Standard Deviation	27.5	23.5
Variance	758.7	553.8
First quartile (Q1)	11	8
Median (Q2)	20	16
Third quartile (Q3)	37	28
Minimum	0	0
Maximum	413	408

V. Evaluation

In this section, we evaluate the proposed formulation for the problem of vRAN VNF placement for energy efficiency. We introduce the method and parameters used in the evaluation, followed by a discussion of the results.

A. Method and Parameters Setup

Topologies – We consider the two classes of RAN topologies commonly deployed by network operators [7], [38]–[40]: T1, in which the nodes are connected in ring structures, and T2, in which the transport network follows a hierarchical tree structure. Both topology classes are represented by instances comprising 50 nodes. In T2, each node's network capacity and computing resources are defined according to its distance (in hops) from the core network. Since links closer to the core may be subject to higher load, they are defined with larger capacity. In T1, we consider two configurations: a High Capacity (HC) scenario, where every link utilizes 100G pluggable transceivers, and a Low Capacity (LC) scenario, with links composed of 10G pluggable transceivers. In a previous work [21], we assessed the impact of heterogeneous hardware topologies on energy consumption and observed that servers with higher energy efficiency are prioritized for activation. To preserve this behavior within our problem instances, we randomly assign the idle power consumption P_w^{idle} from 20% to 25% of the busy power consumption P_w^{busy} for each server $h_w \in \mathcal{H}$.

Paths – The number of paths from each RU to the core network can grow exponentially with the number of nodes in the topology, hindering the ability to solve instances of larger topologies. In this work, we employ an algorithm for k shortest paths to generate the routes for the set $\mathcal{P}_l$. The length of the path is measured by the number of hops. Scenarios, where the parameter k restricts the number of paths generated, may lose the optimal solution. While other algorithms might yield better results from non-optimal solutions, this is beyond the scope of our current work and is left open for future investigation.

Latency – The latency experienced in a given route from the BS to the core network is composed of four components: (i) optical propagation in the fiber, (ii) processing in packet switches, (iii) transmission delay, and (iv) queue delay. Once a solution is expected to remain active during a period of time T, latency constraints must be granted most of the time. To estimate an upper bound for total latency, as in [41], we consider 5 μs/Km of propagation delay, 5 μs of switch processing, a packet of 12368 bits, and an average queue size of two packets. The worst-case latency with these parameters is 26 μs, which does not preclude any evaluated functional splits.

Load Variation – For BS load, we use synthetic data generated by our load generator, as described in Section IV. Since publicly available real-world data are limited to small fragments, we generated synthetic data to model realistic demand variations over time. This allows us to perform a comprehensive evaluation of our model over a 72-hour period, starting on Sunday and ending on Tuesday. The data comprises the hourly state of the network. Therefore, we always solve for a time $T = 1$ hour. However, due to the dynamic nature of the network, T can be adjusted dynamically to suit other scenarios. For example, it is possible to actively monitor the state of a deployed network and dynamically invoke the model whenever a new solution is required.

UE Profiles – We consider four device usage profiles to analyze how different device network requirements

impact the energy consumption of the vRAN. Profile P1 is based on devices requiring Ultra Reliable Low Latency Communication (URLLC) service and 1 Gigabits Per Second (Gbps) of data throughput. Profiles P12, P24, and P53 are all based on Enhanced Mobile Broadband (eMBB) service, with throughput requirements of 12 Gbps, 24 Gbps, and 53 Gbps, respectively.

Functional Splits – Due to data unavailability, we evaluate only 3GPP functional splits 6 and 7.2 with a delay requirement of 250 μs. However, the model is formulated to support additional functional split options, as more data becomes available in the future. When transmitting an eMBB packet (1500 Bytes), the transport network's required bandwidth increases by 1.001 for split 6 and 7.175 for split 7.2 [26]. The bandwidth factors for URLLC packets (128 Bytes) are 1.070 for split 6 and 7.634 for split 7.2.

We consider a transport network based on pluggable DWDM transceivers and Ethernet packet switches, as in [31]. Radio and CR parameters are mainly extracted from [20]. We estimate the VM memory footprint for each VNF based on the memory values provided in [14] and the CPU core usage for each VNF presented in [7]. The parameters utilized in the model evaluation are summarized in Table V.

TABLE V. Evaluation parameters

Parameter	Value
$\|\mathcal{B}\|$	50
ρ^c	0
C_w^{GOPS}	180
P_w^{busy}	94.8 W
P_w^{idle}	20-25% of P_w^{busy}
E_{f_s}	{1795, 242.08, 172.92, 410, 410} MB
e_{ij}^{Cap}	{100, 200, 400, 800, 1000} Gbps
e_{ij}^{L}	(10^{-1}, 10^{-4}) ms
$R_{e_{ij}}^{tr}$	{1, 10, 100} Gbps
$P_{e_{ij}}^{tr}$	{1.0, 2.0, 4.5} W
$P_{v_i}^{s}$	{1.0, 4.2, 14.0} W
T	3600 s
T_s	71.4 μs
N_{used}	1200
N_{bits}	12
N_l	4
SE_0	1.0 bits/s/Hz
τ_c, τ_p	192, 8

We implement the proposed model using Python and the docplex library to perform the evaluation. Next, CPLEX is used to solve each instance of the problem. An instance is represented by a vRAN topology, its load in a given time, and the network usage profile of the devices. We investigate the impact of those three elements in the solution. The experiments were executed in an Intel i7-12700. Considering 50 BSs, instances of topology T2 and topology T1 with low capacity links took a mean time of 300 ms, while instances of topology T1 with high capacity were solved in a mean time of 700 ms.

B. Results

We analyze how the solution reacts to different configurations since the components of the vRAN topology may be structured in various ways. In this context, we consider only topology T2 to analyze the impact of network usage profiles in the vRAN orchestration problem.

Scalability – To evaluate the scalability of the solution, we consider scenarios where all CRs have a co-located RU, and the UEs present a profile P12. For each topology size, i.e., number of CRs, we evaluate 5 instances with different BS loads. The solver is configured with a time limit of 30 minutes and a relative Mixed Integer Programming (MIP) gap tolerance of 10^{-5} (slightly lower than the default value of 10^{-4}), meaning that the solver may stop early and return a solution that can be up to 0.001% worse than a possible optimal solution. Instances not solved within the time limit are not included in the statistics. Figure 4 shows that the solution time grows exponentially with the number of nodes in the topology. The data rate capacity of the transport network also impacts the time it takes to solve an instance. The data indicates that the complexity of an instance increases with the number of nodes in a topology and with the increase of the capacity in the transport network. In addition, we observe that different BS loads may cause greater variation in the solution time, especially as the complexity of the instance increases, resulting in more instances that are not solved within the time limit, as shown by the bar charts which illustrate the stop criteria of the MIP Solver for every evaluated instance. Nonetheless, we obtain a solution for instances with 450 RUs within a time range of 2 to 16 minutes.

To further evaluate the impact of link capacity on the absolute solution time, we present the results in Figure 5 for a topology with 100 CRs and a varying transport network. First, we evaluate a transport network composed entirely of 1 Gbps, 10 Gbps, and 100 Gbps transceivers. Next, in the scenario labeled "Hier.", we assess a transport network where the data rate capacity of the transceivers decreases with the distance from the core. The "Inv. Hier." scenario follows the same logic; however, the transceiver data rate capacity increases with the distance from the core. Finally, in the "Random" scenario, transceivers are randomly assigned to each link. Unlike in the previous evaluation, we now maintain the same overall capacity of the transport network by reducing the number of links between CRs as we increase the data rate capacity of the transceivers. The results indicate

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3568689

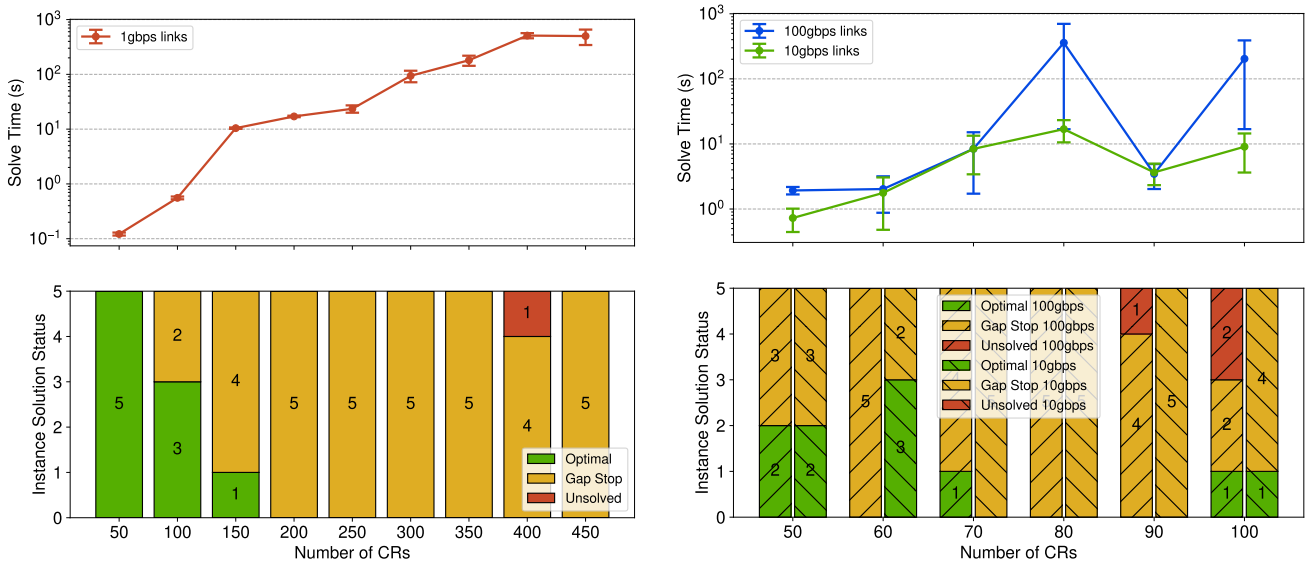


FIGURE 4. **Solution time scalability for topologies with different transport network link capacities and increasing topology size.**

that a transport network with lower capacity transceivers in the links closer to the core tends to present lower solution times.

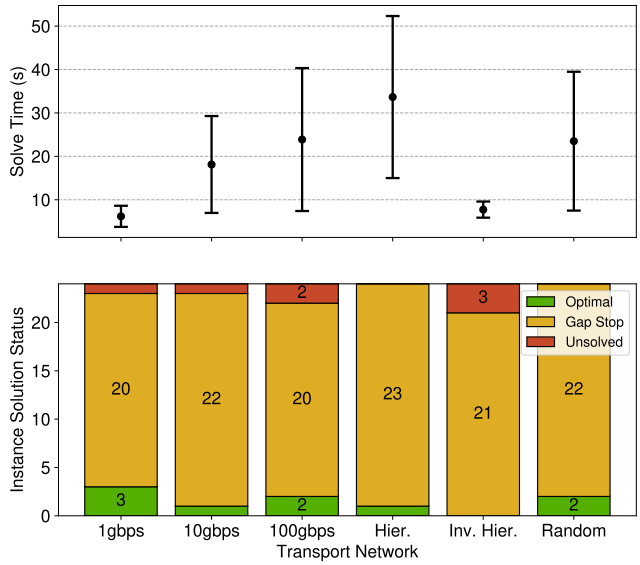


FIGURE 5. **Solution time for topologies with 100 CRs and different transport network link capacities.**

Comparison of different solutions – As shown in Figure 6, for a scenario with 50 CRs, 48 RUs, and UEs with profile P53, our optimization model presents a solution that consumes 22% less energy than C-RAN. This is because C-RAN centralizes all VNFs in a single CR continuously, which limits optimal server allocation, and results in higher utilization of the transport network. If compared with D-RAN, the energy savings increase to 52%. In this case, D-RAN deploys the VNFs from each RU in a co-located server, resulting in under-utilization of active equipment. Since the heuristic solution can assign different split options according to the energy cost of each evaluated situation, it achieves a better result than C-RAN, with the optimal solution presenting 14% energy saving overall if compared to the heuristic. Given the assumptions made in the heuristic, it cannot reach the efficacy of the optimal solution. However, as shown in Section III-D, the heuristic has polynomial complexity and, therefore, has better scalability than the optimal solution.

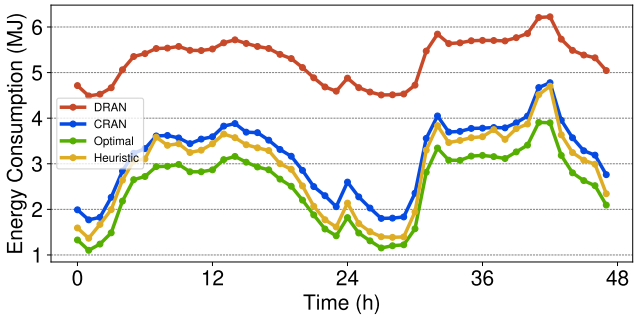


FIGURE 6. **Total energy consumption achieved by different solutions.**

Impact of Topology Structures – Figure 7 shows the empirical distribution of the total energy consumption achieved in different vRAN topologies when our model is employed. Since transceivers with higher throughput capacity are more energetically efficient, topology T1 HC can achieve lower energy consumption and higher centralization rate during low BS load than T1 LC. However, none of them can surpass the energy efficiency of topology T2. Figure 8(a) and 8(b) show the energy consumption by component for each topology under different conditions, i.e., low and high loads.

Figure 9 shows the centralization rate achieved by each topology over time, presented as a percentage of the maximum VNF centralization possible. This maximum central-

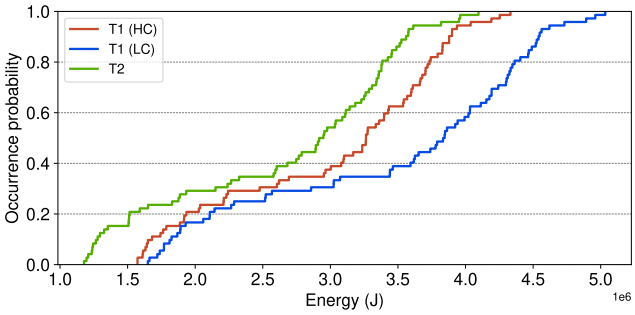


FIGURE 7. **Total energy consumption for different vRAN topologies.**

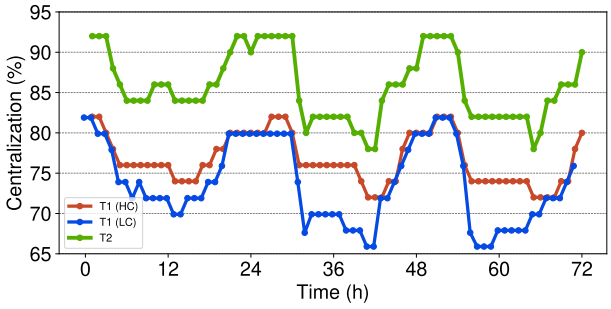


FIGURE 9. **Centralization ratio achieved for profile P53.**

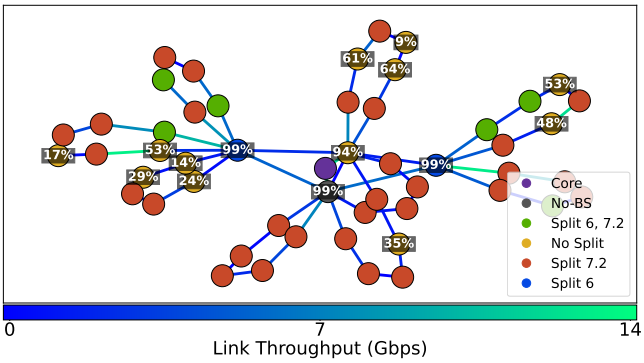
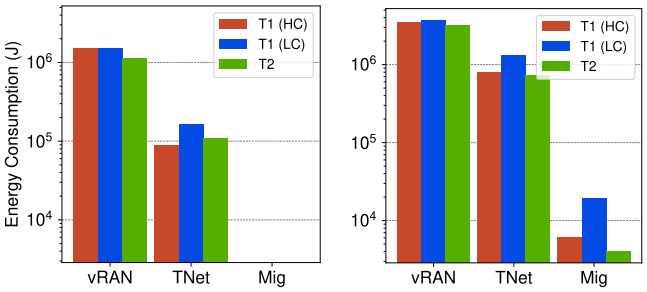


FIGURE 10. **Solution for T1 HC topology during peak load.**

ization occurs when the VNFs of all RUs, implementing functional split 7.2, are deployed in a single CR. We can observe that despite T2 presenting a higher energy usage of the transport network compared to T1 HC, the gains in vRAN processing allow a lower overall energy consumption. This behavior occurs due to the hierarchical organization of the infrastructure components in T2, allowing more effective usage of network equipment, which enables (i) higher centralization of VNFs with lower transport network impact and (ii) most efficient usage of the already turned-on servers in the centralization of CRs during high load. Although the energy consumption of VNF migration is formulated in the objective function, it also indirectly acts as a constraint. As BS load fluctuates over time, it prevents changes in previous associations when the energy overhead in the transport network to deploy a new solution outweighs its energy savings. Beyond that, we could not identify any relation between the number of VNF migrations and the topology structure.



(a) Low load (b) High load

FIGURE 8. **Energy consumption per component: vRAN, TNet, and Mig.**

We illustrate solutions during peak BS load for both topologies in Figure 10 and Figure 11. Topology T1 requires more active CRs, as illustrated by the higher number of percentage tags in the nodes, which inform the turned-on servers CPU load. In topology T2, a slightly higher centralization rate is possible. However, it is not achieved because the model identifies that turning on an additional server in an active CR to aggregate the processing of a few BSs would not be worth the extra energy cost in the transport network.

Impact of Network Usage Profiles – Figure 12 shows an empirical distribution of the vRAN energy consumption achieved by the optimization model for different usage profiles. The throughput generated by the devices connected to the mobile network increases the total energy consumption of vRAN. This increase in consumption is more pronounced during high BS load, as illustrated in Figure 13(a) and Figure 13(b). These figures present the vRAN energy consumption broken down into components in low and high BS load moments. Those figures confirm that VNF migration is the component with the lowest impact, representing less than 2% of the total energy consumption of the topology. This result shows that using a linear approximation model for

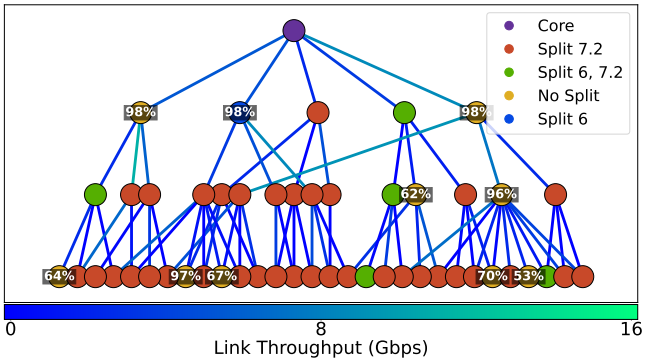


FIGURE 11. **Solution for T2 topology during peak load.**

VNF migration cost does not compromise the robustness of our formulation, as the minimal gain in precision from more complex models does not justify the added complexity. Additionally, the impact of usage profiles on the energy consumed by the CRs running vRAN and VNF migration is minimal. However, the transport network energy consumption is more sensitive to the throughput generated by user devices, further aggravated by the increase in BS load.

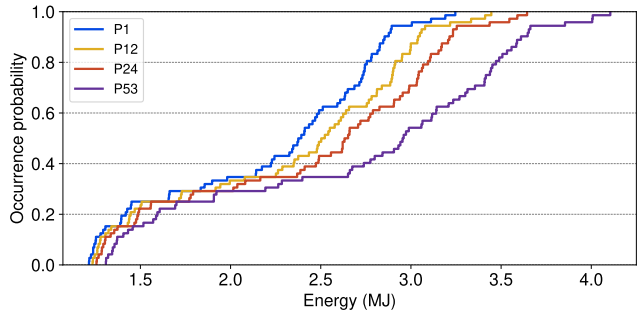
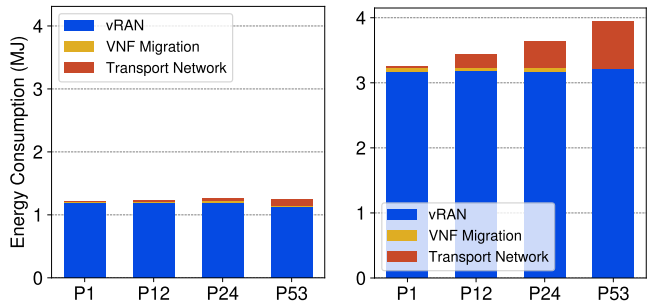


FIGURE 12. Total energy consumption for different network usage profiles.



(a) Low BS load (b) High BS load

FIGURE 13. Energy consumption per component for different net. usage profiles.

The higher demand for the transport network also results in solutions with lower centralization rates, mainly during high BS load, as shown in Figure 14. Once the need for more CRs arises, given the increase in BS load with time, using local servers while performing no functional split becomes preferable to turning on more servers in centralization nodes. This behavior occurs because the energy consumption incurred by the transport network outweighs the energy gains achieved by centralizing the VNFs of multiple BSs into a single node. On the other side, moments with lower BS load can achieve better energy consumption by increasing the centralization rate. In the formulation, we considered that the maximum latency required by each functional split must be granted as a hard constraint. According to the values in [4], all functional splits remain viable in the evaluated topology. However, the latency requirements of the functional splits may vary, as shown in [26]. If the maximum allowed latency is low enough to prevent the choice of some functional split options, it may limit the centralization rate, and we can expect to observe a negative impact on energy consumption.

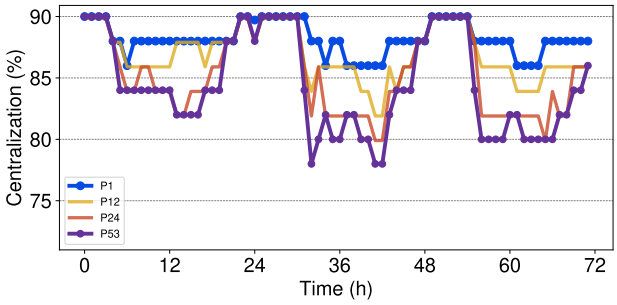


FIGURE 14. Centralization achieved for network usage profiles.

VI. Conclusion

In this article, we addressed the problem of vRAN VNF positioning to minimize the energy consumption of the infrastructure by formulating a flexible and comprehensive MILP model. This linear formulation enables us to solve instances of the problem optimally and quickly (in ms) providing a robust baseline for evaluating non-optimal solutions. We observed that in scenarios with high-throughput demand devices, the transport network imposes a large energy cost on the solution, reducing the achievable centralization rate. For future work, we plan to explore additional functional splits to further assess their impact on energy efficiency. Additionally, given the potential energy savings by entirely disabling an RU, we aim to investigate the problem of association between UEs and BSs, minimizing infrastructure energy consumption while ensuring the quality of service to different service profiles.

Acknowledgment

This work was supported by CAPES, MCTIC/CGI.br/São Paulo Research Foundation (FAPESP) through the Smart 5G Core And MUltiRAn Integration (SAMURAI) project under Grant 2020/05127-2 and the Slicing Future Internet Infrastructures (SFI2) project under Grant 2018/23097-3, by RNP/MCTIC through the Brasil 6G project under Grant 01245.020548/2021-07 and the OpenRAN@Brasil Program.

REFERENCES

[1] J. X. Salvat *et al.*, "Open Radio Access Networks (O-RAN) Experimentation Platform: Design and Datasets," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 138–144, 2023.

[2] 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, October 2018.

[3] ——, "FluidRAN: Optimized vRAN/MEC Orchestration," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, October 2018, pp. 2366–2374.

[4] 3GPP, "3GPP TR 38.801: Study on new radio access technology: Radio access architecture and interfaces," 3GPP, Technical Report, 2017.

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

[6] 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, April 2021.

[7] F. Z. Morais *et al.*, "PlaceRAN: Optimal Placement of Virtualized Network Functions in Beyond 5G Radio Access Networks," *IEEE Transactions on Mobile Computing*, vol. 22, no. 9, pp. 5434–5448, September 2023.

[8] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[9] S. Puthenpura *et al.*, "SMaRT-5G Project," ONF, White Paper, June 2023.

[10] Global System for Mobile Communications Association (GSMA), "Energy Efficiency: An Overview," 2019, https://www.gsma.com/solutions-and-impact/technologies/networks/gsma_resources/energy-efficiency-an-overview/.

[11] J. A. Ayala-Romero *et al.*, "Mean-Field Multi-Agent Contextual Bandit for Energy-Efficient Resource Allocation in vRANs," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, 2024, pp. 1–10.

[12] O.-R. W. G. 1, "O-RAN Architecture Description 13.0," O-RAN Alliance, Technical Report, 2025.

[13] H. Gupta *et al.*, "Apt-RAN: A Flexible Split-Based 5G RAN to Minimize Energy Consumption and Handovers," *IEEE Transactions on Network and Service Management*, vol. 17, no. 1, pp. 473–487, March 2020.

[14] R. Singh *et al.*, "Energy-Efficient Orchestration of Metro-Scale 5G Radio Access Networks," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, July 2021, pp. 1–10.

[15] S. Mollahasani *et al.*, "Energy-Aware Dynamic DU Selection and NF Relocation in O-RAN Using Actor-Critic Learning," *Sensors*, vol. 22, no. 13, July 2022.

[16] N. Sen and A. Franklin A., "Towards Energy Efficient Functional Split and Baseband Function Placement for 5G RAN," in *Proceedings of International Conference on Network Softwarization (NetSoft)*, July 2023, pp. 237–241.

[17] E. Amiri *et al.*, "Energy-Aware Dynamic VNF Splitting in O-RAN Using Deep Reinforcement Learning," *IEEE Wireless Communications Letters*, vol. 12, no. 11, pp. 1891–1895, November 2023.

[18] L. M. Moreira Zorello *et al.*, "Power-Efficient Baseband-Function Placement in Latency-Constrained 5G Metro Access," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 3, pp. 1683–1696, September 2022.

[19] F. Malandrino *et al.*, "An Optimization-Enhanced MANO for Energy-Efficient 5G Networks," *IEEE/ACM Transactions on Networking*, vol. 27, no. 4, pp. 1756–1769, August 2019.

[20] O. T. Demir *et al.*, "Cell-Free Massive MIMO in O-RAN: Energy-Aware Joint Orchestration of Cloud, Fronthaul, and Radio Resources," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 42, no. 2, p. 356–372, January 2024.

[21] W. Pires *et al.*, "Bi-objective Optimization for Energy Efficiency and Centralization Level in Virtualized RAN," in *Proceedings of IEEE International Conference on Communications (ICC)*, August 2022, pp. 1034–1039.

[22] M. A. Habibi *et al.*, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," *IEEE Access*, vol. 7, pp. 70 371–70 421, May 2019.

[23] G. M. Almeida, G. Z. Bruno, A. Huff, M. Hiltunen, E. P. Duarte, C. B. Both, and K. V. Cardoso, "Ric-o: Efficient placement of a disaggregated and distributed ran intelligent controller with dynamic clustering of radio nodes," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 2, pp. 446–459, 2024.

[24] V. H. L. Lopes, G. M. Almeida, A. Klautau, and K. Cardoso, "A coverage-aware vnf placement and resource allocation approach for disaggregated vrans," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 185–190.

[25] V. H. L. Lopes, G. M. Almeida, A. Klautau, and K. V. Cardoso, "O-ran-oriented approach for dynamic vnf placement focused on interference mitigation," in *ICC 2024 - IEEE International Conference on Communications*, 2024, pp. 5479–5484.

[26] F. G. C. Rocha *et al.*, "Optimal Resource Allocation with Delay Guarantees for Network Slicing in Disaggregated RAN," June 2023. [Online]. Available: https://arxiv.org/abs/2305.17321

[27] B. Debaillie, C. Desset, and F. Louagie, "A Flexible and Future-Proof Power Model for Cellular Base Stations," in *Proceedings of IEEE Vehicular Technology Conference (VTC Spring)*, 2015, pp. 1–7.

[28] K. Kazunari, "Approach to Commercial Use of OAI," 2017, https://openairinterface.org/4th-openairinterface-workshop-fall-2017.

[29] X. Fan, W.-D. Weber, and L. A. Barroso, "Power Provisioning for a Warehouse-Sized Computer," *SIGARCH Computer Architecture News*, vol. 35, no. 2, p. 13–23, June 2007.

[30] M. R. Raza *et al.*, "Power and cost modeling for 5G transport networks," in *Proceedings of the 17th International Conference on Transparent Optical Networks (ICTON)*, August 2015, pp. 1–7.

[31] M. Fiorani *et al.*, "Modeling energy performance of C-RAN with optical transport in 5G network scenarios," *Journal of Optical Communications and Networking*, vol. 8, no. 11, pp. B21–B34, November 2016.

[32] H. Liu *et al.*, "Performance and Energy Modeling for Live Migration of Virtual Machines," in *Proceedings of the 20th International Symposium on High Performance Distributed Computing*, June 2011, p. 171–182.

[33] W. Xiang, "Analysis of the time complexity of quick sort algorithm," in *2011 International Conference on Information Management, Innovation Management and Industrial Engineering*, vol. 1, 2011, pp. 408–410.

[34] F. Malandrino *et al.*, "Cellular Network Traces Towards 5G: Usage, Analysis and Generation," *IEEE Transactions on Mobile Computing*, vol. 17, no. 3, pp. 529–542, March 2018.

[35] P. Di Francesco *et al.*, "A Sharing- and Competition-Aware Framework for Cellular Network Evolution Planning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 2, pp. 230–243, June 2015.

[36] H. Ghazzai *et al.*, "Optimized LTE Cell Planning With Varying Spatial and Temporal User Densities," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1575–1589, March 2016.

[37] X. Chen, Y. Jin, S. Qiang, W. Hu, and K. Jiang, "Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 3585–3591.

[38] Z. Zhang, A. Marder, R. Mok, B. Huffaker, M. Luckie, K. C. Claffy, and A. Schulman, "Inferring regional access network topologies: methods and applications," in *Proceedings of the 21st ACM Internet Measurement Conference*, ser. IMC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 720–738. [Online]. Available: https://doi.org/10.1145/3487552.3487812

[39] NGMN, "NGMN Overview on 5 G RAN Functional Decomposition," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:131770760

[40] T. S. S. ITU-T, *Characteristics of transport networks to support IMT-2020/5G*, ITU-T, 2020. [Online]. Available: https://www.itu.int/rec/dologin_pub.asp?lang=s&id=T-REC-G.8300-202005-I!!PDF-E&type=items

[41] N. Tesfalidet and S. Khosravi, "Development of a C-RAN Fronthaul Simulator," Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2023.

William T. Pires-Jr is a M.Sc. student in Computer Science at Universidade Federal de Goiás (UFG). He received his Bachelor degree in Computer Science from UFG in 2024. His research spans wireless networks, virtualization, resource allocation, and performance evaluation.

Gabriel M. Almeida is Ph.D. candidate in Computer Science at Universidade Federal de Goiás (UFG). He received his Bachelor degree in Computer Science (2022) from UFG and his MSc degree in Computer Science also from UFG in 2023. He has been a member of the Laboratory Computer Networks and Distributed Systems (LABORA) since 2018 and his research spans wireless networks, virtualization, resource allocation, and performance evaluation.

Sand L. Corrêa received a bachelor's degree in Computer Science from the Universidade Federal de Goiás (UFG), in 1994. In 1997, she received an M.Sc. degree in Computer Science from the State University of Campinas (Unicamp). She received a D.Sc. degree in Informatics from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), in 2011. Since 2010, she is an associate professor at the Institute of Informatics at UFG.

Cristiano B. Both is a professor of the Applied Computing Graduate Program at the University of Vale do Rio dos Sinos (UNISINOS), Brazil. He coordinates research projects funded by H2020 EU-Brazil, CNPq, FAPERGS, and RNP. His research focuses on wireless networks, next-generation networks, softwarization, and virtualization technologies for telecommunication networks.

Leizer L. Pinto is an associate professor at the Institute of Informatics, Universidade Federal de Goiás (UFG), where he's served since 2010. He earned his Computer Science degree from PUC-GOIÁS (2004) and both his MSc (2007) and Ph.D. (2009) in Systems Engineering and Computer Science from COPPE, Universidade Federal do Rio de Janeiro. His research spans linear programming, multiobjective and combinatorial optimization.

Kleber V. Cardoso is an associate professor at the Institute of Informatics, Universidade Federal de Goiás (UFG), where he's served since 2009. He earned his Computer Science degree from UFG (1997) and both his MSc (2002) and Ph.D. (2009) in Electrical Engineering from COPPE, Universidade Federal do Rio de Janeiro. He spent his sabbaticals at Virginia Tech (2015) and Inria Saclay Research Center, France (2020). His research spans wireless networks, virtualization, resource allocation, and performance evaluation.