






Two-Level Allocation for H-CRAN Architecture Based in Offloading

Mariane de Paula da Silva Gonçalves , Matheus Barros Leto , Rafael Fogarolli Vieira ,
Fabrício José Brito Barros , Diego Lisboa Cardoso 
University Federal of Para
*goncalvesmariane@itec.ufpa.br, matheusleto08@gmail.com, fogarolirafael@gmail.com,
fbarros @ufpa.br, diego@ufpa.br*

Abstract—The accelerated data and apps growth represents significant challenges to the next generation of mobile networks. Amongst them, it is highlighted the necessity for a co-existence of new and old patterns during the transition of architectures. Thus, this paper has investigated solutions for offloading into a hybrid architecture, also known as H-CRAN (Heterogeneous Cloud Radio Access Network Architecture), that centralizes processing and searches a better use of the network resources. The strategy of optimization was analyzed through the evolutive algorithm PSO (Particle Swarm Optimization), in order to find a suboptimal solution to the TLA (Two-Level Allocation) in the H-CRAN architecture and another one based on FIFO (First In, First Out), for benchmarking purposes. SNR (Signal-to-noise ratio) average, Maximum Bit Rate, the number of users with or without connections and number of connections in RRHs and macro were used as performance measurements. Through the results, it was noticed an improvement of approximately 60% in the Maximum Bit Rate when compared to the traditional approach, enabling better service to the users.

Index Terms— Mobile Networks, 5G, H-CRAN, QoS, Offloading, TLA.

I. INTRODUCTION

The increase of the volume of data and apps services joined to the accelerated growth of the demands for wireless access, represent challenges to the next generation of mobile networks (5G). According to Cisco [1], the global month traffic of mobile data until 2021 will be 49 exabytes and shall furnish about 1,000 times more capacity, allied to an economy of up to 90% of energy consumption. Hence, the mobile phone companies are investing more and more in new technologies in order to enhance the capability of the network, mitigate interference and improve the QoS (quality of the service) for the final users.

To improve network data, capacity and coverage rates, the common practice is the densification of the network with small cells or base stations, to enable higher reuse of the band length. Heterogeneous networks (HetNets) were implanted aiming to increase the networks' capacity in dense and high traffic demands areas, filling the blanks existing in the area of coverage and discharging the traffic of

data of the macro stations (MBSs) [2]. However, a dense network brings more Capital Expenditures (CAPEX) and Operational Expenditures (OPEX), due to the implantation and additional maintenances in the network. Also, they cause a rise in Inter-Cell Interference and the coordination of the signal sending [3].

In the last years, several authors have proposed the use of CRAN architecture (Cloud Radio Access Network), which is centered and proposes to be scalable and flexible [3], [4]. Thus, it is aimed to mitigate the problems of non-centered architectures, such as inefficiency in resource usage, mainly in scenarios of high traffic loads. In this new architecture, that moves the processing of the primary signal, function performed in the BBU (Baseband Units) of the eNodesBs, to a centered location, named BBU Pool., which makes it possible for several RRHs to be associated with a single BBU Pool, and the Remote Radio Heads (RRH), with fewer functions, that maintains the radio access in the cell sites, enabling dense implantation at a minimum cost [5].

However, the C-RAN demands many challenges, amongst which it is highlighted the rigorous control of the latency, jitter and the high cost of the infrastructure mainly on what concerns the implantation of the fronthaul, which, due to its limitation of latency, needs to be implanted with high capacity technologies.

This process of migration is not traditionally performed in a subtle way. When a network needs to be updated, almost all the network equipment needs to be substituted and, therefore, there is the necessity of a transition phase in order to satisfy the coexistence of new and old patterns. Then, the H-CRAN comes up, a network architecture of the radio in Heterogeneous Cloud (H-CRAN), where both HetNets (non-centralized) and CRAN (centralized) are combined to permit the implantation of dense and heterogeneous networks, aiming to centralize the workload processing and perform the functions of processing of the cooperative signal in wide scale and the network functions, and then, the development of the SE (Spectral Efficiency) and EE (Energy Efficiency) is substantially improved [6].

As a result, H-CRAN was recognized as a promising network architecture to 5G networks and has gained much attention by the industry and the academics. In spite of all that, there is a series of challenges to be faced inside H-CRAN, including optimized allocation on the restricted frontier, energy consumption, the development of prior versions compatible patterns, high call blocking events and quality of services (QoS) and so on [7].

In [4] it was proposed a load balancing in a CRAN architecture, with dynamic remapping capability, to configure the RRHs in sectors of the BBUs, in a time-variant environment. Key Performance Indicators (KPIs) were used as performance metrics for network analysis. In this work, these performance indicators were used in an H-CRAN architecture, with a better distribution of resources and physical layer analysis, with respect to bandwidth, radius coverage, maximum bit rate,

connected users and disconnected from the network. This, in turn, was not explored in [4], leaving an open gap in the literature.

Offloading on mobile networks is widely used in order to unburden critical points in the network avoiding congestion, increasing coverage and keeping users connected. In this article, an efficient offloading technique is proposed, together with the network load balance and the dynamic mapping between BBU-RRH. A multiobjective solution is proposed for the distribution of network resources, taking into account the load balancing in BBUs, physical layer analysis, with respect to bandwidth, radius coverage, maximum bit rate, connected users and disconnected from the network. Effective management of these users using offloading will free up more network resources while maintaining high levels of transmission.

A. Contributions

In short, the main objective is the development of the architecture of offloading in H-CRAN mobile networks, considering BBU and radio aspects, allied to load balancing of the network, seeking load maximization and to achieve high cooperative gains. In particular, the main contributions of our work can be summarized as follows:

- Efficient offloading technique, which considers resources of radio and BBU in conjunction, for Optimized Allocation of users with better use of resources in H-CRAN Architecture;
- RRH-BBU mapping which considers resources of radio and BBU in conjunction for Optimized Allocation ;
- Study of offloading algorithms applied to hybrid architectures;
- Algorithm of user allocation in two-level H-CRAN architectures;
- Multiobjective Resource Distribution;
- Detailed study of the physical layer of the SUI Propagation Model;
- Implementation of tests to validate the proposal in case studies;

Besides this introductory section, this paper is divided into such a way: Section II, which presents the Related Works; Section III, Model of the System, Section IV presents the Network Parameters, Section V show Analysis Of The Network Through KPIs, in section VI, shows the proposed TLA Algorithm, and the section VII the Results of the paper and ending up at Section VII, with the Conclusion and Future Works.

II. RELATED WORKS

As previously stated, H-CRAN architecture has been one of the promises for 5G mobile networks. Hence, [8], [9] it investigates the benefits and challenges of the workload balance in centralized networks (CRAN), where it is proposed a model of optimized mapping to RRH-BBU, aiming to

maximize the number of connected users and, thus, improving the QoS. The former aims the balancing without worrying about the balance of the sectors, and the latter, besides guaranteeing the KPI, aims to search the load balance amongst the sectors of BBUs, then, optimizing the network balance.

In [10], it is proposed a slicing of the network to H-CRANs multi-tenant, which takes into consideration the resources of the band base, capacities of fronthaul and backhaul, QoS and interference. However, did not explore the efficiency of a balanced allocation in the BBUs sections and the use of offloading in mobile networks, avoiding high traffic congestion in the network and high events of calls blocking.

In [11] the author addresses the problem of wireless backhaul in an H-CRAN network, and an energy allocation is made to multiple cells of different levels, taking into account the energy consumption of different types of cells and wireless backhaul. An iteration algorithm is proposed to determine the maximum number of cells that can be supported in H-CRAN.

The author proposes in [12] a resource allocation algorithm, called H-WDRF (Heterogeneous Weighted Dominant Resource Fairness) to solve the transmission of the energy allocation problem is formulated as a beamforming problem, whose objective function is to minimize the total transmission power with individual computing, resource capacity, SINR, and maximum transmission power constraints.

In [13] the author emphasizes the self-optimization technique for C-RAN architectures aiming at a network's performance improvements and focusing on the modeling of a multi-objective optimization problem. The Cell Differentiation and Integration (CDI) algorithm are used to select the appropriate BBUs and RRHs for activation/deactivation after a fixed CDI cycle, and the dynamic part performs the proper BBU mapping for RRH for network balancing, aiming at the maximum QoS with the minimum possible handovers.

Based on the works cited above, it is noted that there is a guideline for designing the H-CRAN architecture, but the usage of network resources in order to perform tasks is still an unsolved problem. In the survey performed by the authors, it was not found any work that investigates solutions of offloading aiming the network load balance and improvement of the QoS, in H-CRAN architecture. The main contribution of this paper is the proposition of a new schemata for offloading to H-CRAN architecture, focusing the balance of the loads of the BBUs, better use of the network resources, maximization of the capacity, improvement of the coverage and maximization of the QoS of the final users, thus, promoting heterogeneity and, consequently, improving the efficiency of the spectre.

III. SYSTEM MODEL

A. Architecture H-CRAN

The H-CRAN architecture is characterized by functions of SDN (Software Defined Network) and computation in the cloud that may enhance considerably the flexibility of the network architecture, to improve SE system and reduce, significantly, the energy consumption and the operational expenditures. Besides, the QoS to the user may be quite improved due to the reduced distance between RRHs and UEs, as well as the flexibility to allow the association of the UE with different levels [10]. In this paper, it is considered a two-level of the cell network, which consists in a macro/EnodeB and several RRHs, with low energy consumption, that cooperates to one another in centered BBU pool, aiming to reach high cooperative gains. The 5G system, based on this architecture, presents 3 plans: User Plan (U), Control Plan (C) and the Management Plan (M), each one processing different areas of operations and functions [13], as it may be seen in figure 1.

In general, the (U) Plan transfers the real traffic of the user and executes the processing of the related traffic in order to satisfy the various requirements of QoS. The (C) Plan transports the signalization of control and is in charge of the resource allocation and traffic processing, aiming to improve SE and EE. As for Plan (M), it has the function of executing the administration and operation and is responsible to add, exclude, update and modify the interactions with the other plans [13].

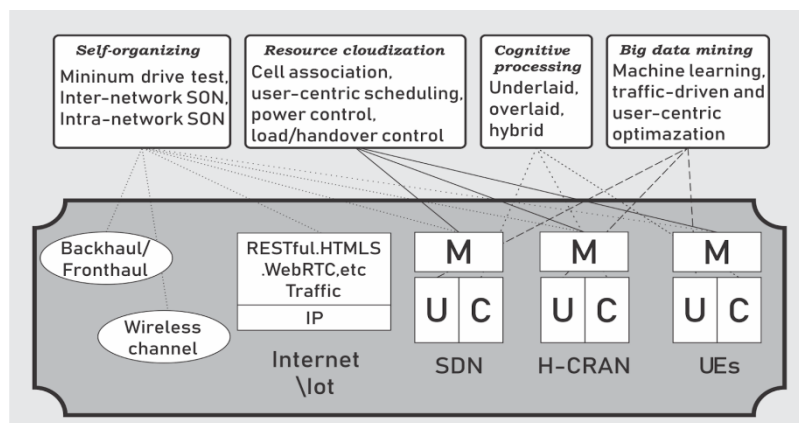


Fig. 1. 5G Components and Features

Fig. 2 shows the layers of a 5G proposed architecture, in which the network layer presents the necessary to the signalization and monitoring of the network. As for the layer of infrastructure, it aims to centralize the load processing, aiming cooperative gains. And the layer of offload that aims to open up critical points of the network, discharging to Macro EnodeB users that were blocked before, reaching the maximization of the QoS, better use of network resources and, consequently, the maximization of the network.

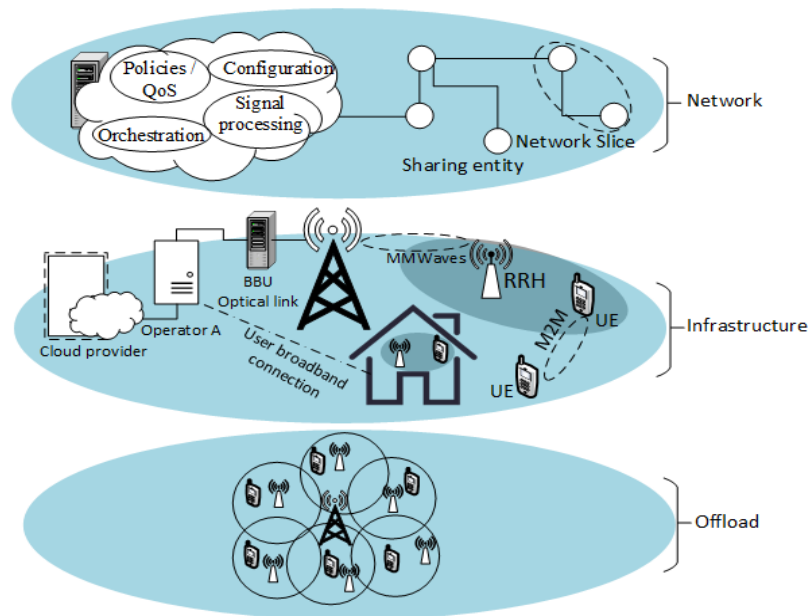


Fig.2. Layers of a 5G network

Considering that the users are uniformly distributed, it is investigated solutions of offloading to the load balancing, in an H-CRAN architecture. The levels of interference and requirements of the QoS are collected aiming to reach the best allocation of the network resources.

In the modeling, it was used 6 RRHs and a macro ENodeB, UEs uniformly distributed in the scenario, according to what is presented in Figure 3. To grant the small cells, it is used 2 BBU Pool, with 2 sectors each, each one with maximum capacity defined as $(HC_{max} = 30)$ [4]. As it was said before, the aim of this proposal is to maximize the coverage and the QoS of the UEs and, to do so, it was developed an evolutive algorithm PSO (Particle Swarm Optimization), in order to find an optimal solution to the allocation of two levels in the H-CRAN architecture.

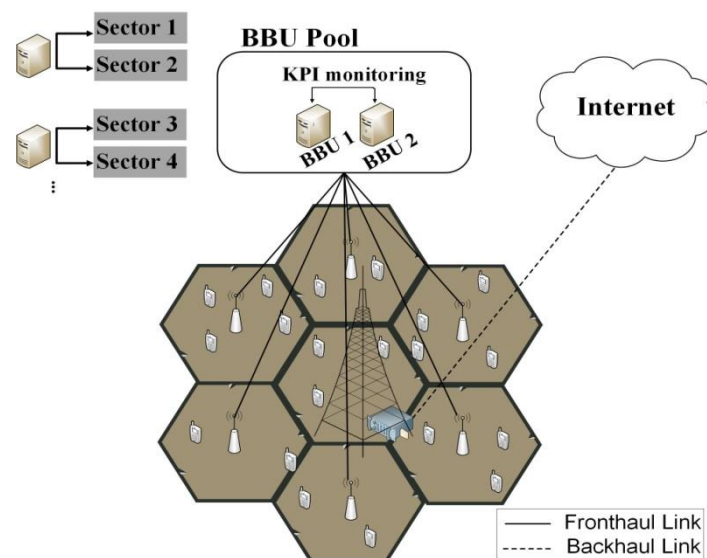


Fig. 3. H-CRAN Architecture Proposed.

IV. NETWORKS PARAMETERS

It was used a SUI propagation model, which is widely used in literature for estimating path loss in cell phone networks operating over 2GHz. Where $PL_{SUI}(d)$ in dB is found by [14] [15] [16]:

$$PL_{SUI}(d) = PL(d_0) + 10n \log_{10} \left(\frac{d}{d_0} \right) + X_{fc} + X_{RX} + X_{\sigma} \quad (1)$$

Where:

$$PL(d_0) = 10 \cdot \log_{10} \left(\frac{4\pi d_0}{\lambda} \right) \quad (1a)$$

$$n = a - b \cdot h_{TX} + \frac{c}{h_{TX}} \quad (1b)$$

$$X_{fc} = 6 \cdot \log_{10} \left(\frac{f_{MHZ}}{2000} \right), f_c > 2GHz \quad (1c)$$

$$X_{RX} = -10.8 \cdot \log_{10} \left(\frac{h_{RX}}{2} \right) \quad (1d)$$

- λ denotes the wavelength of the carrier in meters;
- $PL(d_0)$ represents the loss of way in the free space, in dB at an initial distance of $d_0 = 100$ meters;
- X_{σ} , is the vanishing caused by the shadowing areas, being able to be between $8.2 < \sigma < 10.6$ dB;
- The parameters a, b and c are constants, used to model the types of terrain, It is used A, which is the proper model to mountainous and dense vegetation, with values ($a = 4.6 / b = 0.0075 / c = 12.6$);
- f_{MHZ} is the carrier frequency (fc) in MHz;
- h_{TX} is the height of the base station, being able to be between 10 to 80 meters;
- n is the exponent of losses;

The size of the cells that contain the equipment of H-CRAN architecture was determined considering the maximum loss of propagation obtained through the help of cross-linking of Downlink (DL) and Uplink (UL). To each cross-link, it was obtained the loss of propagation represented by the diminishing of the EIRP (Effective Isotropic Radiated Power) and by the RSL (Received Signal Level), considering the maximum power of the transmitter element and the minimal power of the receptor element. The limiar power of the equipment was considered according to [17]. The maximum propagation loss, considering the cross-links, follows the equation below:

$$PL_{max} = \min[PL_{max(DL)}; PL_{max(UL)}] \quad (2)$$

The cell radius corresponds to the distance of the SUI model when the model is equalized to the maximum loss of propagation given by (2). The equation that represents the radius is given as follows:

$$R = (d0 \cdot 10) \frac{PL_{max} \cdot PL(d0)}{10n} \quad (3)$$

The UEs are only connected if they are inside the ray of coverage of the RRH or macro ENodeB. The power received inside each cell is given by formula (3), where it is performed the dBm conversion to Watts:

$$Pot_w = \frac{10^{\frac{Pot_i - L_{SUI}}{10}}}{1000} \quad (4)$$

Where Pot_w is given in Watts, Pot_i is the initial power or irradiated in dBm and L_{SUI} is the value of the signal vanishing. Up from the value of the power, it is possible to calculate the value of the SNR and, then, the maximum flow of the UE connected, as described below:

$$SNR = \frac{Pot_w}{I} \quad (5)$$

$$V_{max} = B * \log_2(1 + SNR) \quad (6)$$

So that Pot_w is the potency, while I is the value of the white noise at formula (5). To obtain the data rate of the user is used the theorem of Shannon-Hartley [SHANNON, 1949] is used. This theorem specifies the maximum rate, in bits per second, that can be obtained by channel, given a bandwidth B, which is the amount of bandwidth available to users [18].

To connect to the network, each user must achieve a minimum QoS, first established by the flow rate of 200 kbps. Given the bit rate value that the user has achieved in a given channel equation (6), to achieve the established QoS, this user will need N channels, established in equations (7) and (8).

$$X = \frac{QoS}{bitRate * 1024} \quad (7)$$

$$N = \min\{N \in \mathbb{Z} | N \geq X\} \quad (8)$$

Equation (7) shows how many channels the user will need to reach the QoS minimum of 200 kbps [21] since that user has a certain bitRate (multiplying by 1024 to get in the same unit of kilobits). Since the value X can be any real value and it is not possible to provide a non-integer value of channels, equation (8) gives the number of N channels required by users which will be the largest integer greater than or equal to X.

This relation is made for all the users and the antennas, in which they are within range. The allocation algorithm prioritizes the RRHs first since this means a better distribution of the users' attention. After all related, the allocation starts in fact for the best SNR of all relations, so it is guaranteed that the users that occupy the minimum, and the greatest possible number of users is allocated.

When there are no more users to be allocated to the RRHs, either because there are no such channels for the user to achieve the minimum of their connection, because they have reached the limit of their

connected users or all the users can connect to them, the allocation will focus on meeting the users who are still disconnected through the macro.

If a user cannot get enough channels from an antenna to achieve their minimum QoS, this user will be disconnected. After the allocation of antennae users, the algorithm starts to worry about the sectorization in the BBUs, returning to an allocation of RRHs-BBUs. The parameters used in the model are shown in the table below [19], [20]:

TABLE I. SIMULATION PARAMETERS

	Macro EnodeB	RRH
Frequency	3.5 GHZ	2.5 GHZ
Power Transmission	43 dbm	33 dbm
Radius Coverage	1000m	150m
Bandwidth	20 MHZ	

V. ANALYSIS OF THE NETWORK THROUGH KPIS

As the objective function of the implemented PSO, it was used a combination of KPIS proposed in [4], which were adjusted to the scenario proposed and that shall be described below. Weights were assigned to obtain the maximum QoS, using the PSO.

$$\begin{aligned}
 Max\ QoS = & w1 \left[\sum_s \max \left[\left(\left(\sum_i U_i R_{is}^{t+1} \right) - HC_s \right), 0 \right] \right]^{-1} + \\
 & w2 \left[\sum_i \sum_{j \neq i} \frac{U_i}{D_{ij}} \left(1 - \left(\sum_n \sum_{s \in SOS_n} (R_{is}^{t+1} \cdot R_{js}^{t+1}) \right) \right) \right]^{-1} + w3 \\
 & \left[\sum_i \sum_{j \neq i} \frac{U_i}{D_{ij}} \left[\sum_n \sum_{s \in SOS_n} () - \sum_s (R_{is}^{t+1} \cdot R_{js}^{t+1}) \right] \right]^{-1} + w4 \left[\sum_s \sum_i (R_{is}^{t+1} \cdot R_{js}^{t+1}) U_i \right]^{-1}
 \end{aligned} \tag{9}$$

A. Key Performance Indicator of Blocked Users (KPI_{BU})

As the objective function of the implemented PSO, it was used a combination of KPIS proposed in [4], which were adjusted to the scenario proposed and that shall be described below.

The KPI of blocked calls, which aims to evaluate the capacity of the BBUs and its sectors (HCs). When such a value is exceeded, the users are considered blocked. The number of blocked users BU in the net in time t + 1, is given by [4]:

$$BU = \sum_s \max \left[\left(\sum_i U_i R_{is}^{t+1} \right) - HC_s, 0 \right] \tag{10}$$

Where $i = 1, 2, \dots, M$ and $s = 1, 2, \dots, S$. Later, the KPI to blocked users (KPI_{BU}) may be presented as:

$$\begin{cases} KPI_{BU} = \{1 \text{ if } BU = 0\} \\ \frac{1}{BU} \text{ otherwise} \end{cases}$$

According to [4], we may evaluate the index of performance of the network through 3 metrics: Inter-BBU Handovers, Intra-BBU Handovers and Forced Handovers. The formulations used by the author are shown below:

B. Inter – BBU- Necessary transfers of a BBU to another.

$$\begin{aligned} Inter - BBU_{HO} &= \sum_i \sum_{j \neq i} \frac{U_i}{D_{ij}} \left(1 - \left(\sum_n \sum_{s \in SOS_n} (R_{is}^{t+1} \cdot R_{js}^{t+1}) \right) \right) \\ KPI_{inter} &= \{1 \text{ if } Inter - BBU_{HO} = 0\} \\ &\quad \{ [Inter - BBU_{HO}]^{-1} \text{ otherwise} \} \end{aligned} \tag{11}$$

C. Intra – BBU: Are performed when there is the transition of users from one sector to the other inside the same BBU, under the condition that the sectors involved in such transition of the users are treated entirely inside the eNodeB.

$$\begin{aligned} Intra - BBU_{HO} &= \sum_i \sum_{j \neq i} \frac{U_i}{D_{ij}} \left[\sum_n \sum_{s \in SOS_n} () - \sum_s (R_{is}^{t+1} \cdot R_{js}^{t+1}) \right] \\ KPI_{intra} &= \{1 \text{ if } Intra - BU_{HO} = 0\} \\ &\quad \{ [Intra - BBU_{HO}]^{-1} \text{ otherwise} \} \end{aligned} \tag{12}$$

D. Forced Handovers: it is used when the HC limit is exceeded.

$$\begin{aligned} f_{HO} &= \sum_s \sum_i (R_{is}^{t+1} \cdot R_{js}^{t+1}) U_i \\ KPI_f &= \{1 \text{ if } f_{HO} = 0\} \\ &\quad [f_{HO}]^{-1} \text{ otherwise} \end{aligned} \tag{13}$$

VI – ALGORITHM TLA (Two-Level Allocation)

In this section, it is shown the offloading decision process in the proposed algorithm and its constraints.

The main objective of TLA proposed is to maximize network capacity by maintaining minimum levels of QoS with the use of offloading, in order to transport users who were previously blocked in the BBU, offload to the macro EnodeB.

The algorithm seeks to maximize the system's QoS factors while seeking load balancing. For the quantification of QoS, the KPIs in section V are taken into account. Through evaluation, there is a decision-making process for the reorganization of the network. The PSO is applied to reorganize and search for the best result, which has better KPIs and better meets QoS.

The PSO begins with the random generation of particles with a linear distribution, where each particle is a possible solution to the problem and represents the set of RRHs, and their respective sectors, in which they are connected. In the sequence, the paper is evaluated by the objective function, in which the quality of the solution returns. Thus, the values of the particles are updated, searching for the best Gbest, after which, they are in a loop until they reach the stop criteria, defined as 1000 iterations, where each iteration updates its particles and calculates its suitability.

When the current solution is better than the previous one, the particle values are actuated (PBest), otherwise, the solution is discarded and the previous solution is maintained. For each iteration, it is verified if the best solution found remains the same, otherwise, it is updated by the best found (GBest). The decision-making of the offloading is explained in figure 4, where the analyzes of KPIs and QoS are predetermined. One of the restriction rules in the allocation is that an HR cannot have more users connected to it than the maximum capacity of a BBU (HC) sector. That is, it cannot be occupying an industry alone and having users blocked. In this case, all the RRHs are selected to choose the user who will be offloaded, where the UE that has the best SNR with the EnodeB macro is allocated to it.

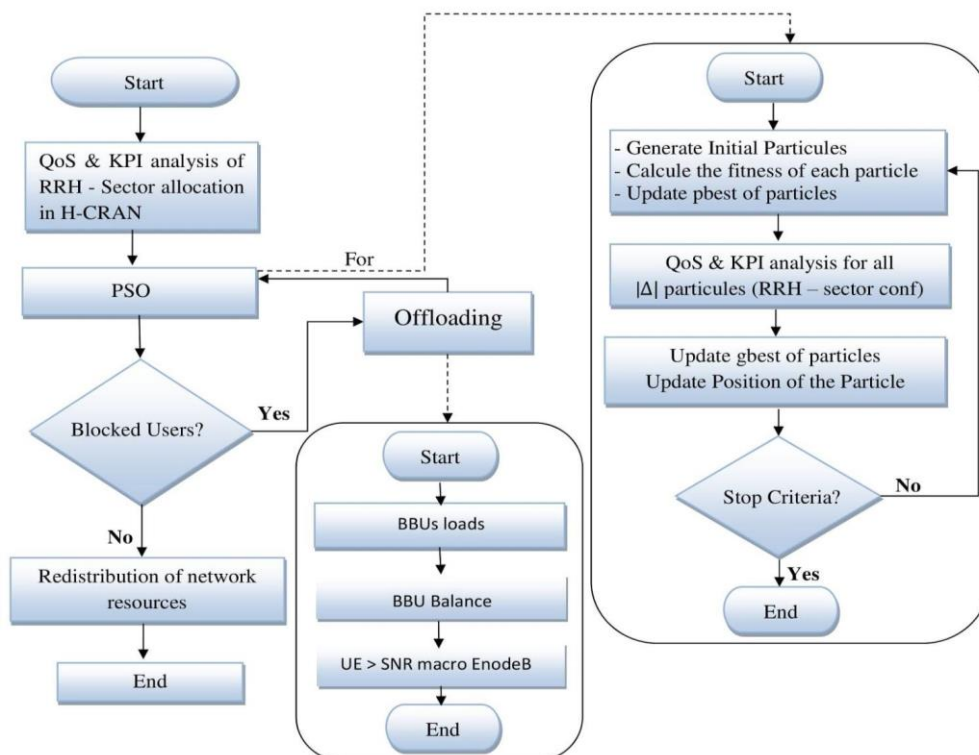


Fig. 4 - Algorithm TLA

However, when the PSO is executed and the network organization still results in blocked users, the system then selects the sectors of the BBUs that had their HC exceeded, choosing, among the antennas present in these sectors, which user has the best SNR with the macro. If it is detected that the user can, with the SNR that he has with the macro, connect to it, attending to his minimum QoS, and then the offloading with it is performed.

When offloading is no longer possible because the macro no longer has how to service the user, the process becomes to choose which user will be disconnected. It is chosen, among the RRH with the largest number of users, which presents worse connection with the antenna that is connected. With this, there is a better use of network resources and maintenance of QoS levels.

A. Allocation

The algorithm starts looking for the best form of allocation, the one in which it obtains more users connected and with better quality. To do this, it generates the users' SNR relations with the antennas (RRHs or macro, in which the user is in the coverage area).

The allocation algorithm prioritizes the connection in the RRHs and, later, for the enodeB macro. After all the users are properly related, the allocation starts with the best SNR of all relations, that is, the user that requires fewer channels, thus, it is guaranteed that the users that are being allocated in the antennas occupy the minimum of possible channels and the number of users can be connected.

When there are no more user allocations in the RRHs, or because of the lack of such channels for the user to achieve the minimum QoS, due to the limit of users connected in the antenna being reached or by all the users in them to be able to connect, the allocation becomes focused to service the users that are still disconnected through the macro, which goes through the same RRH process with the remaining users.

If a user can not get enough channels from an antenna to get their minimum QoS value, this user will be disconnected. After the allocation of users in the antennas, the algorithm starts to treat the sectorization of BBUs.

B. RRH-BBU Balancing B

When the allocation phase is finished, the RRHs with their connected users will have to be allocated in the BBU sectors, obeying the maximum capacity of the sectors (HC). In order for there to be no cases of blocked users, there can initially be no more users connected to the RRHs than the maximum capacity (sum of capacities of all sectors) of the BBU pool.

When the system is in this state, where it is impossible to reach an allocation without getting blocked users, the algorithm tries to perform offloading until the number of users connected in the RRHs becomes equal to the maximum capacity of the BBU pool.

When the scenario is not saturated, the PSO runs to try to map the RRHs in the sectors. Each particle is a vector in which the index represents the antenna, and the value represents the BBU sector.

The results are generated by the objective function according to the KPIs in the formula (9). This setting is tested to see if users are locked. If there is, the algorithm tries to perform offloading, otherwise the configuration is maintained for the resource redistribution phase.

C. Offloading Decision

When users are blocked in the BBU sectors, at least one user cannot get communication, then a user is chosen to be served by the macro, with offloading to free space in the BBU sectors. To select the user in unsaturated scenarios, the RRHs of the sectors of the BBUs that have blocked users are selected. Among these RRHs, those with the highest number of connected users are selected. Finally, of this select group of RRHs, the users with the best SNR with the macro are chosen. However, when starting from a saturated scenario, that is, the number of users in the RRHs is greater than the maximum HC of the BBU pool, all the RRHs are considered, instead of those that are in a sector of the BBU with blocked users, to look for a user who will be offloaded. So, in the end, it is tested if it is possible to complete the offloading of the user, that is if the macro has enough channels to serve it.

VII. RESULTS

It was used the MATLAB® to evaluate the performance of the proposal. As it was said before, as the benchmark, it was used as an approach based on FIFO. In this method, the resources are allocated without any type of priority. The first user to require will be serviced by one available RRH with the best signal-to-noise ratio. As for the TLA proposal, it is prioritized the connection in the RRHs, until the end of the network resources, aiming that the users are serviced with the maximum capacity of the network and maintenance of the QoS levels.

It was generated 13 scenarios containing 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350 users, respectively. All scenarios were simulated 50 times and the results presented refer their average. The confidence interval was 95%.

Fig. 5 (a) and (b) shows the limitations of the equipment during the modeling. The number of disconnected and connected users presents a linear trend due to the addition of users in the scenarios (which is linear). The allocation of the network resources is re-distributed in an efficient way, ensuring a better connection of users in the TLA network.

In Fig. 6, from scenario 6, the TLA loses linearity and FIFO tends to saturation due to network resources, which are underutilized in FIFO. The offloading technique present in TLA, guarantees high levels of QoS, for the best uses of network resources.

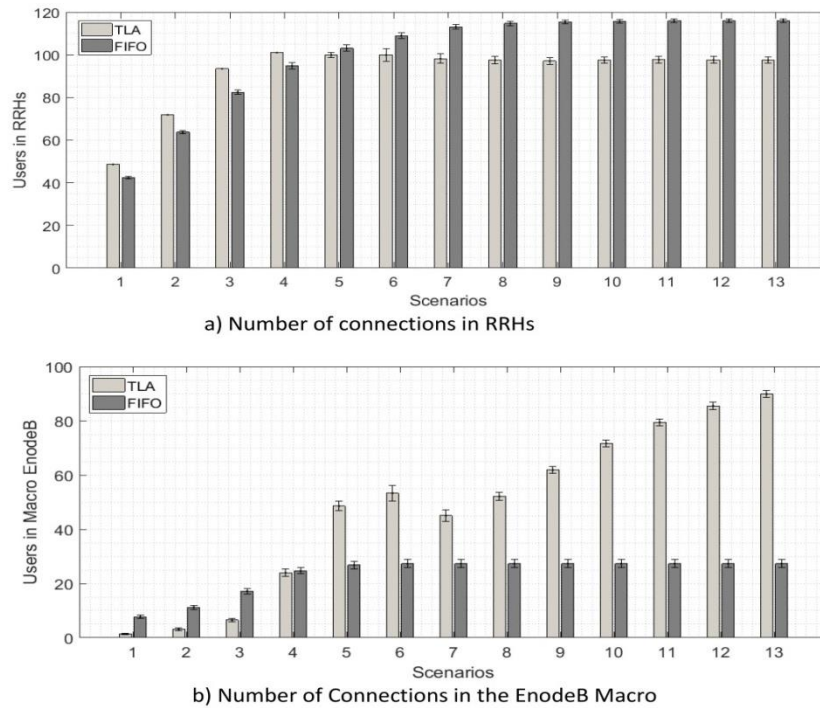


Fig. 5. Number of connections per RAN type

The average of disconnected users is around 30% higher in FIFO, and the average of connected users is 25% higher in TLA, which may be seen in Fig. 6 and Fig. 7 respectively. It happens due to the offloading scheme that opens up the network traffic and discharges to the macro EnodeB the users that had been blocked in the BBU.

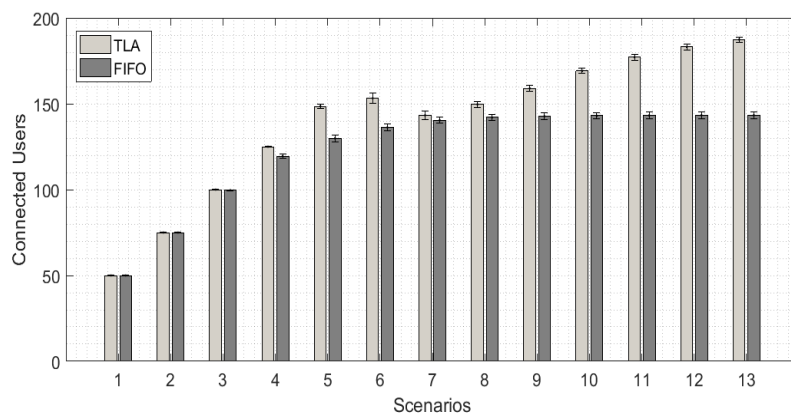


Fig. 6 - Average Connected Users

Fig. 8 presents the maximum data rate, where it is observed the maintenance of the minimum levels of QoS in all scenarios. Where TLA and FIFO have maintained themselves in the average, satisfying

the minimal requirements of QoS. The former has guaranteed a better connection when compared to the latter, where both have maintained themselves constantly, for, in the beginning, the minimal requirements of QoS of the UEs are accomplished and, later, the resources are re-distributed, obtaining a higher number of users along modeling.

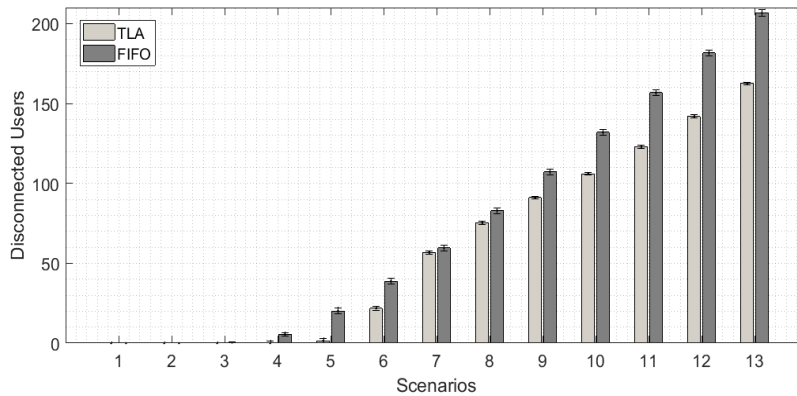


Fig. 7. Average Disconnected Users

Thus, it is a fact that the TLA proposal has obtained a satisfactory performance of about 60% compared to the FIFO, respecting all the established requirements, maximizing its capacity, maintaining the minimal levels of QoS and network coverage.

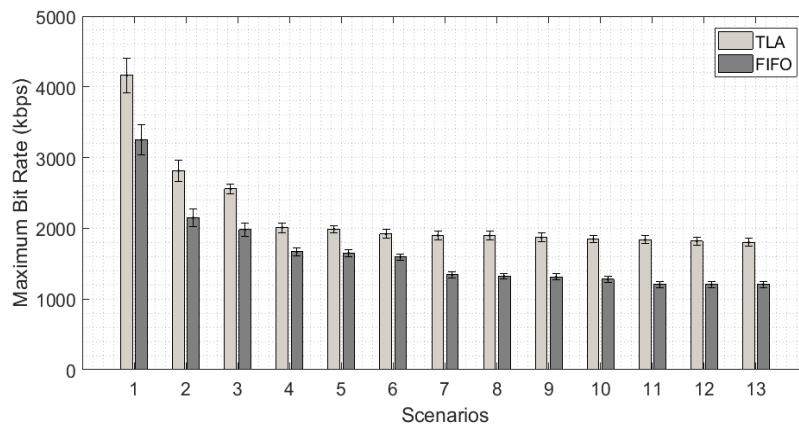


Fig. 8 - Average Maximum Bit Rate (kbps)

The results validate the efficacy of TLA and the impact of the various requirements underlying the resource utilization efficiency in H-CRAN.

VIII. CONCLUSION AND WORKS FUTURES

The raising of the data and apps volume represents significant challenges to the 5G mobile networks. This way, the mobile network operators are investing more and more into new technologies to improve their network capabilities, mitigate interference and improve the QoS of the final users.

In this paper, it was proposed an offloading schema in one H-CRAN architecture, aiming to improve the use of network resources, maintaining its minimal levels of established QoS.

Through the results, it was noticed that the proposal of TLA has obtained a satisfactory performance when compared to FIFO, obtaining improvements of approximately 60% in the Maximum Bits Rate. This in turn provides a better connection to all users of the network, efficiently distributing all the network resources.

As future works, new scenarios will be modeled, with frequency changes, varying the capacity of the BBU, varying the number of antennas and arrangement of them, analyzing other measures of performance.

REFERENCES

- [1] Cisco. 2017. "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update", 20146– 2021.
- [2] X. Duan, X. Wang, "Authentication handover and privacy protection in 5G hetnets using software-defined networking," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 28-35, 2015.
- [3] M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska, P. Monti, "Demonstration of dynamic resource sharing benefits in an optical C-RAN," *Journal of Optical Communications and Networking*, vol. 8, no.8, pp. 621-632, 2016.
- [4] M. Khan, R. S. Alhumaima, H. S. Al-Raweshidy, H. S., "QoS-aware dynamic RRH allocation in a self-optimized cloud radio access network with RRH proximity constraint," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 730-744, 2017.
- [5] J. Wu, Z. Zhang, Y. Hong, Y. Wen, Y., "Cloud radio access network (C-RAN): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35-41, 2015.
- [6] B. Zhang, X. Mao, J. L. Yu, Z. Han, "Resource allocation for 5G heterogeneous cloud radio access networks with D2D communication: a matching and coalition approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 5883-5894, 2018.
- [7] M. A. Marotta, M. Kist, J. A. Wickboldt, L. Z. Granville, J. Rochol, C. B. Both, "Design considerations for software-defined wireless networking in heterogeneous cloud radio access networks," *Journal of Internet Services and Applications*, vol. 8, no.1, pp. 18, 2017.
- [8] M. Khan, R. S. Alhumaima, H. S. Al-Raweshidy, "Quality of service aware dynamic BBU-RRH mapping in cloud radio access network," In 2015 International Conference on Emerging Technologies (ICET), pp. 1-5, 2015.
- [9] E. A. R. da Paixão, R. F. Vieira, W. V. Araújo, D. L. Cardoso, "Optimized load balancing by dynamic BBU-RRH mapping in C-RAN architecture," In 2018 Third International Conference on Fog and Mobile Edge Computing (FMEC), pp. 100-104, 2018.
- [10] Y. L. Lee et al., "Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2146-2161, 2018.
- [11] H. Q. Tran, P. Q. Truong, C. V. Phan, Q. T. Vien, "On the energy efficiency of NOMA for wireless backhaul in multi-tier heterogeneous CRAN," In 2017 International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom), pp. 229-234, 2017.
- [12] M. Khan, Z. H. Fakhri, H. S. Al-Raweshidy, "Semistatic Cell Differentiation and Integration With Dynamic BBU-RRH Mapping in Cloud Radio Access Network," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 289-303, 2018.
- [13] M. Peng, Y. Li, Z. Zhao, C. Wang (2014). System architecture and key technologies for 5G heterogeneous cloud radio access networks. arXiv preprint arXiv:1412.6677.
- [14] Castro, B. S. (2010). Modelo de propagação para redes sem fio fixas na banda de 5, 8 GHz em cidades típicas da região amazônica. Universidade Federal do Pará.
- [15] A. I. Sulyman, A. T. Nassar, M. K. Samimi, G. R. MacCartney, T. S. Rappaport, A. Alsanie, A., "Radio propagation path loss models for 5G cellular networks in the 28 GHz and 38 GHz millimeter-wave bands," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 78-86, 2014.
- [16] A. I.Sulyman, A. Alwarafy, G. R. MacCartney, T. S Rappaport, A. Alsanie, "Directional radio propagation path loss models for millimeter-wave wireless networks in the 28-, 60-, and 73-GHz bands," *IEEE Transactions on Wireless Communications*, vol.15, no. 10, pp. 6939-6947, 2016.
- [17] Dahlman, E., Parkvall, S., & Skold, J. (2013). 4G: LTE/LTE-advanced for mobile broadband. Academic press.
- [18] C. E. Shannon, "A mathematical theory of communication," *Bell Systems Tech. J.*, vol. 27, pp. 379-423, 1948.
- [19] P. Phaiwitthayaphorn, P. Boonsrimuang, P. Reangsuntea, T. Fujii, K. Sanada, K. Mori, H. Kobayashi, "Cell throughput based sleep control scheme for heterogeneous cellular networks," In 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 584-587, 2017.
- [20] B Zhang, X. Mao, J. L. Yu, Z. Han, "Resource allocation for 5G heterogeneous cloud radio access networks with D2D communication: a matching and coalition approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no.7, pp. 5883-5894, 2018
- [21] K. Koutlia, J. Pérez-Romero, R. Agusti, "On enhancing almost blank subframes management for efficient eicic in hetnets," In 2015 IEEE 81st Vehicular Technology Conference (VTC Spring) pp. 1-5. 2015.