Apt-RAN: A Flexible Split Based 5G RAN to Minimize Energy Consumption and Handovers

Himank Gupta, *Student Member, IEEE,* Mehul Sharma, *Student Member, IEEE,*
Antony Franklin A, *Senior Member, IEEE,* Bheemarjuna Reddy Tamma, *Senior Member, IEEE*

***Abstract*—The recent adoption of virtualized technologies in Next Generation Radio Access Network (NG-RAN) has driven a significant impact on energy consumption by subsequently decreasing the number of active base stations. The base station (gNodeB) of 5G is segregated into cost-efficient Central Units (CU) hosted on virtual platforms and cheaper & smaller Distributed Units (DU) present at the cell sites. Multiple CUs are pooled together in a single powerful central cloud, known as CU pool. The logical connection between DU and CU can be dynamically adjusted and can potentially affect the energy consumption of the CU pool. The deployment of NG-RAN imposes strict latency requirements on the fronthaul link that connects DUs to CU. To relax these strict latency requirements, various alternate architectures such as Flexible RAN Functional Splits have been proposed by 3GPP. In this paper, we first evaluate the energy consumption of DU and CU for various functional split options using OpenAirInterface (OAI), a real-time open source software radio solution. We find that lower layer splits have high energy consumption at CU as compared to higher layer split options. We also observe the variation in energy consumption due to traffic heterogeneity. Motivated by the above study, we formulate an optimization model, *Apt-RAN*, that optimizes the energy consumption of the CU pool and the number of handovers, considering different functional splits. To address the computational complexity of solving the optimization model, a lightweight polynomial time heuristic algorithm is proposed. Simulation results demonstrate that our proposed model outperforms existing state-of-art schemes.**

***Index Terms*—Central Unit (CU), Distributed Unit (DU), Flexible functional splits, Handovers, OpenAirInterface (OAI), CU Pool.**

I. Introduction

Over the past few years, the popularity of internet-enabled smartphones and tablets, along with data-intensive high-end applications, have increased to preposterous heights [1]. This has resulted in a colossal increase in data demand and has forced the network operators to upgrade their networks constantly while keeping the costs as low as possible to offer competitive prices. While the current network architecture was not originally designed to cope up with such exponentially growing rate, Next Generation Radio Access Networks (NG-RAN) has been recently introduced as a competent and proficient solution to address the above issues as well as to reduce the deployment cost.

In NG-RAN, the protocol stack of Next Generation Node B (gNB) is split into two components, Central Units (CU) and Distributed Units (DU) [2]. DUs remain at the cell site to provide basic signal transmission functionalities whereas CUs are aggregated in a CU pool where cloud computing and virtualization mechanisms are used to provide significant energy efficiency and multiplexing gains, as shown in Fig. 1. Both CU and DU communicate through a low latency high-bandwidth interface called as fronthaul interface. Specifications from Common Public Radio Interface (eCPRI) [3] and Next Generation Fronthaul Interface (NGFI) [4] are used to carry the IQ samples over the fronthaul link. The bandwidth and latency budget required to run a fully centralized solution is extremely high. As per 3GPP report [5], a fully centralized network considering 5G-NR with 100 MHz and 32 antennas requires a fronthaul bandwidth 157.3 Gbps. Such a high capacity may not be affordable and thus leaves a room for improvement. Therefore, 3GPP proposed the concept of functional splits to have a partially centralized NG-RAN architecture. A functional split determines which gNB functions to be left locally at the cell site and which functions to be moved to the central CU Pool. The functional splits, along with centralization and virtualization technology, provides a higher degree of freedom that can be utilized to make optimized decisions.

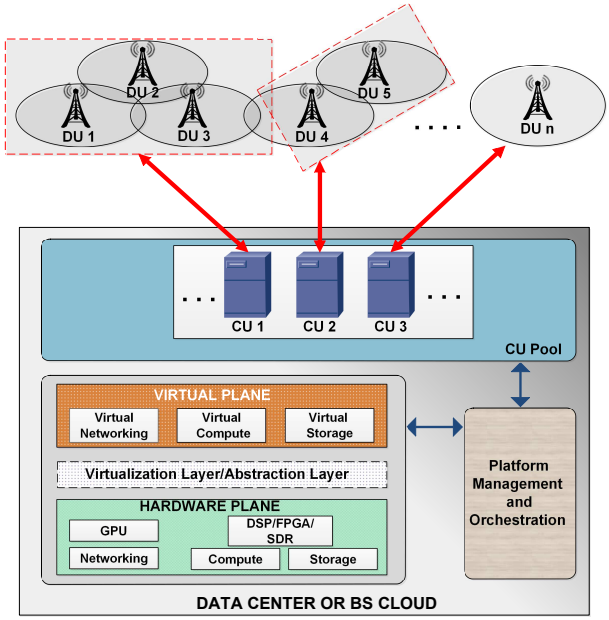

Fig. 1: Programmable & virtualized computing center for CUs in NG-RAN.

The traffic pattern of a cell is observed to be influenced by its geographical location and its neighboring cells, known as spatio-temporal traffic variation or the tidal effect, as shown in Fig. 2. It can be observed that the DU load is more during weekdays as compared to the load during the weekends. It is also noticeable that peak load occurs only for few hours of

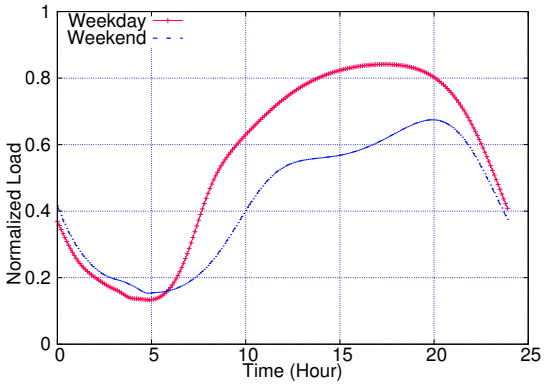

Fig. 2: Time varying normalized mean traffic load at the cell site.

a day. Typically, in NG-RAN, a one-to-one mapping exists between DUs and CUs. However, considering the diverse mobile services and fluctuating traffic patterns, allocating one dedicated CU to each DU leads to extremely poor utilization of CU resources during off-peak hours. The server in which CU is provisioned has to be active all the time even if only few users are served by it. Therefore, allocating one CU to each DU is highly inefficient and inevitably leads to significant wastage of energy and CU resources.

Recent advances in virtualization and cloudification technologies allow operators to deploy instances of CU on top of hypervisors such as dockers, VMs, etc., to reduce power consumption and improve resource utilization [6]–[8]. As shown in Fig. 1, multiple DUs can be mapped to a single CU because of the isolation flexibility of cloud platform, thereby creating a many-to-one deployment relationship between DU and CU. Compared to one-to-one DU to CU mapping, many-to-one mapping provides more flexibility and higher energy savings. More specifically, several VMs can be turned on or off based on traffic heterogeneity to reduce power consumption. The energy consumption can be minimized even further by reducing the extra cost associated with the relocation of DUs. A relocation can occur when the traffic load of DUs hosted on the same CU exceeds beyond CU capacity, thereby degrading Quality of Experience (QoE) for the end users. DU relocations must be triggered in a controlled manner such that the number of user services (e.g., Guaranteed Bit Rate (GBR) applications) affected during the relocation is minimum. Furthermore, while relocating the DUs, the location of DUs should be taken into account such that the neighboring DUs are relocated to the same CU. This will reduce the possibility of inter-DU handovers experienced by the users (discussed later in Section II-B).

The facts mentioned above motivate us to study the DU-CU mapping problem in NG-RAN with the objective to minimize the energy consumption at the CU pool along with the number of handovers while ensuring Quality of Experience (QoE) for the end users. More specifically, the question that is addressed in this paper is: *how does Apt-RAN model, proposed with energy cost in mind, affect QoE of UEs? Apt-RAN* is a novel and innovative model that aims to minimize the total energy consumption at CU pool by reducing the number of active CUs. *Apt-RAN* also reduces the number of handovers by mapping neighboring DUs to the same CU based on the mobility probability of UEs (discussed later in Section III-E). Finally, to ensure better QoE for the users, *Apt-RAN* minimizes the total number of DU relocations and the number of affected GBR flows.

Following are the main contributions in this paper:

1) We implement a real-time and programmable NG-RAN testbed using OpenAirInterface (OAI) [9]. Both CU and DU are realized over virtualized platforms and are connected via a high speed 10 Gbps optical fibre fronthaul link. The DU is connected with a USRP B210 device to transmit and receive radio signals.
2) We use the above implemented NG-RAN testbed to perform extensive experiments to study the variation in energy consumption at CU by varying the load at DU. We also study the effect of various functional splits on energy consumption at CU and DU.
3) Using motivational results from the OAI testbed, we formulate an optimization model, *Apt-RAN*, that minimizes the total energy consumption at CU pool along with the total number of handovers, considering different functional splits.
4) The proposed model improves the QoE by reducing the number of DU relocations and the number of affected GBR flows associated with each user.
5) To address the computational complexity of the optimization model, we propose a lightweight polynomial-time heuristic algorithm. The greedy approach of the heuristic algorithm can find a near optimal solution in the order of seconds. This makes the proposed algorithm an efficient solution for real-world deployments.

The rest of the paper is organized as follows: Section II presents the motivational results obtained from the real-time OAI testbed. The system model and the problem formulation are presented in Section III and Section IV, respectively. In Section V, we propose the heuristic algorithm and evaluate its performance. A comprehensive review of related works on energy and handovers minimization is presented in Section VI. Finally, conclusions and future works are highlighted in Section VII.

II. Motivation

A. Effect of Functional Splits on Energy Consumption

As a part of NG-RAN, 3GPP proposed eight different functional split options between DU and CU [10], as shown in Fig. 3. Some of the benefits of deploying a flexible split based architecture include cost-effectiveness, load management, real-time performance optimization using dynamic split, and reduction in fronthaul bandwidth requirement. The choice of how to split the NG-RAN architecture depends on several factors related to radio network deployment scenarios, traffic constraints, and intended supported services. Some of these factors are QoE (low latency, high throughput), user density, and geographical location of DUs.

Moving from Option 1 to Option 8, computationally costly operations like Fast Fourier Transformation (FFT), Inverse

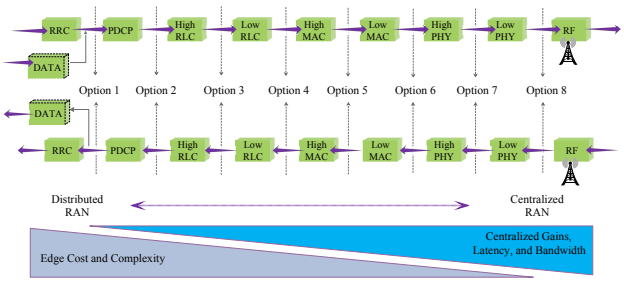

Fig. 3: Split options as defined in 3GPP TR 38.801.

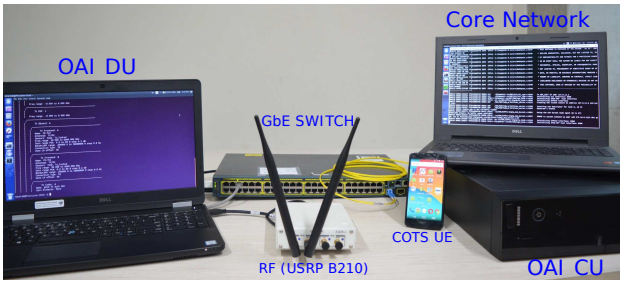

Fig. 4: OAI based NG-RAN testbed for different splits of NG-RAN.

Fast Fourier Transformation (IFFT), Rate Matching, and Turbo encoding/decoding are shifted to CU side, resulting in variation in energy consumption at CU and DU. To validate this claim, an NG-RAN prototype is developed for split Option 2 (PDCP/RLC), Option 7 (Lower PHY/Higher PHY), and Option 8 (PHY/RF) using OAI, as shown in Fig. 4. The CU is deployed on an Intel Xeon x86 machine with 3.4 GHz frequency and connected to DU with the identical configuration through a Gigabit Ethernet (GbE) switch. DU is connected with an RF front-end using USRP B210 device. All the parameters used in the OAI experimental setup are listed in Table I. We use *iPerf3* [11] tool to generate uplink and downlink TCP traffic between a UE and NG-RAN for a fixed duration of 120 seconds. To calculate the energy consumption at both CU and DU, Running Average Power Limit (RAPL) [12] tool is used. RAPL tool is helpful in fetching thread-wise power consumption with the help of Model Specific Register (MSR) [13]. Fig. 5 shows the power consumption at CU and DU for split Option 2, Option 7, and Option 8. One of the key observations from this study is the energy of CU is reduced by nearly 30% when we move from Option 8 to Option 7 as lower PHY layer functions such as FFT and IFFT are moved to DU side. However, higher PHY operations like turbo encoding/decoding operations still reside in CU for both the splits. Similarly, nearly 75% of CU energy is reduced for Option 2 as compared to Option 8. The same prototype is used to study the effect of traffic heterogeneity on energy consumption at CU for split Option 7, as shown in Fig. 6. Traffic heterogeneity is achieved by limiting the number of Physical Resource Blocks (PRBs) in $enb\_scheduler$ module of OAI.

TABLE I: OAI testbed parameters

Parameters	Values
Frequency	2660 MHz (DL)
Bandwidth	10 MHz
Maximum Resource Blocks	50
Number of Connected UEs	2
Mode	FDD Band 7
Fronthaul Connection	10 Gbps Optical cable

Based on the above study, we conclude that there is a significant variation in energy consumption at both DU and CU when different splits are chosen. We also observe that traffic heterogeneity also has a considerable impact on energy consumption. Therefore, we model our RAN topology based on different split options, considering spatio-temporal traffic patterns.

B. Handovers in Virtualized NG-RAN

The handover procedure is one of the critical functions of mobile networks. During the handover, UE changes its association from one DU to another DU. In traditional mobile networks, when UE moves from one cell to another neighboring cell (both part of the same network), ongoing UE's traffic is re-routed to the neighboring cell. Handover is a function of RRC (Radio Resource Control) protocol that is running on UE and CU, without any assistance of DU to which UE is connected to. During the handover procedure, the latency of re-routing of the ongoing traffic could affect QoE anticipated by UEs. The authors of [14] observed that

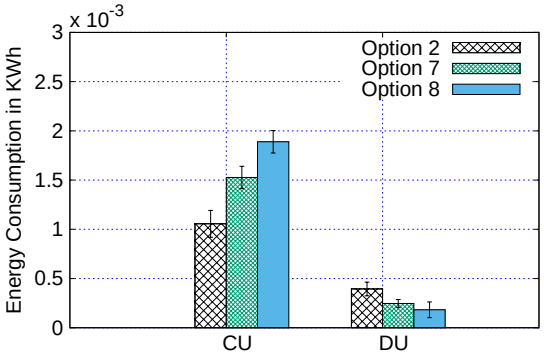

Fig. 5: Energy consumption for 120 seconds for different split options (Option 2, Option 7, and Option 8) at CU and DU.

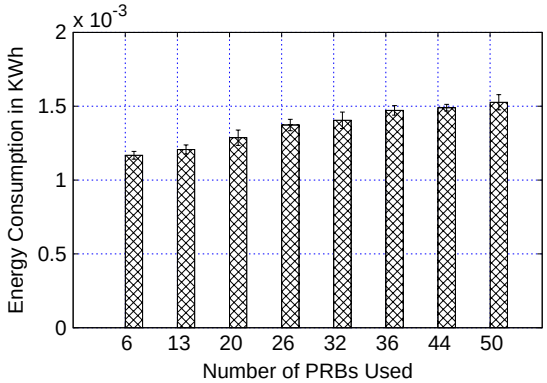

Fig. 6: Variation in energy consumption at CU vs number of PRBs used for 120 seconds for split Option 7.

there is a 10% increase in video session abandonment rate due to increased handover latency. Handovers also affect the QoE of the web traffic. The authors of [15] claimed that most of the web sessions are abandoned in the presence of handovers. The authors of [16] highlighted that signaling overhead is drastically increased during the handover procedure, resulting in increased call holding time.

In the proposed *Apt-RAN* model, we consider many-to-one mapping between DUs and CUs. Therefore, when a user moves from the coverage area of one DU to the coverage area of another DU, it does not result in a handover if both the DUs are mapped to the same CU. Reducing handovers help in satisfying QoE of users in 5G.

III. System Model

Consider $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ as the finite set of DUs and CUs in CU pool, respectively. Since there can be a many-to-one mapping between DUs and CUs, we have $\mid \mathcal{D} \mid \geq \mid \mathcal{C} \mid$. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_k\}$ be the set of k UEs distributed under coverage of $\mid \mathcal{D} \mid$ DUs. All the CUs have a fixed capacity C_c. We define the association between DU $d \in \mathcal{D}$ and CU $c \in \mathcal{C}$ using a binary variable

$$y_{dc} = \begin{cases} 1, & \text{if DU } d \text{ is mapped to CU } c \\ 0, & \text{otherwise} \end{cases}$$

Considering traffic heterogeneity, all CUs may not be active all the time. Therefore, we define a binary variable

$$z_c = \begin{cases} 1, & \text{if CU } c \text{ is active} \\ 0, & \text{otherwise} \end{cases}$$

Similarly, all DUs may not be active if there is no UE connected to it. Such a scenario is highly unlikely but could still be considered in the optimization model. Therefore, we denote another boolean variable

$$z'_d = \begin{cases} 1, & \text{if DU } d \text{ is active} \\ 0, & \text{otherwise} \end{cases}$$

We consider that each UE $u \in \mathcal{U}$ can be served by only one DU. Therefore

$$x_{ud} = \begin{cases} 1, & \text{if UE } u \text{ is served by DU } d \\ 0, & \text{otherwise} \end{cases}$$

A. RAN Topology

For modeling the RAN topology, tools from stochastic geometry, point process, and spatial statistics are proven to be more accurate and realistic. Deployment of UEs and DUs is done based on the stochastic point process theory. Fig. 7 highlights the distribution of 50 DUs and 500 UEs in a region spread over *40 km * 40 km*. The deployment of DUs is accomplished as per the Matern Hard Core Point Process type II (MHCPP II) [17] and the deployment of UEs as per Poisson Point Process (PPP) model. Let, the spatial distribution of DUs and UEs be Φ_{DU} and Φ_{UE}, respectively, where $\{\Phi_{DU}, \Phi_{UE}\} \in \mathbb{R}^2$. To plot and depict the coverage regions, Voronoi tessellation is used. For better precision, propagation loss with shadowing and fading is included in the channel model. Values of SINR and interference from neighboring DUs are computed using 3GPP Pathloss Model [18], and every UE is associated with the DU that gives the highest SINR value.

In this work, we consider three different split options for realizing RAN topology. These splits are PDCP/RLC (Option 2), HighPHY/LowPHY (Option 7), and PHY/RF (Option 8). Each functional split option imposes a certain strict fronthaul bandwidth and latency requirement, as stated by 3GPP in [5]. Authors in [19] studied the datasets of real topologies of different countries and designed a framework to optimally place the RAN functions, considering different functional

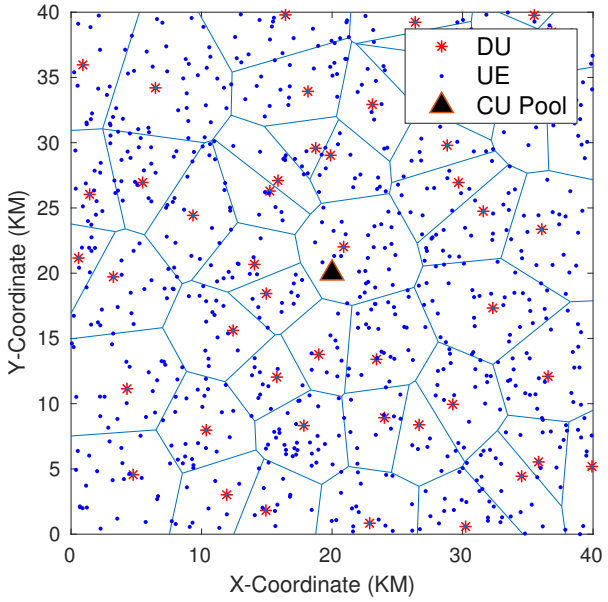

Fig. 7: Spatial distribution of 500 UEs and 50 DUs as per MHCPP II.

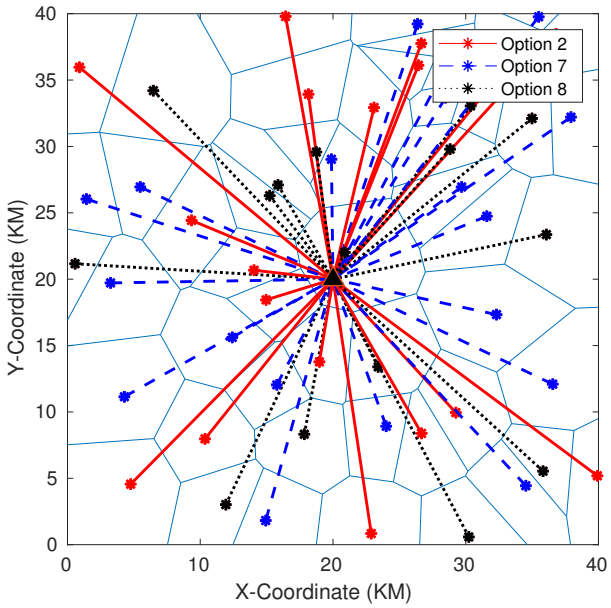

Fig. 8: Split specific assignment of DUs to CU.

splits. We used meta-data of the same topologies to model our RAN topology. The meta-data includes bandwidth and latency of the fronthaul links between DU and CU. The same meta-data is applied on Fig. 7 to decide the split option for each DU-CU pair, as shown in Fig. 8. For each DU, lower-layer split options (Option 7 and Option 8) are preferred over higher layer split options (Option 2), if both latency and bandwidth constraints are satisfied. The main idea behind this preference is to have as many centralization benefits as possible such as Coordinated Multi-Point (CoMP), Inter-Cell Interference Coordination (ICIC), energy efficiency, adaptability to non-uniform traffic, and improved QoE for the users [20].

B. DU Traffic Model: Cell Load

At any given time t, each DU $d \in \mathcal{D}$ is associated with a set of users $U_d(t) \subset \mathcal{U}$ based on maximum received signal strength. Let $\psi^u_{\mathcal{D}}(t) = (\psi^u_1(t), \psi^u_2(t), \ldots, \psi^u_D(t))$ be the complex channel vectors from all the DUs to user $u \in \mathcal{U}$ at time t. In our model, we use the information of instantaneous rate and channel quality of the UEs that are available at the MAC scheduler to estimate the load at cell depending on the split option used by the respective DU-CU pair [21]. Consider, N_u as the total number of PRBs allocated to user u and N be the total number of PRBs available with each DU. Hence, the load $l_d(t)$ at each DU d at time t, is given by,

$$l_d(t) = \sum_{u \in U_d(t)} \frac{N_u}{N} \tag{1}$$

To ensure QoE for the users, the proposed model also aims at supporting Guaranteed Bit Rate (GBR) applications as described in 3GPP TS 23.203 such as conversational videos, voice traffic flows, live streaming videos, and real-time gaming. Let us denote N_c, N_v, N_s, and, N_g as number of conversational videos, voice traffic flows, live streaming videos, and real-time gaming, respectively. The information regarding the type of GBR flows is available at base station (CU) during the establishment of bearers itself. In LTE, different flows can be classified using the QCI (QoS Class Identifier) of the bearers. Similarly, in 5G, QFI (QoS Flow ID) can be used to classify the flows [22]. Let N_{ngbr} be the total number of non-GBR data flows in DU. All the flows are generated randomly using Poisson Distribution. For each DU $d \in \mathcal{D}$, we compute the score metric ws'_d as described in Eqn. (2) that represents the consolidated value of active user flows.

$$\begin{aligned} ws'_d = (w_1 \times N_{ngbr}) + (w_2 \times N_v + \\ w_3 \times N_c + w_4 \times N_g + w_5 \times N_s) \end{aligned} \tag{2}$$

such that, $\sum_{i=1}^{5} w_i = 1$. Individual DU weighted score ws'_d ($ws'_{min} <= ws'_d <= ws'_{max}$) can be normalized within range [0,1] as follows.

$$ws_d = \frac{(ws'_d - ws'_{min})}{(ws'_{max} - ws'_{min})} \tag{3}$$

To prioritize the user flows in a DU, above calculated normalized score is used.

Remark 1. *Several weighted decision variables $(w_1, w_2, w_3, w_4, w_5)$ are considered to calculate the score metric for each DU. However, it is non-trivial to give an analytical proof for finding these values. These parameters are operator dependent and vary from one operator to the other. The proposed Apt-RAN model is flexible enough to be modeled based on any set of values for these variables.*

C. CU Computing Resource Model

For each user $u \in \mathcal{U}$, baseband processing load consists of two major components. First one is user independent static cell specific baseband load. The second one is a dynamic value which is dependent on user and is modeled as a function of Physical Resource Blocks (PRBs) and Modulation and Coding Scheme (MCS) assigned to those channel resources [23]. The baseband processing time per subframe $proc(u,t)$ in microsecond is given by,

$$proc(u,t) = r_{base} + p_{base} + u(mcs, prb) + u(r) \tag{4}$$

where r_{base} and p_{base} are the constant base offsets for the cell and virtualized platform (Docker, VirtualBox, KVM, and etc.), respectively. $u(mcs, prb)$ is the user dependent processing which is a function of allocated PRBs and MCS, and $u(r)$ is the remainder of other user-specific tasks.

Performance of a cloud platform is typically measured in terms of a number of instructions executed per second. However, for scientific formulations and accurate analysis, count of FLoating-point Operations Per Second (FLOPS) is used to measure the cloud platform's performance. E.g., consider a single core Intel CPU with a frequency of 2.5 GHz, capable of performing 4 FLOPS in one cycle. This results in a theoretical performance of $(2.5 \times 10^9 \times 4) = 10$ GFLOPS. Let us denote this maximum value as L_{max}. The compute load $l_c(t)$ for CU $c \in \mathcal{C}$ in Floating-point operations serving multiple DUs $d_1, d_2, \ldots, d_n \in \mathcal{D}$ is given by,

$$l_c(t) = L_{max} \times \left(\sum_{u \in U_d(t)} proc(u,t) \right) \tag{5}$$

D. Consolidation and Relocation Cost

As mentioned in Section II, traffic heterogeneity can lead to over-utilization or under-utilization of CUs. This problem can be tackled by live relocation of DUs from one CU to another CU. However, this relocation incurs additional cost, because it iteratively writes all the active memory pages of DU from a serving CU to the target CU. This extra cost of relocation can be modeled as a linear relationship between compute load and power consumption. Considering P_{idle} as the idle power drawn at 0% compute load and P_{cap} as the maximum power drawn by the CU at 100% compute load, the power consumption of a CU with compute load l_c is given by,

$$P_c = P_{idle} + (P_{cap} - P_{idle}) \times l_c \tag{6}$$

The energy consumption is given by $(P_c \times t_c)$, where t_c represents the time duration for which CU is operated at P_c power.

As per SPEC power benchmark [24], a standard general purpose Intel Xeon X5670 processor consumes nearly 259 Watt on an average when compute load at the server is 100%. The authors in [25] investigated the major factors impacting migration performance and designed a model to evaluate the same. Based on the performance model, a linear model is formulated using linear regression to estimate VM relocation energy. Let E_d be the energy in Watt-second consumed when a DU d are relocated from source to target CU,

$$E_d = \alpha \times B_c + \beta \tag{7}$$

where B_c is the total amount of data volume (in bytes) relocated from source to target and α, β are the regression parameters. The values of regression parameters α and β are derived using linear regression technique and are equal to 0.512 and 20.165, respectively.

E. Probability of User Mobility

Consider Γ as the distance matrix where each entry Γ_{ij} is the geographical distance in meters between DUs d_i and d_j. In a practical scenario, a mobile user is likely to move to one of neighboring DU's region. To simulate this, mobility probability for a user moving from one DU region to other is calculated. Thus, to assign high mobility probability to neighboring DUs, the distance between DUs is normalized. The normalized distance $norm(i)$ for each DU d_i from all other DUs d_j such that j$\in \Gamma_{i,*}$ can be defined as,

$$norm(i) = (max(\Gamma_{i,*}) + 1) - \Gamma_{ij} \tag{8}$$

The mobility probability P_{ij} of a user moving from DU d_i to d_j s.t. j$\in \Gamma_{i,*}$ can be obtained as,

$$P_{ij} = \frac{norm(i)}{\sum_{j=1}^{D} norm(j)} \tag{9}$$

Fig. 9 illustrates the handover based on mobility probability P_{ij} and demonstrates how it differs from the conventional mobile network handover procedure. In this scenario, UE1 is moving from coverage region of DU d_1 at time t to coverage region of DU d_2 at time $t+1$. As a result, UE1 will not confront a handover because d_1 and d_2 are mapped to the same CU. Whereas, UE2 moving from DU d_2 to DU d_3 will encounter a handover as it moves between those DUs which are mapped to different CUs. But in a conventional mobile network, both UE1 and UE2 would undergo handovers.

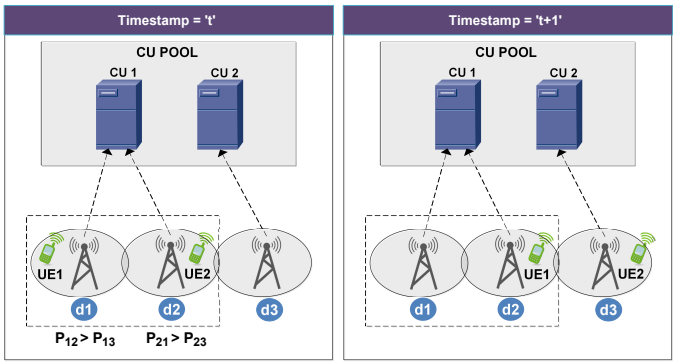

Fig. 9: Illustration of handover process based on mobility probability P_{ij} of UEs moving from DU d_i to DU d_j.

IV. Optimization Model (APT-RAN)

The notation used in this optimization model are listed in Table II.

Objective Function:

$$\min_{x,y,z,z'} : \left(\underbrace{\sum_{c=1}^{\mathcal{C}} (z_c \times cost_c)}_{\text{(A)}} \times \underbrace{\frac{1}{\sum_{\forall i,j \in c} P_{ij}}}_{\text{(B)}} \right) + \left(\underbrace{\sum_{\substack{d \in \mathcal{D}, c \in \mathcal{C}, \\ \text{such that} \\ c \neq A_{(t-1)}(d)}} (ws_d \times y_{dc} \times cost_{dc})}_{\text{(C)}} \right) \tag{10}$$

Constraints :

$$\sum_{d=1}^{\mathcal{D}} (y_{dc} \times l_d) \leq (C_c \times z_c), \quad \forall c \in \mathcal{C} \tag{11}$$

$$\sum_{u=1}^{\mathcal{U}} (x_{ud}(t) \times N_u(t)) \leq N \quad \forall d \in \mathcal{D} \tag{12}$$

$$\sum_{c=1}^{\mathcal{C}} y_{dc} = 1, \quad \forall d \in \mathcal{D} \tag{13}$$

$$\sum_{d=1}^{\mathcal{D}} x_{ud} = 1, \quad \forall u \in \mathcal{U} \tag{14}$$

$$y_{dc} \leq z_c, \quad \forall d \in \mathcal{D}, \forall c \in \mathcal{C} \tag{15}$$

$$x_{ud} \leq z'_d, \quad \forall u \in \mathcal{U}, \forall d \in \mathcal{D} \tag{16}$$

$$x_{ud} \in \{0,1\}, \quad \forall u \in \mathcal{U}, \forall d \in \mathcal{D} \tag{17}$$

$$y_{dc} \in \{0,1\}, \quad \forall d \in \mathcal{D}, \forall c \in \mathcal{C} \tag{18}$$

$$z_c \in \{0,1\}, \quad \forall c \in \mathcal{C} \tag{19}$$

$$z'_d \in \{0,1\}, \quad \forall d \in \mathcal{D} \tag{20}$$

There are three main components of the objective function in the proposed optimization model denoted as A, B, and C in Eqn. (10).

1) Term A minimizes the total energy consumption by minimizing the total number of active CUs in the CU pool.
2) Term B minimizes the number of handovers by mapping DUs to CU whose sum of mobility probability (P_{ij}) with that particular CU is maximum.

TABLE II: Notation used in the *Apt-RAN* Optimization Model

Notation	Definition
l_d	Compute load at DU $d \in \mathcal{D}$
C_c	Capacity of CU $c \in \mathcal{C}$
z_c	1 if CU $c \in \mathcal{C}$ is active; otherwise 0
z'_d	1 if DU $d \in \mathcal{D}$ is active; otherwise 0
y_{dc}	1 if DU d is mapped to CU c; otherwise 0
x_{ud}	1 if UE u is associated to DU d; otherwise 0
A_t	Allocation matrix of all DUs to CUs at time t
$A_t(d)$	Allocation of DU $d \in \mathcal{D}$ to CU c at time t
$cost_c$	Energy cost of operating CU $c \in \mathcal{C}$
$cost_{dc}$	Additional energy cost incurred in relocation of DU $d \in \mathcal{D}$ to CU $c \in \mathcal{C}$
ws_d	Normalized weighted score of $d \in \mathcal{D}$ indicating relocation impact
P_{ij}	Probability of user mobility from DU d_i to d_j
N_u	Total PRBs allocated to user $u \in \mathcal{U}$
N	Fixed number of PRBs available to each DU $d \in \mathcal{D}$

3) Term C minimizes the additional energy incurred due to relocations of DUs while also considering service disruptions to users.

Eqn. (11) ensures that the total CU load does not exceed the maximum rated capacity of the system. Eqn. (12) states that the sum of PRBs allocated to all the users served by a given DU should not exceed the PRB limitation of DU. To be more specific, this equation sets a constraint for DUs in terms of the maximum number of UEs that can be served by it in one transmission Time Interval (TTI). Eqn. (13) ensures that each DU is associated with exactly one CU. Similarly, Eqn. (14) ensures that a UE can only be associated with exactly one DU. Eqn. (15) indicates that DUs can only be mapped with active CUs. Similarly, Eqn. (16) makes sure that a UE can only be associated with an active DU. Eqns. (17-20) indicate that y_{dc}, x_{ud}, z_c, and z'_d are boolean variables.

The above optimization model minimizes the total energy consumption of CU while considering the traffic heterogeneity. It also minimizes the total number of handovers by ensuring that neighboring DUs are mapped to the same CU based on the mobility probability P_{ij} calculated in Eqn. (9). However, the proposed *Apt-RAN* model does not address the problem of minimizing energy consumption at DUs.

A. Experimental Setup

To evaluate the performance of the proposed *Apt-RAN* optimization model, a service region of *40 km * 40 km* is considered with different number of DUs, *i.e.*, 10, 20, 30, 40, and 50 distributed geographically as per MHCPP-II, as described in Section III-A. Large-scale fading and 3GPP Outdoor Path Loss Model [18] are considered to generate the channel gains. For simplicity, the transmission power of each DU is kept as one Watt, shadowing as 10 dB, and fixed the noise Power Spectral Density (PSD) as -184 dBm/Hz. The essential simulation parameters are listed in Table III. User traffic generated during a day is divided into three segments. They are "Low_Load" from 12AM to 8AM, "Medium_Load" from 8AM to 12 Noon and 8PM to 12AM, "High_load" from 12 Noon to 8PM. A total of 240 time series samples are collected using a Gaussian Mixture Model (GMM) as shown in [26] where each sample is generated after a time interval

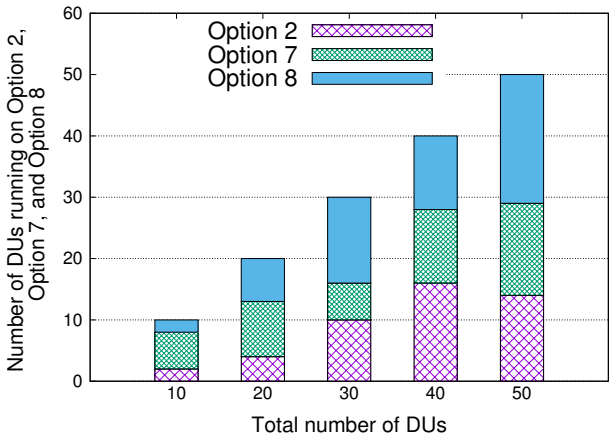

Fig. 10: DUs running on different split options vs total number of DUs.

of 6 minutes. Each DU is associated with a certain number of downlink flows, generated using a Poisson distribution with rate parameter λ calculated as the mean of traffic load during the 6 minutes. The proposed *Apt-RAN* model adopts the concept of flexible functional splits based on the available fronthaul bandwidth and link delay. The RAN splits for all DU-CU pairs are given as inputs to the optimization model for obtaining the optimal mapping of DUs and CUs based on the given traffic load at cells. The number of DUs running on split Option 2, Option 7, and Option 8 is shown in Fig. 10.

B. Performance Analysis of Apt-RAN Model

For the comprehensive performance evaluation, the proposed optimization model *Apt-RAN* is compared with KORA framework developed in [27]. KORA primarily aims at minimizing the CU pool energy and the total number of DU relocations in the CU pool. In KORA, each DU-CU pair is running on only split Option 8 with the assumption that enough fronthaul bandwidth is available to run the split Option 8. However, to maintain such high bandwidth, mobile operators need to provide a dedicated link between each DU-CU pair, which will significantly increase the operational cost. To make a fair comparison with the KORA framework, the *Apt-RAN* model is tuned just for split Option 8 and named as *Apt-RAN-FS8 (Fixed-Split-8)*. All the three models assure QoE by minimizing the total number of affected GBR flows using ws_d factor, shown earlier in Eqn. (3). To compare and contrast KORA, *Apt-RAN-FS8*, and *Apt-RAN*, five sets of experiments are conducted with respect to 1) CU pool energy consumption, 2) the total number of DU relocations, 3) percentage of affected GBR flows due to DU relocations, and 4) the total number of handovers. To ensure better accuracy of results, bar

TABLE III: Simulation parameters

Parameter	Value
Number of DUs	10, 20, 30, 40, and 50
Sampling Interval	6 Minutes
Simulation Duration	24 Hours
Total Generated Samples	240
DU workload Range	Normalized in [0,1]
$[w_1, w_2, w_3, w_4, w_5]$	$[0.10, 0.30, 0.20, 0.25, 0.15]$
Time-varying rate parameter	Gaussian Mixture Model

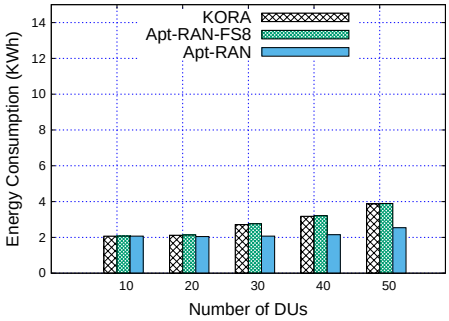

Fig. 11: Total energy consumption (KWh) vs number of DUs during Low_Load (12AM-8AM).

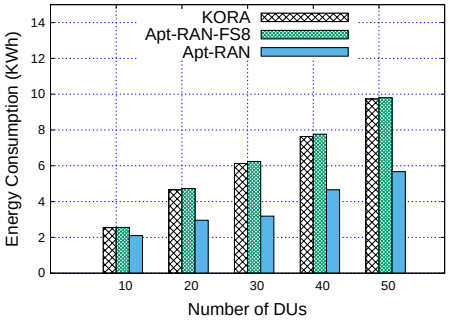

Fig. 12: Total energy consumption (KWh) vs number of DUs during Medium_Load (8AM-12PM & 8PM-12AM).

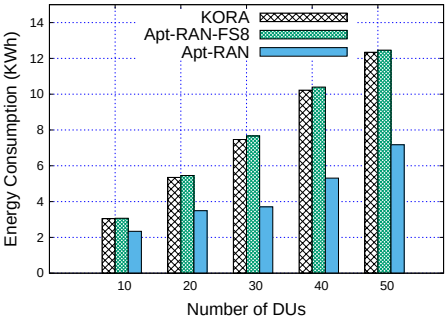

Fig. 13: Total energy consumption (KWh) vs number of DUs during High_Load (12PM-8PM).

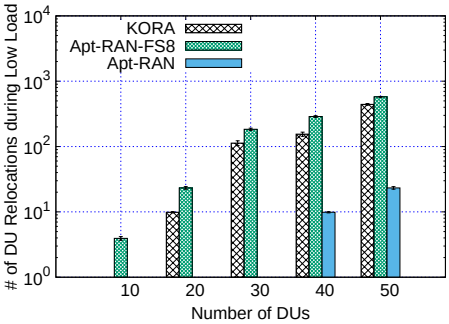

Fig. 14: Total number of DU relocations vs number of DUs during Low_Load (12AM-8AM).

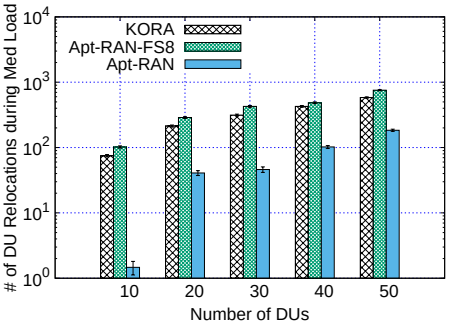

Fig. 15: Total number of DU relocations vs number of DUs during Medium_Load (8AM-12PM & 8PM-12AM).

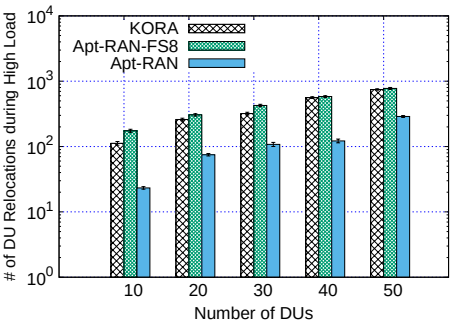

Fig. 16: Total number of DU relocations vs number of DUs during High_Load (12PM-8PM).

plots are plotted with a confidence interval of 99% with 30 seeds.

1) CU Pool Energy Consumption: Figs. 11-13 show the total energy consumption in kWh during Low_load, Medium_Load, and High_Load, respectively. Compared to KORA and *Apt-RAN-FS8*, the proposed *Apt-RAN* models also consider Option 2 and Option 7 due to which effective load at CU is less as some of the key functionalities are shifted to DU side. Therefore, a large number of DUs can be mapped to a single CU, resulting in a less number of active CUs. Moreover, as discussed earlier in Section III-D, CU resides in a General Purpose Processor (GPP) server, which contributes to a significant portion of energy. Therefore, higher number of active CUs results in higher energy consumption. From the obtained results, we observe that *Apt-RAN* consumes nearly 38% and 40% less energy as compared to KORA and *Apt-RAN-FS8*, respectively, for 50 DUs during the High_Load. It is also observed that during the Low_Load, both KORA and *Apt-RAN-FS8* have almost the same energy consumption. However, the energy consumption of *Apt-RAN-FS8* increases slightly ($\sim$1%) more during the Medium_Load and High_Load. Both KORA and *Apt-RAN-FS8* require almost the same number of active CUs in the CU pool. Therefore, the slight increase in energy consumption is due to more number of relocations, resulting in a higher relocation cost.

2) DU Relocations: To ensure better QoE for the users, *Apt-RAN* optimizes the number of DU relocations. Figs. 14-16 show the number of DU relocations during Low_load, Medium_Load, and High_Load, respectively (in log scale). We observe that *Apt-RAN* performs 61% fewer relocations than KORA for 50 DUs during High_Load. As mentioned earlier, *Apt-RAN* maps a large number of DUs in a single CU due to flexible splits. Therefore, only a few DUs are relocated to other CUs. We also observe that *Apt-RAN-FS8* relocates 3.4% higher DUs than KORA as the latter also focuses on mapping neighboring DUs to same CU. Therefore, by relocating a few more DUs, *Apt-RAN-FS8* is able to save more handovers.

3) Affected GBR flows: In the proposed optimization model, the variable ws_d (Eqn. (3)) ensures that the DUs with a minimum number of GBR flows should be considered for relocation to minimize the service disruption. Figs. 17-19 show the percentage of affected GBR flows. We observe that *Apt-RAN* affects 56% fewer GBR flows than KORA, whereas *Apt-RAN-FS8* disrupts 4.8% higher number of GBR flows compared to KORA for 50 DUs during High_Load. In KORA, when CUs underload or overload, a suitable DU is intelligently selected and relocated to another CU to minimize the resource wastage. Whereas in the case of *Apt-RAN-FS8*, a DU is relocated based on the mobility probability P_{ij}. Therefore, *Apt-RAN-FS8* incurs a slightly larger number of affected GBR flows.

4) Handovers: In KORA, mobility probability is not considered in the optimization model, which makes KORA oblivious to handover minimization. To be more specific, KORA does not focus on consolidating neighboring DUs to the same CU to reduce the number of handovers. Whereas in *Apt-RAN*, by exploiting the concept of functional splits, a large number of DUs can be mapped to the same CU. Hence, the probability

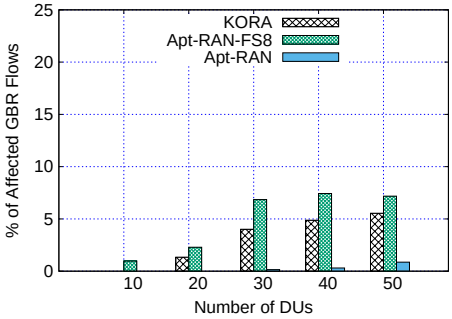

Fig. 17: Percentage of GBR flows affected vs number of DUs during Low_Load (12AM-8AM).

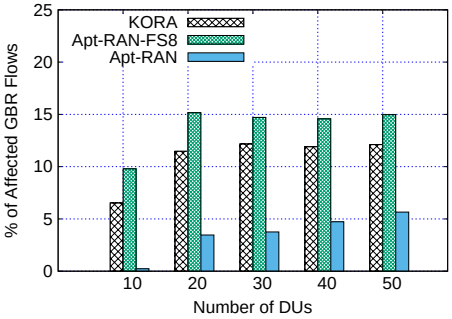

Fig. 18: Percentage of GBR flows affected vs number of DUs during Medium_Load (8AM-12PM & 8PM-12AM).

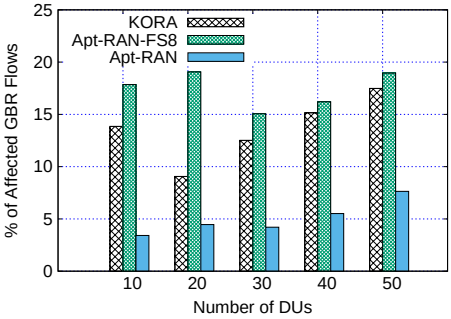

Fig. 19: Percentage of GBR flows affected vs number of DUs during High_Load (12PM-8PM).

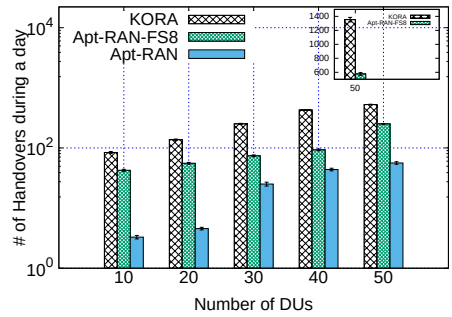

Fig. 20: Total number of handovers vs number of DUs in a day.

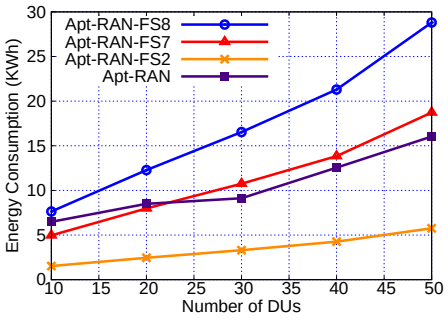

Fig. 21: Total energy consumption (KWh) vs number of DUs for different split options in a day.

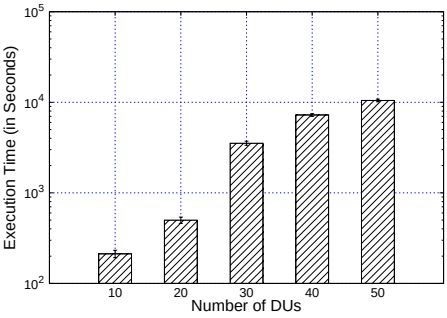

Fig. 22: Execution time vs number of DUs for *Apt-RAN Model*.

of occurring a handover reduces as both source, and target DU may get mapped to the same CU. Fig. 20 depicts that *Apt-RAN* reduces nearly 83% and 92% of the total number of handovers as compared to both *Apt-RAN-FS8* and KORA in a day. *Apt-RAN-FS8* saves 58% of the total number of handovers as compared to KORA for 50 DUs by mapping neighboring DUs to the same CU. Though *Apt-RAN-FS8* and KORA are running on the same split option, we observe that the number of handovers is considerably reduced in *Apt-RAN-FS8* whereas the energy consumption remains comparable in both. To understand the effect of functional splits on energy consumption, we compare the performance of *Apt-RAN* with three static split options, *i.e.*, *Apt-RAN* with only split Option 2 named as *Apt-RAN-FS2*, split Option 7 named as *Apt-RAN-FS7*, and *Apt-RAN-FS8*, in Fig. 21. We observe that *Apt-RAN* consumes 42% and 12% less energy than *Apt-RAN-FS8* and *Apt-RAN-FS7*, respectively. *Apt-RAN-FS2* is more energy efficient than others in terms of energy consumed at CU pool, but it gives fewer centralization advantages to the operators. Whereas *Apt-RAN-FS7* and *Apt-RAN-FS8* provide more centralization benefits to the operators but require a very high fronthaul bandwidth. Therefore, to have higher centralization gains along with slightly relaxed fronthaul constraints and balanced energy consumption, a flexible split based model like *Apt-RAN* should be adopted.

C. Time Complexity Analysis of Apt-RAN Model

Consider a special case of the proposed *Apt-RAN* model where the relocation cost, $cost_{dc}$, associated with relocations of DU is zero. The mobility probability, P_{ij}, is computed based on the fixed geographical locations of the DUs, therefore, P_{ij} becomes constant and can be ignored for the time complexity analysis. We assume that all the CUs are active and have sufficient number of PRBs to serve all the users. Therefore, Eqn. (10) can be rewritten as,

$$\min_{y,z} : \sum_{c=1}^{\mathcal{C}} (z_c \times cost_c) \tag{21}$$

Constraints :

$$\sum_{d=1}^{\mathcal{D}} (y_{dc} \times l_d) \leq (C_c \times z_c), \qquad \forall c \in \mathcal{C} \tag{22}$$

$$\sum_{c=1}^{\mathcal{C}} y_{dc} = 1, \qquad \forall d \in \mathcal{D} \tag{23}$$

Considering the above special case, Eqn (21) represents the classical bin packing problem if CUs are considered as fixed-size bins and DUs are considered as items with different sizes. In the bin packing problem, items with different sizes are packed into a finite number of bins having fixed capacity in such a way that the number of bins used is minimized. The bin packing problem is already proved to be NP-hard problem [28]. Therefore, according to complexity theory, if a special case of the problem is NP-hard, the more generic case ($cost_{dc}$ is not zero) is also NP-hard and thus making it an exponentially solvable problem.

Algorithm 1: Heuristic algorithm for *Apt-RAN*

Input : Previous allocation matrix A_{t-1} and l_d (Based on the split option) for all DUs in $\mathcal{D}$.
Output: Best possible allocation matrix A_t at time epoch t.

1 **while** S_o *is not* $NULL$ **do**
2 $excess \leftarrow \left(\sum_{A_t(d)=c}(l_d)\right) - C_c$;
3 Find eligible DUs for relocation *i.e.,* $l_d > excess$;
4 Compute ζ_d for all eligible DUs;
5 $CandidateDUList \leftarrow$ DU with lowest ζ_d;
6 **end**
7 Sort $CandidateDUList$ in descending order based on l_d;
8 **while** $CandidateDUList$ *is not* $NULL$ **do**
9 Compute P_{ij} $\forall$ CUs $\in S_n$ and $CandidateDU \in CandidateDUList$;
10 Target CU $\beta \leftarrow$ CU with maximum P_{ij};
11 **if** $\exists\beta$ **then**
12 Relocate $CandidateDU$ to β;
13 **else**
14 Instantiate a new CU $\beta^{'}$ as target;
15 Relocate $CandidateDU$ to $\beta^{'}$;
16 **end**
17 **end**
18 **while** S_u *is not* $NULL$ **do**
19 Merge elements of S_u w.r.t. capacity constraint and max(P_{ij});
20 **end**
21 Return the new allocation matrix A_t;

D. Challenges of Apt-RAN Model

To implement *Apt-RAN*, mixed-integer-programming solver called "Gurobi" is employed executing 4 concurrent threads in an Intel Xeon E5620, 3.4 GHz machine running GAMS Version 24 [29] under 64-bit Windows 7. The proposed *Apt-RAN* model always leads to the optimal solution by minimizing the total energy consumption at CU pool, the number of DU relocations, affected GBR flows, and the number of handovers. However, it takes longer execution time, *i.e.*, in order of hours, to converge to a solution. The total execution time taken by the *Apt-RAN* model over 240 iterations is plotted in Fig. 22 for different number of DUs (10 to 50).

In a real-time deployment of NG-RAN, the decision of relocation and consolidation of DUs has to be taken in a very fine granularity of time, *i.e.*, in order of few seconds, to react to real-time tidal traffic variations at DUs. To deal with the high computational cost of the *Apt-RAN* model, a lightweight heuristic algorithm is proposed in the next section.

V. Heuristic/Online Algorithm for Apt-RAN

The motive of heuristic algorithm is to solve the DU-CU mapping problem faster in a greedy manner by sacrificing the actual optimality, accuracy, and precision in order to decrease the execution time. In this section, a greedy heuristic algorithm for *Apt-RAN* is proposed that is computationally feasible in real-time as compared to the optimization model. Proposed heuristic algorithm not only minimizes the total energy consumption by reducing the number of active CUs in CU pool but also minimizes the total number of handovers by consolidating the neighbouring DUs in same CU based on mobility probability P_{ij}. Minimization in number of DU relocations also ensures the QoE by not disrupting the user services frequently. For simplicity, it is assumed that the mobility probability P_{ij} is already computed by the mobile operator based on the geographical location of DUs. An allocation matrix data structure A_t can potentially represent the mapping at time t where each row corresponds to a DU, and each column represents a CU. An entry $A_t(i,j)$ is 1 if DU i is hosted on CU j, else 0. When traffic demand increases or decreases, the allocation matrix $A_t(i,j)$ at time t may not be a good allocation matrix at time $(t+1)$. Let us assume,

- $\mathcal{S}_o$ be the set of overloaded CUs whose utilization exceeded the maximum threshold (say 90% of the total server capacity).
- $\mathcal{S}_u$ be the set of underloaded CUs whose utilization is below the minimum threshold (say 30% of the total server capacity).
- $\mathcal{S}_n$ be the set of non-overloaded CUs whose utilization is below the maximum threshold and above the minimum threshold, *i.e.*, above 30% and below 90%).
- A_t be the allocation matrix of CU at timestamp t.
- P_{ij} be the mobility probability of users moving between DU_i to DU_j.

Algorithm 1 describes the pseudo code of the proposed heuristic algorithm. The algorithm initially identifies a set of overloaded and underloaded CUs in every iteration and performs the following steps thereafter.

I. Identify the most suitable $CandidateDU$ which has to be relocated from overloaded CU.
II. Identify CUs from the set of non-overloaded CUs that can act as target CUs. Instantiate a new CU if no non-overloaded CU can accommodate the $CandidateDU$.
III. Determine the mobility probability P_{ij} of the $CandidateDU$ with all the CUs identified in Step 2.
IV. Copy all the active memory pages of DU to the CU which has the highest mobility probability P_{ij}.

In the initial phase of our algorithm, a $CandidateDU$ is selected (Lines 2-5 in Algorithm 1) based on Minimum Migration Cost (MMC), *i.e.*, relocating a DU $d \in \mathcal{D}$ that has the lowest relocation score (ζ_d). We calculate ζ_d based on weighted score ws_d (Eqn. (3) in Section III-B) and compute load l_d (Eqn. (1) in Section III-C) as follows:

$$\zeta_d = (\alpha \times ws_d) + ((1-\alpha) \times l_d) \mid 0 \leq \alpha \leq 1 \qquad (24)$$

DUs with lower value of ζ_d are preferred for relocation. In Eqn. (24), α plays a decisive role in operator policy planning. The higher value of α guarantees that DUs serving with least number of GBR flows should be prioritized for relocation, hence enhancing QoE for the users. Alternatively, a lower value of α saves the energy cost by relocating a DU which is having less computing load. Depending on whether the operator policy is prioritizing QoE or energy, MMC can be tuned accordingly in the heuristic algorithm.

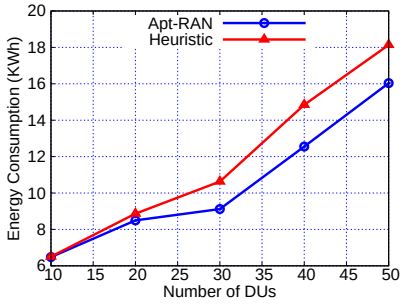

Fig. 23: Total energy consumption vs number of DUs in a day.

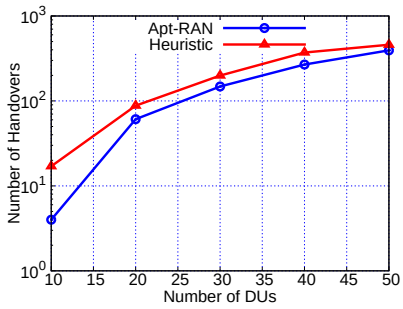

Fig. 24: Total number of handovers vs number of DUs in a day.

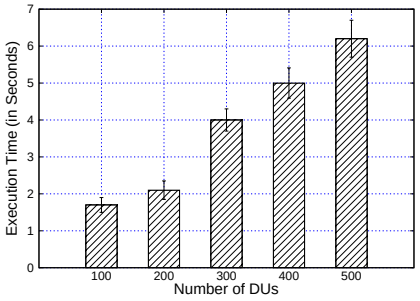

Fig. 25: Run-time performance and scalability of heuristic algorithm.

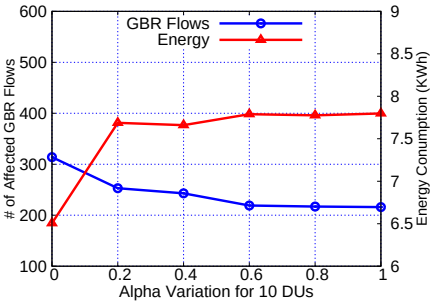

Fig. 26: Effect of α on affected GBR flows and energy consumption in heuristic for 10 DUs.

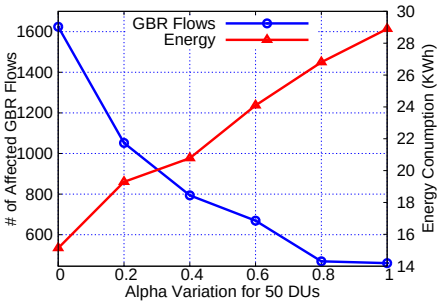

Fig. 27: Effect of α on affected GBR flows and energy consumption in heuristic for 50 DUs.

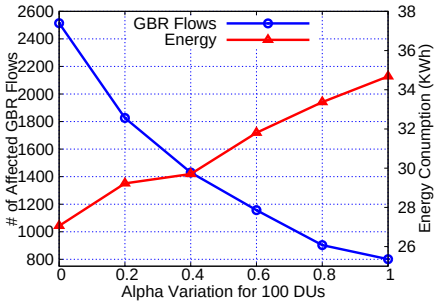

Fig. 28: Effect of α on affected GBR flows and energy consumption in heuristic for 100 DUs.

After selecting Candidate DUs in the first phase, the list of candidate DUs, $CandidateDUList$, is sorted in descending order based on their compute load so that residual utilization at target CU is minimum (Line 7 in Algorithm 1). Identify the target CU from the set of non-overloaded CUs, $\mathcal{S}_n$, that has the maximum mobility probability P_{ij} with the $CandidateDU$ (Lines 9-11 in Algorithm 1). Relocate the $CandidateDU$ to the identified target CU (Line 12 in Algorithm 1). If there is no such existing non-overloaded CUs to accommodate the $candidateDU$, heuristic algorithm instantiates a new CU (Lines 14-15 in Algorithm 1). For each underloaded CU from the set $\mathcal{S}_u$, heuristic algorithm merges one or more underloaded CUs (Line 19 in Algorithm 1) to scale down the number of active CUs used in the CU pool.

A. *Performance Analysis of Heuristic Algorithm*

In this section, we evaluate the performance of the proposed heuristic algorithm with the optimization model *Apt-RAN* and state-of-art *Oracle-type* algorithm proposed in [30].

Compared to *Apt-RAN* model, the heuristic algorithm uses a higher number of active CUs in the CU pool due to its greedy approach. Hence, the heuristic algorithm has 11% higher energy consumption than the *Apt-RAN*, as shown in Fig. 23. On a similar notion, the number of handovers in heuristic is increased by 16% for 50 DUs in a day, as shown in Fig. 24.

Fig. 25 shows the scalability of the proposed heuristic algorithm by varying the number of DUs. In contrast to the execution time of the *Apt-RAN* model shown earlier in Fig. 22, heuristic is light-weight and executes in a few seconds. This feature makes the heuristic algorithm more competent for the realistic deployment of NG-RAN in data centers.

In the proposed heuristic algorithm, by regulating parameter α appropriately (Eqn. (24)), the operator can optimally choose a satisfactory policy for better QoE to the users. The DU with lowest relocation score, ζ_d, is chosen as candidate DU for relocation. When α parameter is 0, the $CandidateDU$ for relocation will be chosen based on minimum DU load irrespective of the number of GBR flows associated with it. This will reduce the relocation energy but at the cost of higher number of affected GBR flows. When α is 1, the relocation score, ζ_d, is calculated based on weighted score, ws_d, which means the DU with minimum number of associated GBR flows will be considered as the $CandidateDU$ for relocation. This will increase the energy consumption as the DU with higher load may become the $CandidateDU$ whereas the number of affected GBR flows will be minimum. Figs. 26 to 28 illustrate the trade-off between energy consumption and the number of affected GBR flows by varying α for different number of DUs (10 to 50). When the number of DUs in the system is less, *i.e.*, 10, beyond a specific value of α, there is no effect on the number of GBR flows and energy consumption. But, when the number of DUs to be served is higher, α should be chosen wisely. For 100 DUs at α = 0, the heuristic algorithm can save 18% energy consumption than that of α = 1, but the number of affected GBR flows is increased by 38%. To evaluate the energy consumption and number of handovers in the simulation setup, suitable value of α is used, *i.e.*, α = 0.12, α = 0.34, and α = 0.42 for 10, 50, and 100 DUs, respectively.

Remark 2. *Based on necessity, the mobile operator should pick α wisely where two contrasting objectives (minimizing energy consumption & minimizing the number of affected GBR flows) are equally considered.*

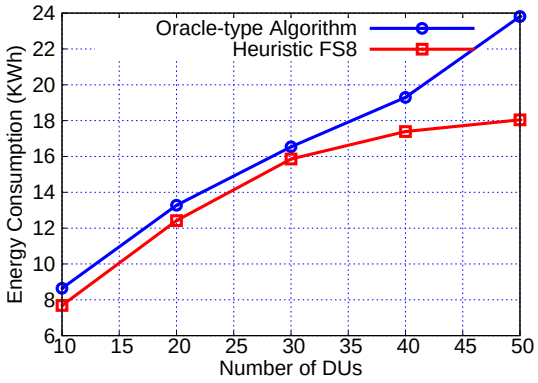

Fig. 29: Comparison between Heuristic FS8 and *Oracle-type* Algorithm for Total Energy Consumption in a day.

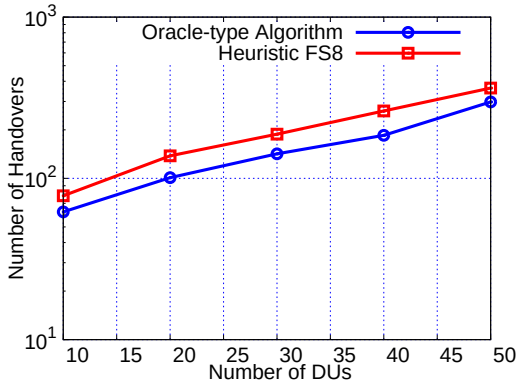

Fig. 30: Comparison between Heuristic FS8 and *Oracle-type* Algorithm for Total Number of Handovers in a day.

Performance of the heuristic algorithm is also compared with state-of-art *Oracle-type* algorithm proposed in [30]. In this work, authors identify the problem of mapping DUs to CU as a classical graph community detection problem [31]. By maximizing the modularity metric of each community, nodes in the graph with similar properties (same neighboring DUs in this case) are clustered in the same community. Authors did not consider any split specific topology in the proposed work. Hence, our proposed heuristic algorithm is also tuned to run only on fixed split Option 8, namely *Heuristic-FS8* for a fair comparison. For 50 DUs, *Heuristic-FS8* consumes 24% less energy than *Oracle-type* algorithm but faces 17% more handovers for 50 DUs, as shown in Figs. 29 and 30. *Oracle-type* algorithm mainly focuses on saving the number of handovers by maximizing the modularity metric for each community, which is an energy oblivious packing. Whereas the heuristic algorithm proposed in this work intelligently maps DUs to CU based on the mobility probability that saves both energy and number of handovers.

B. Time Complexity Analysis of Heuristic Algorithm

Consider $|C|$ as the number of CUs and $|D|$ as the number of DU loads that we want to assign. To identify overloaded and underloaded CUs, the algorithm has to scan the list of CU capacities which takes $O(|C|)$ time. Similarly, to identify the candidate DUs for relocation based on ζ_d parameter, the algorithm takes $O(|D|)$. Since a single CU cannot accommodate all the candidate DUs due to capacity constraint, it has to find the candidate DUs with the maximum load which requires the $CandidateDUList$ to be sorted in descending order. This can be done in $O(|D|log|D|)$ in the worst case. As the locations of DUs are fixed, the distance between all pairs of DUs is computed in advance and stored in a matrix. Therefore, when mapping neighboring DUs to the same CU, the distance can be calculated in constant time, *i.e.*, $O(1)$. Combining all the times, the overall time complexity of proposed heuristic in the worst case is $O(|D|log|D|)$, which can be solved in polynomial time.

VI. Related Work

Both consolidation and relocation mechanisms are extremely vital for efficient resource planning in 5G RAN. Although RAN resource management in a data center is still in its infancy, a few of the existing resource management works are based only on the "bin packing" approach or stochastic modeling to the computational resource consolidation process. In [32], the authors presented a multi-dimensional Markov model to evaluate the statistical multiplexing gain (denotes the extent to which the resources can be shared across multiple parties) of Virtual Base Station (VBS) pools considering the user session level traffic dynamics. Although this model considers the delay-tolerant traffic and expressions for blocking probability, the performance w.r.t. spatio-temporal traffic fluctuations are not considered in the gain calculation. In [33], the authors proposed a bin packing formulation to the BBU to VM packing on an iterative approach to minimize the total number of active BBUs. However, they did not consider the relevance of BBU relocations by tidal traffic variation. The authors in [34] proposed a bin packing solution to consolidate BBUs, which minimizes the energy consumption without factoring the BBU (VM) migration scenario as described before. On similar notion, the authors in [35] highlights a dynamic RRU reassignment algorithm (synonymous with the concept of CU migration) which minimizes the total number of active servers in the cloud platform by considering the spatio-temporal traffic variation, but without factoring migration overhead. Authors in [36] focuses on minimizing total frame delivery time completion time in C-RAN by jointly optimizing user scheduling, transmission rates, and encoded messages of each RRH. The optimization problem is relaxed by an online approach which involves prediction of the completion time resulting in possible associations among users, RRHs, encoded messages, and transmission rate. The authors introduce a rate aware instantly decodable network coding graph (RA-IDNC) and formulate the relaxed version of the optimization problem (Online approach) as maximum weight independent set over the same graph. [37] proposes a hybrid C-RAN architecture where baseband functionalities can be virtualized and split at a different point. Different split options result in two site processing (central and remote site) and introduce a mid-haul in between. The authors propose an optimization framework that jointly minimizes energy and mid-haul bandwidth consumption by developing a constraint programming model,

which finds out a balanced point for optimization of both energy and bandwidth. The model is not scalable when the problem size grows substantially. Authors in [38] present a Virtual Network Embedding (VNE) algorithm, which is formulated as the ILP model that jointly minimizes inter-cell interference among small cells and fronthaul bandwidth utilization by selecting proper functional split dynamically. The model considers requested resources by different Mobile Virtual Network Operators (MVNOs) and allocates resources while optimizing bandwidth and inter-cell interference. Authors in [39], propose a model for virtualized servers where the problem of allocating the optimal number of VMs to the cloud server is addressed. Initially, the number of VMs a server can support is estimated, and then it is optimized using Monte Carlo based evolutionary algorithms to minimize total energy consumption at the cloud server. In this paper, the authors did not study the performance based on functional splits, which is a characteristic of 5G networks. Authors in [40] model the RAN computational resources and evaluate the multiplexing gain for different RAN functional splits. Based on this, authors study the processing savings arising from the consolidation of compute resources. Authors in this model did not consider the variation in energy and multiplexing due to spatio-temporal traffic heterogeneity. Authors in [41] examined the energy consumption in network function virtualization using M/M/c queuing network. Their algorithm saves 40% energy cost while processing 500 flows using MATLAB simulation. In [42], authors have scrutinized the concept of cell differentiation and integration in C-RAN to maximize the network resources without affecting the QoS. Dynamic BBU-RRH mapping has been formulated as a linear integer problem which increases the average throughput by 42% as compared to static mapping. However, the energy consumption of their proposed model is not studied in their work. Authors in [43] also modeled BBU-RRH mapping and UE association problem as an ILP problem. Furthermore, a time-efficient algorithm is proposed, which performs close to the optimal solution. However, the authors did not consider flexible functional split options in their work. The authors in [44] leverage the concept of a virtual base station to form an optimization problem to reduce the number of handovers where future mobility information is known. Authors also proposed a heuristic algorithm when the following mobility information is unknown. Hyebin et al. [45] proposed the Markov-based prediction algorithm to forecast the next location of users in Heterogeneous Cloud Radio Access Network) that lead to a reduction in the number of handovers. The authors in [46] design a novel algorithm named MAPCaching based on mobility aware proactive caching strategy. That significantly outperforms the Greedy and EPC caching strategies.

VII. Conclusions

To conclude, in this work we initially developed a real-time prototype to study the variation in energy consumption due to different functional split options. Based on the key observations from the above prototype, a mathematical model called *Apt-RAN* is developed based on flexible functional splits that minimizes the total energy consumption and number of handovers by efficiently mapping neighboring DUs to same CU. To ensure QoE for the users, *Apt-RAN* also minimizes the total number of relocations and affected GBR flows. The proposed optimization model consumes 38% less energy, saves 92% of total handovers, reduces number of relocations by 61%, and affects 56% less GBR flows as compared to KORA framework proposed in [27]. A lightweight and scalable heuristic algorithm is proposed to reduce the computational complexity of the *Apt-RAN* model. The proposed heuristic algorithm has 11% higher energy consumption and 16% higher number handovers as compared to *Apt-RAN* due to its greedy nature. When compared to *Oracle-type* algorithm that focuses mainly on saving number of handovers, *Heuristic-FS8* saves 24% more energy at the cost of slightly more number of handovers.

VIII. Acknowledgements

This work was supported by the project CCRAN: Energy Efficiency in Converged Cloud Radio Next Generation Access Network, Intel India. We thank Debashisha Mishra for his contribution in this work.

References

[1] Cisco, "White Paper on Visual Networking Index, Global Mobile Data Traffic Forecast Update." https://tinyurl.com/y8kuucvk, 2019.

[2] Y. Yoshida, "Mobile Xhaul Evolution: Enabling Tools For a Flexible 5g Xhaul Network," in *Proc. of IEEE Optical Fiber Communication Conference (OFC)*, December 2018.

[3] CPRI, "New Common Public Radio Interface (eCPRI) for 5G Specification V1.0." http://www.cpri.info/press.html.

[4] C. Mobiles, "White Paper of Next Generation Fronthaul Interface." http://labs.chinamobile.com/cran/wp-content/uploads/2015/09/NGFI-Whitepaper_EN_v1.0_201509291.pdf.

[5] 3GPP, "CU-DU split:Refinement for Annex A (Transport network and RAN internal functional split) ," Tech. Rep. R3-162102, 2016.

[6] H. Gupta, D. Manicone, F. Giannone, K. Kondepu, A. Franklin, P. Castoldi, and L. Valcarenghi, "How Much is Fronthaul Latency Budget Impacted by RAN Virtualisation ?," in *Proc. of IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pp. 315–320, November 2017.

[7] F. Giannone, H. Gupta, K. Kondepu, D. Manicone, A. Franklin, P. Castoldi, and L. Valcarenghi, "Impact of RAN Virtualization on Fronthaul Latency Budget: An Experimental Evaluation," in *Proc. of IEEE Globecom Workshops (GC Workshops)*, pp. 1–5, December 2017.

[8] F. Giannone, K. Kondepu, H. Gupta, F. Civerchia, P. Castoldi, A. Franklin, and L. Valcarenghi, "Impact of Virtualisation Technologies on Virtualised RAN Midhaul Latency Budget: A Quantitative Experimental Evaluation," *IEEE Communications Letters*, vol. 23, pp. 604 – 607, February 2019.

[9] OAI, "OpenAirInterface Software Alliance." http://www.openairinterface.org/.

[10] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network, Study on new radio access technology; radio access architecture and interfaces," Tech. Rep. TR 38.801.

[11] iPERF3, "iPerf3." https://iperf.fr/.

[12] RAPL, "RAPL Tool." https://github.com/kentcz/rapl-tools.

[13] MSR, "Model Specific Register (MSR)." http://man7.org/linux/man-pages/man4/msr.4.html.

[14] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Understanding the Impact of Network Dynamics on Mobile Video User Engagement," in *Proc. of ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, June 2014.

[15] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, and H. Yan, "Modeling Web Quality-of-experience on Cellular Networks," in *Proc. of International Conference on Mobile Computing and Networking (MobiCom)*, September 2014.

[16] H. Zhang, C. Jiang, J. Cheng, and V. C. M. Leung, "Cooperative Interference Mitigation and Handover Management for Heterogeneous Cloud Small Cell Networks," *IEEE Wireless Communications*, vol. 22, pp. 92–99, June 2015.

[17] H. ElSawy *et al.*, "Modeling and Analysis of Cellular Networks Using Stochastic Geometry: A Tutorial," *IEEE Communications Surveys Tutorials*, vol. 19, pp. 167–203, Firstquarter 2017.

[18] 3GPP, "LTE Outdoor Path-loss Model," Tech. Rep. 25.951, 2012.

[19] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith, and G. Iosifidis, "Fluidran: Optimized vRAN/MEC Orchestration," in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, pp. 2366–2374, April 2018.

[20] 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, pp. 146–172, Firstquarter 2019.

[21] ETSI, "5G NR: Medium Access Control (MAC) protocol specification," Tech. Rep. TS 38.321 version 15.3.0.

[22] ETSI, "System Architecture for the 5G System," Tech. Rep. TS 23.501 version 15.3.0.

[23] N. Nikaein, "Processing Radio Access Network Functions in the Cloud: Critical Issues and Modeling," in *Proc. of ACM International Workshop on Mobile Cloud Computing and Services*, pp. 36–43, September 2015.

[24] SPEC, "Standard Performance Evaluation Corporation (SPEC)." http://www.spec.org/power_ssj2008/results/power_ssj2008.html.

[25] H. Liu, H. Jin, C. Xu, and X. Liao, "Performance and Energy Modeling for Live Migration of Virtual Machines," *Cluster Computing*, vol. 16, pp. 249–264, June 2013.

[26] E. Nan, X. Chu, W. Guo, and J. Zhang, "User Data Traffic Analysis for 3G Cellular Networks," in *Proc. of IEEE International Conference on Communications and Networking in China (CHINACOM)*, pp. 468–472, August 2013.

[27] D. Mishra, H. Gupta, B. R. Tamma, and A. A. Franklin, "KORA: A Framework for Dynamic Consolidation and Relocation of Control Units in Virtualized 5G RAN," in *Proc. of IEEE International Conference on Communications (ICC)*, pp. 1–7, May 2018.

[28] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. W. H. Freeman, 1979.

[29] GAMS, "GAMS Solver." https://www.gams.com/optimization-solvers/.

[30] D. Naboulsi, A. Mermouri, R. Stanica, H. Rivano, and M. Fiore, "On User Mobility in Dynamic Cloud Radio Access Networks," in *Proc. of IEEE International Conference on Computer Communications (INFOCOM)*, pp. 1–9, April 2018.

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, October 2008.

[32] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "Statistical Multiplexing Gain Analysis of Heterogeneous Virtual Base Station Pools in Cloud Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 5681–5694, May 2016.

[33] N. Yu, Z. Song, H. Du, H. Huang, and X. Jia, "Multi-Resource Allocation in Cloud Radio Access Networks," in *Proc. of IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2017.

[34] K. Wang, W. Zhou, and S. Mao, "On Joint BBU/RRH Resource Allocation in Heterogeneous Cloud-RANs," *IEEE Internet of Things Journal*, vol. 4, pp. 749–759, June 2017.

[35] D. Mishra, P. Amogh, A. Ramamurthy, A. A. Franklin, and B. R. Tamma, "Load-aware Dynamic RRH Assignment in Cloud Radio Access Networks," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, April 2016.

[36] M. Saif, A. Douik, and S. Sorour, "Rate Aware Network Codes for Cloud Radio Access Networks," *IEEE Transactions on Mobile Computing*, vol. 18, pp. 1898–1910, September 2018.

[37] 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*, vol. 2, pp. 545–555, June 2018.

[38] D. Harutyunyan and R. Riggio, "Flex5G: Flexible Functional Split in 5G Networks," *IEEE Transactions on Network and Service Management*, vol. 15, pp. 961–975, September 2018.

[39] R. Alhumaima, R. Khalf Ahmed, and H. Al-Raweshidy, "Maximising the Energy Efficiency of Virtualised C-RAN Via Optimising the Number of Virtual Machines," *IEEE Transactions on Green Communications and Networking*, vol. 2, pp. 992–1001, August 2018.

[40] M. Shehata, A. Elbanna, F. Musumeci, and M. Tornatore, "Multiplexing Gain and Processing Savings of 5G Radio-Access-Network Functional Splits," *IEEE Transactions on Green Communications and Networking*, vol. 2, pp. 982–991, September 2018.

[41] B. Kar, E. Hsiao-Kuang Wu, and Y.-D. Lin, "Energy Cost Optimization in Dynamic Placement of Virtualized Network Function Chains," *IEEE Transactions on Network and Service Management*, vol. 15, pp. 372–386, December 2017.

[42] M. Khan, Z. H. Fakhri, and H. Al-raweshidy, "Semi-Static Cell Differentiation And Integration With Dynamic BBU-RRH Mapping In Cloud Radio Access Network," *IEEE Transactions on Network and Service Management*, vol. 15, pp. 289–303, November 2017.

[43] J. Yao and N. Ansari, "QoS-aware Joint BBU-RRH Mapping and User Association in Cloud-RANs," *IEEE Transactions on Green Communications and Networking*, vol. 2, pp. 881–889, May 2018.

[44] X. Wang, L. Wang, C. Cavdar, M. Tornatore, G. B. Figueiredo, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Handover Reduction in Virtualized Cloud Radio Access Networks Using TWDM-PON Fronthaul," *Journal of Optical Communications and Networking*, vol. 8, pp. 124–134, December 2016.

[45] H. Park and Y. Lim, "A Markov-Based Prediction Algorithm for User Mobility at Heterogeneous Cloud Radio Access Network," in *Proc. of IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–5, February 2019.

[46] J. Dai and D. Liu, "MAPCaching: A Novel Mobility Aware Proactive Caching over C-RAN," in *Proc. of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, October 2017.

Himank Gupta received his Bachelor of Technology (B.Tech) degree in Computer Science and Engineering from G.B. Pant University of Agriculture and Technology, India, in 2014 and the Master of Technology (M.Tech) degree in Computer Science and Engineering from Indian Institute of Technology Hyderabad (IITH), India, in 2019. He is currently working at MaaS R&D Division in Denso Corporation, Japan. He is a co-recipient of the best Academic Demo Award at COMSNETS 2018. His research interests include 5G, Converged Cloud Radio Access Network, AI in Mobile Networks, Mobility management and Software-Defined Networking.

Mehul Sharma received his Bachelor of Technology (B.Tech) degree in the Department of Computer Science and Engineering from Delhi University (DU), India, in 2017. He is currently pursuing the Master of Technology (M.Tech) degree in Computer Science and Engineering department at Indian Institute of Technology Hyderabad (IITH), India. His main research interests are in Cloud Radio Access Networks (C-RAN), Software Defined Networking, Machine Learning in 5G and beyond, and Vehicular Networks.

Antony Franklin A. received his B.E. degree in Electronics and Communication Engineering from Madurai Kamaraj University, India, in 2000, M.E. degree in Computer Science and Engineering from Anna University, India, in 2002, and Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology Madras, India, in 2010. He is currently working as an Associate Professor in the department of Computer Science and Engineering at Indian Institute of Technology Hyderabad (IITH), India. Before joining IITH, he worked as a Senior Engineer at DMC R&D Center, Samsung Electronics, South Korea between 2012 and 2015 and as a Research Engineer in Electronics and Telecommunications Research Institute (ETRI), South Korea between 2010 and 2012. His current research is on the development of next generation mobile network architectures and protocols which includes Cloud Radio Access Networks (C-RAN), Mobile Edge Computing (MEC), Multi-Radio Aggregation, Internet of Things (IoT), and SDN/NFV. He has published over 50 articles in refereed international journals and conferences. He is a senior member of the IEEE and a member of the ACM. He has received Best Academic Demo Award at COMSNETS 2018 and 2nd Best Paper Award at IEEE ANTS 2017. He has served as TPC co-chair for National Conference on Communications (NCC) 2018, COMSNETS (Posters) 2019, and ADCOM 2019 conferences.

Bheemarjuna Reddy Tamma is an Associate Professor in the Dept. of Computer Science and Engineering at IIT Hyderabad. He obtained his Ph.D. degree from IIT Madras, India in 2007 and then worked as a post-doctoral fellow at the University of California San Diego (UCSD) division of California Institute for Telecommunications and Information Technology (CALIT2) prior to taking up faculty position at IIT Hyderabad, India in 2010. His research interests are in the areas of Converged Cloud Radio Access Networks, SDN/NFV for 5G, Network Security and Green ICT. He has published over 100 articles in refereed international journals and conferences. Dr. Reddy is a recipient of Visvesvaraya Young Faculty Research Fellowship at IIT Hyderabad and iNautix Research Fellowship for his Ph.D. tenure at IIT Madras. He is a co-recipient of Top Cited Article Award from Elsevier publishers, Best Academic Demo Award at COMSNETS 2018, Best Poster Award at ICACCI 2018, 2nd Best Paper Award at IEEE ANTS 2017, and Best Paper award at ICACCI 2015 conferences. He is a senior member of IEEE and a member of ACM and served as a General co-chair for National Conference on Communications (NCC) 2018, TCP co-chair for IEEE ANTS 2015, a TCP vice chair for IEEE ANTS 2014 and a Ph.D. student forum co-chair for IEEE ANTS 2013 conferences.