
Article:
Li, Zhehan, Grace, David orcid.org/0000-0003-4493-7498 and Mitchell, Paul Daniel orcid.org/0000-0003-0714-2581 (2018) Hotspot-oriented green frameworks for ultra-small cell cloud radio access networks. IEEE Transactions on Vehicular Technology. pp. 703-717. ISSN 0018-9545

https://doi.org/10.1109/TVT.2017.2739398

Hotspot-oriented Green Frameworks for Ultra-small Cell Cloud Radio Access Networks

Zhehan Li, David Grace, *Senior Member, IEEE,* and Paul Mitchell, *Senior Member, IEEE*

***Abstract*—Sleep mode operation of base stations aims to switch off some hardware modules to reduce power consumption while not degrading the Quality of Service. In this paper, novel Hotspot-oriented Green Frameworks based on sleep model operation of Remote Radio Heads (RRHs) are proposed for Cloud Radio Access Networks (C-RANs). The trade-off between the reduction in power consumption of RRHs and the increase in transmission power at User Equipment (UE) is first analysed based on realistic models for ultra-small cell C-RANs. In the proposed energy-efficient frameworks, corresponding clustering strategies are adopted to ensure that active RRHs are located as near as possible to hotspot areas for different infrastructure conditions and information availabilities. This reduces the increase in the uplink transmission power while maximising the overall RRH power reduction. The green frameworks are modelled using C-RANs based on random topologies. It is shown that area power consumption can be reduced by more than 79% at a low traffic level compared with no sleep mode operation. One of the frameworks is also compared with a baseline strategy that deals with hotspot areas and shows a 70% reduction in UE transmission power. The pros and cons of applying different frameworks are also investigated and analysed.**

***Index Terms*—Cloud RAN, energy saving, sleep mode operation, ultra-small cell**

I. Introduction

TO meet the continuously increasing demand of data services, industry is developing Radio Access Networks (RANs) with a further 1000× capacity of today [1], [2]. One of the promising paradigms to boost capacity is to have more cells with much reduced sizes in a certain area, i.e. to deploy dense low cost and low power access nodes. This approach enhances frequency reuse in areas of high data traffic densities and is the simplest and most effective way to increase system capacity [3]. With ultra-dense networks, the envisioned Inter-site Distances (ISD) between small cell access nodes range from a few meters in indoor deployments to roughly 50 m in outdoor deployments [4].

The total energy consumption is predicted to increase taking account of the additional development of network infrastructures, which raises the operational expenditure (OPEX), lowers the revenues of operators, and generates greater carbon dioxide emissions. Therefore, when developing dense base stations to boost system capacity, operators are also interested in networking with an energy efficient architecture and involving technologies to lower the overall network power consumption to minimise greenhouse gas emissions and to be more profitable [5].

A novel network architecture, Cloud RAN (C-RAN) is a strong candidate to address the above issues and enable green communications. C-RAN implies centralised processing, cooperative radio, cloud, and clean (green) infrastructure RANs [6]. Since its first introduction, trials have been conducted jointly by operators and academic consortia with the expectation of decreasing interference, increasing resource utilisation, and lowering energy consumption [7]. In a C-RAN, Remote Radio Heads (RRHs) are adopted as radio front ends performing radio functions to provide signal coverage. A Baseband Unit (BBU) pool is placed at a centralised site, where BBUs are aggregated and deployed at multiple general purpose servers [6]–[9]. C-RANs are preferable in hotspot areas since it has co-located BBUs in one pool, where mechanisms to increase spectral efficiency and throughput, such as enhanced Inter-cell Interference Coordination and Coordinated Multi-Point, can be simply implemented [10], [11]. RRHs and the BBU pool can be connected via optical fibres, especially in indoor scenarios, which is depicted in Fig. 1.

As real traffic distributions are temporally variable in reality, the centralisation of the BBUs provides the advantage that some servers can be switched to a sleep mode for energy saving when the overall demand for processing resources is low [10], [12], [13]. Just like base stations in conventional architectures, RRHs can also be switched to low power sleep modes due to temporal and spatial traffic variations in a C-RAN. What is more, the sleep mode operation can also be optimised from a global view of a C-RAN thanks to its centralisation advantage.

With the application of sleep modes at RRHs, there may be fewer active access nodes in a network, which results in a larger average path loss between User Equipment (UE) and a RRH. This forces the average transmission power of UEs to be higher than that without sleep mode application. By introducing much smaller cell sizes, the average transmission power of UEs has been reduced and its increase is minor compared with the significant reduction in network power consumption. On the other hand, the increase in UE transmission power can be reduced by placing the active RRHs nearer to the hotspot areas with high traffic levels. This objective is rarely considered in the designs of sleep mode operation in the literature. In this paper, the trade-off is analysed based on realistic stochastic geometry models. Hotspot-oriented Green Frameworks are proposed, in which different clustering strategies are adopted at central controllers, managing sleep mode operation from a global view and placing active RRHs closer to hotspot areas.

Z. Li was with the Department of Electronics, University of York, York, YO10 5DD, U.K.

D. Grace and P. Mitchell are with the Department of Electronics, University of York, York, YO10 5DD, U.K.

E-mail: {zhehan.li, david.grace, paul.mitchell}@york.ac.uk

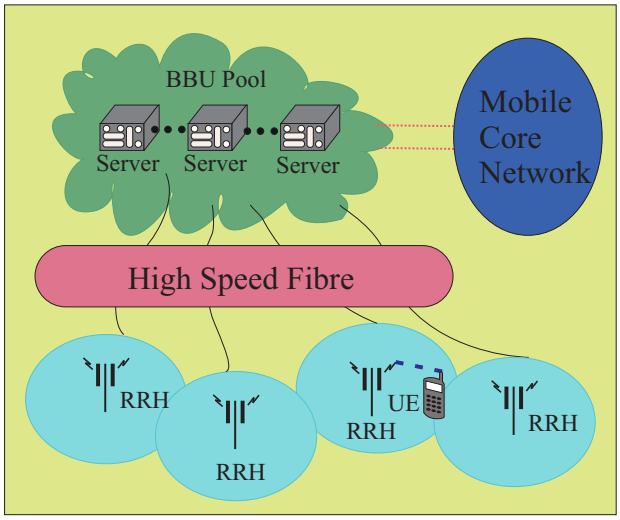


Fig. 1. The C-RAN architecture.

A. Related Work

For conventional base station types, the sleep modes are low power operational states where part or all of the hardware modules of a radio transceiver or a node are shut down. Given the centralisation of BBUs, RRHs only have limited hardware modules, e.g. radio frequency (RF) modules, operating as radio fronts. They are managed by the centralised controller and can be selectively switched off, yielding low sleep mode power [8], [13]. With the exemption of co-located BBUs, ultra-small cell RRHs have different power consumption features. The results of sleep mode operation for conventional base stations in the literature might not be valid and therefore should be re-investigated with appropriate power consumption models as adopted in this paper.

Sleep mode operation algorithms can be classified into static, slow reaction and fast reaction schemes. In a conventional large macrocell, since the number of UEs in the cell is usually big, the traffic volumes in different periods usually conform to a long-term distribution with fewer variations. For this scenario, static (e.g. [14]) or slow reaction (e.g. [5]) sleep mode control schemes may be sufficient to respond to the traffic variations. While in small cell scenarios investigated in this paper, due to the significant reduction in UE quantity in each cell, a fast varying traffic volume can be observed, which may result in a much more unbalanced traffic distribution across the small cells. Thus, static and slow reaction algorithms are too limited in the ultra-small cell C-RANs investigated in this paper. Fast reaction algorithms are required to control base station state transitions, which can be achieved by light weight small cell RRHs.

Considering the fast reaction feature, we noticed that there are some sleep mode operation algorithms that have been already proposed for conventional distributed architectures. For example, in [15]–[17], sleep mode control schemes are designed for conventional cellular networks, where some information formats (such as load information, coverage information and UE handover information) are exchanged among neighbour base stations for decision making. In these schemes, deactivation decisions are made based on the exchanged information, while activation decisions are made by active base stations for base stations in sleep modes. However, the distributed algorithms diminish the centralisation gain provided by C-RANs and thus are not appropriated in our considered situation.

Centralised algorithms in the literature are the fair counterparts to the work presented in this paper. Ideal centralised schemes are expected to be able to keep an appropriated number of active RRHs at locations with high traffic levels. In a recent advancement of centralised algorithms in [18], the minimum active base station density is determined by considering user resource allocations and overload probability based on stochastic geometry models. However, it mainly relies on random selection when choosing sleeping base stations without considering hotspot areas. In [19], authors develop sleep mode strategies to optimise energy efficiency taking hotspot areas into account, but regard hotspots areas as regions with many users instead of considering the actual traffic volumes. This may mislead the placement of active RRHs and yield higher UE transmission power due to larger path loss. The problem of the energy consumption trade-off between base stations and UEs is raised in [20]. However, to the best of our knowledge, there is still a lack of specific sleep mode operation strategies that target the minimisation of the increase in UE transmission power while maximising the network power reduction for ultra-small cell C-RANs under different conditions. This becomes the main objective of the frameworks proposed in this paper.

As the Hotspot-oriented Green Frameworks proposed in this paper are based on sleep mode operation involving clustering strategies, the application of clustering techniques in the wireless communication domain is also briefly reviewed. Clustering is a method of distinguishing objectives into groups. The objectives in the same group have more similar features than those in other groups. In wireless communications, clustering can be used to group network components or other network resources based on their radio environments or parameters for further utilisation. Clustering techniques are widely used in ad-hoc networks to guarantee basic performance achievements [21], [22]. Some work has also been done to transplant the conventional clustering techniques from the ad-hoc network domain to the cellular network domain. In [23]–[25], clustering is applied to femtocell networks, where femtocell base stations are grouped into clusters and further intra-cluster resource management can then be implemented. Clustering techniques have also been applied to distributed architectures for sleep mode operation. In [26]–[28], base stations in a network are formed into clusters so that some formats of information can be exchanged within clusters for sleep mode control. However, there is still no known work done to cluster quantified traffic distributions for sleep mode operation in a fully centralised architecture such as a C-RAN, which is realised in the research presented in this paper.

B. Main Contributions

In sleep mode operation, the number of active RRHs should be just enough to maximise the power reduction while

maintaining the Quality of Service (QoS). It is also important to place active RRHs in hotspot areas to reduce the average UE transmission power. The above objectives are rarely taken account of in combination in the literature, but are considered and guide the main idea of the work presented in this paper, that is utilising clustering techniques and quantified traffic information to put an appropriate number of active RRHs at hotspot areas. The main contributions of the paper are summarised as follows:

- *Problem analysis of hotspot areas:* Sleep mode operation for hotspot areas is analysed with a realistic C-RAN model, based on which the average area power consumption and the average UE transmission power are given. This approach has not been considered to analyse sleep mode operation in the literature. The analysis is dedicated to the objective of reducing the average UE transmission power considering hotspot areas, highlighting the key point of sleep mode operation. It also provides a benchmark for the average UE transmission power. The solution to accommodating the hotspot areas, clustering quantified traffic information, is revealed.
- *Hotspot-oriented Green Frameworks:* Novel frameworks involving sleep mode operation are proposed for C-RANs with different information availabilities. The number of the active RRHs is determined by the central controller to provide just enough radio resources to the C-RAN from a global view. In each framework, a distinct format of information is collected to quantify the local traffic of the RRHs in a C-RAN, which helps the centralised controller put active RRHs where the local traffic levels are high. Clustering techniques are adopted to locate the hotspot areas. In this way, local service demands are satisfied, maintaining the QoS, and the increase in the average UE transmission power is minimised, the combination of which has not been achieved in the state-of-the-art. Through simulation, the comprehensive performance of the frameworks in terms of QoS, power consumption and system overheads are demonstrated and compared with a recent baseline strategy [19] that deals with hotspot areas.

The remainder of the paper is organised as follows. In Section II, the C-RAN model and other system models are presented. The problem of hotspot areas is analysed with a solution proposed in Section III. In Section IV, Hotspot-oriented Green Frameworks involving clustering-based sleep mode operation are proposed. Implementation environments of the frameworks and trade-offs are described. The performance of the frameworks are presented in Section V, where one of the frameworks is also compared with a baseline strategy. Finally, the conclusions are drawn in Section VI.

II. System Models

A. Network Architecture

The C-RAN investigated is assumed to be deployed in a square area $\mathcal{A} \subset \mathbb{R}^2$, where RRHs are connected to a centralised BBU pool. Such types of C-RANs can be deployed in areas with high population densities like shopping malls, airports, train stations and exhibition centres for capacity enhancement. In the practical deployment of ultra-small cells, the access nodes are not usually located on an ideal grid pattern and have random placements instead. Therefore, stochastic geometry models can be utilised to introduce the randomness for the locations of the RRHs. Conventionally, the 2D homogeneous Poisson Point Process (PPP) is adopted to model the spatial distribution of the RRHs in an ultra-small cell C-RAN [29]. It is widely used because of its mathematical tractability, but also limited by its disadvantage that places the RRHs in the process close to each other. To avoid unrealistic modelling, repulsive dependent thinning is applied to the parent homogeneous PPP Φ_p of intensity λ_p, resulting in a type II Matérn Hard-core Point Process (MHPP) Φ_m of intensity λ_m and a hard-core distance δ [30]. Spatial correlations in practical deployments are better captured due to the existence of the hard-core distance in MHPP, which is demonstrated to be a more accurate stochastic model in [31] compared with PPP. An increasing number of interests from researchers yields studies on network performance analysis employing MHPP such as in [32], [33] and there is also work using MHPP to model C-RANs [34]. More comprehensive surveys discussing about difference stochastic models can be found in [35]. An example layout of the RRHs in the C-RAN is shown in Fig. 2. The small circles representing RRHs are generated based on a MHPP with λ_p equal to 20000/km^2 and δ equal to 5 metres, respectively. The side length of the square area $\mathcal{A}$ is set to 100 metres. Other elements of the figure are illustrated in the following subsection. The homogeneous PPP model with the possibility of having arbitrarily close RRHs is also considered to verify the effectiveness of the frameworks on different RRH layouts. Intensity of the PPP should be set to λ_m to obtain a fair comparison.

B. UE Distribution Model

The UE distribution in the C-RAN is modelled in two tiers, the background tier and the hotspot tier, to yield an unbalanced spatial distribution as in reality. The UE density $\lambda_\mathrm{u}(A)$ is thus variable versus A. The average density of all active UEs is denoted by $\overline{\lambda_\mathrm{u}(A)}$, in which active hotspot UEs and active background UEs account for γ and $1-\gamma$, respectively. In the background tier, background UEs are randomly and uniformly scattered in $\mathcal{A}$ while in the hotspot tier, hotspot UEs are equally divided into 5 groups, which are centred at group centres. UEs in each group independently conform to a 2D truncated (into $\mathcal{A}$) normal distribution with 5-metre standard deviations and their mean is defined as the group centre. The group centres are also randomly and uniformly scattered in $\mathcal{A}$ with at least a 5-metre distance to the network boundary and a minimum inter-centre distance of 25 metres. For the purpose of introducing greater hotspot traffic variation, the simulation is equally split into 10 periods with different traffic distributions. During each period, the group centres are re-located and chosen from 10 potential locations. That is equivalent to muting some hotspot UEs and activating some other hotspot groups at different locations in a new period.

The network example in Fig. 2 shows the UE distribution in a period where UEs are generated by setting $\overline{\lambda_\mathrm{u}(A)}$ to

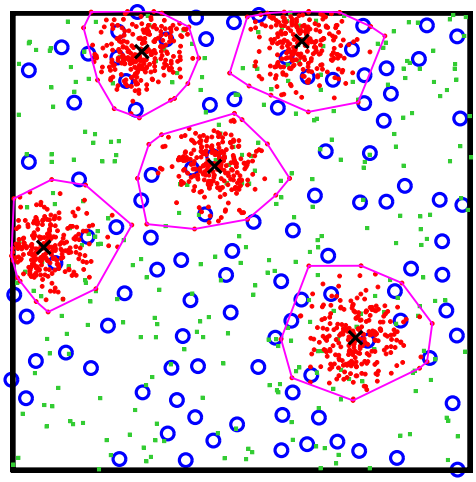


Fig. 2. A C-RAN example: The small circles representing the RRHs are generated according to a MHPP in the square area $\mathcal{A}$ with the side length set to 100 metres. The background UEs are represented by squares and the hotspot UEs are depicted as dots. The convex hulls of hotspot UEs in each group are depicted as polygons and the group centres are denoted by crosses.

150000/km^2 and γ to 0.8. The background UEs are represented by squares and the hotspot UEs are depicted as dots. The convex hulls of hotspot UEs in each group are depicted as polygons and the group centres are denoted by crosses.

C. Traffic Model

The temporal traffic distribution is modelled using FTP traffic model 2 suggested in [36]. Therefore, every active UE has file arrivals following a Poisson process and requests service from its associated RRH when starting to transmit a file. As in [36], the file sizes are all fixed to 4 Mbits. To simulate the system performance at different traffic load levels, the average file arrival rate is varied to achieve various offered traffic levels. The FTP traffic model 1 is also considered for the verification under different types of traffic. In this model, file arrival is assumed to occur in the whole network following a Poisson process, where half of the files are randomly chosen to be 16 Mbits and the others remain 4 Mbits. The optional 16 Mbits file size is also suggested in [36].

D. Spectrum Allocation

As only uplink simulation is used to evaluate the system level performance, the uplink system bandwidth is set to be 20 MHz with a carrier frequency of 3.5 GHz. The uplink system bandwidth is divided into 100 Physical Resource Blocks. Each of them has 180 kHz bandwidth and is mapped to a Virtual Resource Block (VRB) for resource allocation. The type 0 uplink resource allocation as standardised in [37] is selected. All RRHs are allowed to use the whole system bandwidth and the spectrum resources are assumed to be allocated based on a spectrum bundle basis, which consists of 4 consecutive VRBs. Each UE is associated with the nearest active RRH in terms of path loss and requests a resource bundle with the highest signal-to-interference-plus-noise ratio (SINR) for data transmission when a file arrives. A file transmission request is blocked when the associated RRH has no resource bundle having SINR over 5 dB. After a successful resource is acquired, the UE occupies the resource bundle until completing transmitting the file to its associated RRH.

E. Antenna and Link Model

All RRHs are assumed to be deployed with antennas which generate 2D omnidirectional radiation patterns with 5 dB gains while UEs are assumed to have antennas creating isotropic radiation patterns. In the assumed C-RAN deployment scenario such as airports and train stations, the WINNER II B3 model is adopted to model the path loss between a RRH and a UE [38]. Shadowing is considered to be log-normal with the standard deviations of 3 dB and 4 dB for line-of-sight (LOS) and non-line-of-sight (NLOS) paths, respectively, where the corresponding LOS probability model for B3 is applied.

The noise floor is selected to be -96 dBm based on the assumption of 300 K temperature, a 5 dB noise figure at the RRHs, and the whole system bandwidth. Open-loop power control is applied in the uplink data transmission channel, where the target SNR β is set to 25 dB. The link level performance is evaluated using the truncated Shannon bound as recommended in [39].

F. Power Consumption Model

To quantify the energy saving performance, the RRH power consumption model developed in [9] is adopted. The RRH instantaneous power depends on the RRH state, being either in the active or sleep mode. The active RRH power P_B as given in [9], is modelled as

$$P_B = \frac{N_{\text{ant}} \frac{BW}{10[\text{MHZ}]} P_{\text{RF}} + L_b P_{\text{PA-max}}}{(1-\sigma_{\text{DC}})(1-\sigma_{\text{MC}})}, \tag{1}$$

where N_{ant} and BW stand for the number of antennas used by a RRH and the system bandwidth. P_{RF} and $P_{\text{PA-max}}$ denote the RF base consumption and the maximum PA transmission power, respectively. σ_{DC} and σ_{MC} are the main supply losses factor and the DC-DC conversion losses factor. As previously defined in the Hotspot-oriented Green Framework, L_b is the percentage load of a RRH. Realistic values are given in Table I [9].

When a RRH goes into the sleep mode, the RF module and the PA module of the RRH do not need to be active for transceiving, and are assumed able to be turned off instantly. The instantaneous power of a sleeping RRH P_S thus falls to zero based on Equation (1) because the model given in [9] does not consider the power consumption from the backhaul connection and the basic processing unit. Therefore, P_S is modelled according to [40], which is 15% of P_B at zero loads. It is also suggested that the wake-up time is modelled. Since only the RF and PA modules have to be deactivated for a sleep mode, the wake-up time is chosen to be 0.5 seconds as given in [40].

TABLE I
POWER MODEL RELATED PARAMETER VALUES

N_{ant}	2	BW	40 MHz	P_{RF}	0.8 W
$P_{\text{PA-max}}$	0.8 W	σ_{DC}	6.4%	σ_{MC}	7.7%

With the power consumption of a RRH modelled, the equivalent instantaneous area power consumption of a C-RAN, $\widehat{P_{\text{t}}^{N}}$ considered in a period of time τ, can be calculated as

$$\widehat{P_{\text{t}}^{N}} = \frac{\int_0^{\tau} \sum_{b=1}^{|\mathcal{B}|} P_B \mathbf{1}_{\mathcal{B}^{\text{on}}}(B_b) + P_S \mathbf{1}_{\mathcal{B}^{\text{off}}}(B_b) \mathrm{d}t + n_{\text{on}} P_{\text{tran}}^{\text{on}} T_{\text{tran}}^{\text{on}}}{|\mathcal{A}|\tau}, \tag{2}$$

where $\mathbf{1}_{\mathcal{B}}(\cdot)$ is an indicator function of $\mathcal{B}$. $|\mathcal{B}|$ is the total number of the RRHs in a C-RAN, which may vary depending on the aforementioned network layout model. n_{on} denotes the times of switching on in the network during τ. $P_{\text{tran}}^{\text{on}}$ is power consumption of a RRH at the wake-up transient state while $T_{\text{tran}}^{\text{on}}$ represents the wake-up time.

To calculate the average transmission power of UEs in the simulation, $\widehat{P^{\text{u}}}$ is defined as

$$\widehat{P^{\text{u}}} = \frac{\sum_{m=1}^{M} \sum_{f=1}^{F_m} P_{m,f} t_{m,f}}{\sum_{m=1}^{M} \sum_{f=1}^{F_m} t_{m,f}}, \tag{3}$$

where $P_{m,f}$ is the transmission power of UE$_m$ during the transmission of the file f and $t_{m,f}$ is its corresponding elapsed time during file transmission. F_m is the number of files transmitted by UE$_m$ during the simulation and $M = \int_{\mathcal{A}} \lambda_{\text{u}}(A)\mathrm{d}A$ is the number of UEs in $\mathcal{A}$.

III. Problem Analysis and Solution

A. MHPP Features

Φ_{m} resulted from repulsive thinning guarantees that the minimum pairwise distance of RRHs are no less than δ. The retaining probability of the points in Φ_{p} is [41]

$$p = \frac{1 - \exp(-\lambda_{\text{p}} \pi \delta^2)}{\lambda_{\text{p}} \pi \delta^2}. \tag{4}$$

The intensity of Φ_{m}, λ_{m}, is then $\lambda_{\text{p}} p$. The contact distance distribution $F(r)$ of Φ_{m} is defined as the cumulative density function of the distance between a generic point $u \in \mathbb{R}^2$ and a point generated by a Φ_{m}. For type II MHPP, $F(r)$ for $0 \leq r \leq \delta/2$ can be derived [42] while it does not have a closed-form expression for $0 \leq r < \infty$. In [42], [43], the authors resort to approximation approaches to express the full range of $F(r)$. To provide a tractable and straightforward closed-form function for $F(r)$, the approximation of $F(r)$ for $0 \leq r < \infty$ is simplified as

$$F(r) = 1 - \exp(-qp\lambda_{\text{p}} \pi r^2), \tag{5a}$$

where

$$q = 1 + \frac{p}{2(1+p)} \lambda_{\text{p}} \pi \delta r. \tag{5b}$$

For networks modelled by Φ_{m}, the derivative of $F(r)$, $f(r)$, which is the probability density function of the distance between u and its nearest RRH, can capture the distance between a UE of an arbitrary location and a RRH in a C-RAN. The high accuracy approximation of $F(r)$ presented can then be used for calculating the average UE transmission power in the following analysis.

B. Problem of Hotspot Areas

Considering a C-RAN implemented with a two-state sleep mode operation, the RRHs are represented as $\mathcal{B} = \{B_b\} = \mathcal{B}^{\text{on}} \cup \mathcal{B}^{\text{off}}$ ($b \in \mathbb{N}^*, b \leq |\mathcal{B}|$), where active RRHs are denoted as $\mathcal{B}^{\text{on}}$ and $\mathcal{B}^{\text{off}}$ are RRHs in sleep modes. For a C-RAN modelled by Φ_{m}, sleep mode operation can be treated as a second thinning process based on $\Phi_{\text{m}} \cap \mathcal{A}$. The corresponding retaining probability $\eta_t(A)$ at time t depends on the sleep mode operation strategy adopted then and the exact location $A \in \mathcal{A}$. Naturally, it relates to the proportion of the active RRHs and thus area power consumption of the C-RAN $\overline{P_t^{\text{N}}}$ contributed by the RRHs is

$$\overline{P_t^{\text{N}}} = \int_{\mathcal{A}} \frac{\eta_t(A)\lambda_{\text{m}} P_{\text{B}} + \big(1 - \eta_t(A)\big)\lambda_{\text{m}} P_{\text{S}}}{|\mathcal{A}|} \mathrm{d}A, \tag{6}$$

where P_{B} and P_{S} are the RRH power consumption at the active mode and the sleep mode, respectively. As $P_{\text{S}} < P_{\text{B}}$, $\overline{P_t^{\text{N}}}$ can be reduced with lower $\eta_t(A)$. With the sleep mode operation, the resulting $f(r)$ can be adjusted to $f_t\big(r, \eta_t(A)\big)$ and should satisfy $\forall A \in \mathcal{A}, \int_0^{\infty} f_t\big(r, \eta_t(A)\big)\mathrm{d}r = 1$.

A generic UE $u \in \mathcal{A}$ in a typical C-RAN system implemented with the sleep mode operation described by $\eta_t(A)$ can be assumed to be served by its nearest active RRH with open-loop power control for uplink, of which the required signal-to-noise ratio (SNR) is β. For an arbitrary pattern of the active UE distribution knowing the location-dependent intensity $\lambda_{\text{u}}(A)$, the average uplink UE transmission power $\overline{P_t^{\text{u}}}$ can therefore be calculated as

$$\overline{P_t^{\text{u}}} = \frac{\int_{\mathcal{A}} \lambda_{\text{u}}(A) \int_0^{\infty} \beta \sigma^2 \alpha(r) f_t\big(r, \eta_t(A)\big) \mathrm{d}r \mathrm{d}A}{\int_{\mathcal{A}} \lambda_{\text{u}}(A) \mathrm{d}A}, \tag{7}$$

where σ^2 is the noise power and $\alpha(r)$ is the attenuation factor of the propagation path between u and its connecting RRH. For lower $\eta_t(A)$ to reduce $\overline{P_t^{\text{u}}}$, $f_t\big(r, \eta_t(A)\big)$ is skewed to larger r with a longer tail and yields higher expected UE transmission power $P_t^{\text{u}}(A)$ at A by a bigger integral with respect to r. Thus, $\overline{P_t^{\text{u}}}$ is raised owing to the reduced number of active RRHs.

With the application of sleep modes, the first operation target is to ensure that $\forall A \in \mathcal{A}$ the QoS (e.g. blocking probability, delay) at A is not degraded when $\eta_t(A) \neq 1$ compared with the QoS when $\eta_t(A) = 1$. With this constraint, $\overline{P_t^{\text{N}}}$ given by Equation (6) should be maximised. From a global view of a centralised controller, it has to first determine $\overline{\eta_t} = \int_{\mathcal{A}} \eta_t(A) / |\mathcal{A}| \, \mathrm{d}A$, which is the basic objective but usually difficult to solve without knowing the overall QoS when no sleep mode operation is applied.

The problem is more complicated by considering the heterogeneous local service demands that are likely to be caused by a highly unbalanced $\lambda_{\text{u}}(A)$ due to the existence of hotspot areas, $\eta_t(A)$ should accommodate various demands for different A and t even for a fixed $\overline{\eta_t}$. In terms of the average UE uplink

transmission power for a given $\overline{\eta_t} < 1$, the increase in $\overline{P_t^{\mathrm{u}}}$ due to sleep mode operation can be reduced with lower $P_t^{\mathrm{u}}(A)$ for A, where $\lambda_{\mathrm{u}}(A)$ is large, by having $f_t\left(r,\eta_t(A)\right)$ skewed to small r, i.e. with higher $\eta_t(A)$. This means that active RRHs should be put close to hotspot areas.

Owing to various UE distribution patterns, $\overline{P_t^{\mathrm{u}}}$ in a C-RAN cannot be determined if $\lambda_{\mathrm{u}}(A)$ is not accurately modelled. However, considering an arbitrary $\lambda_{\mathrm{u}}(A)$ distribution, the benchmark average UE transmission power $\widetilde{P_t^{\mathrm{u}}}$ can be obtained by replacing $\eta_t(A)$ in Equation (7) with $\overline{\eta_t}$ as

$$\widetilde{P_t^{\mathrm{u}}} = \int_0^\infty \beta\sigma^2\alpha(r)f_t(r,\overline{\eta_t})\mathrm{d}r. \tag{8}$$

It assumes a uniform selection of active RRHs in sleep mode operation, giving equal $P_t^{\mathrm{u}}(A)$ for all A. Taking it as the benchmark, an effective sleep mode operation strategy should make $\eta_t(A)$ accommodate to hotspot areas and achieve $\overline{P_t^{\mathrm{u}}}$ lower than $\widetilde{P_t^{\mathrm{u}}}$.

C. Solution

As analysed, sleep mode operation is required to place active RRHs at the appropriate hotspot places, reducing the average distance between active RRHs and UEs. Clustering techniques can be utilised to achieve the on-demand provisioning of radio resources, ensuring local QoS requirements and minimising the increased average UE transmission power.

Clustering is a technique for partitioning a set of d-dimensional data entries, $\mathcal{X} = \{x_i\}$, into k ($\leq |\mathcal{X}|$) subsets or clusters, $\mathcal{S}_j$ ($i,j \in \mathbb{N}^*, i,j \leq k$). Data entries are some formats of information under consideration and the objective of the clustering process is to find the similarities among such data entries. Then data entries are grouped based on the similarities, and cluster centroids are found to represent them.

Thanks to the above function, clustering can be used to partition the quantified traffic distribution pattern in a C-RAN. The traffic having similarity in geographical location can be represented by a cluster centroid, meaning that the data entry set $\mathcal{X}$ should store location information of the traffic. Cluster formation of quantified traffic helps select the locations of cluster centroids. Then in sleep mode operation, each cluster centroid can be used to determine the location of one closely located active RRH, expected to serve the quantified traffic in the corresponding local area. Therefore, sleep mode operation benefits from the final objective of the clustering process that is to find the proper cluster centroids in the format of locations rather than grouping data entries into clusters. By determining the locations of all active RRHs in this way, a minimised average distance between UEs and RRHs can be achieved and the average UE transmission power can be minimised as well.

As a widely used clustering algorithm, K-means can be adopted and customised to suit a C-RAN architecture for sleep mode operation. In the considered scenario with data representing locations, the purpose of the clustering process is to find the cluster centroids, where the sum of all the within-cluster distances between the cluster members and their respective cluster centroids are minimum. Therefore, by considering the distances as the Euclidean distances, the objective of a clustering process is to minimise the cost function defined as

$$J(\mathcal{S}_1,\ldots,\mathcal{S}_j,\ldots,\mathcal{S}_k,C) = \sum_{j=1}^{k}\sum_{x_i\in\mathcal{S}_j}\|x_i - c_j\|^2, \tag{9}$$

where $C = \{c_j\}$ and c_j is the cluster centroid of the cluster S_j. $\|x_i - c_j\|$ is the Euclidean distance between x_i and c_j.

At the start of the clustering process, a pre-determined number of (k) locations are chosen as the initial cluster centroids $c_j^{(0)}$, where the superscript denotes the iteration number, l ($l \in \mathbb{N}^*$). In iteration 1, each location data entry, x_i, is associated with one $c_j^{(0)}$, which has the smallest distance to x_i. All x_i associated with the same $c_j^{(0)}$ constitute a cluster $S_j^{(1)}$ and become its cluster members. After the cluster formation, the mean (of each dimension) of $x_i \in S_j^{(1)}$ is computed and each $c_j^{(0)}$ is replaced by the calculated mean, becoming $c_j^{(1)}$. To this point, iteration 1 is finished and each $c_j^{(1)}$ becomes the target centroid for association in the next iteration. This process starting from centroid association is repeated for a certain number of iterations or until the locations of the cluster centroids do not have great changes. In successive iterations, the cluster centroids are updated, moving to the locations better representing their corresponding cluster members, and the cost function, $J(\mathcal{S}_1,\ldots,\mathcal{S}_j,\ldots,\mathcal{S}_k,C)$, decreases until a stable level.

Being a promising solution, the application of the clustering technique for sleep mode operation in C-RANs has unsolved issues, which are solved in the proposed Hotspot-oriented Green Frameworks introduced in the next section.

IV. Hotspot-oriented Green Frameworks

As aforementioned, the basic idea of transplanting the clustering technique to sleep mode operation is mapping an active RRH to a cluster centroid, which represents the location of the local traffic. From a global view, the result is that the active RRHs are placed where the traffic is, i.e. closer to hotspot areas. Taking it as the basis, the number of the active RRHs in a C-RAN is equal to the number of the cluster centroids of a clustering process. The Hotspot-oriented Green Frameworks proposed in this section are designed to involve the application of the clustering technique for sleep mode operation, dedicated to C-RANs of different information availabilities. The frameworks solve some critical issues related to transplanting the clustering technique, which are presented as follows.

- *Quantifying the local traffic based on information availabilities:* The local traffic relative to each RRH in a C-RAN should be quantified at discrete locations as data entries of the clustering process. Practical information and infrastructure availabilities should be considered for different C-RAN architectures because the performance of the clustering process depends on the format of the information provided. Generally speaking, the traffic can be better captured with more location information available. The information and infrastructures needed in different environments are considered in the proposed frameworks.

- *Choosing an appropriate clustering frequency:* Clustering frequency refers to the frequency of running the clustering process at the centralised controller to adjust the RRH on-off states in a C-RAN. It also refers to the frequency of evaluating network conditions, e.g. variations of hotspot areas and changes of network load levels, from past experiences. A high cluster frequency usually makes a C-RAN respond to traffic variations more swiftly, while a low cluster frequency requires fewer computing resources and yields less RRH switching on-off overhead. The choices of the clustering frequency and the corresponding trade-offs are presented with the analysis of the simulation results.
- *Determining an appropriate number of cluster centroids:* As the number of cluster centroids in a clustering process is equal to the number of active RRHs in a C-RAN, it needs to be determined properly to reflect the overall traffic level so that just enough radio resources are provided, ensuring a similar QoS and maximum power consumption can be reduced. It can be determined by evaluating the average network load as mentioned later.
- *Determining appropriate initial cluster centroids:* The initial cluster centroids, $c_j^{(0)}$, should be appropriately pre-determined to start the clustering process. Through subsequent iterations in a clustering process, they affect the final choice of cluster centroids and therefore the locations of the active RRHs. An extensive adjustment of RRH states at one time instance may result in network instability and yield large system overheads. It can be avoided by initialising the clustering centroids based on the locations of the current active RRHs with some adjustments. More details will be given in the later explanation.
- *Adjusting cluster centroids to real RRH locations:* Since RRH locations in a C-RAN are spatially discrete while computed cluster centroids are continuous locations, the cluster centroids should be moved to real RRH locations at every iteration during a clustering process so that the clustering process can minimise the cost function based on real RRH locations. Furthermore, the case where multiple cluster centroids are adjusted to one real RRH location should be avoided in order to maintain the number of active RRHs equal to the wanted pre-determined number. To address this, virtual cluster centroids are introduced and are associated with real RRH locations during a clustering process. This process will be described in detail later.

A. Complete Framework

In the *Complete Framework*, the aforementioned local traffic distribution is depicted and quantified based on the UE location information. Due to the spread of Location Based Services, UE location information is much easier to obtain and therefore can be utilised. The location of a UE can be estimated by the Global Navigation Satellite System in outdoor scenarios or other positioning technologies, such as those using Direction of Arrival and Time of Arrival information, in indoor scenarios.

UE location information can be reported and collected at the central controller to perform the clustering process because of the centralisation of a C-RAN. In the case where each UE has its location information available when requesting services, data entries $x_i \in \mathcal{X}$ of a clustering process are just the locations of the service requests. The clustering process then partitions the locations into clusters and the locations correspondingly become the cluster members. The location information of the RRHs, $\mathcal{H} = \{h_b\}$, is acquired through manual measurements by operators and recorded in the controller when deployed. In a self-organised network, they can also be self-measured during self-configuration and may be self-optimised through a long-term measurement.

The clustering frequency is defined as $1/T$, where T is the clustering period and the central controller runs the clustering algorithm every T seconds. During the clustering period, the location information of UEs is collected by the active RRHs delivering services only when UEs request services and is stored at the corresponding RRHs. Meanwhile, load variations are monitored at each RRH. At the end of a clustering period, the average load $\overline{L_b}$ of $B_b \in \mathcal{B}^{\text{on}}$ in the last T seconds is estimated as

$$\overline{L_b} = \int_0^T \frac{L_{b,t}}{T} \mathrm{d}t, \tag{10}$$

where $L_{b,t}$ is the load of B_b at time t. Then, $\overline{L_b}$ of all active RRHs are sent to the central controller as well as each stored location, which is treated as a data entry, x_i, for the next clustering process. The average overall load $\overline{L}$ of the C-RAN in the last T seconds is estimated as

$$\overline{L} = \sum_{B_b \in \mathcal{B}^{\text{on}}} \frac{\overline{L_b}}{|\mathcal{B}^{\text{on}}|}. \tag{11}$$

The number of the cluster centroids of the next clustering process, k, i.e. the number of the active RRHs on completion of the next clustering process, depends on $\overline{L}$, the number of the current active RRHs in the C-RAN, $|\mathcal{B}^{\text{on}}|$, and a RRH load reference, L_{ref}, based on the principle as

$$k \leftarrow \min(\lceil \frac{\overline{L}\,|\mathcal{B}^{\text{on}}|}{L_{\text{ref}}} \rceil, |\mathcal{B}|). \tag{12}$$

The principle yields a larger k when $(\overline{L} > L_{\text{ref}}) \wedge (|\mathcal{B}^{\text{on}}| < |\mathcal{B}|)$ and a smaller k when $(\overline{L} < \frac{|\mathcal{B}^{\text{on}}|-1}{|\mathcal{B}^{\text{on}}|} L_{\text{ref}}) \wedge (|\mathcal{B}^{\text{on}}| > 1)$. In other cases, k remains the same. This is equivalent to the fact that more RRHs will be active when the overall load of the RRHs is above a threshold and some RRHs will be switched to sleep mode when the overall load of the RRHs is below a threshold. If the calculated average load of the clustering period is between these two thresholds, the number of the active RRHs does not change. No matter which case it is, the locations of the active RRHs after a clustering process may change based on the data entries provided.

When the clustering starts with the pre-determined k clusters (cluster centroids), the locations of the current $\mathcal{B}^{\text{on}}$ are considered as the candidate cluster centroids. If the number of the active RRHs are intended to increase, $k - |\mathcal{B}^{\text{on}}|$ more cluster centroids, which are the locations of $\mathcal{B}^{\text{off}}_{k-|\mathcal{B}^{\text{on}}|}$, are randomly chosen from the locations of the current $\mathcal{B}^{\text{off}}$. Together with

the candidate cluster centroids, they form the initial cluster centroids $C^{(0)}$. On the contrary, if some RRHs are to be deactivated, $|\mathcal{B}^{\text{on}}| - k$ cluster centroids are randomly chosen and removed from the candidate cluster centroids. The rest of them, the locations of $\mathcal{B}_k^{\text{on}}$ (as a subset of $\mathcal{B}^{\text{on}}$), form $C^{(0)}$. If the number of the active RRHs remains the same, the candidate centroids are selected as $C^{(0)}$ for the cluster centroid updates through the iterations in a clustering process. Thus, the locations of the active RRHs on completion of the clustering process are based on the locations of the active RRHs before the clustering process to avoid a significant adjustment of the RRH on-off states and additional system overheads.

As RRHs in a C-RAN are given discrete spatial locations $\mathcal{H} = \{h_b\}$ when deployed, the updated cluster centroids of continuous locations should be linked to $\mathcal{H}$ of the C-RAN. To ensure that the clustering process is minimising the cost function, $J(\mathcal{S}_1,\ldots,\mathcal{S}_j,\ldots,\mathcal{S}_k,C)$, based on the actual C-RAN layout, h_b should be taken into consideration for updating the cluster centroids c_j in every clustering iteration l. Thus, instead of directly replacing $c_j^{(l-1)}$ in iteration l with the computed within-cluster means of all cluster members x_i in a cluster $\mathcal{S}_j^{(l)}$, they are treated as the virtual cluster centroids, $v_j^{(l)} \in \mathcal{V}^{(l)}$ as

$$v_j^{(l)} = \frac{\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i}{\left|\mathcal{S}_j^{(l)}\right|}. \tag{13}$$

For each $v_j^{(l)}$ of $\mathcal{S}_j^{(l)}$, the closest (in terms of the Euclidean distance) h_b is assigned as $c_j^{(l)}$ of iteration l. If there are multiple $v_j^{(l)}$ sharing the same h_b, $c_j^{(l-1)}$ are not updated to $c_j^{(l)}$ in iteration l to avoid merging multiple cluster centroids into one. Otherwise, there will be fewer active RRHs than the pre-determined number of cluster centroids. For example, if at iteration 1, the closest RRH locations of the virtual centroids, $v_3^{(1)}$ and $v_5^{(1)}$, are both h_7, $c_3^{(1)}$ and $c_5^{(1)}$ roll back to $c_3^{(0)}$ and $c_5^{(0)}$, respectively. In this way, at the start of every iteration, the association of data entries to the clusters is also based on real RRH locations. As a result, the cluster formation process is essentially grouping quantified traffic around RRH physical locations, which are determined by traffic data entries in the format of service request locations.

Before the end of iteration l in a clustering process, the cost function, $J^{(l)}(\mathcal{S}_1^{(l)},\ldots,\mathcal{S}_j^{(l)},\ldots,\mathcal{S}_k^{(l)},C^{(l)})$, is calculated and compared with the cost function of iteration $l-1$, $J^{(l-1)}$. The clustering iteration may cease at this point if the difference between two cost functions, $|J^{(l)} - J^{(l-1)}|$, is smaller than 1% of $J^{(l-1)}$, or if $l = 10$. This is to guarantee that the clustering process has reached the state where the average distance between data entries and their cluster centroids are reduced to a relatively stable level. Through simulation, it is observed that more iterations do not significantly decrease the final values of the cost function.

After the clustering process, deactivation signals are sent to the active RRHs if the final cluster centroid set does not contain their locations, and then the corresponding active RRHs are switched to sleep modes. Activation signals are transferred to the sleeping RRHs if their locations belong to the final cluster centroid set, and the corresponding sleeping RRHs are switched to active modes. The data entry memory is cleared for the next clustering period and the timer is set to zero. For the active RRHs to be deactivated, it stops admitting UEs and remains active until the data transmissions of all the associated UEs are finished. After state transitions, all the active RRHs expand or shrink their cell sizes to ensure coverage and the UEs re-associate to the closest active RRHs in terms of path loss.

Through such cycles, the centralised controller dynamically adapts the RRH states to the overall RRH loads and variations of the hotspot areas. The clustering process is run based on a massive amount of data collected, resulting in a highly accurate inference of the hotspot areas with a large number of service requests in a clustering period. The active RRHs can always be placed where needed because of the complete information collected, reducing the average distance between UEs and RRHs, which can be regarded as an upper bound of the reduction in the increased average UE transmission power. The framework is recommended to be used when the location information of UEs are available when requesting services. The pseudo-code of the *Complete Framework* is given in **Algorithm 1**.

B. Load Weighted Framework

In some scenarios such as indoor environments, the complete availability of accurate UE locations does not always exist considering the current technologies and the protocols. In these cases, a C-RAN can tolerate the inaccuracy or the incompleteness in UE location information. Otherwise, leaving the *Complete Framework* as a strong future candidate, a weighting strategy is considered in the *Load Weighted Framework* under the circumstance where only the locations of RRHs in the C-RAN are available. The *Load Weighted Framework* differs from the *Complete Framework* in its computation process of cluster centroids and the rest, including the determination of the number of active RRHs using the average overall load, remains the same.

Although sleep mode operation benefits from the proper locations of cluster centroids rather than grouping UEs into clusters, the loss of UE location information affects the local traffic quantification, which should be realised with other available formats of information to capture the traffic variations. In the *Load Weighted Framework*, $\mathcal{X}$ are formed by the locations of the active RRHs in the current clustering period as $\{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$. Since the average load of a RRH in a period of time $\overline{L_b}$ indicates the number of service requests received by the RRH, it can be used to represent its local traffic level during that time. By weighting the locations of $\mathcal{B}^{\text{on}}$ with their respective average load levels, the distribution of the traffic in $\mathcal{A}$ served by $\mathcal{B}^{\text{on}}$ can be roughly quantified. Taking this into consideration at the beginning of a clustering process, each $x_i \in \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$ is assigned a weight, $w_i \in \mathcal{W}$. The weight value is equal to $\overline{L_b}$ of B_b, which is defined in Equation (10) and collected by the centralised controller at the end of a clustering period.

Representing the traffic levels at areas served by $\mathcal{B}^{\text{on}}$, $\mathcal{W}$ fundamentally biases a clustering process, moving cluster

Algorithm 1 Complete Framework

for all $B_b \in \mathcal{B}^{\text{on}}$ **do**
  Report $\overline{L_b} \leftarrow \int_0^T \frac{L_{b,t}}{T} \mathrm{d}t$ to the central controller when required
  Report the locations of received service requests to the central controller when required
  if Receive deactivation signal from the controller **then**
    Switch to the sleep modes
  end if
end for
for all $B_b \in \mathcal{B}^{\text{off}}$ **do**
  Do nothing
  if Receive activation signal from the controller **then**
    Switch to the active modes
  end if
end for
for the central controller **do**
  Run the timer
  if T seconds has elapsed **then**
    Request $\overline{L_b}$ from $B_b \in \mathcal{B}^{\text{on}}$
    Request the locations of service requests from $B_b \in \mathcal{B}^{\text{on}}$ and record as x_i, $\mathcal{X} \leftarrow \mathcal{X} \cup \{x_i\}$
    $\overline{L} \leftarrow \sum_{B_b \in \mathcal{B}^{\text{on}}} \frac{\overline{L_b}}{|\mathcal{B}^{\text{on}}|}$
    $k \leftarrow \min(\lceil \frac{\overline{L}|\mathcal{B}^{\text{on}}|}{L_{\text{ref}}} \rceil, |\mathcal{B}|)$
    $l \leftarrow 0$
    if $k > |\mathcal{B}^{\text{on}}|$ **then**
      $\mathcal{C}^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in (\mathcal{B}^{\text{on}} \cup \mathcal{B}^{\text{off}}_{k-|\mathcal{B}^{\text{on}}|})\}$
    else if $k < |\mathcal{B}^{\text{on}}|$ **then**
      $\mathcal{C}^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}_k\}$
    else
      $\mathcal{C}^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$
    end if
    $J^{(l)} \leftarrow \infty$
    Do nothing
    repeat
      $l \leftarrow l + 1$
      $\forall x_i \in \mathcal{X}, x_i \mapsto \arg\min_{c_j^{(l-1)} \in \mathcal{C}^{(l-1)}} \|x_i - c_j^{(l-1)}\|^2$
      $\mathcal{S}_j^{(l)} \leftarrow \{x_i \in \mathcal{X} : x_i \mapsto c_j^{(l-1)}\}$
      $v_j^{(l)} = \frac{\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i}{|\mathcal{S}_j^{(l)}|}$
      $c_j^{(l)} \leftarrow \arg\min_{h_b \in \mathcal{H}} \|v_j^{(l)} - h_b\|^2$
      $\mathcal{C}^{(l)} \leftarrow \{c_j^{(l)}\}$
      $\forall c_j^{(l)} \in (\mathcal{C}^{(l)} - c_j^{(l)}), c_j^{(l)} \leftarrow c_j^{(l-1)}$
      $J^{(l)}(\mathcal{S}_1^{(l)}, \ldots, \mathcal{S}_j^{(l)}, \ldots, \mathcal{S}_k^{(l)}, \mathcal{C}^{(l)})$
      $\leftarrow \sum_{j=1}^{k} \sum_{x_i \in \mathcal{S}_j^{(l)}} \|x_i - c_j^{(l)}\|^2$
    until $\frac{|J^{(l)} - J^{(l-1)}|}{J^{(l-1)}} < 1\% \vee l \geq 10$
    Send activation signals to $\{B_b \in \mathcal{B}^{\text{off}} : h_b \in C^{(l)}\}$
    Send deactivation signals to $\{B_b \in \mathcal{B}^{\text{on}} : h_b \notin C^{(l)}\}$
    $\mathcal{X} \leftarrow \varnothing$
    Set the timer to 0
  end if
end for

centroids closer to the areas assigned with higher weights. This is achieved by computing a virtual cluster centroid $v_j^{(l)}$ in iteration l considering $\mathcal{W}$ after the formation of the cluster $\mathcal{S}_j^{(l)}$ as

$$v_j^{(l)} = \frac{\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i w_i}{\sum_{w_i \in \{w_i : x_i \in \mathcal{S}_j^{(l)}\}} w_i}. \tag{14}$$

In iteration l, all $c_j^{(l)}$ are chosen based on the biased $v_j^{(l)}$, increasing the probability of placing active RRHs at the areas with high traffic levels and reducing the average UE transmission power. It is worth mentioning that when the number of active RRHs is to increase after a clustering process, the iterations are skipped after the generation of $C^{(0)}$ due to more clustering centroids than the data entries. However, the location optimisation can be processed at the ends of the following clustering periods as long as $k \leq |\mathcal{B}^{\text{on}}|$. In this case, clusters are formed with locations of active RRHs as data entries, biased by their loads as weights.

The *Load Weighted Framework* does not require extra information as load variations are already obtained for the determination of the number of the active RRHs. It can compensate the loss of UE location information in some scenarios, where no additional traffic information except the RRH loads is available. The pseudo-code for the *Load Weighted Framework* is given in **Algorithm 2** with changes made based on the *Complete Framework*.

Algorithm 2 Load Weighted Framework (Changes)

3: Do nothing
18: $\mathcal{X} \leftarrow \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$
23: $\mathcal{C}^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in (\mathcal{B}^{\text{on}} \cup \mathcal{B}^{\text{off}}_{k-|\mathcal{B}^{\text{on}}|})\}$
  Go to Operation 41
30: $\mathcal{W} \leftarrow \{\overline{L_b} : B_b \in \mathcal{B}^{\text{on}}\}$
35: $v_j^{(l)} \leftarrow (\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i w_i)/(\sum_{w_i \in \{w_i : x_i \in \mathcal{S}_j^{(l)}\}} w_i)$

C. Energy Weighted Framework

Due to the loss of UE location information and the utilisation of a rough method of traffic quantification, the locations of the active RRHs generated by the *Load Weighted Framework* are sometimes inaccurate. To compensate for the loss of the information, power of the uplink transmission bands of the C-RANs from the ambient background of each RRH can be measured and recorded in a clustering period. Each RRH can calculate the integral of the discrete power samples with respect to time at the end of the clustering period, obtaining

$$E_b = \int_0^T P_{b,t}^{\text{r}} \mathrm{d}t, \tag{15}$$

where $P_{b,t}^{\text{r}}$ is the received uplink power of B_b sampled at time t. Considering that UE transmission power attenuates rapidly versus distance, E_b can represent the uplink energy virtually accumulated from the links in its local area during a period of time. Therefore, a large number of local links results in more energy accumulated, which indicates a high traffic level in its local area.

This strategy needs all RRHs to sample the uplink power at all times and report the quantified local traffic in the format of energy at the end of a clustering period, which require sleep modes to be designed with more functionality and may lead to a little higher power consumption. Alternatively, the accumulated energy can be realised with extra practical hardware modules such as energy harvesters. RF energy harvesters usually consist of passive receivers and power management modules to convert ambient RF power to stored electricity or power low-power devices. The application of RF energy harvesting has been widely considered for low power devices, such as wireless sensors [44], and may be generalised to cellular network devices [45]. However, the application of energy harvesters does not mainly aim to provide power source to RRHs, but to quantify the traffic levels in their local areas, such that it is not important whether the accumulated energy can support continuous operation of a RRH or not. With an energy harvester attached at each RRH, E_b is effectively the uplink energy scavenged at B_b during a clustering period, at the end of which E_b will be reported to the central controller.

Taking the same weighting strategy in the *Load Weighted Framework* without changing other parts, $\mathcal{E} = \{E_b\}$ is used for weighting the locations of the RRHs, $\mathcal{H}$, which are treated as $\mathcal{X}$ and form the clusters. Therefore, it is named the *Energy Weighted Framework*, where the weights of the data entries exactly take the value of the energy virtually accumulated or harvested at the corresponding RRHs, at the beginning of a clustering process. Similarly, the computation of virtual cluster centroids in Equation (14) is biased by the assigned weights. With more data entries enriching the information for the update of cluster centroids in every iteration, the final locations of active RRHs are usually more accurate than that in the *Load Weighted Framework*.

Although this solution needs additional functionality of sleep modes or extra hardware modules, most parts of the RRH hardware can still be deactivated and the energy calculated or harvested only needs to be reported when required at the end of a clustering period, which only introduce minor system overheads and power consumption. This framework is suitable when UE location information is lost, but with the aforementioned hardware functionality enabled. The changes of the pseudo-code based on that of the *Complete Framework* is given in **Algorithm 3**.

Algorithm 3 Energy Weighted Framework (Changes)

3: Report the energy virtually accumulated or harvested to the central controller when required
9: Report the energy virtually accumulated or harvested to the central controller when required
18: $\mathcal{X} \leftarrow \{h_b \in \mathcal{H}\}$
30: Require $\mathcal{B}$ to report $\mathcal{E} = \{E_b\}$, $\mathcal{W} \leftarrow \mathcal{E}$
35: $v_j^{(l)} \leftarrow (\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i w_i)/(\sum_{w_i \in \{w_i : x_i \in \mathcal{S}_j^{(l)}\}} w_i)$

D. Random Framework

Considering the simplest strategy without local traffic quantification, $\mathcal{X}$ is then not needed to calculate new cluster centroids. Therefore, cluster centroids are not determined via sequential iterations. After the initial cluster centroids are determined by the same method in the *Complete Framework* if the number of active RRHs are to be changed, the clustering process ceases at this point and outputs the initial cluster centroids as the final cluster centroids, i.e. $C^{(0)}$ are regarded as the locations of the active RRHs for the state adjustment. The rest of the scheme remains unchanged and it is named the *Random Framework* due to the random choices of cluster centroids. The clusters are randomly formed with only the cluster centroid itself in each cluster. It needs the central controller to collect only $\overline{L_b}$ at the end of a clustering period, without any extra information exchange. The pseudo-code is similar to that of the *Complete Framework* except the change listed in **Algorithm 4**.

In this framework, the central controller only needs to know the locations of the RRHs deployed in the C-RAN, the same as all the other proposed frameworks. It requires the least computational resource and information, but can only guarantee just enough active RRHs. Moreover, it relies on randomness to place active RRHs, which loses the advantage of placing active RRHs in areas with high service demands. It is presented only as a benchmark because it still features the power consumption reduction by sleep mode operation at the RRH side. However, the *Load Weighted Framework* is always preferred in terms of reducing UE transmission power when only the RRH load information is available at the central controller so that the *Random Framework* should not be used.

Algorithm 4 Random Framework (Changes)

3: Do nothing
18: Do nothing
30: **Go to** Operation 41
44: Do nothing

V. Performance Evaluation

The system level performance of the proposed Hotspot-oriented Green Frameworks are evaluated using Monte Carlo simulation based on the models introduced. To average the effect of the varying number of RRHs due to MHPP and stochastic traffic modelled by FTP model 2, 100 instances of C-RANs are taken for the simulation. The arguments of the frameworks are varied and appropriate values are chosen by observing the simulation results, where the QoS should be maintained similar to the system without sleep mode operation. The performance in reduction in area power consumption is then evaluated and analysed. The average UE transmission power and the generated system overheads of the frameworks are investigated. The effectiveness of the proposed frameworks is also validated based on the PPP model for the RRH distribution and FTP model 1 for traffic, showing wide applicability.

A. Comparisons with the Baseline Strategy

The *Complete Framework* is compared with a recent baseline scheme, a strategic sleeping scheme for the scenario with no UE movements presented in [19], which relies on

idealised models to optimise energy efficiency for determining the number of active nodes N_{on}. To deal with hotspot areas, it requires all RRHs to be temporally active when selecting active RRHs, which is based on the number of covered UEs with qualified SINR. As an analytical approach, it does not consider the practical implementation issues, e.g. the frequency of optimising node states. When transplanting it to the considered ultra-small cell C-RAN, it is assumed to run the optimisation every T seconds, the same as the proposed frameworks. Due to the complex models used in the simulation and the radio resource management in a different scenario, the optimisation of N_{on} in [19] is not applicable and is chosen to be equal to k. By doing these, the comparison can benefit from the controlled system variables and be focused on the mechanism dealing with hotspot areas.

To first ensure QoS, the *Complete Framework* and the baseline strategy are varied by setting $L_{\text{ref}} = 6\%$ and $T = 20$ s to have the same blocking probability and similar delay compared with the system without sleep mode operation as shown in Fig. 3. Although with the same average proportion of active RRHs, $\overline{\eta}$, by setting the same $L_{\text{ref}} = 6\%$, the *Complete Framework* outperforms the baseline strategy in delay because it better manages hotspot areas by putting active RRHs closer to them. This advantage provides more radio resource to areas in demand and yields a higher probability that a UE can acquire radio resources with higher SINRs, which accelerates the transmission rates and yields less delay.

The above superiority is echoed in the average UE transmission power as presented in Fig. 4 (a), where the *Complete Framework* reduces the UE transmission power by 70% (5.3 dB difference) at low traffic levels compared with the baseline strategy. The average path loss between UEs and their serving RRHs is reduced due to the same reason, which creates lower uplink transmission power on condition of open-loop power control and prolongs the battery lifetimes of mobile devices in a C-RAN with sleep mode operation. Although UE transmission power is increased compared with the no sleep mode operation, it is already very low when introducing ultra-small cells to C-RANs and increased power consumption at the UE side is minor compared with the large reduction in power consumption at the RRH side. As can be seen from Fig. 4 (b), the average area power consumption stemming from the RRHs can be reduced by more than 79% at traffic levels below 0.1 Gbps/km^2/MHz compared with the C-RAN with no sleep mode operation. The dashed line is the lower bound of the reduction in the total RRH power consumption, where all RRHs are turned into sleep modes.

In Fig. 4 (b), the baseline strategy consumes slightly higher power even with the same $\overline{\eta}$, this stems from the mechanism that all RRHs are required to be temporarily switched on to count the covered UEs in their local areas when the central controller is to decide the selection of active RRHs. As energy consumption per switch-on is suggested to be quantified, it is considered when calculating area power consumption and its proportion is shown in Fig. 5 (a), where the effect becomes severe at lower traffic levels and causes non-negligible wasted power. The *Complete Framework*, however, does not need sleeping RRHs to execute any other operations. It keeps the switching on-off frequency relatively low and stable as revealed in Fig. 5 (b). Therefore, it does not pose significant extra area power consumption and system overheads accompanying the switching on-off behaviours.

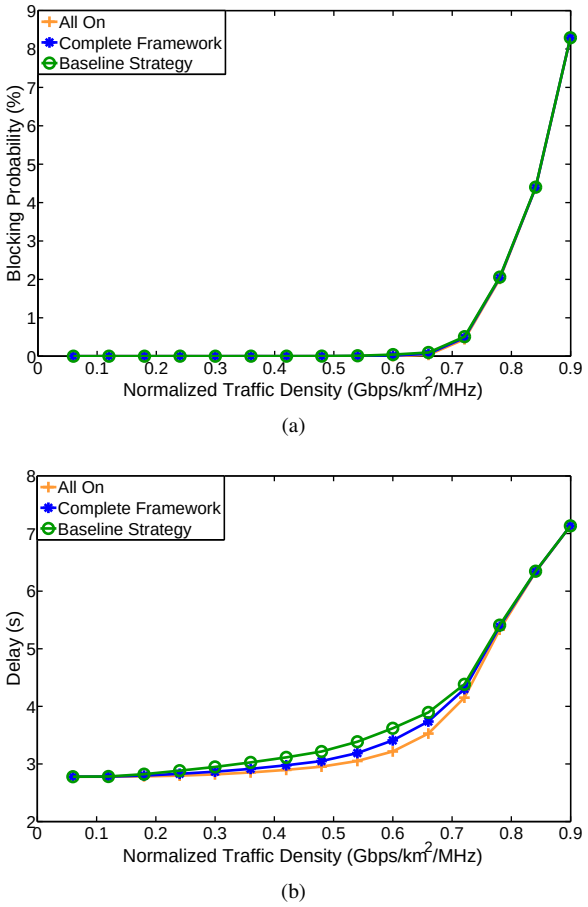
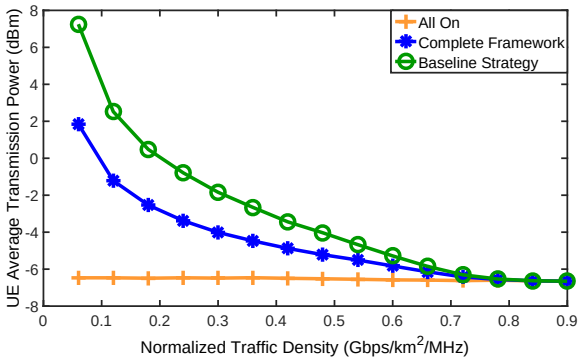


Fig. 3. The comparison of QoS between the *Complete Framework* and the baseline strategy. (a) Average blocking probability. (b) Average delay.

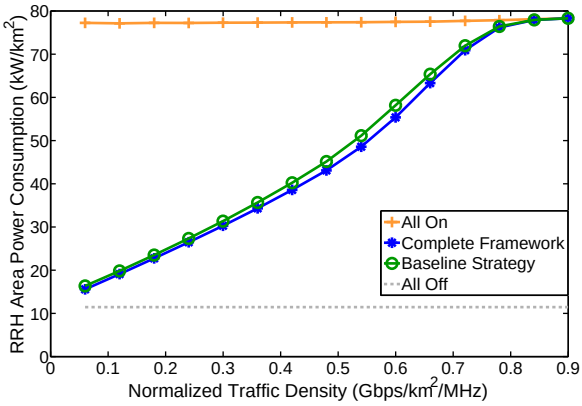
B. Trade-offs in Framework Implementation

In each of the proposed Hotspot-oriented Green Frameworks, only two parameters, L_{ref} and T, are to be determined. The fundamental objective of sleep mode operation is to maintain QoS while switching off some RRHs to save energy. If there are more active RRHs in the C-RAN, there are more radio resources in a certain area, which may lead to better QoS, but less power reduction. The RRH load reference, L_{ref}, is a parameter that can be selected to balance the trade-off. Fig. 6 and Fig. 7 show the trade-off between delay and area power consumption, selecting the *Load Weighted Framework* and $T = 20$ s as examples. As the blocking probability is maintained the same as the system with all RRHs active, it is not shown.

In Fig. 6, when L_{ref} is decreased, the delay result can be improved to be the same as the C-RAN without sleep mode operation, which is the upper bound of the QoS performance applying the proposed frameworks. The improvement is because the lower L_{ref} indicates a lower load level per RRH, which brings about more available spectrum resources at each

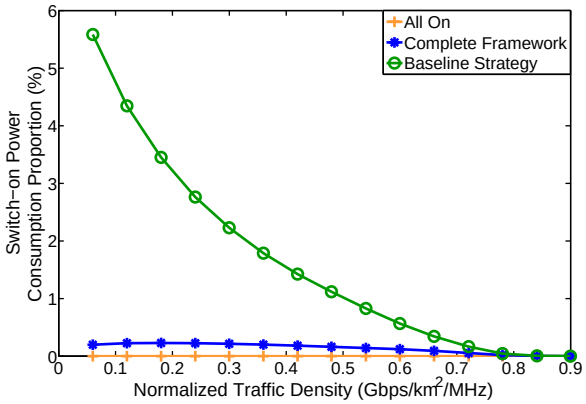


(a)

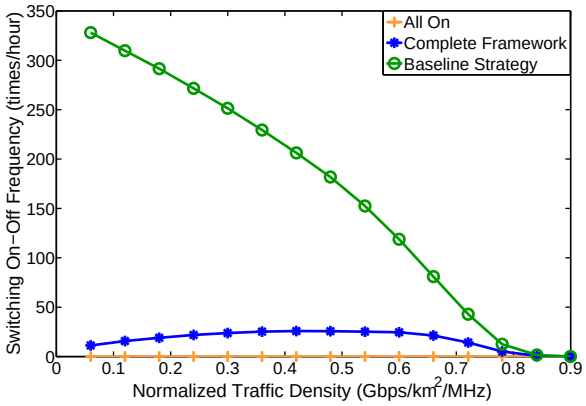


(b)

Fig. 4. The comparison of power consumption between the *Complete Framework* and the baseline strategy. (a) Average UE transmission power. (b) Average area power consumption.



(a)



(b)

Fig. 5. The comparison of switching effects between the *Complete Framework* and the baseline strategy. (a) Average proportion of the switch-on power consumption. (b) Average switching on-off frequency.

RRH. The probability of acquiring the spectrum resource of a higher SINR for an admitted UE is usually higher in this case, yielding a higher transmission rate and lower delay.

On the other hand, the average area power consumption contributed by RRHs drawn in Fig. 7 shows that a little sacrifice in delay with a slightly higher L_{ref}, and fewer active RRHs in the C-RAN benefit from a large reduction in area power consumption especially at medium traffic levels. Therefore, the trade-off between QoS and the power consumption reduction is clear that the QoS improvement brought by keeping more active RRHs in a C-RAN for a certain traffic level is not apparent compared with the power consumption reduction brought by just enough active RRHs.

Another parameter, the clustering period, T, determines the frequency of processing clustering to adjust the RRH states in a C-RAN. If a smaller T is chosen, the switching on-off times of a RRH in a certain period of time is increased as shown in Fig. 8 (taking the *Load Weighted Framework* with $L_{\text{ref}} = 6\%$ as an example). This enables a C-RAN to adapt to fast variations of network load levels and traffic distributions, but may yield more overheads. The advantage of fast responses is reflected in reducing the average UE transmission power, which is shown in Fig. 9. A more frequent adjustment of RRH states makes the central controller exploit more timely traffic information and increases the probability of putting active RRHs at the current hotspot areas. The accompanying compromise is that it will

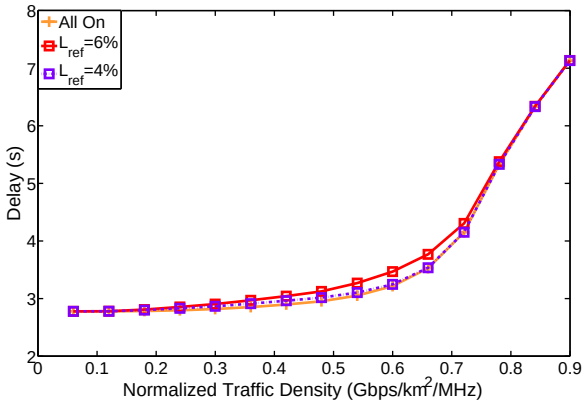


Fig. 6. The impact of different L_{ref} in average delay for the *Load Weighted Framework* at $T = 20$ s.

generate more switching on-off overhead and require more frequent information exchange (every T) between the central controller and RRHs. However, it can be seen that it is small enough to have T equal to 20 seconds. Traffic information sizes required by different frameworks are also acceptable, with only load information needed at a minimum. The low frequency and small size required yield light information exchange overhead.

To ensure a fair comparison among the proposed four frameworks, they are all simulated by setting $L_{\text{ref}} = 6\%$

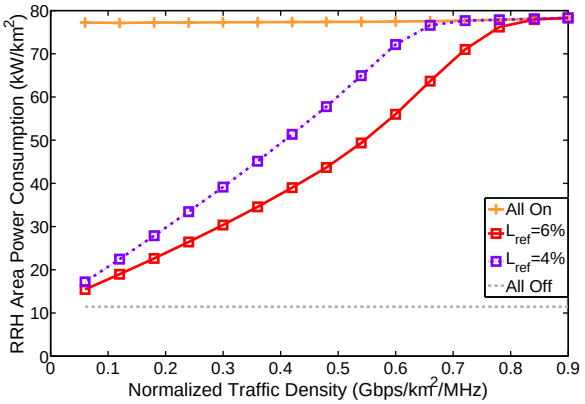


Fig. 7. The impact of different L_{ref} in the average area power consumption contributed by RRHs for the *Load Weighted Framework* at $T = 20$ s.

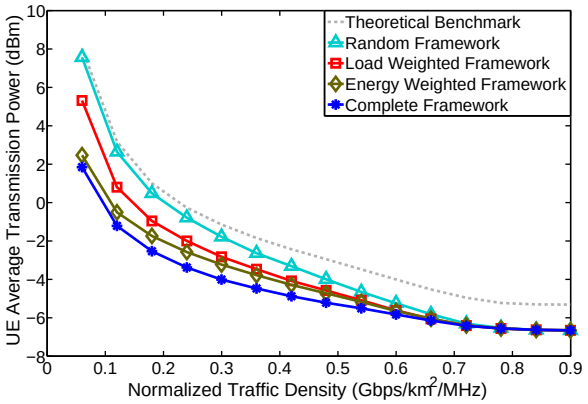


Fig. 10. The impact of different frameworks in the average UE transmission power.

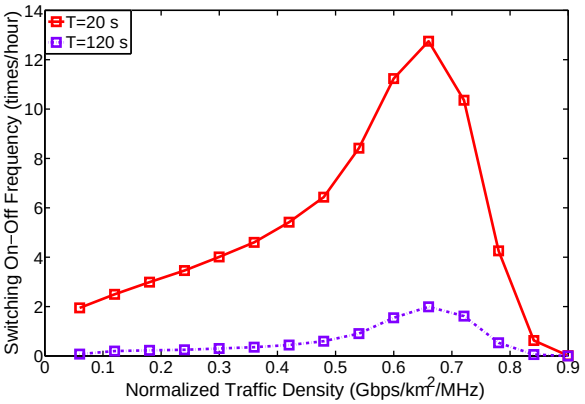


Fig. 8. The impact of different T in average RRH switching frequency for the *Load Weighted Framework*.

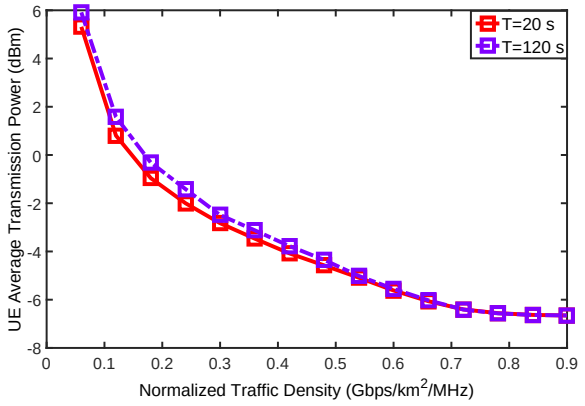


Fig. 9. The impact of different T in average UE transmission power for the *Load Weighted Framework*.

and $T = 20$ s. The main compromise is that better traffic quantification based on richer information availability yields more appropriate locations of active RRHs, reducing UE transmission power and better meeting local service demands, but has a higher requirement in terms of either information or infrastructure availability.

For the *Complete Framework*, a significant amount of UE location information is needed, which provides the most accurate deduction of hotspot areas. It can be seen from Fig. 10 that the *Complete Framework* contributes the largest reduction in UE transmission power, where 6 dB difference is observed compared with the benchmark at a low traffic level. If such UE locations are not supplied or not reliable, the weakness can be compensated to some extent by enabling sleep modes to have the function of measuring uplink power or deploying energy harvesters at RRHs as in the *Energy Weighted Framework*, which achieves the second best UE transmission power reduction. The performance difference from the *Complete Framework* comes from the smaller amount of traffic information available. Another promising solution requiring no extra information is the *Load Weighted Framework*, which guarantees some reduction in UE transmission power and performs better than the *Random Framework* thanks to the weighting strategy using RRH load levels. The *Load Weighted Framework* is appropriate to be used when UE location information and the aforementioned hardware functionality are not available. The dashed line is plotted based on Equation (8) presented before and the average proportion of active RRHs at the corresponding traffic levels, which predicts the average UE transmission power by assuming an uniform active RRH selection. It should be close to the results of the *Random Framework*, which adopts random active RRH selection. The difference from the analytical prediction comes from the minor approximation deviation of the distance distribution function used and the association policy assumed that is based on the smallest path loss instead of the Euclidean distance. The impact of different hotspot handling mechanisms in QoS is parallel to UE transmission power that a better active RRH placement renders better delay. This is similar to what is demonstrated before and thus neglected. Comparatively, the *Complete Framework* including more traffic information orients active RRH placement more to hotspot areas and mitigates most effect on UE battery lives by reducing their transmission power. The *Energy Weighted Framework* and the *Load Weighted Framework* can be treated as strong candidates in aforementioned scenarios depending on different information availabilities.

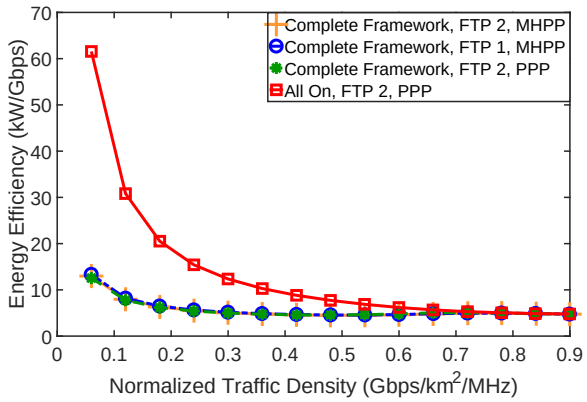


Fig. 11. Energy efficiency on different models.

C. Validation on Different Models

To verify the effectiveness of the proposed frameworks in various scenarios, the frameworks are simulated on different models. FTP traffic model 1 [36] is compared and traffic of mixed types is generated involving different file sizes. The system performance based on the PPP model for the RRH distribution is also investigated. The expected density of RRHs modelled by the PPP is set the same as that of the MHPP. To show the system performance from another angle, energy efficiency is chosen as a representative, which is defined as the ratio of total network power consumption to network throughput. It indicates how much power is consumed to deliver a unit amount of data.

As can been seen from Fig. 11, the proposed frameworks (the *Complete Framework* as an example) are not sensitive to various system models, i.e. energy efficiency is not affected by the type of traffic and the topology of RRHs. This is due to the mechanism of the frameworks that RRH state transitions are dependent on the traffic levels. Compared to the system without the application of the frameworks, the power consumption is relatively linear to the amount of traffic delivered with the *Complete Framework*, yielding a stable energy efficiency. The performance proves that the proposed frameworks are robust and applicable in different scenarios.

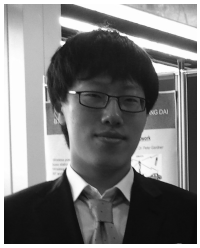
VI. Conclusion

This paper has proposed novel Hotspot-oriented Green Frameworks involving sleep mode operation for ultra-small cell C-RANs, where RRHs can be switched to sleep modes for power consumption reduction when there are more than enough radio resources in the C-RANs. The frameworks employ clustering techniques to choose active RRHs based on distinct formats of quantified local traffic for different information availabilities and infrastructures. The Hotspot-oriented Green Frameworks aim to provide global on-demand radio resources to reduce area power consumption while placing active RRHs at hotspot areas to reduce UE transmission power. The simulation results show that the *Complete Framework* can reduce area power consumption by more than 79% at low traffic levels compared with no sleep mode operation and over 70% in UE transmission power compared with a baseline strategy when the traffic level is below 0.1 Gbps/km^2/MHz. The implementation adaptability of the frameworks are analysed with the trade-offs presented. The robustness and applicability are demonstrated with different models.

References

[1] Qualcomm Incorporated. (2013, June) The 1000× data challenge. [Online]. Available: https://www.qualcomm.com/media/documents/files/1000x-mobile-data-challenge.pdf
[2] Ericsson, "5G radio access," Ericsson, Report, June 2014.
[3] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.*, vol. 6, no. 1, pp. 37–43, Mar. 2011.
[4] R. Baldemair, T. Irnich, K. Balachandran, E. Dahlman, G. Mildh, Y. Selén, S. Parkvall, M. Meyer, and A. Osseiran, "Ultra-dense networks in millimeter-wave frequencies," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 202–208, Jan. 2015.
[5] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.
[6] C. M. R. Institute. (2013, December) C-RAN: The road towards green RAN. [Online]. Available: http://labs.chinamobile.com/cran/wp-content/uploads/2014/06/20140613-C-RAN-WP-3.0.pdf
[7] C.-L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on C-RAN centralization and cloudification," vol. 2, pp. 1030–1039, Sep. 2014.
[8] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): a primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.
[9] D. Sabella, A. De Domenico, E. Katranaras, M. 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," vol. 2, pp. 1586–1597, Jan. 2014.
[10] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, Mar. 2015.
[11] C. Chen, J. Huang, J. Wang, Y. Wu, and G. Li, "Suggestions on potential solutions to CRAN," NGMN, Report, January 2013.
[12] L. Budzisz, F. Ganji, G. Rizzo, M. Marsan, M. Meo, Y. Zhang, G. Koutitas, L. Tassiulas, S. Lambert, B. Lannoo, M. Pickavet, A. Conte, I. Haratcherev, and A. Wolisz, "Dynamic resource provisioning for energy efficiency in wireless access networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2259–2285, Nov. 2014.
[13] R. Wang, H. Hu, and X. Yang, "Potentials and challenges of C-RAN supporting multi-RATs toward 5G mobile networks," vol. 2, pp. 1187–1195, Oct. 2014.
[14] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Multiple daily base station switch-offs in cellular networks," in *Proc. 4th Int. Conf. Commun. and Electron.*, Aug. 2012, pp. 245–250.
[15] A. S. Alam, L. S. Dooley, and A. S. Poulton, "Traffic-and-interference aware base station switching for green cellular networks," in *Proc. IEEE 18th Int. Workshop Comput. Aided Modeling and Design of Commun. Links and Networks*, 2013, Conference Proceedings, pp. 63–67.
[16] K. Samdanis, D. Kutscher, and M. Brunner, "Self-organized energy efficient cellular networks," in *Proc. IEEE 21st Int. Symp. Personal, Indoor and Mobile Radio Commun.*, 2010, Conference Proceedings, pp. 1665–1670.
[17] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, May 2013.
[18] K. Zhang, T. Lv, and H. Gao, "A stochastic geometry based two-stage energy consumption minimization strategy via sleep mode with QoS constraint," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2016, pp. 87–92.
[19] C. Liu, B. Natarajan, and H. Xia, "Small cell base station sleep strategies for energy efficiency," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1652–1661, Mar. 2016.
[20] Y. Jiang, G. Yu, J. Wu, and R. Yin, "Energy consumption tradeoff between network and user equipment in small cell networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2013, pp. 396–401.
[21] J. Yu and P. Chong, "A survey of clustering schemes for mobile ad hoc networks," *IEEE Commun. Surveys Tuts.*, vol. 7, no. 1, pp. 32–48, 1st Qtr. 2005.
[22] S. AlMheiri and H. AlQamzi, "MANETs and VANETs clustering algorithms: A survey," in *Proc. IEEE 8th GCC Conf. and Exhibition*, 2015, pp. 1–6.

[23] K. Hosseini, H. Dahrouj, and R. Adve, "Distributed clustering and interference management in two-tier networks," in *Proc. IEEE Global Commun. Conf.*, 2012, pp. 4267–4272.
[24] A. Abdelnasser, E. Hossain, and D. I. Kim, "Clustering and resource allocation for dense femtocells in a two-tier cellular OFDMA network," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1628–1641, Mar. 2014.
[25] R. Estrada, H. Otrok, and Z. Dziong, "Clustering and dynamic resource allocation for macro-femtocell networks," in *Proc. 16th Int. Telecommun. Netw. Strategy and Planning Symp.*, 2014, pp. 1–6.
[26] W. Li, W. Zheng, Y. Xie, and X. Wen, "Clustering based power saving algorithm for self-organized sleep mode in femtocell networks," in *Proc. 15th Int. Symp. Wireless Personal Multimedia Commun.*, 2012, pp. 379–383.
[27] G. Lee, H. Kim, Y.-T. Kim, and B.-H. Kim, "Delaunay triangulation based green base station operation for self organizing network," in *Proc. IEEE Int. Conf. Green Comput. and Commun. and IEEE Internet of Things and IEEE Cyber, Physical and Social Comput.*, Aug. 2013, pp. 1–6.
[28] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-aho, "Dynamic clustering and sleep mode strategies for small cell networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst.*, 2014, pp. 934–938.
[29] H. Wang, X. Zhou, and M. Reed, "Coverage and throughput analysis with a non-uniform small cell deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2047–2059, Apr. 2014.
[30] M. Haenggi, J. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.
[31] A. Guo and M. Haenggi, "Spatial stochastic models and metrics for the structure of base stations in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5800–5812, Nov. 2013.
[32] M. Haenggi, "Mean interference in hard-core wireless networks," *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 792–794, Aug. 2011.
[33] S.-R. Cho and W. Choi, "Coverage and load balancing in heterogeneous cellular networks with minimum cell separation," *IEEE Trans. Mobile Comput.*, vol. 13, no. 9, pp. 1955–1966, Sep. 2014.
[34] H. He, J. Xue, T. Ratnarajah, F. A. Khan, and C. B. Papadias, "Modeling and analysis of cloud radio access networks using matérn hard-core point processes," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4074–4087, Jun. 2016.
[35] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 996–1019, Jul. 2013.
[36] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA): Further advancements for E-UTRA physical layer aspects*, 3GPP Standard TR 36.814, Rev. V9.1.0, Jan. 2017.
[37] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA): Physical layer procedures*, 3GPP Standard TS 36.213, Rev. V12.0.0, Dec. 2013.
[38] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V. Holappa, M. Alatossava, R. Bultitude, Y. Jong, and T. Rautiainen, "IST-4-027756 WINNER II D1.1.2 V1.2 WINNER II channel models," Report, Feb. 2008.
[39] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA): Radio Frequency (RF) system scenarios*, 3GPP Standard TR 36.942, Rev. V11.0.0, Sep. 2012.
[40] W. Vereecken, I. Haratcherev, M. Deruyck, W. Joseph, M. Pickavet, L. Martens, and P. Demeester, "The effect of variable wake up time on the utilization of sleep modes in femtocell mobile access networks," in *Proc. 9th Annu. Conf. Wireless On-demand Netw. Syst. and Services*, Jan. 2012, pp. 63–66.
[41] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 3rd ed., ser. Wiley Series in Probability and Statistics. Chichester, UK: John Wiley & Sons Ltd, Aug. 2013.
[42] G. Alfano, M. Garetto, and E. Leonardi, "New directions into the stochastic geometry analysis of dense CSMA networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 2, pp. 324–336, Feb. 2014.
[43] A. Al-Hourani, R. J. Evans, and S. Kandeepan, "Nearest neighbor distance distribution in hard-core point processes," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1872–1875, Sep. 2016.
[44] H. Visser and R. Vullers, "RF energy harvesting and transport for wireless sensor network applications: Principles and requirements," *Proc. IEEE*, vol. 101, no. 6, pp. 1410–1423, Jun. 2013.
[45] X. Lu, P. Wang, D. Niyato, and E. Hossain, "Dynamic spectrum access in cognitive radio networks with RF energy harvesting," *IEEE Wireless Commun. Mag.*, vol. 21, no. 3, pp. 102–110, Jun. 2014.



Zhehan Li received his Ph.D. from University of York in 2017. His Ph.D. research was on energy-efficient radio resource provisioning for small cell networks. He joined Huawei in 2017 after working as a research associate for Department of Electronics at University of York. Current research interests include cognitive green networks, radio resource management and applications of machine learning to wireless networks.

David Grace (S'95-A'99-M'00-SM'13) received his PhD from University of York in 1999, with the subject of his thesis being 'Distributed Dynamic Channel Assignment for the Wireless Environment'. Since 1994 he has been a member of the Department of Electronics at York, where he is now Professor (Research) and Head of Communications and Signal Processing Research Group. He is also a Co-Director of the York - Zhejiang Lab on Cognitive Radio and Green Communications, and a Guest Professor at Zhejiang University. Current research interests include aerial platform based communications, cognitive green radio, particularly applying distributed artificial intelligence to resource and topology management to improve overall energy efficiency; 5G system architectures; dynamic spectrum access and interference management. He is currently a lead investigator on H2020 MCSA 5G-AURA. He was a one of the lead investigators on FP7 ABSOLUTE and focussed on extending LTE-A for emergency/temporary events through application of cognitive techniques. He was technical lead on the 14-partner FP6 CAPANINA project that dealt with broadband communications from high altitude platforms. He is an author of over 220 papers, and author/editor of 2 books. He is the former chair of IEEE Technical Committee on Cognitive Networks for the period 2013/4. He is a founding member of the IEEE Technical Committee on Green Communications and Computing. In 2000, he jointly founded SkyLARC Technologies Ltd, and was one of its directors. He is currently a Non-Executive Director of a technology start-up company.

Paul Mitchell (M'00-SM'09) received the M.Eng. and Ph.D. degrees from the University of York, York, U.K., in 1999 and 2003, respectively. His Ph.D. research was on medium access control for satellite systems, which was supported by British Telecom. He has been a member of the Department of Electronics at York since 2002, and is currently Senior Lecturer. He has gained industrial experience at BT and QinetiQ. Research interests include medium access control and routing, wireless sensor networks, cognitive radio, traffic modelling, queuing theory, satellite and mobile communication systems. Dr Mitchell is an author of over 90 refereed journal and conference papers and he has served on numerous international conference programme committees. He was general chair of the International Symposium on Wireless Communications Systems which was held in York in 2010. He is an Associate Editor of the *IET Wireless Sensor Systems* journal. Dr Mitchell is a Senior Member of the IEEE, a member of the IET and a Fellow of the Higher Education Academy.