Energy Efficiency using Cloud Management of LTE Networks Employing Fronthaul and Virtualized Baseband Processing Pool

Anwer Al-Dulaimi, *Member, IEEE,* Saba Al-Rubaye, *Member, IEEE,* and Qiang Ni, *Senior Member, IEEE*

Abstract—The cloud radio access network (C-RAN) emerges as one of the future solutions to handle the ever-growing data traffic, which is beyond the physical resources of current mobile networks. The C-RAN decouples the traffic management operations from the radio access technologies, leading to a new combination of a virtualized network core and a fronthaul architecture. This new resource coordination provides the necessary network control to manage dense Long-Term Evolution (LTE) networks overlaid with femtocells. However, the energy expenditure poses a major challenge for a typical C-RAN that consists of extended virtualized processing units and dense fronthaul data interfaces. In response to the power efficiency requirements and dynamic changes in traffic, this paper proposes C-RAN solutions and algorithms that compute the optimal backup topology and network mapping solution while denying interfacing requests from low-flow or inactive femtocells. A graph-coloring scheme is developed to label new formulated fronthaul clusters of femtocells using power as the performance metric. Additional power savings are obtained through efficient allocations of the virtualized baseband units (BBUs) subject to the arrival rate of active fronthaul interfacing requests. Moreover, the proposed solutions are used to reduce power consumption for virtualized LTE networks operating in the Wi-Fi spectrum band. The virtualized network core use the traffic load variations to determine those femtocells who are unable to transmit to switch them off for additional power savings. The simulation results demonstrate an efficient performance of the given solutions in large-scale network models.

Index Terms—Cloud computing, energy-efficient computing, network function virtualization, resource management, virtual machine

1 Introduction

The adoption of C-RAN, or increased network virtualization, by mobile operators is a new path to improve LTE network efficiency including architecture planning, networks operations, and backhaul management; most importantly, it helps to address the growing traffic requirements. The current LTE is approaching the upper bound limits for spectrum utilization, leaving only one option to meet the increasing traffic requirements, which is deploying more femtocells. These small sites under the umbrella of evolved node B (eNB) macrocells are denoted in the 3rd Generation Partnership Project (3GPP) terminology as home eNB (HeNB). In general, femtocells are interconnected with the backbone using landline networks through a cable or fiber. However, deploying more femtocells creates a more complex network architecture that increases the intercell interference and cost levels. Therefore, the C-RAN emerges as a prominent solution to scale the legacy network architecture and map data to the fronthaul in order to reduce unnecessary resource consumption.

- *A. Al-Dulaimi is with EXFO Electro-Optical Engineering Inc., Toronto, Canada.*
  E-mail: anwer.al-dulaimi@exfo.com
- *S. Al-Rubaye is with Quanta Technology, Toronto, Canada.*
  E-mail: salrubaye@quanta-technology.com
- *Q. Ni is with the School of Computing and Communications, Lancaster University, United Kingdom.*
  E-mail: q.ni@lancaster.ac.uk

Manuscript Submitted for Publication on May 20, 2015.
Revision Submitted on Jan. 30, 2016.
Accepted for Publication on Apr. 16, 2016

The C-RAN changes the basic access point technique as it moves the main signal processing functions performed by the digital baseband units to the cloud while maintaining the radio access at the cell sites in the form of remote radio heads (RRHs) technology. This imposes various challenges such as optimal utilization of the processing resources, efficient connectivity with distributed RRHs, and central management of transmitted signals [1]. The virtualized baseband pooling allows a mobile network to efficiently aggregate its processing resources across cells and introduces dynamic architecture management that can be adjusted subject to capacity and power requirements. The requests from fronthaul are instantiated as virtual machines (VMs) on generic servers at a central location within 40 Km of the fronthaul due to the limitations coming from the processing and propagation delays. Each VM consists of a variable number of BBUs with a variable number of external interfaces to RRHs. A BBU pool can be described as a virtualized cluster of segment purpose processors that perform baseband (PHY/MAC) processing operations. The interface X2 between eNBs is replaced with a new form namely X2+ to organize the inter-virtual cluster communications [2]. Therefore, efficient scaling and pooling of resources enable statistical dimensioning of multiple network sites that consist of macro and femto cells. For example in [3] and [4], significant energy savings are achieved by tracking the dynamic routing of traffic between the fronthaul and the centralized BBUs in order to switch off some BBUs during the time durations of low network load.

In a virtualized LTE network, the cloud management is represented by the virtualized evolved packet core (vEPC) that is interconnected to the VMs servers. As an example, Fig. 1 shows

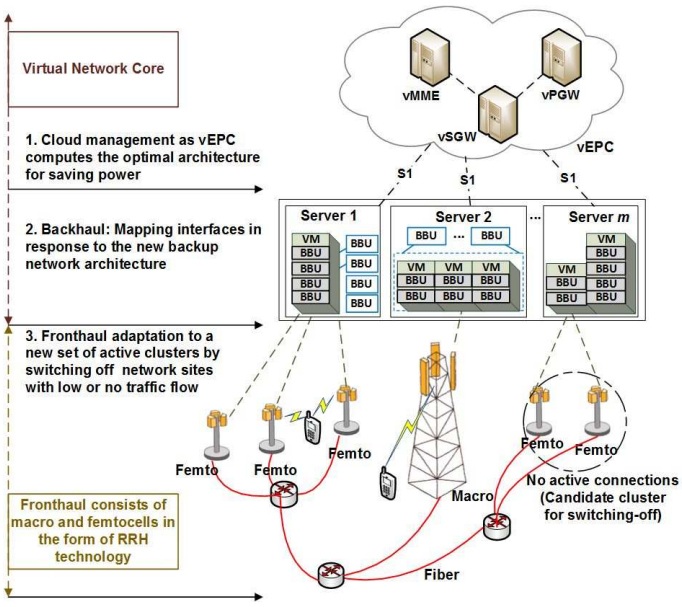

Fig. 1. Cloud management of LTE network consisting of macro and femto cells.

a multilevel network topology where the fronthaul and backhaul links are interconnected to enable baseband processing in the cloud and radio transmissions at the fronthaul. The provisioning of the network structure is performed centrally at the virtualized network core (Backhaul servers and vEPC). At the fronthaul, frequency resources are shared between these cells using the intercell interference coordination (ICIC) technique as defined in Release 8 [5]. In such scheme, a mobile user experiencing interference can report back to the associated fronthaul access point, which can be either a macro or femto cell. The X2+ backhaul interface is used by the vEPC to coordinate resources between neighboring cells to avoid multiple transmissions using the same subcarriers used by that particular mobile [6]. Therefore, adopting convergent network sites using a virtualized management approach means a more consistent service experience for the end user and greater power savings for operators.

In this paper, we present an efficient C-RAN model for mapping an LTE network overlaid with femto cells as an approach to achieve optimal network structure to reduce power consumption. There are two major steps in this approach: Firstly, we apply a traffic-aware graph-based power saving scheme that identifies the active and inactive femtocells. With the combination of C-RAN and network structure, the femtocells with relatively low traffic flow will be switched off and the traffic load will be diverted to neighbor femtocell clusters or macrocell. The graph coloring method is used to label the activity of each femtocell using certain color set for further decisions. Secondly, the LTE virtualized network core operations are extended to connect with the virtualized evolved packet data gateway (vePDG) to allow sharing information with the Wi-Fi. This framework bound the control and data planes of the two technologies for efficient management of operations. Specifically, it allows an efficient coordination of the spectrum access as this reduces the number of transmission collisions when the two technologies compete to access the same channel. Also, this coordination of the spectrum can be used to optimize the LTE consumed power as the core network process only the attainable interfaces that are anticipated to connect successful transmissions while other femtocells are switched off.

The rest of this paper is organized as follows: Section 2 presents a survey of the related work. Section 3 illustrates the motivation, problem formulation, and simulation results for the proposed graph-coloring algorithm. Section 4 illustrates the BBU optimization problem and analysis to manage resources under varying traffic loads. Section 5 illustrates the cloud interface of LTE and Wi-Fi, power optimization, and analysis. Finally, the paper is concluded in Section 6.

2 Related Work

The telecom industry is going through a major transformation by pursuing network functions virtualization and cloud management that should, over time, reduce the network consumed power because of the software-driven minimized hardware architecture. In the view of the definition of the fifth generation (5G) wireless communication standards, the cloud-based design is an important architectural solution, especially in terms of efficient usage of network resources [7], [8]. The C-RAN context inline with the network function virtualization (NFV) concept provides the necessary mechanism to activate the appropriate volume/type of functional components [9], [10], [11]. As a result, the central processing of C-RAN imposes an additional time delay to the signaling between the BBUs and RRHs. Therefore, distributed clouds allow contents to be processed closer to the users, resulting in more power savings. This approach to having servers closer to the user is a significant alteration of the concept of C-RAN to bring the network closer to users. A mixed integer linear programming (MILP) model combined the aforementioned technique to optimize energy consumption in the core network and jointly optimize the virtualization of the network functions, processing, and storage to minimize the power consumption [12].

The authors in [13] derive a tight upper bound of the C-RAN block error rate (BLER) to propose adaptive transmission schemes for power optimization. The BLER and pair-wise error probability (PEP) were used to minimize the consumed energy at the RRHs while meeting the predefined quality of service (QoS) constraint. A self-optimization method for C-RAN is proposed in [14], [15], considering a large-scale multiple-input multiple-output (MIMO). The power is allocated under the QoS requirement for each mobile user considering the asymptotic approximations of signal-to-interference-plus-noise (SINR) and RRH transmit power. The method includes network dimensioning and antenna clustering. Other advantages are easing the process for dimensioning and optimization, requiring fewer antennas, and less demand for controller and baseband computing power due to the characteristic of low complexity. The optimal power allocations were also investigated in [16] considering a fixed fronthaul rate allocation over orthogonal subcarriers. This means that a threshold-based policy is adopted depending on the channel power of a subcarrier, i.e., no power is allocated to a subcarrier if the channel power is below the threshold. Using the same scheme, a subcarrier with the highest channel power may receive the least transmit power, and vice versa.

The energy efficient coordinated transmission design for downlink transmission in C-RAN with special consideration to fronthaul capacity and user QoS constraints was presented in [17]. Specifically, the baseband signals are delivered to RRHs equipped

with multiple antennas over fronthaul links for transmissions to single-antenna users. The design aims to determine the set of RRHs that can efficiently provide end users with the requested service. This is performed by adjusting the power levels for downlink transmissions while maintaining the fronthaul capacity and user QoS constraints. Toward this end, the paper considers two problems, namely pricing-based total power and fronthaul capacity tradeoff (PFT), and fronthaul-constrained power minimization (FCPM) problems. The concave approximation and gradient search methods were employed to solve the PFT problem for the given pricing coefficients, which capture the power and fronthaul capacity tradeoff. A new algorithm to address the FCPM problem was also proposed by iteratively solving the PFT problem while intelligently updating the pricing coefficients. The efficient power control in small cell networks (SCNs) was also studied in [18]. In particular, the authors propose a power control mechanism for efficient power allocation in SCNs in order to control the system interference while guaranteeing user QoS. The methodology includes a priority grouping in which home users in the topology are assigned to one of the available groups with different priorities in terms of power requirements and requested traffic load. The mechanism dynamically updates the small cell power setting based on real-time home users' requirements. The mechanism provided better protection (in terms of interference) for both macro users and home users. The authors in [19] formulate a joint RRH selection and power minimization beamforming problem, where the transport network power consumption is determined by the set of active RRHs, while the transmit power consumption of the active RRHs is minimized through coordinated beamforming, which is NP-hard. To reduce the complexity, a novel group sparse beamforming method is proposed by inducing the group-sparsity of beamformers using the weighted $\ell 1/\ell 2$ -norm minimization, where the group sparsity pattern indicates the RRHs that can be switched off. In addition, relay nodes (RNs) were deployed to enhance the performance of edge users as a new relay assisted C-RAN in [20]. The beamforming matrices at the RRHs and RNs are jointly optimized as a non-convex optimization problem to make the design suitable for large-scale networks. Toward this end, these solutions consider improving power efficiency by adapting RRH transmissions using wireless environment factors of performance without any direct involvement of the C-RAN in managing power allocations.

In terms of other related work, there have been many attempts to optimize the energy consumption in cloud radio access networks, but only over a single dimension each time, e.g., the energy savings in BBU pool is studied in [21]. The authors stated that changing the processor frequency of the computational backhaul minimizes the energy consumption while still meeting the quality targets of the communication system. Markov model analysis show the trade off between the lost packets and the energy consumption of the cloud-server that has been mapped to the channel packet throughput. On the other hand, [22] propose a BBUs virtualization scheme that minimizes the power consumption with a linear computational complexity order. The scheme is based on a heuristic simulated annealing (HSA) algorithm. The given results show that the proposed HSA effectively decreases system power consumption when compared to standard approaches. The NFV is used to set a criterion to bundle multiple functions of a virtualized evolved packet core in a single physical device or a group of adjacent devices in [23]. The analysis shows that the proposed grouping can reduce the network control traffic by 70%. On the other hand, the authors in [24] propose a virtualized core network architecture which decouples the control plane from the user plane by using software-defined network (SDN) technology while running the user plane forwarding function on low-cost IT hardware. The major outcomes of this work are decreasing the expenditure of operating services in LTE networks. Literature, such as [25], [26], [27], [28], and [29] studied a traffic-aware graph-based dynamic frequency reuse scheme. The graph-coloring method is used to allocate different lengths of bandwidth to cells in different tiers based on their traffic demands. This type of scheme also considers the time-varying traffic and can adjust the frequency allocation based on the change of traffic demands in each cell without re-performing the whole scheme. Employing this scheme in heterogeneous C-RAN, improves not only the spectrum utilization and efficiency, but also reduces the energy consumption through reducing the number of active BBUs. Despite the detailed procedures of the given literature, these solutions do not provide the interactions between the BBUs and RRHs for joint operations of power management. In other words, there is a need to provide a high-level management of operations between the backhaul and the fronthaul for efficient power savings.

From the aforementioned literature and the fact that there are few studies available on power management in C-RAN employing NFV, it is clear that the dimensioning of the network sites using the cloud has not yet considered network structure adaptations for power savings. In this paper, we further extend the C-RAN scope by using the cloud to control the network configurations in response to changes in traffic load with the main goal of reducing the consumed power. The contributions of this paper are different from the literature in the sense of using the cloud as controller for the fronthaul architecture and transmissions. Our new model allows to combine the power savings obtained from the beamforming and mapping BBU for a reduced data processing into more centrally managed network operations. Our proposed cloud management framework is also designed to interface the LTE and Wi-Fi through virtualized functions. In this case, LTE power savings are obtained through sophisticated transmissions that consider spectrum utilization and prior knowledge of Wi-Fi activities.

3 Graph-Coloring Model for C-RAN

Considering an LTE network that employ a fronthaul technology consisting of macro and femtocells, we propose a cloud-managing technique that can adapt the network structure subject to certain metrics. These metrics combine the power consumption, femto density, and traffic load variations. The adaptation actions are performed at the fronthaul by switching off certain clusters of femtocells. As a result, it is reasonable to have multi-computing operations of actions at the cloud in response to the changes in the fronthaul. Specifically, BBUs are re-allocated and interconnected to reduce the operational power at the host servers when the data processed drops due to a reduction in the number of active femtocells. This dynamic resource provision creates a flexible network architecture with a variable number of interfaces to enable a fast response to changes in the load demand as well as a short transition time to normal steady state. Fig. 1 shows a layout of high-level dynamic provision architecture for cloud management, which shows the interfaces between the vEPC, the backhaul computational pool, and fronthaul physical units. The cloud provides the necessary computational resources that can

process the fronthaul operations virtually. To do so, we use the performance per watt (PPW) metric, which is a measure of the energy efficiency of a particular computer architecture or computer hardware, to evaluate the power analysis for the C-RAN [30]. The main challenge comes from having dense femtocells that share the capacity of physical resources using the underlying hardware. Tailoring this to a virtualized application, the computational resources pool will need to allocate sufficient resources to each VM to process the configuration requests. However, tuning the virtualized resources may need a significant response time according to the instant requirements of the different network sites. Therefore, cloud provisioning can only be efficient with fewer computational duties for better traffic management. This can be achieved jointly with power reduction by considering the cluster switch-off model that is studied in this section.

3.1 The Problem Formulation

The energy efficiency of the C-RAN is represented by the problem of cloud resource reservation using graph coloring and the PPW metric. The virtual resources are represented by the basic units of BBUs that provide the central segment of the C-RAN model. Therefore, a virtual BBU can be identified as the computational, memory, and communication unit that is allocated to a fronthaul RRH with a temporarily interface through a VM. Each BBU has a unique color in the graph coloring problem and is characterized by the PPW of its hosting system. This also means that a BBU pool is divided into graphs that reflect on their involvement in VM operations and cloud-based data processing. Referring to Fig. 1, each macro or femto (or simply RRH) k has a connection request represented by R_k, which consists of a user demand processed at a n_k virtual machine (VMs are usually in the range of VM_1 to VM_{n_k}). A VM is denoted as the guest system that is characterized by a set of BBUs at the hosting server, where a fronthaul RRH base station has to be allocated to at least one BBU to be considered as active. Therefore, a graph G can be created with vertices that represent RRHs associated with BBUs, as shown in Fig. 2. The goal is to find the optimal mapping (using graph coloring) between the requested resources given by the RRHs and the processing capacity that is virtually provided by BBUs. Given a graph $G = (V, E)$ with an integer k, where a different k color is given to each vertex (V of graph G) such that no two adjacent vertices are labeled with the same color. In a graph G, one RRH is mapped onto only one BBU at a time while overlapped requests are never assigned to the same color. With the arrival of new set of requests, we have $G = G_1 \cup G_2 ... \cup G_{n_k}$ and $G_k \cap G_j = \varnothing$, for $\forall i \neq j$ where n_k is the number of VMs, which is also representative for the number of sub-graphs. Thus, the network model is transferred onto color-based active VMs, as shown in Fig. 2.

The BBUs are normally created once a VM is launched at a physical server. Therefore, a server may contain various BBUs according to the network size, capacity management, and vendor-designed features. These virtually located BBUs are considered as the computational, memory, and communication units associated with the fronthaul service requests. In order to provide the clustering model for the BBUs and associated RRHs, we assume that one color is reserved for each $\xi_{s,j}$, where s refers to the server that is hosting a j BBU. In this way, a cluster of colors ζ_s is the set of colors (BBUs) that belongs to the same server but not necessarily to the same virtual machine. This means that each server has its own power considerations, which can be modeled by assigning a

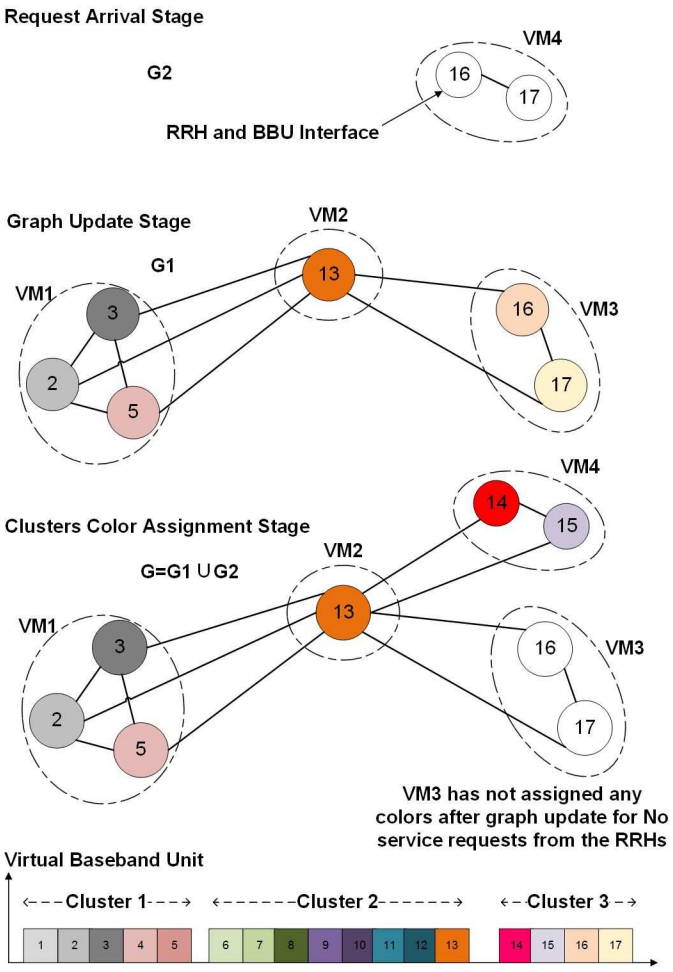

Fig. 2. Clusters formulation and color assignment based on RRHs requests.

special weight variable w_s to each server s to evaluate the PPW of the ζ_s. A fronthaul RRH is considered to be inactive if there are no R_k request to the cloud. In this way, the associated BBUs will be inactive and there will be no color reservations. Referring to Fig. 2, a new cluster ζ_3 abbreviated as VM3 is considered as color empty due to the absence of any requests initiated by the RRHs. This means that the more active and dynamic traffic load at the fronthaul, the more colored BBUs and active clusters at the backhaul. Therefore, we describe the state of a color that belongs to a certain server (cluster) as

$$y_s = \begin{cases} 1 & \text{if a color } \xi_{s,j} \text{ is assigned a RRH;} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The arbitrary requests of the RRHs present a demand of n_k for VMs, or $R_k = \{VM_1, ..., VM_{n_k}\}$. This means that each VM_{n_k} requires a specific amount R_k to reserve the necessary number of BBUs and maintain a clustering model of RRH to ease the resource management. The interface between any RRH and BBU is only reserved for the duration of the service request and will be evaluated afterwards to allow for more efficient processing of other active requests. However, the same server data center should be responsible for a set of RRHs positioned at a certain geographical location. This is one method to achieve a distributed cloud that provides more flexibility in re-structuring the RRH

clusters according to the received connection requests. Therefore, RRHs can be clustered if

$$z_{k,j} = \begin{cases} 1 & \text{if a RRH } k \text{ is assigned to a BBU or a color } j; \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Consequently, all requests belonging to the same VM can be constructed as a subgraph G_{n_k} with vertex set V_k (or a set of RRHs) and a set of edges E_k, similar to [31]. Overlapped RRHs are labeled the next color in the row that refers to the allocation of the next BBU resource within the same VM or the host server. If a new request arrives (another R_k), the colorized sub-graph will be updated and the change will lead to a new global graph as $(G = G \cup G_{n_k})$. These sub-graphs are the virtual core representation of BBUs reflecting on RRHs clusters of the fronthaul. Similarly, if a R_k departs or reaches the end of service, the sub-graph associated with that R_k will be retrieved and will be modified or deleted from the global graph G. The dynamic color adaptation of the graph G is associated with our goal to maximize the average power efficiency by maximizing

$$PPW = \frac{\sum_{s=1}^{m}\sum_{j=1}^{|E|}\zeta_s \mathcal{X}_{\xi_{s,j}}(\text{MFlops})}{\text{Power}(\sum_{s=1}^{m}\sum_{j=1}^{|E|} w_s \mathcal{X}_{\xi_{s,j}})(\text{in Watt})}. \quad (3)$$

where m refers to the overall number of used servers in cloud. The system performance is impacted by w_s, which is used in evaluating all active RRHs or colors associated with a server s. Moreover, the number of colors is preferred to be kept at minimum to ease the burden of load traffic and match the number of active RRHs. Therefore, our solution is based on assigning distinct colors to distinct vertices using chromatic number $\mathcal{X}(G)$ to produce a proper coloring with lower power requirements, similar to [32].

3.2 Energy Color Graph Optimization

The color reservation problem is formulated as integer linear programming (ILP). As shown in (1), $y_s = 1$ indicates that color j is reserved to a RRH within cluster ζ_s, otherwise $y_s = 0$. We use $d_{k,j}$ to define the demand of any RRH k for a BBU or simply a color j. Therefore, we use the variable $z_{k,j}$ from (2) to indicate whether RRH k receives color j. The conditions for color reservation with minimized power consumption are explained as follows

$$\min \sum_{k=1}^{n}\sum_{j\in J} \xi_{s,j} z_{k,j} + \sum_{s}^{m} w_s y_s \quad (4)$$

$$s.t. \quad \sum_{k}^{n} z_{k,j} = d_{k,j},\ (k = 1, 2, ..., n;\ j = 1, 2, ..., J) \quad (4a)$$

$$\sum_{k}^{n}\sum_{j\in J} z_{k,j} - y_s\left(\sum_{k}^{n}\sum_{j\in J} d_{k,j}\right) \leq 0 \quad (4b)$$

$$z_{k,j} \geq 0\ (k = 1, 2, ..., n;\ j = 1, 2, ..., J)) \quad (4c)$$

$$y_s = 0 \ \text{ or } \ 1\ (k = 1, 2, ..., n) \quad (4d)$$

$$\sum_{k\neq n} z_{k,j} \leq \sum_{j\in J} \xi_{s,j},\ \forall (k, j) \in \zeta_s,\ s = \{1, ..., m\} \quad (4e)$$

The algorithm and performance evaluations of the proposed color-graph solution are given in the following subsection.

3.3 Evaluation Results

In C-RAN, VMs are likely to accommodate more than one RRH request with the compatible BBU to the requested service. Once the RRHs are colorized, they are assigned to different clusters using their VM identifications. In Algorithm 1, we solve the problem of minimizing the power consumed by different color sets considering weight w_s of the host server s. This algorithm creates two lists for BBUs: one for the approved colored RRH, and one for the unused colors or simply the switched off RRHs. Once a color is approved, it is included in the allocated colors list and cannot be assigned again until it becomes empty. These colors mean that the corresponding BBU and RRH interfaces can be either colored (active) or not colored (switched-off). The empty list may be used to re-allocate resources when further network adaptations are required by the C-RAN.

Algorithm 1 Color Reservation Algorithm

Activate RRHs k send n_k to BBUs j
Input Graph G and a set of colors J
Operation A VM_{n_k} allocated subject to w_s
Output Colored ζ_s in the form of $\xi_{s,j}$ colors

```
for VM_{n_k} ∈ V do
    if color j is available in ξ_{s,j} then
        Add to the associated cluster ζ_s
    else No color allocated, add to switched-off RRHs
    end if
end for
for ξ_{s,j} in the list ζ_s do
    if (VM_{n_k}, ζ_s) = ζ_s − ξ_{s,j} then
        True: return ξ_{s,j}
    else False: Reallocate unused colors between s according to their PPW
    end if
end for
for VM_{n_k} ∈ V do
    if color j is available in ζ_s then
        Assign color ξ_{s,j} according to w_s
    else No color allocated, add to switched off RRHs
    end if
end for
```

The simulation parameters are given in Table 1. We create a large-sized network model for a better assessment of the proposed model. Moreover, the simulation include one macro and many femto cells in the form of RRHs where the processing operations are performed at a central server. The cloud server is interconnected with the RRHs using a fiber network.

Fig. 3 shows the performance of the proposed graph-coloring algorithm versus various service requests k, which is coming from different densities of RRHs. The figure shows the obtained PPW values for normal interfacing connections compared with the graph-coloring model for different sets of (10 and 20) RRHs. The results show a significant improvement in the PPW when using the proposed model. This is because a more sophisticated allocation of colors is performed with the new model. As a result, efficient resources allocations are achieved by identifying the power requirements of various RRH clusters.

Although the results shown in Fig. 4 are similar to the results obtained in Fig. 3, Fig. 4 is more related to the graph-coloring algorithm as it evaluates the clusters' capacity to provide

TABLE 1
Parameters for Simulation

Parameter	Assumption
Network layout	1 macrocell as a RRH (9, and19) femtocells as RRHs
Macrocell radius	400 meters
Macro distance dependent pathloss	123.1 + 37.6log10(R) dB, R in km
Femto distance dependent pathloss	140.7 + 36.7log10(R) dB, R in km
Maximum macro transmitted power	46dBm
Maximum femto transmitted power	30dBm
Landline	fiber
Cloud BBU	Sliver server

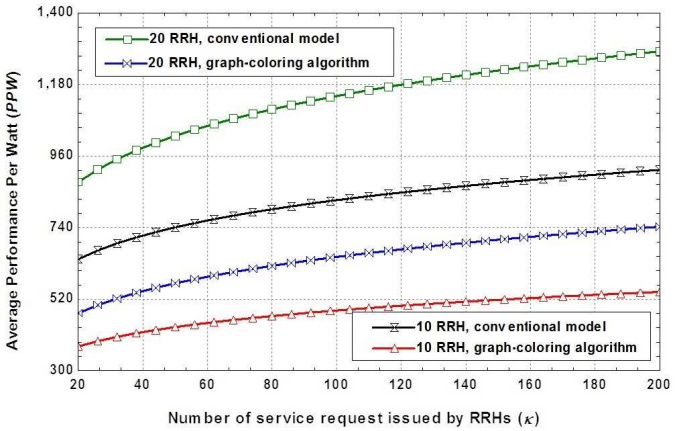

Fig. 3. Average performance per watt vs. number of service requests arriving to the cloud from different density of RRHs.

enough colors to the connection requests. It can be seen that the proposed model is more efficient in reserving colors at higher PPW compared with the other simulated scenario. This bigger graph size allows more colors to formulate new clusters that can be constructed according to their traffic load. Assigning colors to RRHs can happen only when they are active, otherwise they will be considered inactive and will not be colored. In this case, an inactive cluster is a combination of inactive BBU-RRH interfaces. For a cluster of inactive nodes, the virtual machine can be switched off and power savings are obtained from the fronthaul and backhaul, simultaneously.

In this section, we proposed a graph-coloring model for allocating clusters of active and inactive BBUs. This model provides the necessary functionalities to achieve power savings using cloud core management. In real applications, this can be used to switch off selected fronthaul RRHs if they are not requesting any reservations of the BBU pool for a considerable time interval. The other active neighbored RRHs can continue to operate as usual to provide local alternatives for communications. However, the switched off RRHs can be woken at any time by owners, mobile users, or the C-RAN in case of additional load arrival. In the next sections, we provide optimization schemes in support of the BBU pool for more power savings with a special focus on the cloud side.

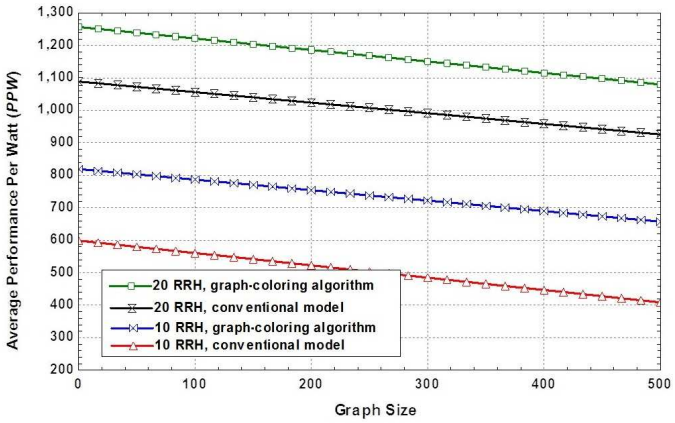

Fig. 4. Average performance per watt vs. graph size under different density of RRHs.

4 BBU Pool Optimization

A new mechanism for switching off inactive BBUs can be enabled by analyzing the clusters load using the connection request rate. Therefore, it is necessary to compute the number of users that can be handled at each site with consideration to the requested QoS. Assume T to be the desired end-to-end response time for a virtualized system to maintain network operations without performing architectural adaptation. Then, $\tau_r(1), \tau_r(2), ..., \tau_r(N)$ are denoted as the per user end-to-end response time durations, so that $\sum_{j=1}^{J} \tau_r(j) = T$. As mentioned earlier, R_k is the arrival rate of requests to a BBU j from an RRH k. Assume that not all requests are processed by VMs as some may be ignored for having overlapped reservation time, the probability of relative request arrival rate is given by R_k/R_{k-1}= $\varphi_{k-1} \leq 1$, $\forall k \geq 2$, such that $R_2/R_1 = \varphi_1$, $R_3/R_2 = \varphi_2$, ..., $R_K/R_{K-1} = \varphi_{K-1}$. The variables $\varphi_1, \varphi_2, ..., \varphi_{K-1}$ are set randomly during the simulation. With the arrival of these requests to the cloud interfaces, they get processed with the same criteria, as $\nu_{1,j} = \nu_{2,j} = \nu_{3,j} = ... = \nu_{n,J}$. This means that the overall processing time for arrived requests is given by $\sum_{k=1}^{n} \nu_{n,J}$, the processing rate of each BBU is $\sum_{k=1}^{n} \nu_{n,J}/n_K$, while the processing rate of a certain RRH request denoted as $(k = 1)$ is given as $\sum_{j=1}^{J} \nu(j)$, where $\nu(j) = \nu_{1,1} + \nu_{1,2} + ... + \nu_{1,J}$.

4.1 Problem Definition

In the studied model, cloud processing time durations is assumed to be drawn from a known fixed network architecture. Therefore, the BBU processing utilization request is $\lambda_j = R_k/\nu(j) \leq 1$, $0 \leq \lambda_j \leq 1$, where λ_j refers to the utilization of processing speed for BBU j.

As we employ multiple BBUs to process requests from the RRHs and these BBUs are contained at multiple servers, then we model the queuing system as an M/M/c system, using the same method given in [33]. The end-to-end average response time for the RRH requests is given as

$$\tau_r(j) = \frac{R_k.(\nu(j) + J - J\nu(j))}{\Pi_{k=1}^{n} \lambda_j(k)(1-\nu(j))^2}.\gamma_{0,j} + \frac{1}{R_k}.\sum_{j=0}^{J-1}(k.\gamma_{k,j}) \quad (5)$$

where $\gamma_{0,j}$ is the probability that a RRH request leaves BBU j immediately after processing at the backhaul, and $\gamma_{k,j}$ is the probability that a RRH k is connected to a BBU j.

For a BBU j delivers its response to RRH k at a rate $\mu_{k,j}(2 \leq j \leq J,\ 1 \leq k \leq n_k)$, the end-to-end average response time on the cloud side is given as

$$\tau_r(j) = \frac{1}{\nu_{1,j} - \mu_{1,j}} = \frac{1}{\nu_{2,j} - \mu_{2,j}} = ... = \frac{1}{\nu_{n_k,j} - \mu_{n,J}} \quad (6)$$

Once the BBUs are mapped to the RRHs, the network structure is adapted to include the active sites only. Therefore, Λ_{pt} is defined as the optimization function to minimize the weighted interfaces of the system as some of the clusters are switched off. This is given as

$$\min\{\Lambda_{pt} = f(R_1, \nu_{1,1}, ..., \nu_{n,1}; R_k, \nu_{1,j}, ..., \nu_{1,J}; R_k, \nu_{1,J}, ..., \nu_{n,J})\} \quad (7)$$

$$s.t. \quad \sum_{j=1}^{J} \tau_r(j) \leq T_0 \quad (7a)$$

$$\sum_{k=1}^{J} \nu_{k,j} > R_k,\ j = \{1, ..., J\};\ k = \{1, ..., n\} \quad (7b)$$

where T_0 is a given response time for the optimization operations, which is obtained from the backhaul.

The above optimization problem is solved using Algorithm 2 that computes RRHs requests to BBUs using the connection request rate to evaluate Λ_{pt}.

Algorithm 2 BBU Pool Management Algorithm

Activate RRHs k request to service by BBUs j
Input Relative request arrival rate R_k/R_{k-1}= φ_{k-1}
Operation Evaluate the response time $\tau_r(j)$
Output BBU optimization function Λ_{pt}

for $j \leq J-1$ **do**
  if $\tau_r(j) > T_0$ **then**
    $\gamma_{k,j} = 0$
  else $\mu_{k,j}$ is sent to RRH k
  end if
end for
for $\mu_{k,j} \in s$ **do**
  if $\lambda_j < 1$ **then**
    Re-calculate $\tau_r(j),\ \forall(2 \leq j \leq J,\ 1 \leq k \leq n_k)$
  else Use equation (7)
  end if
end for
for $k \in n_k$ **do**
  if $\nu_{k,j} > R_k$ **then**
    Calculate Λ_{pt}
  else Remove $\mu_{k,j}$
  end if
end for

4.2 Simulation Results

In order to evaluate the concept of BBU pool re-configuration in response to traffic and response time delays by the BBUs to RRHs requests, we compute the number of BBUs and the performance of their interfaces to the fronthaul RRHs besides several other input parameters, such as $\tau_r(j)$ = 0.11 sec, $\nu_{k,j}$ = 370 requests/sec, and T_0 = 0.83 sec, using Algorithm 2.

The simulation results show that the BBU response delay time increases as the number of service requests issued by RRHs increases, as shown in Fig. 5. Although the number of BBUs increases simultaneously with the increase in the number of RRHs, the system still experiences an increasing delay time due to the exhausted physical processing resources of the hosting servers. Moreover, additional times delays are required for launching application instances in VMs with multi interfaces to RRHs. Therefore, the system have better performance when the network adaptation reduces the number of active interfaces to the cloud. This provide VMs with more physical processing resources that can be amplified through NFV especially when these interfaces have defined capacity channels. The time delay could increase further in a busy wireless environment as there is more data in transmission and more time consumed while processing requests arrive at the backhual. In fact, Fig. 5 confirms the need for the scheme proposed in Section 3.2 scheme that can not only save power but also maintain high network performance.

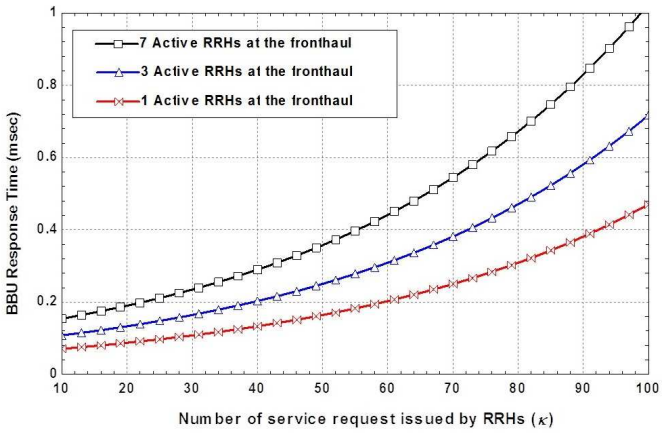

Fig. 5. The BBUs response time vs. the number of service requests arriving to the cloud from different density of RRHs.

Fig. 6 shows the optimization function performance for an increasing load reflected as service requests from RRHs. The results confirm that the system is able to reach a steady performance and semi-stable architecture when there is enough flexibility to adapt the architecture in response to the network performance metrics. It can be seen that the system requires additional time before reaching the steady performance as launching new VMs and creating new BBUs become slow when there is a limited number of servers. There is also an additional time delay needed for creating the virtual network interfaces between the requested RRHs and the newly created BBUs. However, once a RRH stops sending requests for service, the network architecture is adapted to switch off that RRH. Subsequently, all associated BBU and virtual network connectivity will also removed from the cloud because of service cancelation and power efficiency requirements. Migrating VMs between servers can also save more power by switching

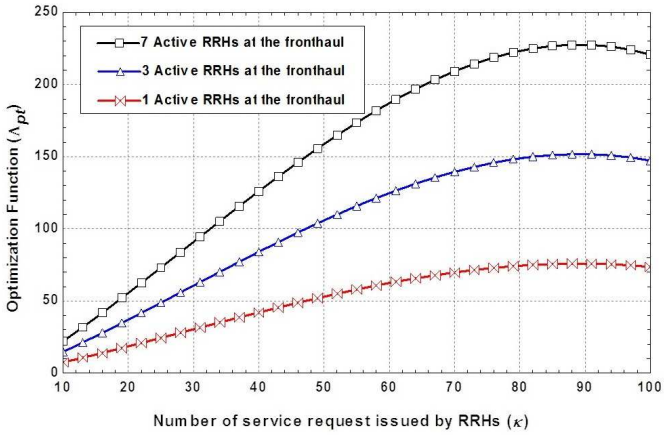

Fig. 6. Optimization functionality performance vs. the number of service requests arriving to the cloud from different density of RRHs.

off the offline servers. However, the backhaul resources can be switched on if there are new interface requests due to the change in traffic load at the fronthaul side.

5 Fronthaul Power-Related Spectrum Optimization

The cloud control of RRHs is the enabler for an efficient C-RAN solution that interconnects various operator technologies especially at enterprise locations. Such architecture may be operated over standard Ethernet switches and cabling or fiber to provide local connectivity and radio frequency planning. The recent proposals to extend the LTE to unlicensed band raise a new challenge to establish cloud management between the LTE RRHs and the Wi-Fi for efficient spectrum sharing [34], [35]. A virtualized network can be the enabler for an efficient coexistence between multiple systems over the unlicensed spectrum to achieve the highest value of spectrum utilization. This coexistence between LTE and Wi-Fi can be achieved by creating a central control of spectrum access and information sharing between these different systems. Tailoring this to power consumption, more RRHs deployed in unlicensed band cause a significant increase in the power values that are required for static and dynamic operations. In fact, the extension of LTE to the unlicensed band makes it more difficult to reduce the network energy expenditure. In this section, we propose a scenario for the cloud control of RRHs where any RRH may be switched off whenever it fails to access the spectrum for a certain number of times. The proposed model uses fronthaul information to make the necessary actions of adapting various interfaces at the fronthaul and remove the allocated virtual resources at the backhaul.

5.1 Cloud Interfacing Between LTE and Wi-Fi

Controlling the Wi-Fi through a centralized integrated core network requires employing NFV entities as an enabler for any management between both LTE and Wi-Fi. The NFV was also considered by the 3GPP RAN Sharing Enhancements (RSE) study, which allows a common spectrum network sharing for a number of operators that decide to pool their allocated spectrum [36]. This virtualized controlled-based architectures supports separating the data plane from the control plane for efficient centralized management of the RAN. Oppositely, data are forwarded in a distributed model between base stations and mobile users in order to meet the relatively large volume of data while competing for the scarce spectrum. As for having a separated data plane, traffic is either offloaded locally or tunneled to other access points, enabling efficient and flexible spectrum access [35], [37]. The proposed cloud model virtualizes the evolved packet data gateway (ePDG) and trusted WLAN access gateway (TWAG) functions for secure access to the untrusted and the trusted Wi-Fi networks through cloud integration. This kind of heterogenous networks provides the control access to the Wi-Fi gateway beyond what is defined in 3GPP TS23.402 [38], [39]. The vePDG is responsible for interworking between the vEPC and untrusted non-3GPP networks that require secure access, such as Wi-Fi access networks. The vePDG function supports rapid packet processing and significant memory resources for processing control signaling between the backhaul and the fronthaul technologies. Fig. 7 shows the control plane for vePDG and vEPC to improve control and adaptability between fronthaul and backhaul considering the fromthaul transmissions. This virtualization within the control plane provides the necessary features for authentication, interference detection and avoidance, transmission power and channel assignment, coverage hole detection and correction, and load balancing.

The distributed architecture incorporates a high-performance network that can intelligently adapt fronthaul operations. Call control and packet forwarding paths are separated on different control and data planes, reducing the number of traffic-flow inefficiencies, which diminishes latency and accelerates call setup time and handoffs. The power allocation is implemented by retrieving the radio access experience from the fronthaul network. This is performed when the core network starts intelligently scanning the delivery of various multimedia services that are performed by the fronthaul. The mechanism is based on a binary exponential backoff algorithm that is used to space out repeated retransmissions of the same block of data, often as part of network congestion avoidance. The failure of a RRH to transmit for many attempts enables the vEPC network-based traffic optimization actions that re-divert the traffic to other neighbor fronthaul RRHs in order to achieve a higher degree of power optimization and spectrum management. Therefore, core network operations can increase the network scalability through employing only successful transmissions that reduce the overall power consumption.

5.2 Spectrum Access

The main objective of this subsection is to analyze the power consumption subject to the changes in traffic at the fronthaul. We assume that the interconnection of LTE and Wi-Fi virtualized functionalities will allow for prior knowledge about the collisions that may occur between the LTE and Wi-Fi in the unlicensed band. This should enable the BBU pool to take the necessary actions in order to save power. Considering a C-RAN model of a fronthaul distributed system, the BBUs are contained within a server s that process the arrival of RRHs packets at a rate of α_s. The RRHs are normally connected to the BBUs through interfaces that have a bandwidth of $B_{k,j}$. Then, the traffic processed by a server s consists of many BBUs can be given as $\sum_{k=1}^{n}\sum_{j=1}^{J} d_{k,j} = D_s$ and $0 \leq d_{k,j} \leq D_s$. Therefore, the power consumption for a BBU is $P_j = P_j^{\theta} + P_j^{0}$; where P_j^{θ} denotes the power consumption due to the traffic at j, and P_j^{0} denotes the idle power of the same BBU

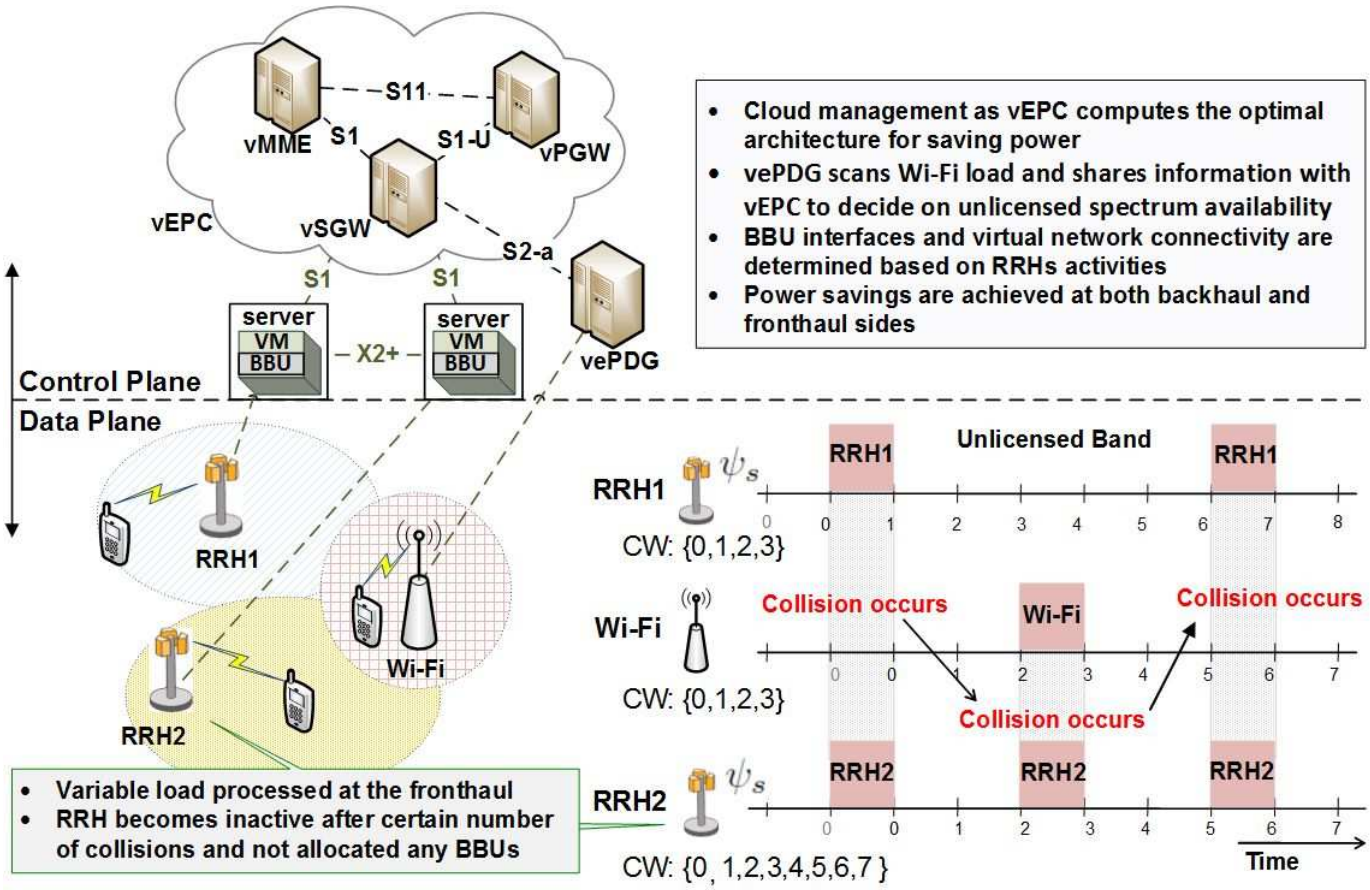

Fig. 7. Cloud interfacing between RRH and Wi-Fi technologies. The control and data planes are decoupled using virtualized entities that are able to adapt network structure in response to the wireless resources changes. Exchange of Information is performed using the shared backhaul interfaces between the two technologies.

j. The average power consumption for all BBUs within s is given as

$$P_{BBU} = \sum_{s=1}^{m} \sum_{j=1}^{J} \frac{P_j}{\xi_{s,j}} \tag{8}$$

The normalized load for a server s is given as $\psi_s = \frac{\theta_s}{\alpha_s}$. This means that a load balancing condition can be given as

$$\theta_s = \frac{\alpha_s \eta_s}{\rho} = \frac{\alpha_s}{\rho} D_s = \sum_{k=1}^{n} \sum_{j=1}^{J} \frac{\alpha_j}{\rho} d_{k,j} \tag{9}$$

where η_s denotes the system throughput at server s, ρ denotes the number of interfaces between RRHs and the BBUs contained by server s and can be given as $\rho = \sum_{k=1}^{n} \sum_{j=1}^{J} z_{k,j}$

Since the traffic load is changing dynamically, then the load balance can be given as

$$\beta = \sum_{s=1}^{m} \sum_{j=1}^{J} \frac{1}{\xi_{s,j}} (\psi_s - \Upsilon)^2 = \sum_{s=1}^{m} \sum_{j=1}^{J} \frac{1}{\xi_{s,j}} (\psi_s - \frac{\eta_s}{\rho})^2 \tag{10}$$

where $\Upsilon = \frac{\eta_s}{\rho}$ denotes the normalized mean load for all BBUs. Therefore, less variation in the load results in a more stable system with fewer adaptations at the fronthaul side.

To combine the load variation with the average power consumed by BBUs in server s, the weighted sum method is used to jointly analyze (10) and (12) as

$$\begin{aligned} \min d_{k,j}(J) &= \min d_{k,j}\big((1-w_f)\tilde{P}_{BBU} + w_s\beta\big) \\ &= \min d_{k,j}\Big(\sum_{s=1}^{m}\sum_{j=1}^{J} \frac{(1-w_f)}{\xi_{s,j}} \big(\frac{P_j}{P^{max}} + w_s\beta\big)\Big) \end{aligned} \tag{11}$$

$$s.t. \quad \theta_s \leq \alpha_s, \ \forall j = \{1, ..., J\} \tag{11a}$$

$$\sum_{k=1}^{n} \sum_{j=1}^{J} d_{k,j} = D_s, \ \forall j = \{1, ..., J\}, \ \forall k = \{1, ..., n\} \tag{11b}$$

$$0 \leq D_s \leq \beta, \ \forall s = \{1, ..., m\} \tag{11c}$$

where $0 \leq w_f \leq 1$ denotes the weighting factor between power consumption and load balancing. $\tilde{P}_{BBU}$ is the normalized mean power.

Considering server s, the problem of load variation impact on power is evaluated using Algorithm 3.

Algorithm 3 Activate RRHs Transmissions Algorithm

Activate P_j is allocated to each $\xi_{s,j}$
Input Generate data traffic rate of D_s
Operation Evaluate load for a server s
Output β and P_{BBU}

for $P_j - P_j^{\theta} > P_j^0$ **do**
  if BBU j process α_j for RRH k **then**
    Perform equation (10)
  else switch off RRH k
  end if
end for
for $z_{k,j} \in \rho_s$ **do**
  if $\theta_s \leq \alpha_s$ **then**
    Calculate P_{BUU} and β for s
  else switch off BBU j in server s
  end if
end for

5.3 Simulation Results

We conduct a simulation that investigates the power changes due to the load rate scheduling. The studied system model consists of 7 RRHs and 10 BBUs that provide the backhaul for the cloud system. The system capacity α_s is assumed to be 4.1 Gbps and power consumption P_j^{θ} is assumed to be 30 nJ/b, while other simulation parameters are specified as in Table 1. On the fronthaul side, the RRHs share the band with another 3 Wi-Fi units to mimic the challenge of dynamic load delivery due to the collision in transmissions when the two technologies operate over the same band. The evaluations are performed using Algorithm 3.

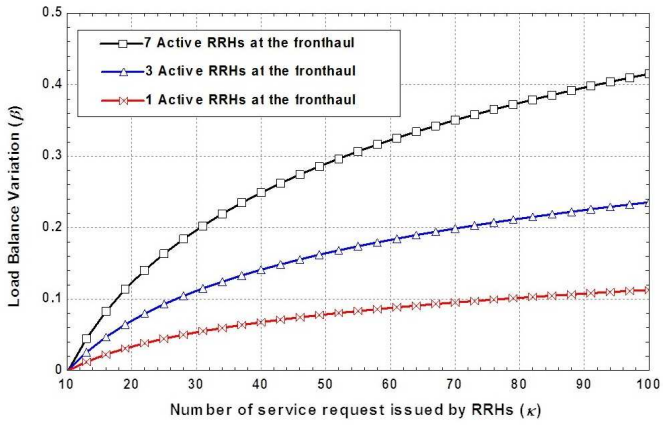

Fig. 8. Load balance variation vs. the number of service requests arriving to the cloud from different density of RRHs.

Fig. 8 shows the simulation results for load balance variation versus the number of received service requests by the cloud. Obviously, the load variation increases as the number of service requests despite the number of active RRHs. This is due to the fact that more interfaces between the RRHs and BBUs are required to process the transferred load. This means that the continuous increase in the number of users across the structure of the network causes more traffic and most importantly causes more power consumption. Therefore, it is necessary for a cloud managed network to set up the most appropriate number of active RRHs and virtual machines to maintain an efficient power network structure. Also, the power saving adaptations should be performed with a minimum change in the network structure to maintain connections with mobile users and the fronthaul from one side and between the fronthaul and the backhaul on the other side.

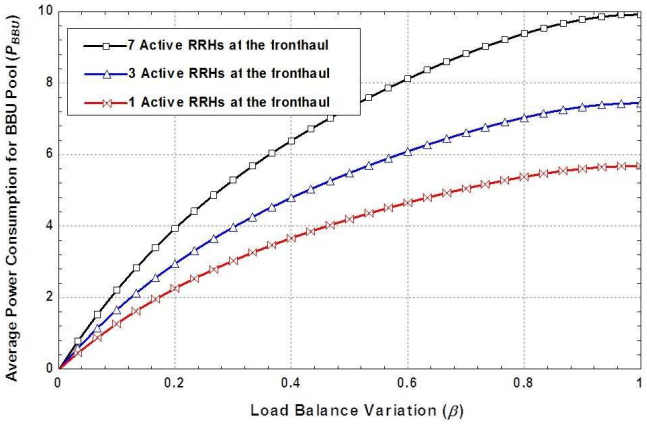

Fig. 9. Average BBUs power consumption vs. the load balance variation under different density of RRHs.

In Fig. 9, we simulate the overall average power consumption experienced by the BBUs subject to the increase in the load and the number of deployed RRHs. It can be seen that the power consumption increases jointly with the increase in the number of deployed RRHs and the variation in traffic load. This shows that the cloud is necessary to enable an efficient allocation of resources that can allow switching off the network nodes that are not required to operate for certain periods of time. This can significantly reduce the energy burden on mobile operators and help to maintain large-scale dense network models at low power requirements. However, even though a cloud model network is simulated here, it worthwhile noting that the backhaul interfaces could be another source for increasing the overall network power consumption. Therefore, mobile operators should employ the most efficient power scheme for interfacing the various network sites with the cloud along with reducing the power required for processing that huge data at the backhaul. The computing process of the data at the virtualized core should always remain low. Furthermore, it is necessary to develop the control process for the cloud to allow for more smooth controlling of non-trusted radio interfaces other than LTE. These can be incorporated into the cloud backhaul that can further increase power management as well as improve the network performance characteristics.

In this section, we presented a power analysis for an LTE RRHs operating in the same unlicensed band with Wi-Fi in order to improve the vision of cloud management while analyzing the power requirements.

6 Conclusion

This paper proposed a cloud provisioning model for an LTE network consisting of one macrocell and variable numbers of femtocells in the form of RRHs. The new proposed C-RAN model allowed for a smooth coordination of densely heterogeneous networks through central core management. We focused on the challenge of adapting the virtual baseband units interfaces at the cloud in response to the variation of traffic load at the fronthaul. Our contributions were organized into three parts. In the first part,

we provided a graph-coloring model that switches off the virtual machines clusters with no or low traffic, enabling significant power savings that are obtained from both fronthaul and backhaul sides. In the second part, we studied the concept of BBU pool configuration in response to traffic and the round-time delays required to perform network adaptation. The analysis showed the number of BBUs and the required capacity of their interfaces to meet the fronthaul connection requests. Finally, the proposed C-RAN model was extended to couple the Wi-Fi with the cloud system for more power savings through prior knowledge on traffic load transmitted over various unlicensed channels. The power analyses were conducted using ILP and the simulation results confirmed the novelty of the given solution.

REFERENCES

[1] I. Chih-Lin, J. Huang, R. Duan, C. Cui, J.X. Jiang, and L. Li, "Recent Progress on C-RAN Centralization and Cloudification," *IEEE Access*, vol. 2, pp. 1030 - 1039, 2014.

[2] A. Checko, H.L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M.S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks? A Technology Overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405 - 426, First quarter 2015.

[3] I. Sadooghi, J. H. Martin,T. Li, K. Brandstatter, Y. Zhao, K. Maheshwari, T. Pais Pitta de Lacerda Ruivo, S. Timm, G. Garzoglio, and I. Raicu, "Understanding the Performance and Potential of Cloud Computing for Scientific Applications," *IEEE Transactions on Cloud Computing*, vol. PP, no. 99, pp. 1 - 14, 2015.

[4] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Inter-Cluster Design of Precoding and Fronthaul Compression for Cloud Radio Access Networks," *IEEE Wireless Communications Letters*, vol. 3, no. 4, pp. 369 - 372, Aug. 2014.

[5] V. Pauli, J. D. Naranjo, and E. Seidel, "Heterogeneous LTE Networks and Inter-Cell Interference Coordination," *Nomor Research White Paper GmbH*, Munich, Germany, Dec. 2010.

[6] S. Luo, R.i Zhang, and T. J. Lim, "Downlink and Uplink Energy Minimization Through User Association and Beamforming in C-RAN," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 494 - 508, Jan. 2015.

[7] D. Sabella, A. De Domenico, E. Katranaras, M.A. Imran, M. di Girolamo, U. Salim, M. Lalam, K. Samdanis, and A. Maeder, "Energy Efficiency Benefits of RAN-as-a-Service Concept for a Cloud-Based 5G Mobile Network Infrastructure," *IEEE Access*, vol. 2, pp. 1586 - 1597, 2014.

[8] A. de la Oliva, J.A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 152 - 159, Feb. 2016.

[9] P. Demestichas, A. Georgakopoulos, K. Tsagkaris, and S. Kotrotsos, "Intelligent 5G Networks: Managing 5G Wireless/Mobile Broadband," *IEEE Vehicular Technology Magazine*, vol. 10, no. 3, pp. 41 - 50, Sept. 2015.

[10] https://www.ict-earth.eu/

[11] H. Hawilo, A. Shami, M. Mirahmadi and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network*, vol. 28, no. 6, pp. 18-26, Nov.-Dec. 2014.

[12] "Reducing the Net Energy Consumption in Communications Networks by up to 98% by 2020," *White Paper*, GreenTouch Foundation, v. 1, Jun. 2015.

[13] T.X. Vu, T. Nguyen, and T.Q.S. Quek, "Power Optimization with BLER Constraint for Wireless Fronthauls in C-RAN," *IEEE Communications Letters*, vol. PP, no. 99, pp. 1 - 1, Dec. 2015.

[14] P.-R. Li, T.-S. Chang, and K.-T. Feng, "Energy-efficient Power Allocation for Distributed Large-scale MIMO Cloud Radio Access Networks," *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1856 - 1861, Apr. 2014.

[15] C.H. Tang, Y.-K.i C. Chen, and L.-C. Wang, "Self-optimized Cloud RAN based Smart Zone," *16th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 1 - 5, Sept. 2014.

[16] L. Liu, S. Bi, and R. Zhang, "Joint Power Control and Fronthaul Rate Allocation for Throughput Maximization in OFDMA-Based Cloud Radio Access Network," *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 4097 - 4110, Nov. 2015.

[17] V.N. Ha, L. B. Le, and N.-D. Dao, "Energy-efficient Coordinated Transmission for Cloud-RANs: Algorithm design and trade-off," *48th Annual Conf. on Info. Sciences and Systems (CISS)*, pp.1 - 6, Mar. 2014.

[18] A. Alexiou, D. Billios, and C. Bouras, "A Power Control Mechanism Based on Priority Grouping for Small Cell Networks," *Eighth International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA)*, pp. 170 - 176, Oct. 2013.

[19] Y. Shi, J. Zhang, and K.B. Letaief, "Group Sparse Beamforming for Green Cloud-RAN," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2809 - 2823, May 2014.

[20] Z. Fei, C. Xing, N. Li, D. Zhu, and M. Lei, "Leakage-based Distributed Minimum-mean-square Error Beamforming for Relay-assisted Cloud Radio Access Networks," *IET Communications*, vol. 8, no. 11, pp. 1883 - 1891, Jul. 2014.

[21] R. Wang, H. Hu, and X. Yang, "Potentials and Challenges of C-RAN Supporting Multi-RATs Toward 5G Mobile Networks," *IEEE Access*, vol. 2, pp. 1187 - 1195, 2014.

[22] M.i Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband Processing Units Virtualization for Cloud Radio Access Networks," *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 189 - 192, Apr. 2015.

[23] H. Hawilo,A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the Art, Challenges, and Implementation in Next Generation Mobile Networks (vEPC)," *IEEE Network*, vol. 28, no. 6, pp.18 - 26, Nov.-Dec. 2014.

[24] X. An, W. Kiess, and D. Perez-Caparros, "Virtualization of Cellular Network EPC Gateways based on a Scalable SDN Architecture," *IEEE Global Comms Conf. (GLOBECOM)*, pp. 2295 - 2301, Dec. 2014.

[25] W. Cheng, X. Zhang, and H. Zhang, "QoS-Aware Power Allocations for Maximizing Effective Capacity Over Virtual-MIMO Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 10, pp. 2043 - 2057, Oct. 2013.

[26] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud Radio Access Network (C-RAN): a primer," *IEEE Net*, vol. 29, no. 1, pp. 35 - 41, Jan.-Feb. 2015.

[27] A.W. Dawson, M.K. Marina, and F.J. Garcia, "On the Benefits of RAN Virtualisation in C-RAN Based Mobile Networks," *Third European Workshop on Software Defined Networks (EWSDN)*, pp. 103 - 108, Sept. 2014.

[28] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. di Girolamo,and G. Giuliani, "RAN as a service: Challenges of Designing a Flexible RAN Architecture in a Cloud-based Heterogeneous Mobile Network," *Future Network and Mobile Summit*, pp. 1 - 8, Jul. 2013.

[29] L. Zhou, X. Hu, E. Ngai, H. Zhao, S. Wang, J. Wei, and V. Leung, "A Dynamic Graph-based Scheduling and Interference Coordination Approach in Heterogeneous Cellular Networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1 - 13, 2015.

[30] J.-M. Pierson, *Large-Scale Distributed Systems and Energy Efficiency: A Holistic View*, John Wiley & Sons Ltd., 2015.

[31] H. Elghazel, H. Kheddouci, V. Deslandres, and A. Dussauchoy, "A Graph b-coloring Framework for Data Clustering," *Journal of Mathematical Modelling and Algorithms*, vol. 7, pp. 389 - 423, 2008.

[32] P. Mani and D. Petr, "Clique Number vs. Chromatic Number in Wireless Interference Graphs: Simulation Results," *IEEE Communications Letters*, vol. 11, no. 7, pp. 592 - 594, Jul. 2007.

[33] J. McKenna, "A Generalization of Little's Law to Moments of Queue Lengths and Waiting Times in Closed, Product Form Queueing Networks, *Journal of Applied Probability*, no. 26, pp. 121 - 133, 1989.

[34] F.M. Abinader, E.P.L. Almeida, F.S. Chaves, A.M. Cavalcante, R.D. Vieira, R.C.D. Paiva, A.M. Sobrinho, S. Choudhury, E. Tuomaala, K. Doppler, and V.A. Sousa, "Enabling the Coexistence of LTE and Wi-Fi in Unlicensed Bands," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 54 - 61, Nov. 2014.

[35] A. Al-Dulaimi, S. Al-Rubaye, Q. Ni, and E. Sousa, "5G Communications Race: Pursuit of More Capacity Triggers LTE in Unlicensed Band," *IEEE Vehicular Technology Magazine*, vol. 10, no. 1, pp.43 - 51, Mar. 2015.

[36] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio Access Network Virtualization for Future Mobile Carrier Networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27 - 35, Jul. 2013.

[37] X. Ding, C. Liu, L. Wang, and X. Zhao, "Coexisting Success Probability and Throughput of Multi-RAT Wireless Networks With Unlicensed Band Access," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 4 - 7, Feb. 2016.

[38] S. Hamalainen, H. Sanneck, and C. Sartori, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*, John Wiley Sons Ltd, Chippemham, Great Britain, 2012.

[39] 4G Americas, "Integration of Cellular and Wi-Fi Networks," *White Paper*, Sept. 2013.

Anwer Al-Dulaimi (M11) received the Ph.D. degree in electrical and computer engineering from Brunel University London, U.K., in 2012. In 2013 - 2014, he was a Postdoctoral Research Fellow with the department of electrical and computer engineering, Ryerson University, Toronto, Canada. In 2014 - 2015, he was a Postdoctoral Research Fellow with the department of electrical and computer engineering, University of Toronto, Toronto, Canada. Currently, he is with the R&D department at EXFO Electro-Optical Engineering Inc., Toronto, Canada. He is member of the IEEE 5G standards action and Networld2020 working groups. His research interests lie in the area of 5G wireless systems with special focus on dynamic spectrum access, network functions virtualization, and alternative routing algorithms that consider the energy savings and information exchange between peer radios. He was the recipient of the 2013 Worldwide Universities Network Cognitive Communications Consortium best paper for outstanding research in cognitive communications for his book entitled “Self-organization and Green Applications in Cognitive Radio Networks”.

Saba Al-Rubaye (M10) received her Ph.D. degree in Electrical and Electronic Engineering from Brunel University London, U.K., in 2013. She is currently working with Renewable Energy Group at Quanta Technology, Toronto, Canada. Her current research focuses on energy efficient networks, smart grid communications, microgrid, and wireless communications. Dr. Al-Rubaye registered as a Chartered Engineer (CEng) by Engineering Council in the U.K. and recognized as Associate Fellow of the British Higher Education Academy. Dr. Al-Rubaye is a recipient of the best paper award twice from Wireless World Research Forum published in IEEE Vehicular Technology in 2011 and 2015, respectively. She published over twenty journal and conference papers in the areas of energy efficiency and telecommunications. She acts as a reviewer for IEEE Transections on Vehicular Technology and IEEE Transactions on Control Systems Technology. Also, she has served on technical program committees and organizing committees for leading conferences of several international conferences such as IEEE GLOBECOM, IEEE PIMRC, IEEE ICC, IEEE WCNC and IEEE CIT. She is a member of the IEEE, IET and European Technology Platform Photonics 21.

Qiang Ni (M04-SM08) is a Professor and the Head of Communication Systems Group at the School of Computing and Communications, Lancaster University, InfoLab21, Lancaster, U.K. Previously, he led the Intelligent Wireless Communication Networking Group at Brunel University London, U.K. He received the B.Sc., M.Sc., and Ph.D. degrees from Huazhong University of Science and Technology, China, all in engineering. His main research interests lie in the area of future generation communications and networking, including Green Communications and Networking, Cognitive Radio Network Systems, Heterogeneous Networks, Small Cell and Ultra Dense Networks, 5G, SDN, Cloud Networks, Energy Harvesting, Wireless Information and Power Transfer, IoTs and Vehicular Networks in which areas he had already published over 120 papers. He was an IEEE 802.11 Wireless Standard Working Group Voting member and a contributor to the IEEE Wireless Standards.