Energy-Efficient Orchestration of Metro-Scale 5G Radio Access Networks

Citation for published version:
Singh, R, Hasan, C, Foukas, X, Fiore, M, Marina, MK & Wang, Y 2021, Energy-Efficient Orchestration of Metro-Scale 5G Radio Access Networks. in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications.* IEEE, 2021 IEEE International Conference on Computer Communications, Virtual Conference, 10/05/21. https://doi.org/10.1109/INFOCOM42981.2021.9488786

Digital Object Identifier (DOI):
10.1109/INFOCOM42981.2021.9488786

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
IEEE INFOCOM 2021 - IEEE Conference on Computer Communications

Energy-Efficient Orchestration of Metro-Scale 5G Radio Access Networks

Rajkarn Singh*, Cengis Hasan†, Xenofon Foukas‡, Marco Fiore§, Mahesh K. Marina* and Yue Wang¶
*The University of Edinburgh, †University of Luxembourg, ‡Microsoft, §IMDEA Networks Institute, ¶Samsung Electronics UK
r.singh@ed.ac.uk, cengis.hasan@gmail.com, xefouk@microsoft.com, marco.fiore@imdea.org,
mahesh@ed.ac.uk, yue2.wang@samsung.com

***Abstract*—RAN energy consumption is a major OPEX source for mobile telecom operators, and 5G is expected to increase these costs by several folds. Moreover, paradigm-shifting aspects of the 5G RAN architecture like RAN disaggregation, virtualization and cloudification introduce new traffic-dependent resource management decisions that make the problem of energy-efficient 5G RAN orchestration harder. To address such a challenge, we present a first comprehensive virtualized RAN (vRAN) system model aligned with 5G RAN specifications, which embeds realistic and dynamic models for computational load and energy consumption costs. We then formulate the vRAN energy consumption optimization as an integer quadratic programming problem, whose NP-hard nature leads us to develop `GreenRAN`, a novel, computationally efficient and distributed solution that leverages Lagrangian decomposition and simulated annealing. Evaluations with real-world mobile traffic data for a large metropolitan area are another novel aspect of this work, and show that our approach yields energy efficiency gains up to 25% and 42%, over state-of-the-art and baseline traditional RAN approaches, respectively.**

I. Introduction

The telecommunication industry currently consumes 2-3% of global energy and energy consumption constitutes 20-40% of the operating expenditure (OPEX) for mobile network operators [1]. As we head to 5G, the energy consumption is expected to increase further by 2-3 times due to the infrastructure growth needed to cope with the mobile data traffic surge [2], [3]. Over 90% of the operators have expressed concerns about the rise in energy costs [4]. Base stations (BSs) and consequently the radio access network (RAN) have traditionally been the source of major energy consumption in cellular networks [5]. This is expected to be the case also in 5G systems [1]. Developing RAN solutions that achieve high energy efficiency is thus crucial towards 5G sustainability.

In light of the above, this paper focuses on the optimization of energy consumption in the 5G RAN context. The 5G RAN architecture marks a paradigm shift from earlier generations of RAN architectures, and presents both opportunities and challenges from an energy efficiency perspective [6]. Aiming at greater flexibility, cost reduction and easier evolution, 5G RAN embraces the concept of *virtualized RAN (vRAN)* that combines RAN disaggregation, virtualization and cloudification. By running RAN functionalities as Virtual Network Functions (VNFs) over commodity hardware on a (edge) cloud infrastructure, it provides resource pooling and multiplexing gains, and enables coordinated processing [7], [8].

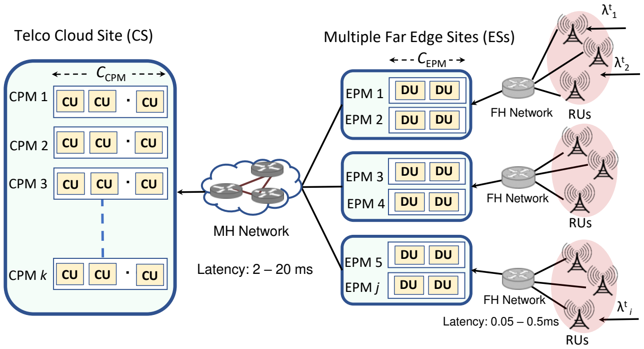

Fig. 1. Schematic of a virtualized RAN (vRAN) reflecting the 5G RAN architecture. A Telco cloud site (CS) consists of a set of Telco Cloud physical machines (CPMs) that implement the central units (CUs) of different base stations (BSs). Similarly, each far-edge site (ES) includes multiple far-edge physical machines (EPMs) that realize distributed units (DUs).

Fig. 1 shows a schematic of the vRAN architecture in line with the 5G RAN specifications [9]. Base station (BS) processing in a vRAN is disaggregated into three units: radio unit (RU) co-located with the antennas and dedicated to PHY functions; distributed unit (DU) closer to the RU; and a central unit (CU). A fronthaul (FH) network connects RUs with DUs, whereas DUs and CUs are connected through a midhaul (MH) network. CUs and DUs can be hosted either at far-edge sites (ES), *e.g.*, on premises close to the radio cell sites, or on Telco cloud sites (CS), *e.g.*, located in Central Offices or operator-owned local exchange sites. This gives rise to a hierarchical and multi-tier vRAN architecture [10], [11], with different latency from the RUs as illustrated in Fig. 1. Distribution of BS processing between CU and DU and their placement on the underlying compute servers, which we refer to as physical machines (PM), is determined by the choice of functional split [12]. For example, the various 3GPP defined functional split options for 4G/LTE case are shown in Fig. 2. Depending on the functional split selected, a range of different RAN configurations can be realized – all the way from a distributed RAN (with no function realized at the CS, for latency sensitive applications) to a fully centralized RAN (all functions at the CS, for latency insensitive applications), with several hybrid alternatives in between with functions split between CS and ES.

The problem of minimizing energy consumption in the vRAN outlined above is markedly different from that in traditional cellular networks. The problem is more challenging in the vRAN setting, due to a number of factors, including the added degrees of freedom in the form of multiple functional split options, and the interdependent CU and DU

‡X. Foukas was with the University of Edinburgh during this work.

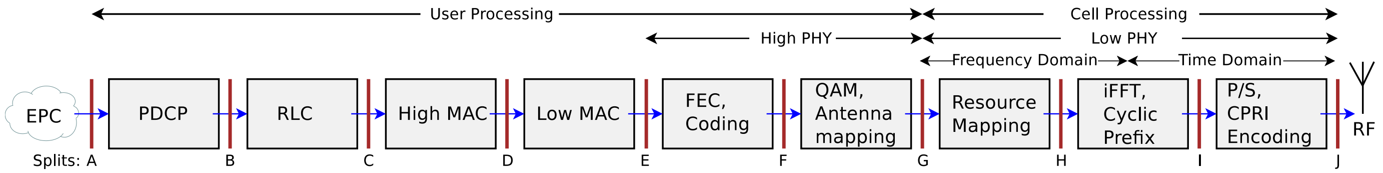

Fig. 2. 3GPP specified RAN functional splits in the downlink direction [13]. Split G onwards to the right typically at RU. Other splits distribute RAN functions between CU and DU. For example, split B implies that PDCP layer realized at CU and layers between B and G realized at DU.

placement choices at far-edge and Telco cloud sites. Here, complicating factors abound: (i) there are correlations between the traffic load of each BS and its overall RAN processing requirements (and energy consumption) [14], hence RAN orchestration decisions must be updated over time as the BS traffic loads vary; (ii) the processing load across all BSs in the RAN influences the optimal choice of functional splits and CU/DU placements for each BS, since the overall energy consumption of a vRAN depends on how processing demands of different BSs are mapped to PMs at CS/ES; and, (iii) RAN reorchestration entailing VNFs migration between PMs in the CS/ES also affect the energy consumption. *To the best of our knowledge, the energy-efficient vRAN orchestration problem as outlined above has not been tackled in its entirety, and not at all in large-scale real-world deployment scenarios.*

Motivated by the above, we aim at an energy-efficient 5G RAN orchestration targeting metro-scale scenarios. We consider the generic vRAN architecture as depicted in Fig. 1 that covers the full range of 5G RAN configurations. Our system model also embeds detailed (measurement-based) models for all aspects pertinent to the problem, including RAN function processing loads, memory footprints and various energy consumption costs. In particular, we account for both main energy consumption costs outlined above: (1) the processing-induced and traffic-dependent energy consumption cost; (2) the energy consumption overhead of seamless VNF migration [15], [16]. We then formulate the vRAN energy consumption optimization problem that minimizes the sum of these two costs across CS and ES sites as an integer quadratic programming (IQP) problem that finds: (i) an optimal CU-DU functional split for each BS; (ii) an optimal DU-EPM association; and (iii) an optimal CU-CPM association. In view of the NP-hard nature of this optimization problem, we propose a novel multi-phase decomposition based distributed solution named `GreenRAN`. Specifically, `GreenRAN` leverages Langrangian Relaxation [17] to decompose the overall energy consumption optimization problem into multiple sub-problems at the CS and ES sites, which are then solved via Simulated Annealing [18]. We extensively evaluate the energy efficiency benefits of `GreenRAN` over baseline and state of the art approaches such as those in [14], [19], using data traffic measurements collected in a production, metropolitan-scale mobile network.

In summary, we make three key contributions:

- The comprehensive nature of our vRAN system model (§III) sets our work apart from the literature (§II). Compared to works like [14], we consider a three-level disaggregated and cloudified RAN that is aligned with 5G specifications and industry trends [20]. In contrast with [19] and other prior works, we employ realistic and detailed computational load and energy consumption models that account for crucial aspects such as traffic dependence and VNF migration overhead.
- We formulate the problem of vRAN energy consumption optimization in the aforementioned system model as an IQP (§III), and propose `GreenRAN` (§IV), a novel distributed solution based on Lagrangian decomposition and simulated annealing that is efficient and scalable.
- In contrast to any other related prior work, we use real-world mobile network traffic dataset for a large metropolitan city to conduct evaluations of `GreenRAN` comparing it with alternative approaches (§V and §VI). Our results show that `GreenRAN` yields energy efficiency gains by up to 25% and 42% compared to the state-of-the-art and traditional distributed RAN configuration, respectively. They also provide insights on how edge cloud configurations and MH bandwidth influences the optimal functional split as well as the processing load distribution between CU and DU.

II. Related Work

Energy efficiency in a traditional distributed cellular RAN architecture has been extensively studied [5]. Most of these works focus on techniques for BS sleep modes (*e.g.*, [21]) that in some cases rely on traffic prediction to forecast intervals of low traffic while in others consider extending the coverage of neighboring BSs through a technique called cell zooming [22].

The follow on work has considered a centralized RAN (C-RAN) scenario, where only the RF functionality of BSs stays distributed at remote radio heads (RRHs) while the rest of the BS functionality performing baseband processing, i.e., the baseband unit (BBU), is realized centrally, thereby consolidating processing of multiple BSs. Consequently, unlike in traditional cellular networks, BSs are physically interdependent as they share part of the infrastructure. When BBUs are virtualized and realized over a cloud infrastructure, the same scenario is referred to as *cloud RAN*. Representative examples of prior work assuming the above outlined C-RAN scenario include: [23], where BBU placement across PMs in the CS is modeled as a bin-packing problem and a simulated annealing based heuristic is proposed to resolve it; [24], [25], which improve this system model further by considering co-location and correlation among RRHs; and [26], which demonstrates how traffic estimations at different BSs aid in dynamic switching on/off RRHs. Another recent work in the C-RAN category is [27], which takes an end-to-end perspective on energy efficiency (including the core network functions) and considering activation of PMs, placing and using VNFs on them, and forwarding traffic between them.

The above described C-RAN scenario, however, reflects one fixed lowest-level functional split of BSs, between BBUs and RRHs with very high FH datarate requirements and consequently very high deployment cost. More recent work has considered the impact of flexible and dynamic functional splits [12]. For example, [8] analytically models the computational resource and power savings with different functional splits, and shows that 25-30% savings relative to a traditional distributed RAN can be achieved by leveraging the functional splits. [19] considers a vRAN model like ours with a three-level disaggregation of RAN processing across CU, DU and RU, in line with the 3GPP specifications [9]. Their focus is on jointly optimizing MH bandwidth consumption and overall system energy efficiency, whereas we have MH bandwidth consumed as a constraint. However, more crucially and unlike our work, [19] adopts an unrealistic traffic-invariant power consumption model as do all other prior C-RAN works [23]–[25], [27]. Also unlike our work, none of these above works consider the energy consumption for seamless migration of VNFs across PMs in CS/ES. This migration overhead [15], [16] has been shown to have non-negligible impact on energy consumption in data center networks [28], [29] but has largely been ignored in the energy-efficient C-RAN/vRAN literature.

Apt-RAN [14] is a recent work that experimentally shows the dependence of CU/DU energy consumption on the BS traffic demand and functional split chosen. However, in contrast to our work, Apt-RAN does not distinguish between DU and RU, and instead assumes they both are realized together via a dedicated PM, thereby overlooking the potential multiplexing gains with edge clouds through co-location of DUs of different BSs. Finally, different from the above mentioned works, we evaluate our proposed solution `GreenRAN` at metro-scale, driven by real-world mobile network traffic data.

III. System Model and Problem Formulation

A. Generic Virtualized RAN Model

We consider a 2-tier vRAN model *i.e.*, with two levels of edge clouds as shown in Fig 1 which consists of a set of BSs ($\mathcal{B}$), a set of far-edge clouds (referred as *edge clouds* henceforth for simplicity) at the far-edge site ($\mathcal{E}$) and a set of *Telco clouds* ($\mathcal{C}$). Let $\mathcal{F}$ represent the set of 5G RAN functions and $\mathcal{S}$ be the set of functional splits. We adopt 3GPP CU-DU functional split model where different RAN VNFs are placed either at ES or at CS [13]. Various RAN functions with the corresponding functional splits are shown in Fig 2.

Let $\mathcal{P}_E$ be the set of compute servers or edge PMs (EPMs) deployed across all far-edge sites and $\mathcal{P}_C$ be the set of compute servers deployed across all Telco cloud sites *i.e.*, Telco cloud PMs (CPMs). Throughout the text, we use indices i, j and k for BS (and its CU/DU), EPM and CPM respectively. Each edge, $e \in \mathcal{E}$, is connected to Telco cloud, $c \in \mathcal{C}$, via a dedicated MH link. Let $C_{\texttt{epm}}$ and $C_{\texttt{cpm}}$ be the processing capacity of EPMs and CPMs respectively (in *Hz*), and $C_{\texttt{mh,e}}$ be the bandwidth (maximum data transfer capacity) of MH links from edge $e \in \mathcal{E}$ to Telco cloud $c \in \mathcal{C}$ (in $Mbps$).

Each edge cloud is responsible for the processing of VNFs for a specific set of mutually exclusive BSs (Fig. 1). Thus, the DU of a BS can associate with the PMs of only one edge cloud and its CU with the PMs of only one Telco cloud. A binary matrix $\mathbf{V}_{\mathcal{B}\times\mathcal{E}}$ represents the allowed associations between BSs $i \in \mathcal{B}$ and edge cloud $e \in \mathcal{E}$. For instance, $v_{i,e} = 1$ means that DU i can be placed at any of the EPM of edge cloud e. Similarly, we use binary matrices $\mathbf{W}_{\mathcal{B}\times\mathcal{P}_E}$ and $\mathbf{X}_{\mathcal{B}\times\mathcal{P}_C}$ to represent the set of BSs that can be associated with EPMs (in all edge clouds) and CPMs (in all Telco clouds), respectively. $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{X}$ are the network configuration variables pre-defined by the mobile network operator to reflect vRAN topological constraints.

B. 5G RAN Function Processing and Functional Splits

Full centralization of RAN functions leads to maximum energy savings [30], however placing all RAN VNFs in the Telco cloud, or in core datacenter as envisioned in a fully centralized cloud RAN, is not always feasible. For example, lower physical layer (`lowPHY`) RAN functions such as FFT, Cyclic Prefix, P/S (from split option G to the right, Fig. 2) have strict latency constraints as well as incur huge datarate demand on the limited transport capacity of FH (MH) network if implemented on DU (CU). These functions are therefore typically placed at RUs, *e.g.*, as per the latest O-RAN 7.2 functional split specifications [31]. The stringent latency constraints also apply to higher physical layer (`highPHY`) and MAC/RLC functions. In LTE, MAC uses synchronous uplink HARQ thus imposing a strict round-trip latency budget of 3ms (including processing) [32]. However, with the adoption of fully asynchronous uplink HARQ in 5G [33], the MAC latency is no longer a stringent constraint and response time is determined by the service class of the traffic being served *i.e.*, URLLC, eMBB, mMTC [34]. Moreover, the latency tolerance for the MAC/RLC layer can further increase through the use of techniques like HARQ prediction considered in the context of non-ideal fronthauls [35], [36]. All of the above means that different functions can be served by different compute sites depending on the deployment scenario. The RAN functions that we include in our model are presented in Table I along with representative relative CPU processing load or demand (d_p) [37], [38] and latency requirements [10], [13], [34], [36].

In our vRAN model, we consider the four practical functional splits listed in Table II, which are responsible for the most significant changes of load across DU and CU [12], [13], and are standardized and used in operational networks [39]–[41]. A functional split $s_{i,p}$ is performed at BS i if all VNFs above and including f_p are executed at CU while VNFs below f_p are executed at DU. Therefore, for a split $s_{i,p} \in \mathcal{S}$, CPU processing load at CU (δ_p), is equal to the cumulative sum of processing load of all VNFs above and including f_p *i.e.*, $\delta_p = \sum_{i \geq p} d_i$. While our model considers processing load for downlink traffic, it can be easily extended to uplink scenarios with appropriate changes to the figures in Table I.

TABLE I

CPU load and latency requirements for various RAN VNFs. CPU load is expressed as the processing gain of moving a VNF from DU to CU.

RAN VNFs ($\mathcal{F}$)	CPU load (d_p)	Latency (σ_p)
f_1: `highPHY`	65%	0.25 — 1 ms
f_2: `MAC,RLC`	15%	0.25 – 30 ms
f_3: `PDCP, RRC`	20%	10 – 50 ms

TABLE II

CU-DU placement of RAN VNFs for different functional splits.

Split p at BS i	Split Type	RAN Functions at ES $\leftrightarrow$ CS	Standard
$s_{i,1}$	G: No split, all at CS	$\leftrightarrow f_1, f_2, f_3$	-
$s_{i,2}$	E: `highPHY - MAC,RLC`	$f_1 \leftrightarrow f_2, f_3$	nFAPI
$s_{i,3}$	B: `MAC,RLC-PDCP,RRC`	$f_1, f_2 \leftrightarrow f_3$	F1
$s_{i,4}$	A: No split, all at ES	$f_1, f_2, f_3 \leftrightarrow$	-

C. Decision Variables

1) DU-EPM association matrix: The placement of DUs at EPMs at each time step t is represented by the binary association matrix $\mathbf{A}^t \in \{0,1\}^{|\mathcal{B}|\times|\mathcal{P}_E|}$ such that each element $a^t_{i,j} \in \mathbf{A}^t$ represents

$$a^t_{i,j} = \begin{cases} 1, & \text{DU of BS } i \text{ is associated with EPM } j, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

2) CU-CPM association matrix: The placement of CUs at CPMs is represented in a similar manner *i.e.*, by the binary association matrix $\mathbf{B}^t \in \{0,1\}^{|\mathcal{B}|\times|\mathcal{P}_C|}$ with each element $b^t_{i,k} \in \mathbf{B}^t$ given by

$$b^t_{i,k} = \begin{cases} 1, & \text{CU of BS } i \text{ is associated with CPM } k, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

3) CU-DU split matrix: Functional split between CU and DU of a BS is represented by the binary matrix $\mathbf{S}^t \in \{0,1\}^{|\mathcal{B}|\times|\mathcal{S}|}$ where each element $s^t_{i,p} \in \mathbf{S}^t$ represents

$$s^t_{i,p} = \begin{cases} 1, & \text{BS } i \text{ performs functional split } p, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

We consider that the above variables are decided over time at intervals or epochs of fixed length (T) or of dynamically determined variable length. Either way, we henceforth refer to these decision points in time as *resource reorchestration intervals*.

D. Computational Load at Edge and Telco Cloud

Let $\hat{\lambda}^t_i$ and $\overline{\lambda}^t_i$ be the peak and average input traffic to the BS i at time epoch t *i.e.*, during the time interval $[t, t+1)$. CPU computation load (*i.e.*, number of CPU cores required per second) incurred at CU and DU of each BS depends upon its input traffic demand. Mathematically,[1]

$$l^t_{\text{du},i} = \lambda^t_i \sum_{p\in\mathcal{S}} s^t_{i,p}(\delta_1 - \delta_p), \quad (4)$$

$$l^t_{\text{cu},i} = \lambda^t_i \sum_{p\in\mathcal{S}} s^t_{i,p}\delta_p. \quad (5)$$

The peak, $\hat{l}^t_{\text{epm},j}$, and average, $\overline{l}^t_{\text{epm},j}$, computation load of an EPM j at time epoch t are the sum of the computation load of their associated DUs as per matrix $\mathbf{A}^t$, hence

$$l^t_{\text{epm},j} = \sum_{i\in\mathcal{B}} a^t_{i,j} l^t_{\text{du},i}, \ \forall j \in \mathcal{P}_E, \ a^t_{i,j} \in \mathbf{A}^t. \quad (6)$$

Similarly, the load at CPM k is given as the sum of the computation load of its associated CUs, as

$$l^t_{\text{cpm},k} = \sum_{i\in\mathcal{B}} b^t_{i,k} l^t_{\text{cu},i}, \ \forall k \in \mathcal{P}_C, \ b^t_{i,k} \in \mathbf{B}^t. \quad (7)$$

E. Energy Consumption Related to Processing

Our energy consumption model builds on the earlier work from the well-known EARTH project [42] and adapts it to the vRAN setting, accounting for various aspects of energy costs such as static and traffic-dependent dynamic energy consumption [43]. More specifically, the energy consumed by an EPM at epoch t over a period of length T is a function of average CPU load ($\bar{l}$) and can be modeled as[2]:

$$E^p_{\text{epm},j}(t) = \Big(\mathbb{I}_{\left(\overline{l}^t_{\text{epm},j}>0\right)} P_{\text{epm}} + P'_{\text{epm}} \cdot \frac{\overline{l}^t_{\text{epm},j}}{C_{\text{epm}}}\Big)T, \quad (8)$$

where P_{epm} is the static power (in *Watts*) of EPM and accounts for fixed costs of running a server (cooling, power amplification, network switches etc.) for the duration it is kept switched-on. P'_{epm} (also in *Watts*) is the dynamic (*i.e.*, load dependent) power consumption that increases linearly with the machine's load. $\mathbb{I}_{\left(\overline{l}^t_{\text{epm},j}>0\right)}$ is the indicator variable which takes value 1 when the condition $\overline{l}^t_{\text{epm},j} > 0$ is $True$ (representing EPM j is busy) or 0 when EPM j is idle.

Similarly, the energy consumption of a CPM k is given as:

$$E^p_{\text{cpm},k}(t) = \Big(\mathbb{I}_{\left(\overline{l}^t_{\text{cpm},k}>0\right)} P_{\text{cpm}} + P'_{\text{cpm}} \cdot \frac{\overline{l}^t_{\text{cpm},k}}{C_{\text{cpm}}}\Big)T, \quad (9)$$

where P_{cpm}, P'_{cpm} are static and dynamic power consumptions.

F. Energy Consumption Related to Function Migration

To ensure reliable and uninterrupted service delivery during reorchestration, live VM migration needs to be employed. However, this incurs additional energy consumption due to an increased amount of memory and processing information being transferred between the two PMs in question. This additional network transfer consumes additional storage, processing as well as networking resources [44]. We refer to the energy consumption cost due to the above as the migration cost [45]. We model this cost based on experimentally derived power consumption modelling in [46]. Specifically, the migration cost of DU and CU at BS i, respectively, is modelled as:

$$E^m_{\text{du},i}(t) = \alpha V^m_{\text{du},i} + \beta \quad (10)$$

$$E^m_{\text{cu},i}(t) = \alpha V^m_{\text{cu},i} + \beta \quad (11)$$

where $V^m_{\text{du},i}$ ($V^m_{\text{cu},i}$) is the volume of data transfer incurred while migrating DU (CU) i. α and β are the coefficients that map network traffic to energy consumption [46]. Migration volume depends upon the split p and can be calculated from the cumulative memory footprint of VNFs (ω_p) (see Table III). Authors in [15] report that actual volume of data transferred between two VMs for a VNF p, ω_p, increases by a factor τ.

[1] The peak and average load are computed in similar manner, hence we use a neutral notation in Eq. (4) and (5) to denote DU and CU peak/average load.

[2] Note that $E^p(t)$ depends on the average load during interval T, as the power changes with load. However, CPU resources for CU/DU are reserved based on the peak load during T in order to avoid service disruption.

This is because some memory pages are transferred multiple times as they become dirty during the migration period. The total migration cost at EPM j can therefore be given as,

$$E^m_{\text{epm},j}(t) = \left(\sum_{i \in \mathcal{B}} (a^t_{i,j} - a^{t-1}_{i,j})^2 a^t_{i,j} E^m_{\text{du},i}(t) \right) \tag{12}$$

where $(a^t_{i,j} - a^{t-1}_{i,j})^2$ tell us that association of DU i changed w.r.t EPM j from the previous interval. Product with $a^t_{i,j}$ implies that DU i moved to EPM j from some other EPM. Since $a^{t-1}_{i,j}$ takes binary values, Eq. (12) can be rewritten as,

$$E^m_{\text{epm},j}(t) = \left(\sum_{i \in \mathcal{B}} (1 - a^{t-1}_{i,j}) a^t_{i,j} E^m_{\text{du},i}(t) \right) \tag{13}$$

Similarly, total migration cost at CPM k is,

$$E^m_{\text{cpm},k}(t) = \left(\sum_{i \in \mathcal{B}} (1 - b^{t-1}_{i,k}) b^t_{i,k} E^m_{\text{cu},i}(t) \right) \tag{14}$$

It is worth noting that we consider above the migration cost with respect to one previous interval, however a better decision may be taken by accounting for the traffic load over multiple epochs. To explore this opportunity, we later extend our model by adapting the length of the resource reorchestration interval to the fluctuations in the network traffic (§IV-E).

G. Overall Optimization Framework

Our aim is to minimize the processing as well as resource orchestration energy consumption at each time epoch t by finding the optimal DU-EPM association matrix ($\mathbf{A}^t$), CU-CPM association matrix ($\mathbf{B}^t$) and the optimal functional split matrix ($\mathbf{S}^t$). Formally,

$$\min_{\mathbf{A}^t, \mathbf{B}^t, \mathbf{S}^t} E_{tot}(t) = \min_{\mathbf{A}^t, \mathbf{B}^t, \mathbf{S}^t} \left(E_{\text{ES}}(t) + E_{\text{CS}}(t) \right) \tag{15}$$

$$= \min_{\mathbf{A}^t, \mathbf{B}^t, \mathbf{S}^t} \sum_{j \in \mathcal{P}_E} \left(E^p_{\text{epm},j} + E^m_{\text{epm},j} \right) + \sum_{k \in \mathcal{P}_C} \left(E^p_{\text{cpm},k} + E^m_{\text{cpm},k} \right)$$

subject to

$$a^t_{i,j} \leq w_{i,j}, \ \forall i \in \mathcal{B}, j \in \mathcal{P}_E \tag{16}$$

$$b^t_{i,k} \leq x_{i,k}, \ \forall i \in \mathcal{B}, k \in \mathcal{P}_C \tag{17}$$

$$\hat{l}^t_{\text{epm},j} \leq C_{\text{epm},j}, \ \forall j \in \mathcal{P}_E \tag{18}$$

$$\hat{l}^t_{\text{cpm},k} \leq C_{\text{cpm},k}, \ \forall k \in \mathcal{P}_C \tag{19}$$

$$\sum_{i \in \mathcal{B}} \sum_{p \in \mathcal{S}} v_{i,e} s^t_{i,p} \mu^t_{i,p} \leq C_{\text{mh},e}, \ \forall e \in \mathcal{E} \tag{20}$$

$$\sum_{j \in \mathcal{P}_E} a^t_{i,j} = \sum_{k \in \mathcal{P}_C} b^t_{i,k} = 1, \ \forall i \in \mathcal{B} \tag{21}$$

$$\sum_{p \in \mathcal{S}} s^t_{i,p} = 1, \ \forall i \in \mathcal{B}. \tag{22}$$

Constraints (16) and (17) restrict the framework to only associate a BS with those EPMs and CPMs where the connection to BS is possible given by matrices $\mathbf{W}$ and $\mathbf{X}$. Constraints (18) and (19) mean that the loads on any physical machine (EPM or CPM) should not exceed its processing capacity, C_{epm} or C_{cpm}, respectively. Constraint (20) means that data transferred over MH link ($\mu^t_{i,p}$) from all BSs associated with edge e should not exceed the maximum bandwidth of MH link, $C_{\text{mh},e}$. Constraint (21) means that CU and DU of each BS, i (represented by row i in matrices $\mathbf{A}^t$ and $\mathbf{B}^t$) can be placed at only one EPM and CPM, respectively, during an orchestration interval. Constraint (22) commends that a BS can only use one functional split in one time interval. Note that the focus of our study is on the energy gains, hence we do not consider CoMP-related benefits of centralization, which could be included as an additional constraint into our formulation. Moreover, the formulation assumes an eMBB use case only; for other service classes, diverse latency-related constraints need to be added. The constrained optimization problem in Eq. (15)–(22) is an IQP, which is NP-hard. To solve it in a computationally efficient and near-optimal way, we propose a distributed algorithm, introduced next.

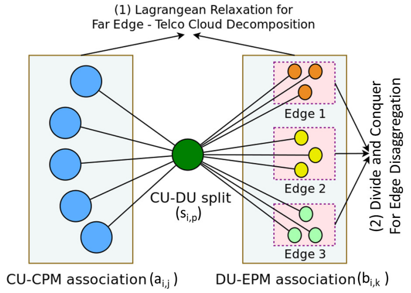
(1) Lagrangean Relaxation for
Far Edge - Telco Cloud Decomposition
Edge 1
Edge 2
Edge 3
(2) Divide and Conquer
For Edge Disaggregation
CU-DU split
(s_{i,p})
CU-CPM association (a_{i,j})
DU-EPM association (b_{i,k})

Fig. 3. Visualization of inter-dependence between association and split variables, and decomposition during different phases of heuristic.

IV. GreenRAN: A Multi-phase Distributed Solution

The CU/DU placement on the cloud is equivalent to a bin-packing problem and is known to be NP hard [23]. With the vRAN model involving distributed far-edge and Telco clouds used based on functional split employed, we face placement decisions at both ES and CS. The centralized problem is equivalent to solving for local optima among $|\mathcal{S}|^{|\mathcal{B}|}$ bin-packing problems making it computationally intractable as the network components increase. To overcome this, we propose a distributed solution called `GreenRAN` that takes a divide-and-conquer approach in that it divides the main problem into multiple sub-problems and then conquers the smaller sub-problems via a metaheuristic algorithm [47].

More formally, `GreenRAN` consists of three phases, each derived from well-known approximation solution: (a) decomposing the problem into two sub-problems, one for Telco cloud in the CS and the other for all far-edge clouds at the ES, using Lagrangean Relaxation, (b) dividing the formulation on the ES into further independent sub-problems, each corresponding to a different far-edge cloud, and (c) solving the decomposed problems using Simulated Annealing [18]. Each decomposition targeting different parts of the vRAN is depicted in Fig 3. The phases of the solution methodology are presented next.

A. Phase 1: Lagrangian Relaxation for vRAN Far Edge – Telco Cloud Decomposition

Dual Decomposition or Lagrangian Relaxation (LR) method is used to change the centralized nature of the problem into a distributed one, by means of the introduction of linear constraints [17]. This enables finding independent solutions to the decomposed sub-problems while agreeing with each other based on the constraints added [48].

As seen in Fig. 3 and Eq. (15), DU-EPM association $\mathbf{A}^t$ cannot be seen as independent of CU-CPM association $\mathbf{B}^t$, since they are interlinked via the functional split variable $\mathbf{S}^t$. Building on this observation, we use LR to make the two association variables independent by introducing new split variables for EPM ($\mathbf{S}_1$) and CPM ($\mathbf{S}_2$), and adding the constraint $\mathbf{S}_1 = \mathbf{S}_2$. This enables our objective to be divided into two independent parts, executing at ES and CS respectively:

$$f_1(\mathbf{A}^t, \mathbf{S}_1^t) = E_{\text{ES}}(t) = \sum_{j\in\mathcal{P}_E} \Big(E^p_{\text{epm},j}(t) + E^m_{\text{epm},j}(t)\Big), \quad (23)$$

$$f_2(\mathbf{B}^t, \mathbf{S}_2^t) = E_{\text{CS}}(t) = \sum_{k\in\mathcal{P}_C} \Big(E^p_{\text{cpm},k}(t) + E^m_{\text{cpm},k}(t)\Big), \quad (24)$$

such that $\mathbf{S}_1^t = \mathbf{S}_2^t$. Enumerating over all possible combinations of $\mathbf{S}$ is computationally not feasible, therefore we relax this constraint by introducing a Lagrange multiplier ($\boldsymbol{\theta}$) and obtain two independent dual problems, as follows.

$$f_1(\mathbf{A}^t, \mathbf{S}_1^t) = E_{\text{ES}} = \min_{\mathbf{A}^t,\mathbf{S}^t} f_1(\mathbf{A},\mathbf{S}) + \sum_{i\in\mathcal{B}}\sum_{p\in\mathcal{S}} \theta_{i,p} s_{i,p}, \quad (25)$$

$$f_2(\mathbf{B}^t, \mathbf{S}_2^t) = E_{\text{CS}} = \min_{\mathbf{B}^t,\mathbf{S}^t} f_2(\mathbf{B},\mathbf{S}) - \sum_{i\in\mathcal{B}}\sum_{p\in\mathcal{S}} \theta_{i,p} s_{i,p}. \quad (26)$$

The value of $\boldsymbol{\theta}$ is updated using gradient ascent (as we are now solving the dual problem) such that as $\boldsymbol{\theta} \to 0$ when $\mathbf{S}_1 = \mathbf{S}_2$.

B. Phase 2: Divide-and-Conquer for Edge Disaggregation

An additional advantage of using LR to separate the ES energy computation from CS is that the DU-EPM association of BSs belonging to one edge cloud become independent of DU-EPM associations of other edges. Motivated by this observation, we further decompose the ES optimization problem (Eq. (25)) into $|\mathcal{E}|$ independent sub-problems (as shown in Fig. 3).

To adopt our optimization framework for distributed edge setting, our formulation of energy equations (Eq. (25) and (26)) remains the same, with only a minor difference *i.e.*, instead of all BSs $\mathcal{B}$, all EPMs $\mathcal{P}_E$ and all CPMs $\mathcal{P}_C$, for each optimization corresponding to one edge $e \in \mathcal{E}$ and one Telco cloud $c \in \mathcal{C}$, only a subset of BSs $\mathcal{B}_e$ is used where $v_{i,e} = 1$. Similarly, a subset of EPMs $\mathcal{P}_e$ and CPMs $\mathcal{P}_c$ is used that belong to the far-edge and Telco cloud considered during that iteration. Also, the LHS of Eq. (25) and (26) now refer to the energy obtained by edge e ($E_e(t)$) and Telco cloud c ($E_c(t)$).

C. Phase 3: Efficient Placement via Simulated Annealing

Simulated Annealing (SA) is a well-known probabilistic heuristic method for approximating global optimum for an optimization problem [18] and has been widely used in problems with discrete and very large problem spaces, like ours. We perform SA for each edge cloud (and Telco cloud) separately, as their optimization variables are now independent following Phases 1 & 2 above.

The SA algorithm starts with a random initialization of decision variables[3], $\Omega^{(t)} = \{a^t_{i,j}, b^t_{i,k}, s^t_{i,p}\}$. The current energy value $E_x(t)$ is computed for any $x \in \{e, c\}$, and the current solution is slightly shifted to new values $\Omega^{(t)}_{new}$ through randomization. The SA algorithm parameters, *i.e.*, temperature (γ), annealing parameter (k) and fitness ($\rho = E_{x,new}(t) - E_x(t)$) determine the probability of adopting a new random solution, $\Omega^{(t)}_{new}$. At higher initial temperatures, the probability of selecting worse solutions is higher (exploration stage which helps the algorithm not to get stuck in local optima), while as γ reduces, the algorithm goes into exploitation stage settling in the neighborhood of current best solution. The pseudocode of our solution framework, GreenRAN, is shown in Alg. 1.

D. Computational Complexity Analysis of GreenRAN

Let K_{SA} be the number of iterations taken by SA to converge. Then the time complexity of finding the solution for BSs $\mathcal{B}_e$ belonging to one edge $e \in \mathcal{E}$ having $\mathcal{P}_E$ EPMs can be given as $\mathcal{O}(K_{\text{SA}}|\mathcal{B}_e||\mathcal{P}_E||\mathcal{S}|)$, where $|\cdot|$ represents the count. Similarly, the computation time at the Telco cloud will be $\mathcal{O}(K_{\text{SA}}|\mathcal{B}_c||\mathcal{P}_C||\mathcal{S}|)$. Let K_{LR} be the number of iterations taken by Lagrangean Relaxation to converge, then the time complexity of running GreenRAN over the entire network will be $\mathcal{O}\Big(K_{\text{LR}}K_{\text{SA}}|\mathcal{S}|\ \big(|\mathcal{B}_e||\mathcal{P}_e||\mathcal{E}|+|\mathcal{B}_c||\mathcal{P}_c||\mathcal{C}|\big)\Big)$. Given the independent and distributed solution nature of GreenRAN, its time complexity is polynomial in the number of network parameters and does not increase exponentially with the network scale, as is the case with combinatorially NP-hard problems like these.

E. Adaptive Resource Reorchestration Intervals

The traffic load varies significantly over time, so does the rate at which traffic changes. Hence, performing reorchestration at fixed interval may be inefficient, causing, *e.g.*, unnecessary migrations when traffic changes slowly or resource underutilization when the demands shift suddenly. To address this issue, we propose an enhancement to GreenRAN that adaptively determines the next reorchestration interval duration. This value is computed by considering the rate of change in the traffic load.

At the start of a generic epoch t_i, we calculate the difference between the traffic at t_i and that at successive intervals $t_{i+k,\ k=\{1,2,\ldots\}}$, until the absolute difference $|\hat{\lambda}_{t_{i+k}} - \hat{\lambda}_{t_i}|$ becomes larger than a system parameter Δ_λ. The start of first epoch $i + k$ for which the Δ_λ threshold is exceeded is selected as the next point in time for orchestration.

V. Evaluation Methodology

We evaluate the quality of our solution with real-world traffic loads observed at the radio access network of a metropolitan-scale mobile network. The measurement data was collected by a major operator in a wide urban region in Europe, and consists of the aggregate downlink demands accommodated by each of 450 4G BSs at every 5 seconds during 24 consecutive hours of a typical weekday.

[3] We initialize variables using a *first-fit* bin-packing strategy, for faster and near-optimal convergence.

Algorithm 1: `GreenRAN` pseudocode

```
procedure GREENRAN ($\mathcal{B}, \mathcal{P}_E, \mathcal{P}_C, \mathcal{F}, \mathcal{S}, C_{EPM}, C_{CPM}, \lambda^t$)
    Initialize Lagrange multiplier, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$
    Decompose system into ES and CS using LR (Eq. (25), (26))
    Divide edge site further into per edge components
    while True, do
        foreach $e \in \mathcal{E}$ do
            $\Omega_e^{(t)}, E_e(t) =$ SimAnneal($\mathcal{B}_e, \mathcal{P}_e, \mathcal{F}_e, e$)
            $\Omega^{(t)} = \Omega^{(t)} \cup \Omega_e^{(t)}$
        end
        foreach $c \in \mathcal{C}$ do
            $\Omega_c^{(t)}, E_c(t) =$ SimAnneal($\mathcal{B}_c, \mathcal{P}_c, \mathcal{F}_c, c$)
            $\Omega^{(t)} = \Omega^{(t)} \cup \Omega_c^{(t)}$
        end
        $\boldsymbol{\theta}_{new} = \boldsymbol{\theta} - \eta(\mathbf{S}_1 - \mathbf{S}_2)$ # Gradient Ascent
        if $(\boldsymbol{\theta}_{new} - \boldsymbol{\theta}) \leq$ threshold, then
            break
        end
        $\boldsymbol{\theta} = \boldsymbol{\theta}_{new}$
    end
    Calculate $E_{\text{ES}}(t) = \sum_{e \in \mathcal{E}} E_e(t)$ and $E_{\text{CS}}(t) = \sum_{c \in \mathcal{C}} E_c(t)$
    Calculate total energy, $E_{tot}(t)$, using Eq. (15)
    return $\Omega^{(t)}, E_{tot}^{(t)}$
end
procedure SimAnneal($\mathcal{B}_x, \mathcal{P}_x, \mathcal{F}_x, x$)
    Initialize: $\Omega^{(t)} = \{x_{i,j}^t, s_{i,p}^t\}$
    Initialize temperature and annealing parameters: $\gamma = \gamma_0, k = 1$
    Calculate EPM (or CPM) energy, $E_x(t)$, using Eq. (25) (or (26))

    while $\gamma > 0$ do
        Update variables: $\Omega_{new}^{(t)}$
        Calculate $E_{x,new}(t)$ with $\Omega_{new}^{(t)}$ variables
        Calculate fitness, $\rho = E_{x,new}(t) - E_x(t)$
        $p = \exp^{\left(\frac{\rho}{\gamma}\right)}$
        if $\rho > 1$ or $random(0,1) \leq p$, then
            $\Omega^{(t)} = \Omega_{new}^{(t)}$
            $E_x(t) = E_{x,new}(t)$
        end
        $k = k + 1, \gamma = \frac{\gamma_0}{\log(k)}$
    end
    return $\Omega^{(t)}, E_x(t)$
end
```

TABLE III
SIMULATION PARAMETERS.

Parameter	Value
Base Stations	450 RUs in total, 25 per Edge Cloud
Cloud Configuration	1 Central Cloud with 30 CPMs
C_{cpm}	64 RCs
$P_{\text{cpm}}, P'_{\text{cpm}}$	200 W
Edge Configuration	18 Edge Clouds with 15 EPMs each
C_{epm}	$\{12, 16, 32, 48, 64\}$ RCs
$P_{\text{epm}}, P'_{\text{epm}}$	$\{40, 60, 120, 180, 240\}$ W
CPU Load $\{d_1, d_2, d_3\}$	$\{3.25, 0.75, 1.00\}$ RCs
Memory $\{\omega_1, \omega_2, \omega_3\}$	$\{1795, 415, 820\}$ MB
Latency $\{\sigma_1, \sigma_2, \sigma_3\}$	$\{0.25, 2, 10\}$ ms
Processing interval, T	$\{1, 3, 5, 8\}$ hours and adaptive intervals
Migration params α, β, ω	0.512, 20.165, 3
Heuristic params $\boldsymbol{\theta}_0, \gamma_0$	1, 100

A. Network Infrastructure Configuration

We consider a single Telco cloud located in the CS and 18 far-edge clouds located in the ES. We map each of 450 4G BSs to a RU[4] site, and consider that each group of 25 geographically close BSs are associated to the same edge cloud. We estimate CPU load, d_p, of each VNF, $f_p \in \mathcal{F}$, as the number of Reference CPU cores (RC)[5] required to execute 1 Gbps of input traffic [37], [38], [49], [50]. This is in-line with the assumption that CPU utilization can be estimated as a linear function of the maximum downlink datarate [51]. While we consider uniform capacity of cloud servers *i.e.*, each CPM is a 64 RC machine, we vary the capacity of each EPM from 12 to 64 RCs to evaluate different edge cloud configurations.

Our simulation parameters are carefully taken from multiple references: (i) a hierarchical cloud network topology at metropolitan scale [11], [52], (ii) a power consumption of different sized EPMs and CPMs based on cloud's Power Usage Effectiveness[6] (PUE) [43], (iii) a power consumption proportional to the processing load with a static baseline of $50 - 70\%$ of the total server energy consumption at peak load [53], and, (iv) MH bandwidth requirement for various splits ($\mu_{i,p}^t$) obtained from Appendix C of [13] and [54].

We obtain the memory footprint of different RAN functions (ω_p) by carrying out measurements at different PRB utilization. The experiments were performed on an OpenAirInterface (OAI) [55] testbed using the latest F1 interface implementation supporting CU/DU splits [56].To compute adaptive reorchestration interval lengths, we set the threshold $\Delta_\lambda = 0.2 * \max_i \hat{\lambda}_{t_i}$ upon extensive parametric analyses. This gives us six reorchestration epochs at 00:00, 01:00, 02:30, 07:00, 08:00, and 13:30 during our reference one-day scenario. Various vRAN topology and configuration parameters, along with our heuristic solution parameters are presented in Table III.

B. Comparison Benchmarks

We compare our solution against the following benchmarks.

1) Traditional distributed RAN (`D-RAN`): This is a baseline approach where all baseband processing occurs at DUs in the ES, in line with upcoming 5G settings where dedicated servers will be typically placed in EPMs at the edge aggregation sites. All the energy consumption is at the ES and no Telco or central cloud is involved. We employ a well-known bin-packing strategy (*first-fit*) to associate DUs of each BS with EPMs: each DU is selected from an ordered sequence of BSs and associated with first EPM with enough remaining CPU capacity from an ordered sequence of EPMs. Thus, the order of association of BSs to EPMs is fixed.

2) Greedy centralized RAN (`Greedy`): At the opposite end of the spectrum of solutions from `D-RAN`, we have a fully centralized C-RAN that performs all processing in the CU. However, this is not a viable approach in realistic settings, due to strict latency constraints at the PHY layer as well as the limited capacity of the MH between DU and CU. Instead, we consider a practical version of the C-RAN approach that aims at moving to the CU as many functions as possible, while considering the limits imposed by latency and link capacity

[4] Each RU is 2×2 MIMO enabled and configured with 20 MHz bandwidth transmitting at full capacity *i.e.*, 100 PRBs with MCS index 28 [49].

[5] An RC as considered in [37] is a single Intel Haswell i7-4770 3.40GHz.

[6] PUE, a measure of data center energy efficiency, is the ratio of total data center annual energy consumption to the total compute related annual energy consumption.

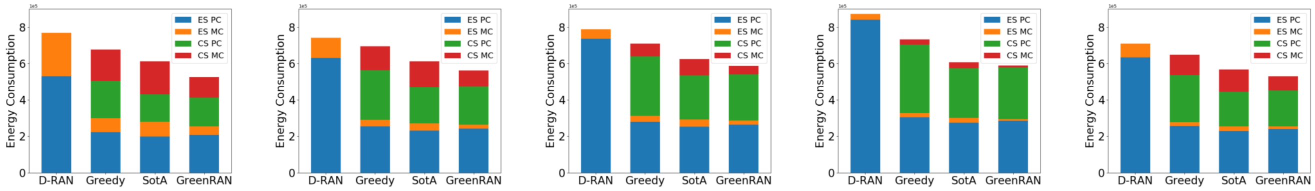

Fig. 4. Energy consumption (KJoules) of different solutions for interval lengths of 1, 3, 5, 8 hours and adaptive interval length (plots in the order from left to right). ES configuration consists of 6 EPMs per edge, each equipped with a 32 core server. MH bandwidth is 10 Gbps.

constraints. Only the functions that cannot be moved to the CU are left at the DU. This happens in two phases. First, based on the premise that higher the level of centralization, lower is the energy consumption, it greedily chooses the lowest possible functional split for each BS that meets MH capacity constraints in Eq. (20). Second, it adopts again a *first-fit* approach for DU-EPM and CU-CPM association like in `D-RAN`.

3) State-of-the-art Cloud RAN (`SotA`): A vast majority of solutions proposed for energy-efficient RAN orchestration neglect consumptions due to VM migrations. Moreover, the energy models of most related works are not as comprehensive as ours, so a direct comparison is not possible. Therefore, we benchmark `GreenRAN` against an equivalent solution where the contributions of Eq. (13) and (14) are neglected. This is equivalent to performing an optimization of the association and split variables at each epoch independent of previously made decisions, which, as mentioned above, is a common assumption in current state-of-the-art (`SotA`) works. This benchmark comprehensively models state-of-the-art works on energy-efficient vRANs, *e.g.*, [19] and [14].

VI. Results

We run experiments by orchestrating the large-scale RAN infrastructure outlined in Section V-A under realistic mobile data traffic demands. Our results not only compare the energy efficiency gains of `GreenRAN` over the benchmarks, but also provide insight into resource utilization of servers at ES and CS. We also analyze energy consumption under different vRAN configurations, thereby enabling informed decisions by mobile operators about infrastructure deployment. We also note that the runtime performance of `GreenRAN` is significantly better than solving IQP (with IBM ILOG CPLEX solver [57]). As an instance, for a single edge cloud with 25 RUs, optimization over a single reorchestration interval using CPLEX solver (branch and cut algorithm) takes more than 3 hours. Whereas, `GreenRAN` takes less than 5 minutes for the same scenario. Moreover, our heuristic, for most of the reorchestration intervals, is within 2% of the optimal IQP solution. We omit detailed comparisons with IQP due to space restrictions.

A. Energy Consumption

We start by looking at the main metric, *i.e.*, the overall energy consumption of RAN operation. Fig. 4 shows the result attained by the various solutions. Each plot in Fig. 4 refers to a different resource orchestration interval (from 1 hour to 8 hours, plus the case where adaptive interval lengths are used). In all cases, `GreenRAN` outperforms all other solutions, reducing daily energy costs by up to 14% with respect to the best competitor, and up to 33% to a traditional fully distributed `D-RAN` processing approach. When looking at gains during individual reorchestration intervals, savings can reach 25% over `SotA`, and 42% over `D-RAN`. The rightmost plot shows that adaptive intervals yield minimum energy consumption: in this case, `GreenRAN` achieves a 22000-*KWh* power consumption for the whole network under consideration.

A closer analysis reveals the origin of gains obtained by `GreenRAN`. Each cost bar is indeed split into the contributions to total energy consumption due to processing cost (PC) and migration cost (MC), at both ES and CS. Interestingly, we observe that all strategies that consider both ES and CS for processing tend to take advantage of both edge and Telco cloud resources. Therefore, and contrary to common opinion, it is not always best to offload as many RAN functions as possible to CS, if an edge cloud is available. The reason is that some EPMs are always active to process PHY functions: as such EPMs have sufficient capacity to accommodate higher layer functions as well, they are a more convenient option than activating new CPMs. For this same reason `Greedy` suffers from inefficient resource utilization, as it tries to offload all functions to CS.

Further, the energy cost breakdown demonstrates that legacy `SotA` approaches that are oblivious to migration costs do yield a reduced PC compared to `GreenRAN` because of their close-to-full-capacity server consolidation which is an artifact of considering each reorchestration interval in isolation. However, their advantage disappears when factoring in the cost of migrations: in other words, awareness of previous orchestrations can lead to sacrificing on optimal packing to reduce the overall energy consumption.

B. Impact of Reorchestrating over Adaptive Intervals

Fig. 4 lets us observe the advantage of employing adaptive resource reorchestration intervals. As noted earlier, smaller intervals may cause excessive migrations, while longer intervals cause CPU resources to stay unnecessarily reserved at PMs even when the load decreases in time. An adaptive selection of the reorchestration interval solves the problem, and leads to substantial energy consumption reductions, as shown in the rightmost plot in Fig. 4. Note that the weaknesses of the benchmarks outlined previously persist with adaptive intervals.

C. Analysis of Associations and Splits Chosen

We now delve deeper in our analysis, by providing insights on what associations and splits are chosen by `GreenRAN` and why.

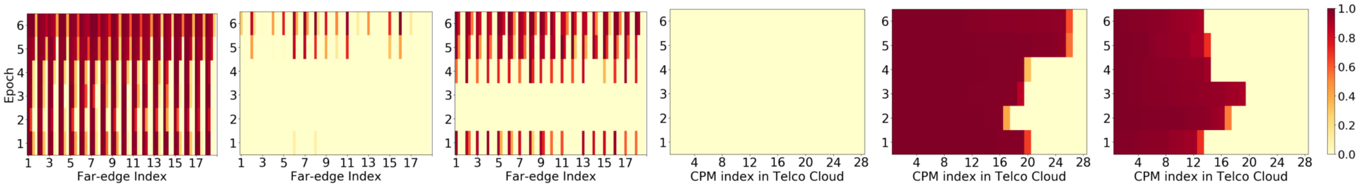

Fig. 5. Heatmap showing fractional CPU load at 18 different edges in ESs (left three) and 1 Telco cloud at CS. (right three). Benchmarks include `D-RAN` (left), `Greedy` (middle) and `GreenRAN` (right) at 6 different epochs of adaptive interval length. Each edge at ES consists of 4 EPMs with $C_{\text{epm}} = 32$ RCs. The Telco cloud consists of 28 CPMs with $C_{\text{cpm}} = 64$ RCs. Also, $C_{\text{mh}} = 10$ Gbps. Darker colors denote higher loads; light yellow are turned-off PMs.

1) Load at ES EPMs and CS CPMs: Fig. 5 shows the number of PMs that are turned on, and their CPU load, for various approaches at different epochs. In the left three plots, `D-RAN` shows higher load at EPMs while `Greedy` yields the lowest load at EPMs; these results are expected as the two solutions have opposing goals. A more efficient strategy does not try to push all processing to the CS, hence EPMs show a higher utilization under `GreenRAN` than with `Greedy`; in other words, `GreenRAN` favors DU-EPM associations over CU-CPM associations. EPMs with zero load (denoted in light yellow) are turned off, hence consume no energy. An opposite trend is observed in CPMs, in the right three plots of Fig. 5, where a reduced number of CPMs are active under `GreenRAN`: our solution tries to maximize usage of an EPM or CPM once turned on, while keeping other PMs powered off.

2) Selected Split: To better understand the splits chosen by `GreenRAN`, we plot heatmaps of the fraction of vRAN functions that are offloaded to CS in Fig 6 (left). Each row refers to a different orchestration approach. Our solution adopts an intermediate strategy between the extremes represented by `D-RAN` and `Greedy`: it prefers processing more functions at ES even when the PUE of CS is better than that of ES [43]. The behavior is especially evident when offloading a new VNF to CS may lead to turning on a CPM: instead of switching on a new CPM, an association with an EPM is preferred, subject to multiplexing opportunities available at the EPM (Fig. 5). In conclusion, these results confirm our previous intuition that greedily offloading as much computation as possible to the CS is not necessarily the best approach if edge clouds are available.

D. Effect of Different Edge Cloud and MH Configurations

To conclude our performance evaluation, we explore how `GreenRAN` behaves under a combination of different edge site configurations (*i.e.*, number and capacity of EPMs) and of MH link capacities. This analysis is helpful to mobile operators in deploying vRAN infrastructure with a reduced redundancy. Fig. 6 (right) shows the heatmap of energy consumption at various EPM sizes and MH bandwidths. We clearly see that the MH bandwidth plays a crucial role in determining the fraction of VNFs processed at the edge. In presence of little MH bandwidth, since most of the processing stays at the edge, small sized EPMs cannot benefit from the multiplexing opportunity at the edge, hence consume higher energy. As the MH bandwidth increases, the overall energy consumption reduces, due to VNFs moving towards the more efficient Telco cloud. Moreover, under large MH provisioning, smaller EPMs turn more energy efficient as they operate close to full capacity.

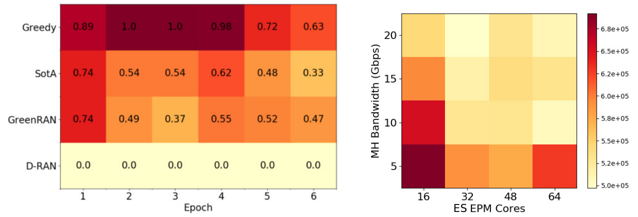

Fig. 6. Left: Heatmap showing fraction of vRAN functions processed at CS at 6 different adaptive intervals on 32 core EPMs and 15 Gbps MH bandwidth. Right: Heatmap showing total energy consumption ($E_{tot}(t)$, in *KJoules*) for different combinations of ES configurations and MH bandwidths.

On the other hand, a large EPM size surprisingly consumes more energy in presence of a low MH bandwidth. We ascribe the effect to the fact that multiple EPMs can be turned on even if some of them are underutilized, owing to heavy processing at the edge. This leads to an increase in the static energy consumption. The phenomenon disappears as the MH bandwidth increases up to a certain level. However, further increments in the MH bandwidth do not reduce energy consumption: unlike `Greedy`, the `GreenRAN` solution does not use the available MH bandwidth entirely to offload all processing to CS. To summarize, mobile operators opting for high capacity MH links or large EPM sizes (which entail high CAPEX) may not necessarily be making the best choice.

VII. Conclusions

We modeled energy consumption problem in the virtualized RAN setting as an integer quadratic program, jointly optimizing processing and migration energy consumption. Our solution, `GreenRAN`, is grounded on well-known heuristics, and achieves considerable energy savings over relevant benchmarks. The distributed nature of our algorithm enables fast and efficient computation of optimal CU-DU functional splits and BS-PM associations for a metro-scale vRAN scenario. A detailed analysis of our results showed that, in a vRAN setting, it is not always efficient to greedily offload all processing to the Telco or central clouds, and that the best strategy processes most functions in the far-edge whenever available. Among multiple far-edge cloud configurations and MH capacities, our results showed that deploying a high capacity MH link plus high-performance (many-core) edge servers is not always the best combination, as the two resources serve opposing purposes.

Acknowledgments

We thank Jon Larrea for providing the RAN VNF memory footprint measurements. R. Singh is supported in part by a PhD studentship under the EPSRC Centre for Doctoral Training in Pervasive Parallelism at the University of Edinburgh. M. Fiore is supported by the European Union Horizon 2020 research and innovation programme under grant agreement no.101017109.

References

[1] "Energy Efficiency: An Overview," (accessed Aug, 2020), https://www.gsma.com/futurenetworks/wiki/energy-efficiency-2/.

[2] "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," *White Paper*, Feb 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html

[3] A. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030," *Challenges*, vol. 6, no. 1, Apr 2015.

[4] https://www.mobileeurope.co.uk/press-wire/more-than-90-of-operators-concerned-about-rising-energy-costs-for-5g-and-edge.

[5] J. Wu, Y. Zhang, M. Zukerman, and E. K. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, 2015.

[6] C. I. *et al.*, "Toward green and soft: a 5G perspective," *IEEE Communications Magazine*, vol. 52, February 2014.

[7] J. Liu *et al.*, "On the statistical multiplexing gain of virtual base station pools," in *IEEE Global Communications Conference*, Dec 2014.

[8] M. Shehata *et al.*, "Multiplexing gain and processing savings of 5G Radio-Access-Network functional splits," *IEEE Tran. on Green Communication and Networking*, vol. 2, Dec 2018.

[9] 3GPP, "Summary of Rel–15 work items," *TR 21.915–Rel. 15*, June 2019.

[10] S. Partners, "Building Telco Edge Infrastructure: MEC, Private LTE VRAN," August 2020, https://telco.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/microsites/telco/vmware-building-telco-edge-infrastructure.pdf.

[11] Fujitsu, "New Transport Network Architectures for 5G RAN," https://www.fujitsu.com/us/Images/New-Transport-Network-Architectures-for-5G-RAN.pdf.

[12] 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, 2019.

[13] Small Cell Forum, "R6.0 small cell virtualization functional splits and use cases," *159.07.02, Release 6*, Jan 2016.

[14] 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, 2020.

[15] S. Akoush *et al.*, "Predicting the performance of virtual machine migration," in *2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, 2010.

[16] Z.-H. Zhan *et al.*, "Cloud computing resource scheduling and a survey of its evolutionary approaches," *ACM Comput. Surv.*, vol. 47, Jul 2015.

[17] K. Jörnsten and M. Näsberg, "A new lagrangian relaxation approach to the generalized assignment problem," *European Journal of Operational Research*, vol. 27, no. 3, pp. 313 – 323, 1986.

[18] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[19] A. Alabbasi, X. Wang, and C. Cavdar, "Optimal processing allocation to minimize energy and bandwidth consumption in Hybrid CRAN," *IEEE Transactions on Green Communications and Networking*, June 2018.

[20] "O-RAN Use Cases and Deployment Scenarios," February 2020, https://tinyurl.com/yxa2b7xz.

[21] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-On/Off strategies for green cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, May 2013.

[22] Z. Niu, "TANGO: Traffic-aware network planning and green operation," *IEEE Wireless Communications*, vol. 18, no. 5, Oct 2011.

[23] M. Qian *et al.*, "Baseband processing units virtualization for cloud radio access networks," *IEEE Wireless Communications Letters*, April 2015.

[24] X. Wang *et al.*, "Energy-efficient virtual base station formation in optical-access-enabled Cloud-RAN," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, May 2016.

[25] B. J. R. Sahu *et al.*, "Energy-efficient BBU allocation for green C-RAN," *IEEE Communication Letters*, vol. 21, July 2017.

[26] N. Saxena, A. Roy, and H. Kim, "Traffic-aware Cloud RAN: A key for green 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, April 2016.

[27] 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, 2019.

[28] A. Verma *et al.*, "PMapper: Power and migration cost aware application placement in virtualized systems," in *Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware*, 2008.

[29] K. Zheng, X. Wang, L. Li, and X. Wang, "Joint power optimization of data center network and servers with correlation analysis," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014, pp. 2598–2606.

[30] C. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, 2014.

[31] A. Umesh, T. Yajima, and T. U. S. Okuyama, "Overview of o-ran fronthaul specifications," *NTT Docomo Technical Journal*, vol. 21, no. 1, Jul 2019. [Online]. Available: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/bn/vol21_1/vol21_1_007en.pdf

[32] "4G/5G RAN architecture: how a split can make the difference," (accessed Jan, 2021). [Online]. Available: https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/4g5g-ran-architecture-how-a-split-can-make-the-difference

[33] K. I. Pedersen *et al.*, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, 2016.

[34] "Report ITU-R M.2410-0: Minimum requirements related to technical performancefor IMT-2020 radiointerface(s)," November 2017, https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf.

[35] "Learnings from virtualized RAN technology trials over non-ideal fronthaul," *Telecom Infra Project*, 2019. [Online]. Available: https://telecominfraproject.com/wp-content/uploads/vRAN_FH_WP2.pdf

[36] "Creating an ecosystem for vRANs supporting non-ideal fronthaul," *Telecom Infra Project*. [Online]. Available: https://telecominfraproject.com/wp-content/uploads/VRAN_WP_final.pdf

[37] A. Garcia-Saavedra *et al.*, "FluidRAN: Optimized vRAN/MEC Orchestration," in *IEEE INFOCOM 2018*.

[38] C. Y. Yeoh *et al.*, "Performance study of lte experimental testbed using openairinterface," in *International Conference on Advanced Communication Technology*, Jan 2016.

[39] Small Cell Forum, "FAPI and nFAPI Specifications," *Release 9.0*, February 2017.

[40] ——, "About 5G nFAPI," *Document 226.1.0*, September 2020.

[41] ETSI TS, "5G; NG-RAN; F1 Application Protocol (F1AP)," *3GPP TS 38.473 V 15.3.0 Rel. 15*, October 2018.

[42] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, vol. 18, no. 5, October 2011.

[43] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communication Surveys & Tutorials*, 2016.

[44] J. W. Jiang *et al.*, "Joint VM placement and routing for data center traffic engineering," in *2012 Proceedings IEEE INFOCOM*, 2012.

[45] Q. Fan, N. Ansari, and X. Sun, "Energy driven avatar migration in green cloudlet networks," *IEEE Communications Letters*, vol. 21, July 2017.

[46] H. Liu, H. Jin, C.-Z. Xu, and X. Liao, "Performance and energy modeling for live migration of virtual machines," *Cluster Computing*, vol. 16, 2013.

[47] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual comparison," *ACM Comput. Surv.*, Sept. 2003.

[48] S. Boyd, L. Xiao, A. Mutapcic, and J. Mattingley, "Notes on Decomposition Methods," (2015).

[49] A. Garcia-Saavedra, J. X. Salvat, X. Li, and X. Costa-Perez, "WizHaul: On the centralization degree of cloud ran next generation fronthaul," *IEEE Transactions on Mobile Computing (TMC)*, vol. 17, Oct 2018.

[50] N. Nikaein, "Processing radio access network functions in the cloud: Critical issues and modeling," in *Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services*, 2015.

[51] T. X. Tran *et al.*, "Understanding the computational requirements of virtualized baseband units using a programmable cloud radio access network testbed," in *IEEE International Conference on Autonomic Computing (ICAC)*, 2017.

[52] https://www.rcrwireless.com/20190806/5g/rakuten-deploy-4000-edge-servers-virtualized-mobile-network-report.

[53] B. Yu *et al.*, "An energy-aware algorithm for optimizing resource allocation in software defined network," in *IEEE GLOBECOM*, 2016.

[54] X. Wang *et al.*, "Centralize or distribute? A techno-economic study to design a low-cost cloud radio access network," in *IEEE International Conference on Communications*, May 2017.

[55] "OpenAirInterface: 5G Software Alliance for Democratising Wireless Innovation," (accessed Aug, 2020), https://www.openairinterface.org/.

[56] "OpenAirInterface 5G Wiki - F1 interface," (accessed Aug, 2020), https://gitlab.eurecom.fr/oai/openairinterface5g/-/wikis/f1-interface.

[57] "IBM ILOG CPLEX optimization studio, Version 12.9," *IBM*, 2019. [Online]. Available: https://www.ibm.com/products/ilog-cplex-optimization-studio