A Traffic Load Balancing Framework for Software-defined Radio Access Networks Powered by Hybrid Energy Sources

Tao Han
Nirwan Ansari

TR-ANL-2014-005
June 25, 2014

Advanced Networking Laboratory
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
A Traffic Load Balancing Framework for Software-defined Radio Access Networks Powered by Hybrid Energy Sources

Tao Han, Student Member, IEEE, and Nirwan Ansari, Fellow, IEEE
Advanced Networking Laboratory
Department of Electrical and Computer Engineering
New Jersey Institute of Technology, Newark, NJ, 07102, USA
Email: {th36, nirwan.ansari}@njit.edu

Abstract

Dramatic mobile data traffic growth has spurred a dense deployment of small cell base stations (SCBSs). Small cells enhance the spectrum efficiency and thus enlarge the capacity of mobile networks. Although SCBSs consume much less power than macro BSs (MBSs) do, the overall power consumption of a large number of SCBSs is phenomenal. As the energy harvesting technology advances, base stations (BSs) can be powered by green energy to alleviate the on-grid power consumption. For mobile networks with high BS density, traffic load balancing is critical in order to exploit the capacity of SCBSs. To fully utilize harvested energy, it is desirable to incorporate the green energy utilization as a performance metric in traffic load balancing strategies. In this paper, we have proposed a traffic load balancing framework that strives a balance between network utilities, e.g., the average traffic delivery latency and the green energy utilization. Various properties of the proposed framework have been derived. Leveraging the software-defined radio access network architecture, the proposed scheme is implemented as a virtually distributed algorithm, which significantly reduces the communication overheads between users and BSs. The simulation results show that the proposed traffic load balancing framework enables an adjustable trade-off between the on-grid power consumption and the average traffic delivery latency, and saves a

This work was supported in part by NSF under grant no. CNS-1218181 and no. CNS-1320468.
considerable amount of on-grid power, e.g., 14%, at a cost of only a small increase, e.g., 3%, of the average traffic delivery latency.

I. INTRODUCTION

Proliferation of wireless devices and bandwidth greedy applications drive the exponential growth of mobile data traffic that leads to a continuous surge in capacity demands across mobile networks. Heterogeneous network (HetNet) is one of the key technologies for enhancing mobile network capacity to satisfy the capacity demands [1]. In HetNet, low-power base stations referred to as small cell base stations (SCBS) are densely deployed to enhance the spectrum efficiency of the network and thus increase the network capacity. Owing to the disparate transmit powers and base station (BSs) capabilities, traditional user association metrics such as the signal-to-interference-plus-noise ratio (SINR) and the received-signal-strength-indication (RSSI) may lead to a severe traffic load imbalance [1]. Hence, user association algorithms should be well designed to balance traffic load and thus to fully exploit the capacity potential of HetNet.

In order to maximize network utilities, balancing traffic load requires coordination among BSs. The dense deployment of BSs in HetNet increases the difficulty on coordinating BSs. To address this issue, software-define radio access network (SoftRAN) architecture [2] has been proposed. SoftRAN enables coordinated radio resource management in the centralized control plane with a global view of network resources and traffic load. The user association algorithm leveraging the SoftRAN architecture is desired for future mobile networks with an extremely dense BS deployment.

Owing to the direct impact of greenhouse gases on the earth environment and the climate change, the energy consumption of Information and Communications Technology (ICT) is becoming an environmental and thus social and economic issue. Mobile networks are among the major energy hoggers of communication networks, and their contributions to the global energy consumption increase rapidly. Therefore, greening mobile networks is crucial to reducing the carbon footprints of ICT. Although SCBSs consume less power than macro BSs (MBSs), the number of SCBSs will be orders of magnitude larger than that of MBSs for a wide scale network deployment. Hence, the overall power consumption of such a large number of SCBSs will be phenomenal. Greening HetNets have thus attracted tremendous research efforts [3], [4].
As energy harvesting technologies advance, green energy such as sustainable biofuels, solar and wind energy can be utilized to power BSs [5]. Telecommunication companies such as Ericsson and Nokia Siemens have designed green energy powered BSs for mobile networks [6]. By adopting green energy powered BSs, mobile network operators (MNOs) may further save on-grid power consumption and thus reduce their $CO_2$ emissions. However, since the green energy generation is not stable, green energy may not be a reliable energy source for mobile networks. Therefore, future mobile networks are likely to adopt hybrid energy supplies: on-grid power and green energy. Green energy is utilized to reduce the on-grid power consumption and thus reduce the $CO_2$ emissions while on-grid power is utilized as a backup power source.

In HetNets with hybrid energy supplies, the utilization of green energy should be integrated into user association metrics to optimize the green energy usages. For instance, while balancing traffic loads, MNOs may enable BSs with sufficient green energy to serve more traffic load while reducing the traffic load of BSs consuming on-grid power [7]. The traffic load balancing with the consideration of green energy may not maximize network utilities such as the network capacity and the traffic delivery latency. Therefore, a trade-off between the green energy utilization and network utilities should be carefully evaluated in balancing traffic load among BSs. In addition, as a result of the trade-off, users’ utilities such as data rates and the service latency may be decreased because of the consideration of green energy in the traffic load balancing. Thus, users may not cooperate in the traffic load balancing. For example, a distributed user association algorithm may involve multiple interactions between users and BSs and require users to report their measurements to BSs [8], [9]. Seeking to improve their own utilities, they may not report the correct information to BSs. Therefore, it is desirable to hide BSs’ energy information from users to avoid counterfeit reports.

In this paper, we propose a virtually distributed user association scheme that leverages the SoftRAN concept. The proposed scheme, in determining user association, allows an adaptable trade-off between network utilities, e.g. the average traffic delivery latency and the green energy utilization. In the proposed scheme, users only report their downlink data rates calculated based on perceived SINRs via an associating BS to the radio access networks controller (RANC) where the user association decisions are made. In the RANC, virtual users and virtual BSs (vBSs) are generated to simulate the iterative user association adjustments between users and BSs. The optimal user association is derived when the iteration converges. Then, the RANC
informs individual users about the user association decision. The virtualization benefits the user association from two aspects. First, instead of iteratively exchanging information via the air interface, the interactions between virtual users and vBSs rely on a wired link, e.g., a message bus. This significantly reduces communication overhead over the air interface. Second, the virtualization avoids leaking energy information to users because all the iterations are simulated in the RANC. Based on the above features, we name the proposed user association scheme as vGALA: virtualized Green energy Aware and Latency Aware user association.

The rest of the paper is organized as follows. In Section II, we briefly review related works. In Section III, we define the system model and formulate the user association problem. Section IV presents the vGALA scheme. Section V discusses the energy-latency trade-off and the integration of admission control mechanisms. Section VI shows the simulation results, and concluding remarks are presented in Section VII.

II. RELATED WORKS

Balancing traffic load in HetNet has been extensively studied in recent years. In mobile networks, traffic load among BSs is balanced by executing handover procedures. In the LTE system, there are three types of handover procedures: Intra-LTE handover, Inter-LTE handover, and Inter-RAT (radio access technology) handover. There are two ways to trigger handover procedures. The first one is “Network Evaluated” in which the network triggers handover procedures and makes handover decisions. The other one is “Mobile Evaluated” in which a user triggers handover procedures and informs the network about the handover decision. Based on the radio resource status, the network decides whether to approve the user’s handover request. In 4G and LTE networks, a hybrid approach is usually implemented where a user measures parameters of the neighboring cells and reports the results to the network. The network makes the handover decision based on the measurements. Here, the network can decide which parameters should be measured by users.

Aligning with the above procedures, various traffic load balancing algorithms have been proposed to optimize the network utilities. The most practical traffic load balancing approach is the cell range expansion (CRE) technique that biases users’ receiving SINRs or data rates from

---

1The initial idea about green energy aware and latency aware user association was presented at GLOBECOM 2013.
some BSs to prioritize these BSs in associating with users [12]. Owing to the transmit power difference between MBSs and SCBSs, a large bias is usually given to SCBSs to offload users to small cells [11]. By applying CRE, a user associates with the BS from which the user receives the maximum biased SINR or data rate. Although CRE is simple, it is challenging to derive the optimal bias for BSs. Singh et al. [13] provided a comprehensive analysis on traffic load balancing using CRE in HetNet. The authors investigated the selection of the bias value and its impact on the SINR coverage and the downlink rate distribution in HetNet.

The traffic load balancing problem can also be modeled as an optimization problem and solved by convex optimization approaches. Ye et al. [8] modeled the traffic load balancing problem as a utility maximization problem and developed distributed user association algorithms based on the primal-dual decomposition. Kim et al. [14] proposed an $\alpha$-optimal user association algorithm to achieve flow level load balancing under spatially heterogeneous traffic distribution. The proposed algorithm may maximize different network utilities, e.g., the traffic latency and the network throughput, by properly setting the value of $\alpha$. In addition, game theory has been exploited to model and solve the traffic load balancing problems. Aryafar et al. [15] models the traffic load balancing problem as a congestion game in which users are the players and user association decisions are the actions.

The above solutions, though effectively balance the traffic load to maximize the network utilities, do not consider the green energy utilization as a performance metric in balancing traffic load. As green energy technologies advance, powering BSs with green energy is a promising solution to save on-grid power and reduce the carbon footprints [5]. It is desirable to recognize green energy as one of the performance metrics when balancing the traffic load. Zhou et al. [16] proposed a handover parameter tuning algorithm for target cell selection, and a power control algorithm for coverage optimization to guide mobile users to access the BSs with renewable energy supply. Considering a mobile network powered by multiple energy sources, Han and Ansari [7] proposed to optimize the utilization of green energy for cellular networks by optimizing BSs’ transmit powers. The proposed algorithm achieves significant on-grid power savings by scheduling the green energy consumption along the time domain for individual BSs, and balancing the green energy consumption among BSs. The authors have also proposed a user association algorithm that jointly optimizes the average traffic delivery latency and the green energy utilization [9].
III. SYSTEM MODEL AND PROBLEM FORMULATION

In this paper, we consider a HetNet with multiple MBSs and SCBSs as shown in Fig. 1. Both the MBSs and SCBSs are powered by on-grid power and green energy. We consider solar power as the green energy source. We focus on balancing the downlink traffic load among BSs by designing the green energy and latency aware user association scheme. We adopt a software-defined radio access network (SoftRAN) architecture in which all BSs are controlled by the RAN controller (RANC). The RANC has a global view of BSs’ traffic load and green energy. The user association decisions are made by the RANC. The specific design of the RANC is beyond the scope of this paper.

Fig. 1: A HetNet powered by hybrid energy sources: on-grid power and green energy.

A. Traffic model

Denote $\mathcal{B}$ as a set of BSs including both the MBS and SCBSs. We assume that the traffic arrives according to a Poisson process with the arrival rate per unit area at location $x$ equaling to $\lambda(x)$, and the traffic load has a general distribution with average traffic load of $\nu(x)$. Assuming a mobile user at location $x$ is associated with the $j$th BS, then the user’s data rate $r_j(x)$ can be generally expressed as a logarithmic function of the perceived SINR, $\text{SINR}_j(x)$, according to the Shannon-Hartley theorem [14],

$$ r_j(x) = W_j \log_2(1 + \text{SINR}_j(x)), $$

(1)

where $W_j$ is the total bandwidth in the $j$th BS.

$$ \text{SINR}_j(x) = \frac{P_j g_j(x)}{\sigma^2 + \sum_{k \in \mathcal{I}_j} I_k(x)}. $$

(2)

Here, $P_j$ is the transmission power of the $j$th BS, $\mathcal{I}_j$ represents the set of interfering BSs which is defined as the set of BSs whose transmission interferes the $j$th BS’s transmission toward a
user at location \( x \), \( I_k(x) \) is the average interference power seen by a user at location \( x \) from the \( k \)th BS, \( \sigma^2 \) denotes the noise power level and \( g_j(x) \) is the channel gain between the \( j \)th BS and the user at location \( x \). Here, the channel gain reflects only the slow fading including the path loss and the shadowing. We assume the channel gain is measured at a large time scale, and thus fast fading is not considered.

In HetNet, the total bandwidth in a BS is determined by the network’s frequency planning. Different frequency reuse strategies result in different inter-BS interference. In this paper, we assume the network’s frequency reuse strategy is given and static. Thus, \( I_j \) contains the set of BSs who share the same spectrum with the \( j \)th BS. We assume users experience a roughly static interference from the interfering BSs. Although the inter-BS interference in HetNet varies depending on the activities in the interfering BSs, the interference can be well coordinated via time domain techniques, frequency domain techniques and power control techniques [17]. Therefore, the inter-BS interference can be reasonably modeled as a static value for analytical simplicity. The static inter-BS interference model has also been adopted in previous works for modeling the user association problem [14], [18].

The average traffic load density at location \( x \) in the \( j \)th BS is

\[
\rho_j(x) = \frac{\lambda(x) \nu(x) \eta_j(x)}{r_j(x)} \tag{3}
\]

Here, \( \eta_j(x) \) is an indicator function. If \( \eta_j(x) = 1 \), the user at location \( x \) is associated with the \( j \)th BS; otherwise, the user is not associated with the \( j \)th BS. Assuming mobile users are uniformly distributed in the area and denoting \( A \) as the coverage area of all the BSs, the traffic load in the \( j \)th BS can be expressed as

\[
\rho_j = \int_{x \in A} \rho_j(x) dx. \tag{4}
\]

The value of \( \rho_j \) indicates the fraction of time during which the \( j \)th BS is busy.

We assume that traffic arrival processes at individual locations are independent. Since the traffic arrival toward a user at a location is a Poisson process, the traffic arrival in the \( j \)th BS, which is the sum of the traffic arrivals toward all users in its coverage area, is also a Poisson process. The required service time for a user at location \( x \) in the \( j \)th BS is

\[
\gamma_j = \frac{\nu(x)}{r_j(x)}. \tag{5}
\]
Since $\nu(x)$ follows a general distribution, the user’s required service time is also a general distribution. Hence, a BS’s service rate follows a general distribution. Therefore, a BS’s downlink transmission process realizes a M/G/1 processor sharing queue, in which multiple users share the BS’s downlink radio resource [19].

In mobile networks, various downlink scheduling algorithms have been proposed to enable proper sharing of the limited radio resource in a BS [20]. These algorithms are designed to maximize the network capacity, enhance the fairness among users, or provision QoS services. According to the scheduling algorithm, users are assigned different priorities on sharing the downlink radio resource. We assume that during the user association process, users’ data rates do not change. As a result, users in different priority groups perceive different average waiting time. Since traffic arrives a BS according to Possion arrival statistics, the allowed variation in the average waiting times among different priority groups is constrained by the Conservation Law [19]. The integral constraint on the average waiting time in the $j$th BS can be expressed as

$$\bar{L}_j = \rho_j E(\gamma_j^2) \frac{2(1 - \rho_j)}{2(1 - \rho_j)}.$$  

This indicates that given the users’ required service time in the $j$th BS, if the scheduling algorithm gives some users higher priority and reduces their average waiting time, it will increases the average waiting time of the other users. Therefore, $\bar{L}_j$ generally reflects the $j$th BS’s performance in terms of users’ average waiting time. Since $E(\gamma_j^2)$ mainly reflects the traffic characteristics, we assume that $E(\gamma_j^2)$ is roughly constant during a user association process and define $\vartheta_j = \frac{E(\gamma_j^2)}{2}$. Thus, we adopt

$$L(\rho_j) = \frac{\vartheta_j \rho_j}{1 - \rho_j},$$  

as a general latency indicator for the $j$th BS. A smaller $L(\rho_j)$ indicates that the $j$th BS introduces less latency to its associated users.

### B. Energy model

In the network, both MBSs and SCBSs have their own solar panels for generating green energy. Therefore, BSs are powered by hybrid energy sources: on-grid power and green energy. If green energy generated by its solar panel is not sufficient, the BS consumes on-grid power. Since MBSs usually consume more energy than SCBSs, we assume that MBSs are equipped with larger solar panels that have a higher energy generation capacity than that of a SCBS. A
reference design of a hybrid energy powered BS [5] is shown in Fig 2. The charge controller optimizes the green energy utilization based on the solar power intensity, the power consumption of BSs, and energy prices on power grid. Here, the green energy utilization is optimized over time horizon. For example, the charge controller may predict the solar power intensity and mobile traffic loads in a BS over a certain period of time, e.g., 24 hours. The prediction can be based on statistical data and real time weather forecasts. The charge controller according to the prediction determines how much green energy should be utilized to power a BS during a specific time period, e.g., the time duration between two consecutive traffic load balancing procedures.

In this paper, instead of investigating how to optimize the green energy utilization over the time horizon, we aim to study how to balance traffic load among BSs to save on-grid energy within the duration of a traffic balancing procedure. Therefore, we assume that the amount of available green energy for powering a BS is a constant within this duration as determined by the charge controller. It is reasonable to assume that the available green energy is constant because the traffic load balancing process is at a time scale of several minutes [14] while solar power generation is usually modeled at a time scale of a hour [21]. Denote $e_j$ as the amount of green energy for powering the $j$th BS in a traffic load balancing procedure. If the power consumption of the $j$th BS is larger than $e_j$, the BS consumes on-grid power. Otherwise, the residual green energy will be either stored in battery for future usage or upload to power grid via the smart meter. Since we are not focusing on optimizing the green energy utilization over the time horizon, we simply model the BS’s on-grid energy consumption is zero when the BS’s power consumption is less than $e_j$. In other words, we do not consider the redistribution of the
residual green energy in our model.

The BS’s power consumption consists of two parts: the static power consumption and the dynamic power consumption [22]. The static power consumption is the power consumption of a BS without carrying any traffic load. The dynamic power consumption refers to the additional power consumption caused by traffic load in the BS, which can be well approximated by a linear function of the traffic load [22]. Denote \( p_j^s \) as the static power consumption of the \( j \)th BS. Then, the \( j \)th BS’s power consumption can be expressed as

\[
p_j = \beta_j \rho_j + p_j^s.
\]

(8)

Here, \( \beta_j \) is the load-power coefficient that reflects the relationship between the traffic load and the dynamic power consumption in the \( j \)th BS. The BS power consumption model can be adjusted to model the power consumption of either MBSs or SCBSs by incorporating and tweaking the static power consumption and the load-power coefficient. The on-grid power consumption in the \( j \)th BS is

\[
p_j^o = \max(p_j - e_j, 0).
\]

(9)

C. Problem formulation

In determining the user association, the network aims to strive for a trade-off between network utilities, e.g., the average traffic delivery latency and the on-grid power consumption. In this paper, we focus on designing user association algorithm to enhance the network performance by reducing the average traffic delivery latency in BSs as well as to reduce the on-grid power consumption by optimizing the utilization of green energy.

On the one hand, to reduce the average traffic delivery latency, the network desires to minimize the summation of the latency indicators of BSs. On the other hand, since BSs are powered by both green energy and on-grid power, the network seeks to minimize the usage of on-grid power by optimizing the utilization of green energy. According to Eq. (9), on-grid power is only consumed when green energy is not sufficient in the BS. When \( p_j > e_j \), to alleviate on-grid power consumption, the \( j \)th BS has to reduce its traffic load. We define the green traffic capacity as the maximum traffic load that can be supported by green energy. Denote \( \bar{\rho}_j \) as the green traffic capacity of the \( j \)th BS. Then,

\[
\bar{\rho}_j = \max(\epsilon, \min(\frac{e_j - p_j^s}{\beta_j}, 1 - \epsilon)).
\]

(10)
Here, \( \epsilon \) is an arbitrary small positive constant to guarantee \( 0 < \tilde{\rho}_j < 1 \). To reduce traffic loads from \( \rho_j \) to \( \tilde{\rho}_j \), the \( j \)th BS has to shrink its coverage area. As a result, its traffic load is offloaded to its neighboring BSs and may lead to traffic congestion in the neighboring BSs. The traffic congestion increases the average traffic delivery latency of the network. To achieve a trade-off between the average traffic delivery latency and the on-grid power consumption, we define the energy-latency coefficient in the \( j \)th BS as \( \theta_j \). We further define the desired traffic load in the \( j \)th BS after the energy-latency trade-off as

\[
\bar{\rho}_j = (1 - \theta_j)\rho_j + \theta_j\tilde{\rho}_j.
\]  

(11)

Here, \( 0 \leq \theta_j \leq 1 \). If \( \theta_j \) is set to zero, the \( j \)th BS’s desired traffic load is its actual traffic load without considering green energy. In this case, we consider the \( j \)th BS being latency-sensitive; otherwise, if \( \theta_j \) equal to one, the \( j \)th BS’s desired traffic load is dominated by its green traffic capacity and thus the BS is energy-sensitive. We assume \( \theta_j \) remains constant within the duration of a user association process.

Considering both the average traffic delivery latency and the on-grid power consumption, the user association (UA) problem is formulated as

\[
\min_{\rho} \sum_{j \in B} w_j(\rho_j)L(\rho_j)
\]

subject to:

\[
0 \leq \rho_j \leq 1 - \epsilon.
\]

(12)

(13)

Here, \( \rho = (\rho_1, \rho_2, \cdots, \rho_{|B|}) \), and

\[
w_j(\rho_j) = \frac{e^{\rho_j}}{(1 - \theta_j)\rho_j + \theta_j\tilde{\rho}_j} = e^{\kappa \theta_j (\rho_j - \tilde{\rho}_j)}
\]

(14)

In the objective function, \( w_j(\rho_j) \) indicates the weight of the \( j \)th BS’s latency indicator. If the \( j \)th BS has sufficient green energy (\( \tilde{\rho}_j \geq \rho_j \)), \( 0 < w_j(\rho_j) \leq 1 \); otherwise, \( w_j(\rho_j) > 1 \). This is because when the amount of available green energy in the \( j \)th BS is sufficient, the green traffic capacity, \( \tilde{\rho}_j \), is larger than \( \rho_j \). Then, \( \tilde{\rho}_j > \rho_j \) and \( w_j < 1 \). With a large weight, the latency ratio of the \( j \)th BS has a high priority while minimizing Eq. (12) as compared with those of the BSs having a small weight. Therefore, as compared with \( w_j(\rho_j) \leq 1 \), \( w_j(\rho_j) > 1 \) enables the \( j \)th BS to achieve a smaller latency ratio. Since

\[
\frac{dL(\rho_j)}{d\rho_j} = \frac{\theta_j}{(1 - \rho_j)^2} > 0,
\]

(15)
a smaller latency indicator means less traffic load in the \( j \)th BS, which is desirable for saving on-grid power in the \( j \)th BS. Thus, introducing the weights for BSs’ latency indicator in the objective function enables the green energy aware and traffic delivery latency aware user association. \( \kappa \) is a parameter that further adjusts the value of the weight according to that of the traffic latency indicator and enables the network to control the trade-off between the on-grid power consumption and the average traffic delivery latency.

IV. vGALA: A Green Energy and Latency Aware Load Balancing Scheme

In this section, we present the vGALA scheme and prove its properties.

A. Load balancing procedures

The vGALA scheme includes three phases. The first phase is the initial user association and the network measurement. When a user is powered on, it requests to attach to the mobile network. In this phase, no load balancing is performed. The user may attach to the BS with the highest perceived SINR. After the network attachment, a user is able to communicate with the mobile network. The user measures the downlink data rates, e.g., \( r_j(x) \), of its surrounding BSs and reports the measurements to the RANC via its attaching BS. BSs also report its green energy status, e.g., \( \theta_j, \tilde{\rho}_j, \beta_j, \) and \( p^s_j \), to the RANC. The second phase is the user association optimization. The RANC based on the collected measurements calculates the optimal user association for individual users. The third phase is the user handover procedure. According to the user association derived in the second phase, the RANC initializes user handover procedures to balance traffic load among BSs. Note that the other events, e.g., user mobility, green energy generation variations and traffic load changes, can also trigger the traffic load balancing procedure. These events may trigger the network measurement updates. Based on the updates, the RANC follows the second and third phase of the vGALA scheme to balance the traffic load among BSs. How to optimize the event triggering thresholds is beyond the scope of this paper. Instead, we focus on the second phase of the vGALA scheme that designs the user association algorithm.

B. The vGALA user association scheme

In the first phase of the vGALA scheme, users and BSs measure and report their downlink data rates and green energy statuses to RANC, respectively. Based on the measurements, the
RANC optimizes the user association. Leveraging the SoftRAN architecture, the RANC has a global view of the traffic load and the availability of green energy in the network, to facilitate the user association optimization. However, owing to the large number of users and BSs, the user association algorithm if not well designed may be time consuming and incurs excessive delays. In order to efficiently optimize the user association, the vGALA scheme divides the user association algorithm into two parts: the user side algorithm and the BS side algorithm. The user side algorithm calculates the user’s BS selection. The BS side algorithm updates the BS’s operation status including the green traffic capacity and the traffic load. Based on the updates, the user side algorithm re-calculates the BS selection. The user association algorithm iterates until it converges.

The information exchanges over the air interface between users and BSs may introduce additional communication overhead and incur extra power consumption. Leveraging cloud computing and virtualization, the vGALA scheme generates virtual users and virtual BSs (vBSs) in the RANC. The user side algorithm runs on virtual users while the BS side algorithm runs on vBSs. In this way, instead of exchanging information over the air interface, the virtual users and vBSs can iteratively update their BS selection and measurements via a wired link, e.g., a message bus. Fig. 3 illustrates the design of the user association algorithm. Here, the virtualization only virtualizes the computation resources for BSs and users rather than virtualizing all their functions.

![Figure 3: The illustration of the user association algorithm.](image)

1) **The user side algorithm:** We define the time interval between two consecutive BS selection updates as a time slot. At the beginning of the $k$th time slot, vBSs send their operation statuses
to virtual users. The $j$th vBS’s operation status in the $k$th time slot is defined as
\[ \phi_j(\rho_j(k)) = \frac{d w_j(\rho_j(k)) L(\rho_j(k))}{d \rho_j(k)} = \frac{\partial_j e^{\kappa \theta_j(\rho_j(k) - \bar{\rho}_j)} (\kappa \theta_j \rho_j(k) - \kappa \theta_j \rho_j(k)^2 - 1)}{(1 - \rho_j(k))^2}. \] (16)

Here, the $j$th vBS maps to the $j$th BS in the mobile network.

The BS selection rule for a virtual user that maps to a user at location $x$ can be expressed as
\[ b^k(x) = \arg\max_{j \in \mathcal{B}} \frac{r_j(x)}{\phi_j(\rho_j(k))} \] (17)

Here, $b^k(x)$ is the index of the vBS selected by the virtual user at location $x$ in the $k$th time slot.

2) The BS side algorithm: Upon receiving BSs’ operation status updates, virtual users select vBSs according to the user side algorithm. Then, the coverage area of the $j$th vBS in the $k$th time slot is updated as
\[ \tilde{A}_j(k) = \{x | j = b^k(x), \forall x \in \mathcal{A}\} \] (18)

Then, given $\rho(k) = (\rho_1(k), \rho_2(k), \cdots, \rho_{|\mathcal{B}|}(k)), \theta = (\theta_1, \theta_2, \cdots, \theta_{|\mathcal{B}|}),$ and $\tilde{\rho} = (\tilde{\rho}_1, \tilde{\rho}_2, \cdots, \tilde{\rho}_{|\mathcal{B}|}),$ the $j$th vBS’s perceived traffic load in the $k$th time slot is
\[ M_j(\rho(k), \theta, \tilde{\rho}) = \min \left( \int_{x \in \tilde{A}_j(k)} \phi_j(x) dx, 1 - \epsilon \right). \] (19)

Since $\theta$ and $\tilde{\rho}$ are assumed not to change within the duration of a user association process, $M_j(\rho(k), \theta, \tilde{\rho})$ evolves based only on $\rho(k).$ Thus, we use $M_j(\rho(k))$ instead of $M_j(\rho(k), \theta, \tilde{\rho})$ for simplicity in the following analysis.

The perceived traffic load in the $j$th vBS evolves as follows: after vBSs have updated their operation statuses, virtual users select their associating vBSs according to the user side algorithm; based on the user association, vBSs measure their perceived traffic load $M_j(\rho(k)).$ After having derived the perceived traffic load, the $j$th vBS updates its traffic load as
\[ \rho_j(k + 1) = \delta \rho_j(k) + (1 - \delta) M_j(\rho(k)). \] (20)

Here, $0 < \delta < 1$ is an exponential averaging parameter. In the $k + 1$th time slot, the $j$th BS’s operation status is $\phi_j(\rho_j(k + 1))$. 


C. The convergence of vGALA

In order to prove the convergence of vGALA, we first prove that the vBSs’ traffic load vector converges. The feasible set for the UA problem is

\[
\mathcal{F} = \{ \rho | \rho_j = \int_{x \in A} \rho_j(x) dx, \\
0 \leq \rho_j \leq 1 - \epsilon, \sum_{j \in B} \eta_j(x) = 1, \\
\eta_j(x) = \{0, 1\}, \forall j \in B, \forall x \in A \}
\]

(21)

Since \( \eta_j(x) = \{0, 1\} \), \( \mathcal{F} \) is not a convex set. Thus, the traffic updates in Eq. (20) cannot guarantee the updated traffic load is in the feasible set. In order to show the convergence of vGALA, we first relax the constraint to let \( 0 \leq \eta_j(x) \leq 1 \) and then prove the traffic load vector converges to the traffic load vector that is in the feasible set. Define

\[
\tilde{\mathcal{F}} = \{ \rho | \rho_j = \int_{x \in A} \rho_j(x) dx, \\
0 \leq \rho_j \leq 1 - \epsilon, \sum_{j \in B} \eta_j(x) = 1, \\
0 \leq \eta_j(x) \leq 1, \forall j \in B, \forall x \in A \}
\]

(22)

as the relaxed feasible set.

**Lemma 1.** The relaxed feasible set \( \tilde{\mathcal{F}} \) is a convex set.

**Proof:** The lemma is proved by showing that the set \( \tilde{\mathcal{F}} \) contains any convex combination of the traffic load vector \( \rho \).

Let

\[
\psi(\rho) = \sum_{j \in B} w_j(\rho_j) L(\rho_j).
\]

(23)

**Lemma 2.** \( \psi(\rho) \) is a convex function of \( \rho \) when \( \rho \) is defined in \( \tilde{\mathcal{F}} \).

**Proof:** The lemma is proved by showing \( \nabla^2 \psi(\rho) > 0 \).

**Lemma 3.** When \( M(\rho(k)) \neq \rho(k) \), \( M(\rho(k)) \) provides a descent direction of \( \psi(\rho) \) at \( \rho(k) \).

**Proof:** Since \( \psi(\rho) \) is convex function, proving the lemma is equivalent to prove

\[
\langle \nabla \psi(\rho) |_{\rho = \rho(k)}, M(\rho(k)) - \rho(k) \rangle < 0.
\]

(24)
Let $\eta_j^m(x)$ and $\eta_j(x)$ be the user association indication of the $j$th BS that result in the traffic load $M_j(\rho(k))$ and $\rho_j(k)$, respectively.

\[
\langle \nabla \psi(\rho)|_{\rho=\rho(k)}, M(\rho(k)) - \rho(k) \rangle
\]
\[=
\sum_{j\in B}(M_j(\rho(k)) - \rho_j(k))\phi_j(\rho_j(k))
\]
\[=
\sum_{j\in B} \int_{x\in A} \lambda(x)\nu(x)(\eta_j^m(x) - \eta_j(x))dx
\]
\[=
\int_{x\in A} \lambda(x)\nu(x) \sum_{j\in B} \frac{\eta_j^m(x) - \eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j(k))} dx.
\]

Since
\[\eta_j^m(x) = \begin{cases} 
1, & \text{for } j = b^k(x) \\
0, & \text{for otherwise},
\end{cases}
\]

\[
\sum_{j\in B} \frac{\eta_j^m(x) - \eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j(k))} \leq 0.
\]

Because $M(\rho(k)) \neq \rho(k)$, there exists $j \in B$ such that $\eta_j^m(x) \neq \eta_j(x)$, $x \in A$. Hence,

\[
\sum_{j\in B} \frac{\eta_j^m(x) - \eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j(k))} < 0,
\]

and $\langle \nabla \psi(\rho)|_{\rho=\rho(k)}, M(\rho(k)) - \rho(k) \rangle < 0.$

\[\text{Theorem 1. The traffic load vector } \rho \text{ converges to the traffic load vector } \rho^* \in \mathcal{F}.
\]

\[\text{Proof:}
\]
\[\rho(k+1) - \rho(k)
\]
\[= \delta \rho(k) + (1 - \delta)M(\rho(k)) - \rho(k)
\]
\[= (1 - \delta)(M(\rho(k)) - \rho(k)).
\]

Since $M(\rho(k))$ gives a descent direction of $\psi(\rho)$ at $\rho(k)$ and $0 < \delta < 1$, $\rho(k+1)$ also provides a descent direction of $\psi(\rho)$ at $\rho(k)$.

Define $\rho_j(k+1) = \delta \rho_j(k) + (1 - \delta)M_j(\rho_j(k))$. Assume there are sufficient users in the network and $M(\rho(k))$ can be recognized as a continuous function of $\rho(k)$. Since $M_j(\rho_j(k))$, $\rho_j(k+1)$ and $\rho_j(k)$ belong to $[0, 1 - \epsilon]$, according to the fixed point theorem, there exists $\delta$ and
\[ \rho_j^* \in [0, 1 - \epsilon] \] such that the mapping, \( \rho_j(k+1) = \delta \rho_j(k) + (1 - \delta) M_j(\rho_j(k)) \), can converges to \( M_j(\rho_j^*) = \rho_j^* \).

Therefore, there exists \( \rho^* \) such that \( M(\rho^*) = \rho^* \). Hence, the traffic load vector converges to \( \rho^* \). Since \( M(\rho^*) \) is derived based on the user side algorithm where \( \eta_j^*(x) = \{0, 1\}, \forall j \in B, \forall x \in A \), \( \rho^* \) is in the feasible set \( F \).

**Corollary 1.** The vBSs’ operation status \( \phi_j(\rho_j), \forall j \in B \), converges to \( \phi_j(\rho_j^*) \).

**Proof:** Within the duration of a user association process, \( \vartheta_j, \theta_j, \) and \( \tilde{\rho}_j \) are constant. Thus, \( \phi_j(\rho_j) \) is only determined by \( \rho_j \). Since \( \rho_j \) converges to \( \rho_j^* \), \( \phi_j(\rho_j) \) converges to \( \phi_j(\rho_j^*) \).

**D. The optimality of vGALA**

Since the vBSs’ traffic load vector converges to \( \rho^* \), we show that the corresponding user association minimizes \( \psi(\rho) \).

**Theorem 2.** Suppose \( F \) is not empty and the traffic load vector converges to \( \rho^* \), the user association corresponding to \( \rho^* \) minimizes \( \psi(\rho) \).

**Proof:** Denote \( \eta^* = \{\eta_j^*(x)|\eta_j^*(x) = \{0, 1\}, \forall j \in B, \forall x \in A \} \) and \( \eta = \{\eta_j(x)|\eta_j(x) = \{0, 1\}, \forall j \in B, \forall x \in A \} \) as the user association corresponding to \( \rho^* \) and any other traffic load vector \( \rho \in F \), respectively.

Let \( \Delta \rho^* = \rho - \rho^* \). Since \( \psi(\rho) \) is a convex function over \( \rho \), proving the theorem is equivalent to prove

\[ \langle \nabla \psi(\rho)|_{\rho=\rho^*}, \Delta \rho^* \rangle \geq 0. \quad (30) \]

\[ \langle \nabla \psi(\rho)|_{\rho=\rho^*}, \Delta \rho^* \rangle \]

\[ = \sum_{j \in B} (\rho_j - \rho_j^*) \phi_j(\rho_j^*) \]

\[ = \sum_{j \in B} \int_{x \in A} \lambda(x) \nu(x) (\eta_j(x) - \eta_j^*(x)) dx \]

\[ = \sum_{j \in B} \int_{x \in A} \lambda(x) \nu(x) \frac{\eta_j(x) - \eta_j^*(x)}{r_j(x) \phi_j^{-1}(\rho_j^*)} dx. \]

\[ = \sum_{j \in B} \int_{x \in A} \lambda(x) \nu(x) \frac{\eta_j(x) - \eta_j^*(x)}{r_j(x) \phi_j^{-1}(\rho_j^*)} dx. \]
According to the user side algorithm,
\[
\eta^*_j(x) = \begin{cases} 
1, & \text{for } j = \arg \max_{i \in B} \frac{r_i(x)}{\phi_i(\rho^*_i)} \\
0, & \text{for otherwise}
\end{cases}
\]

(32)

Therefore,
\[
\sum_{j \in B} \eta^*_j(x) r_j(x) \phi_j^{-1}(\rho^*_j) \leq \sum_{j \in B} \eta_j(x) r_j(x) \phi_j^{-1}(\rho^*_j).
\]

(33)

Hence, \( \langle \psi(\rho) |_{\rho=\rho^*}, \Delta \rho^* \rangle \geq 0. \)

\[\text{Lemma 4.} \]

If \( f(\rho_j) \) is positive, convex and non decreasing over \( \rho_j, \forall j \in B, \tilde{\psi}(\rho) = \sum_{j \in B} w_j(\rho_j) f(\rho_j) \) is convex over \( \rho \in \tilde{F}. \)

\[\text{Proof:}\] Since \( f(\rho_j) \) is positive, convex and non decreasing, \( f(\rho_j) > 0, f''(\rho_j) \geq 0, \) and
Because $w_j''(\rho_j) \geq 0$, $w_j'(\rho_j) \geq 0$, and $w_j(\rho_j) \geq 0$,
\[
\frac{\partial^2}{\partial \rho_j^2} \sum_{j \in B} w_j(\rho_j) f(\rho_j) = w_j''(\rho_j) f(\rho_j) + 2w_j'(\rho_j) f'(\rho_j) + w_j(\rho_j) f''(\rho_j) \geq 0.
\] (36)

Since
\[
\frac{\partial^2}{\partial \rho_j \partial \rho_i} \sum_{j \in B} w_j(\rho_j) f(\rho_j) = 0, \forall i \neq j,
\] (37)
\[\nabla^2 \tilde{\psi}(\rho) \geq 0.\] Therefore, $\tilde{\psi}(\rho)$ is convex over $\rho$, $\rho \in \tilde{\mathcal{F}}$.

**Theorem 3.** If the $j$th BS’s network performance metric, $f(\rho_j)$, is positive, convex and non decreasing over $\rho_j$, $\forall j \in B$, the UAG problem can be solve by the vGALA scheme.

**Proof:** In order to guarantee the convergence and the optimality of the vGALA scheme, $\tilde{\psi}(\rho)$ has to be convex over $\rho \in \tilde{\mathcal{F}}$. According to the above lemma, if $f(\rho_j)$ is positive, convex and non decreasing, $\tilde{\psi}(\rho)$ is convex. Thus, the vGALA framework can be utilized to solve the UAG problem in which $f(\rho_j)$ is the $j$th BS’s network performance metric.

V. **THE ENERGY-LATENCY TRADE-OFF ADAPTATION AND ADMISSION CONTROL**

**A. The energy-latency trade-off adaptation**

The vGALA scheme provides two parameters for adapting the trade-off between the on-grid power consumption and the average traffic delivery latency. The parameters are $\theta$ and $\kappa$. $\theta$ is the energy-latency coefficient of a BS. It reflects individual BSs’ operation strategies. A BS with a large $\theta$ ($\theta \to 1$) indicates that the BS is energy-sensitive. When a BS chooses a small $\theta$ ($\theta \to 0$), the BS is latency-sensitive. Therefore, by choosing the value of $\theta$, a BS adapts its sensitivity about the on-grid power consumption and the average traffic delivery latency. Hence, $\theta$ is chosen by individual BSs based on their operation strategies.

$\kappa$ is chosen by the RANC based on the global view of green energy status and the mobile traffic demands. Fig. 4 shows the impact of the value of $\kappa$ on $w_j(\rho_j)$. Let $\theta_j = 0.6$ and $\tilde{\rho}_j = 0.5$. Here, $\tilde{\rho}_j$ is determined by the available green energy. $w_j(\rho_j)$ grows exponentially as the traffic demand increases. For a large $\kappa$, i.e. $\kappa = 5$, $w_j(\rho_j)$ grows faster than it does with a small $\kappa$, i.e., $\kappa = 1$. This indicates that the vGALA scheme is more energy-sensitive when $\kappa$ is assigned a
larger value. When $\kappa$ keeps increasing, the vGALA scheme will become a solely energy-aware user association scheme. On the other hand, when $\kappa = 0$, the vGALA scheme is a solely latency-aware user association scheme. In addition, since $0 \leq \theta_j \leq 1$, $0 \leq \theta_j \leq \kappa$. Thus, the value of $\kappa$ restricts the individual BSs’ capability in adapting the energy-latency trade-off. The adaptation of $\kappa$ can be triggered by either green energy changes or the mobile traffic demands changes. For example, when the network experiences heavy traffic load, the RANC will focus on balancing the traffic load to reduce the network congestion. In this case, the RANC may choose a small $\kappa$ to give a high priority to the latency awareness in balancing the traffic load. On the other hand, if the network experiences light traffic load while available green energy reduces, the RANC may increase $\kappa$ to emphasize the utilization of green energy.

B. Admission control mechanism

The necessary condition for the convergence and optimality of the vGALA scheme is that the UA problem is feasible. In other words, the BSs’ traffic load should be within the feasible set defined in Eq. (21). When the traffic load is beyond the network capacity, the UA problem is no longer feasible. As a result, the properties of the vGALA scheme will not hold. Therefore, the admission control mechanism is necessary for the vGALA scheme to ensure the feasibility of the UA problem.

Denote $\mu(x)$ as the admission control coefficient for a user located at $x$. $0 \leq \mu(x) \leq 1$ indicates the probability that a user at location $x$ is admitted to the network. $\mu = \{\mu(x), \forall x \in \mathcal{A}\}$ represents the vector of the admission control coefficient for all users in the network. Here, we
assume that users are at different locations. This assumption is reasonable when \( x \) is a high resolution location. With this assumption, the RANC can specify admission control coefficients for individual users. Thus, the admission control mechanism is not limited to location based admission control. For example, the admission control coefficient can be assigned based on a user’s service priority that is defined in the user’s service agreement with the network. Moreover, the network may also adapt a user’s admission control coefficient to enforce fairness, e.g., proportional fairness, among all users. However, how to select \( \mu(x) \) is beyond the scope of this paper. The following paragraph presents how the admission control is integrated into the vGALA scheme.

For the admission control mechanism, the RANC assigns \( \mu(x) \) for individual users. A user’s admission control coefficient, e.g., \( \mu(x) \), does not depend on its BS selection. In other words, no matter which BS is selected by a user, the user’s admission control coefficient does not change. Thus, integrating admission control mechanism does not change the BS selection rule in the virtual user. The coverage area of a BS, e.g., \( \hat{A}_j(k) \), is still calculated by Eq. (18). Owing to the admission control, the traffic load measurement in the \( j \)th vBS is revised as

\[
M_j(\rho(k)) = \min \left( \int_{x \in \hat{A}_j(k)} \mu(x) \varrho_j(x) dx, 1 - \epsilon \right).
\]

The vBS updates its traffic load based on Eq. (20).

With the admission control, the RANC is able to restrict the traffic load in the network to ensure the UA problem being feasible. The relaxed feasible set for the UA problem with admission control is

\[
\tilde{F} = \{ \rho | \rho_j = \int_{x \in A} \mu(x) \varrho_j(x) dx, \\
0 \leq \rho_j \leq 1 - \epsilon, \sum_{j \in B} \eta_j(x) = 1, \\
0 \leq \eta_j(x) \leq 1, \forall j \in B, \forall x \in A \}
\]

Since \( 0 \leq \mu(x) \leq 1 \) is a constant, Lemma 1 still holds, which means that \( \tilde{F} \) is convex set. Integrating admission control does not change the objective function of the UA problem. Thus, Lemma 2 also holds. In order to prove the convergence of the vGALA scheme under the admission control, we introduce the following lemma.
**Lemma 5.** With the admission control, when $M(\rho(k)) \neq \rho(k)$, $M(\rho(k))$ provides a descent direction of $\psi(\rho)$ at $\rho(k)$.

**Proof:** Since $\psi(\rho)$ is a convex function, proving the lemma is equivalent to prove

$$\langle \nabla \psi(\rho)|_{\rho=\rho(k)}, M(\rho(k)) - \rho(k) \rangle < 0. \quad (40)$$

Let $\eta^m_j(x)$ and $\eta_j(x)$ be the user association indication of the $j$th BS that result in the traffic load $M_j(\rho(k))$ and $\rho_j(k)$, respectively.

$$\langle \nabla \psi(\rho)|_{\rho=\rho(k)}, M(\rho(k)) - \rho(k) \rangle = \sum_{j \in B} (M_j(\rho(k)) - \rho_j(k)) \phi_j(\rho_j(k))$$

$$= \sum_{j \in B} \int_{x \in A} \frac{\lambda(x) \nu(x) \mu(x)(\eta^m_j(x) - \eta_j(x))}{r_j(x) \phi_j^{-1}(\rho_j(k))} dx$$

$$= \int_{x \in A} \frac{\lambda(x) \nu(x) \mu(x)}{r_j(x) \phi_j^{-1}(\rho_j(k))} \sum_{j \in B} \eta^m_j(x) - \eta_j(x) dx.$$ 

Since

$$\eta^m_j(x) = \begin{cases} 1, & \text{for } j = b^k(x) \\ 0, & \text{for otherwise,} \end{cases} \quad (42)$$

$$\sum_{j \in B} \frac{\eta^m_j(x) - \eta_j(x)}{r_j(x) \phi_j^{-1}(\rho_j(k))} \leq 0. \quad (43)$$

Because $M(\rho(k)) \neq \rho(k)$, there exists $j \in B$ such that $\eta^m_j(x) \neq \eta_j(x), x \in A$. Hence,

$$\sum_{j \in B} \frac{\eta^m_j(x) - \eta_j(x)}{r_j(x) \phi_j^{-1}(\rho_j(k))} < 0, \quad (44)$$

and $\langle \nabla \psi(\rho)|_{\rho=\rho(k)}, M(\rho(k)) - \rho(k) \rangle < 0.$

**Theorem 4.** The traffic load vector $\rho$ converges to $\rho^* \in F$ by the vGALA scheme with admission control.

**Proof:** Since integrating the admission control does not change the traffic load update method of vGALA and $M(\rho(k))$ gives a descent direction of $\psi(\rho)$ at $\rho(k)$, the proof of the theorem is the same as that of Theorem 1.
Theorem 5. With the admission control, the traffic load vector converges to $\rho^* \in \mathcal{F}$ that minimizes $\psi(\rho)$.

**Proof:** Denote $\eta^* = \{\eta_j^*(x)|\eta_j^*(x) = \{0, 1\}, \forall j \in \mathcal{B}, \forall x \in \mathcal{A}\}$ and $\eta = \{\eta_j(x)|\eta_j(x) = \{0, 1\}, \forall j \in \mathcal{B}, \forall x \in \mathcal{A}\}$ as the user association corresponding to $\rho^*$ and any other traffic load vector $\rho \in \mathcal{F}$, respectively.

Let $\Delta \rho^* = \rho - \rho^*$. Since $\psi(\rho)$ is a convex function over $\rho$, proving the theorem is equivalent to prove

\[
\langle \nabla \psi(\rho)|_{\rho=\rho^*}, \Delta \rho^* \rangle \geq 0. \tag{45}
\]

\[
\langle \nabla \psi(\rho)|_{\rho=\rho^*}, \Delta \rho^* \rangle \tag{46}
\]

\[
= \sum_{j \in \mathcal{B}} (\rho_j - \rho_j^*) \phi_j(\rho_j^*)
\]

\[
= \sum_{j \in \mathcal{B}} \int_{x \in \mathcal{A}} \frac{\lambda(x)\nu(x)\mu(x)(\eta_j(x) - \eta_j^*(x))dx}{r_j(x)\phi_j^{-1}(\rho_j^*)}
\]

\[
= \int_{x \in \mathcal{A}} \lambda(x)\nu(x)\mu(x) \sum_{j \in \mathcal{B}} \frac{\eta_j(x) - \eta_j^*(x)}{r_j(x)\phi_j^{-1}(\rho_j^*)}dx.
\]

According to the user side algorithm,

\[
\eta_j^*(x) = \begin{cases} 
1, & \text{for } j = \arg \max_{i \in \mathcal{B}} \frac{r_i(x)}{\phi_i(\rho_i^*)} \\
0, & \text{for otherwise,}
\end{cases} \tag{47}
\]

Therefore,

\[
\sum_{j \in \mathcal{B}} \frac{\eta_j^*(x)}{r_j(x)\phi_j^{-1}(\rho_j^*)} \leq \sum_{j \in \mathcal{B}} \frac{\eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j^*)}. \tag{48}
\]

Hence, $\langle \nabla \psi(\rho)|_{\rho=\rho^*}, \Delta \rho^* \rangle \geq 0$. \hfill \blacksquare

Based on the above theorems, under the admission control, the vGALA scheme still enables convergence of the traffic load and obtains the optimal user association in terms of minimizing $\psi(\rho)$.

VI. Simulation Results

We set up system level simulations to investigate the performance of the vGALA scheme for the downlink traffic load balancing in HetNet. In the simulation, three MBSs and seven SCBSs are randomly deployed in a $2000m \times 2000m$ area. The static power consumption of the MBS
TABLE I: Channel Model and Parameters

| Parameters | Value |
|------------|-------|
| $PL_{MBS}$ (dB) | $PL_{MBS} = 128.1 + 37.6 \log_{10}(d)$ |
| $PL_{SCBS}$ (dB) | $PL_{SCBS} = 38 + 10 \log_{10}(d)$ |
| Rayleigh fading | 9 dB |
| Shadowing fading | 5 dB |
| Antenna gain | 15 dB |
| Noise power level | -174 dBm |
| Receiver sensitivity | -123 dBm |

Fig. 5: The coverage areas of different user association schemes.

and the SCBS are 750 W and 37 W, respectively [22]. The load-power coefficient of the MBS and the SCBS are 500 and 4, respectively [22]. The solar cell power efficiency is 17.4% [23]. We assume that the weather condition is the standard condition which specifies a temperature of 25 °C, an irradiance of 1000 W/m², and an air mass of 1.5 spectrum. Thus, the green energy generation rate is 174 W/m². The solar panel sizes are randomly selected but ensure the green power generation capacity of MBSs from 750 w to 1300 w while that of SCBS from 37 w to 48 w. BSs’ energy-latency coefficients are set to be the same. The total bandwidth is 10 MHz and the frequency reuse factor is one. The channel propagation model is based on COST 231 Walfisch-Ikegami [24]. The model and parameters are summarized in Table I Here, $PL_{MBS}$ and $PL_{SCBS}$ are the path loss between the users and MBSs and SCBSs, respectively. $d$ is the distance between users and BSs.
A. Performance comparison

We compare the vGALA scheme with a green energy aware (GA) user association scheme and a latency aware (LA) user association scheme. The GA scheme solves the green energy aware problem (GAP) formulated as

$$\min_{\rho} \sum_{j \in B} \max(\rho_j - e_j, 0)$$

subject to : \hspace{1em} 0 \leq \rho_j \leq 1 - \epsilon. (49) (50)

The LA scheme solves the latency aware problem (LAP) as

$$\min_{\rho} \sum_{j \in B} L(\rho_j)$$

subject to : \hspace{1em} 0 \leq \rho_j \leq 1 - \epsilon. (51) (52)

As shown in Figs. 5 different user association schemes result in different traffic load distribution among BSs. In the figure, the coverage areas of different BSs are filled with different colors. A larger coverage area indicates the BS serves more traffic load. The first, second and third BSs are MBSs and the other BSs are SCBSs. Taking the coverage area of the 9th BS as an example, as compared with the GA scheme (Fig. 5a), the LA scheme significantly reduces the BS’s coverage area as shown in Fig. 5a. The 9th BS has sufficient green energy. Therefore, the GA scheme will redirect more traffic load to the BS to minimize the on-grid power consumption.

2The white color indicates the coverage area of the second BS.
The LA scheme, which does not consider the energy usage, balances the traffic load among BSs to minimize the average traffic delivery latency. As a result, the LA scheme limits the traffic load in the BS. Considering both the power consumption and the average traffic delivery latency, the vGALA scheme slightly reduces the BS’s coverage area as shown in Fig. 5c to obtain a trade-off between the on-grid power consumption and the average traffic delivery latency.

Fig. 6 shows the trade-off achieved by the vGALA scheme between the on-grid energy consumption and the average traffic delivery latency. Fig. 6a shows the on-grid power consumption of the LA, the vGALA, and the GA schemes, respectively. As compared with the vGALA scheme, the LA scheme consumes 14% more on-grid power while the GA scheme costs 43% less on-grid power. Fig. 6b shows that the average traffic delivery latency of the vGALA scheme
is only 3% more than that of the LA scheme and the average traffic delivery latency of the GA scheme is about 80% more than that of the vGALA scheme. The above observation indicates that the vGALA scheme achieves a preferable trade-off: saving 14% on-grid power at the cost of 3% increase in the average traffic delivery latency.

In addition, as shown in Fig. 9, the vGALA scheme requires more than a hundred iterations to converge to the optimal solution. If vGALA scheme is implemented as a distributed algorithm, one iteration indicates one user-BS interaction over the air interface. These iterations introduce significant communication overheads and may consume a considerable amount of energy. The vGALA scheme avoids the communication overhead over the air interface by virtualizing users and BSs in the RANC to simulate the interactions between users and BSs. As a result, the vGALA scheme only requires one interaction per user over the air interface in which the user reports its downlink data rate measurement and the BS sends the user the association decision.

### B. Performance adaptation

The trade-off between the on-grid power consumption and the average traffic delivery latency can be adapted by adjusting $\kappa$ and $\theta$ in the vGALA scheme. Fig. 7 shows the performance of the vGALA scheme with different $\kappa$. By varying $\kappa$, the vGALA scheme may act as the LA scheme when $\kappa \to \infty$ and performs like the GA scheme when $\kappa \to 0$. As shown in Fig. 8, given $\kappa$, adjusting $\theta$ has a limited performance adaptation. In other words, $\kappa$ defines a performance adaptation range and adjusting $\theta$ can only adapt the performance within the range.
C. Green energy generation rate evaluation

The amount of green energy in BSs impacts the performance of the vGALA scheme. In Fig. 9, the axis is the solar cell power efficiency. As the solar cell power efficiency enhances, the amount of green energy in BSs will increase. As shown in Fig. 9a, the on-grid power consumption of BSs decreases as the solar cell power efficiency increases. This is because more green energy is available in BSs. With the increase of the solar cell power efficiency, the performance on the average traffic delivery latency can be divided into four regions as shown in Fig. 9b. In the first region (R1), all BSs do not have sufficient green energy to offset their static power consumption. As a result, BSs’ green traffic capacities are zero. In this condition, the vGALA scheme performs like the LA scheme. In the second region (R2), the green traffic capacities of BSs start to impact the traffic load balancing. The traffic load will be directed to BSs that have sufficient green energy. Meanwhile, the vGALA scheme avoids to excessively increase the average traffic delivery latency. In the third region (R3), as the solar cell power efficiency further increases, the traffic load balancing becomes more flexible with respect to the green energy constraint, which enables the vGALA scheme further reduces the average traffic delivery latency. In both region R2 and R3, the vGALA scheme determines the trade-off between the on-grid power consumption and the average traffic delivery latency. In the fourth region (R4), all BSs have sufficient green energy to operate with full traffic load. In other words, the green traffic capacities of all the BSs equal to one. Thus, green energy is no longer a concern in balancing the traffic load and the vGALA scheme acts as the LA scheme.
D. CRE evaluation

The cell range expansion (CRE) approach is proven to have similar performance as optimal traffic balancing schemes in terms of maximizing network utilities [11], [8]. This simulation evaluates the traffic balancing performance of a two-tier data rate bias algorithm in terms of on-grid power consumption and the average traffic delivery latency. We assume that BSs in the same tier have the same cell bias. In the simulation, MBSs are in the first tier while SCBSs are in the second tier. The cell bias of a MBS is one. We vary the cell bias of a SCBS to investigate the performance of the algorithm. In the data rate bias algorithm, a user select the BS to maximize the biased data rate.

\[
b(x) = \arg \max_{j \in B} Z_j r_j(x).
\] (53)

Here, \(b(x)\) and \(Z_j\) are the index of the selected BS and the cell bias of the \(j\)th BS, respectively. Fig. 10 shows the performance of the data rate bias algorithm. As the small cell data rate bias increases, more traffic load will be directed to SCBSs. This reduces the on-grid power consumption of BSs because SCBSs consume less power than MBSs. Meanwhile, the average traffic delivery latency is convex versus the small cell data rate bias. When the data rate bias algorithm minimizes the average traffic delivery latency \(\sum_{j \in B} L(\rho_j) = 7.24\), the on-grid power consumption is 455 \(w\). With the same simulation setting, the vGALA scheme achieves an on-grid power consumption of 389 \(w\) and \(\sum_{j \in B} L(\rho_j) = 6.76\). This indicates the tier-based data rate bias algorithm may not perform well on jointly optimizing the utilization of green energy and the network utilities.

VII. Conclusion

In this paper, we have proposed a traffic load balancing framework referred to as vGALA. During the procedure of establishing user association, the vGALA scheme not only considers the network performance, e.g., the average traffic delivery latency, but also adapts to the availability of green energy. Various properties, in particular, convergence of vGALA, have been proven. The vGALA scheme reduces the on-grid power consumption with a little sacrifice of the average traffic delivery latency. The trade-off between the network performance and the on-grid power consumption is adjustable in individual BSs and controllable by the radio access network controller. The vGALA scheme includes both the user side algorithm and the BS side
algorithm. To avoid the extra communication overheads, the vGALA scheme, leveraging the SoftRAN architecture, introduces virtual users and vBSs to simulate the interactions between users and BSs and significantly reduces the information exchanges over the air interface. The excessive simulation results have validated the performance of the vGALA scheme.

REFERENCES

[1] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, “An overview of load balancing in hetnets: old myths and open problems,” Wireless Communications, IEEE, vol. 21, no. 2, pp. 18–25, April 2014.
[2] A. Gudipati, D. Perry, L. E. Li, and S. Katti, “Softran: Software defined radio access network,” in Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking, Hong Kong, China, 2013, pp. 25–30.
[3] T. Han and N. Ansari, “On greening cellular networks via multicell cooperation,” IEEE Wireless Communications Magazine, vol. 20, no. 1, pp. 82–89, 2013.
[4] Z. Hasan, H. Boostanimehr, and V. Bhargava, “Green cellular networks: A survey, some research issues and challenges,” IEEE Communications Surveys and Tutorials, vol. 13, no. 4, pp. 524–540, 2011.
[5] T. Han and N. Ansari, “Powering mobile networks with green energy,” Wireless Communications, IEEE, vol. 21, no. 1, pp. 90–96, February 2014.
[6] Ericson Inc., “Sustainable energy use in mobile communications,” Aug. 2007, white Paper.
[7] T. Han and N. Ansari, “On optimizing green energy utilization for cellular networks with hybrid energy supplies,” Wireless Communications, IEEE Transactions on, vol. 12, no. 8, pp. 3872–3882, August 2013.
[8] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, “User association for load balancing in heterogeneous cellular networks,” Wireless Communications, IEEE Transactions on, vol. 12, no. 6, pp. 2706–2716, June 2013.
[9] T. Han and N. Ansari, “Green-energy aware and latency aware user associations in heterogeneous cellular networks,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM’12), Atlanta, GA, USA, Dec. 2013.
[10] L. Wang and G.-S. Kuo, “Mathematical modeling for network selection in heterogeneous wireless networks: A tutorial,” Communications Surveys Tutorials, IEEE, vol. 15, no. 1, pp. 271–292, First 2013.
[11] “LTE; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (3GPP TS 23.401 version 11.9.0 Release 11).” [Online]. Available: http://www.etsi.org/deliver/etsi_ts/123400_123499/123401/11.09.00_60/ts_123401v110900p.pdf
[12] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, “A survey on 3gpp heterogeneous networks,” Wireless Communications, IEEE, vol. 18, no. 3, pp. 10–21, June 2011.
[13] S. Singh, H. Dhillon, and J. Andrews, “Offloading in heterogeneous networks: Modeling, analysis, and design insights,” Wireless Communications, IEEE Transactions on, vol. 12, no. 5, pp. 2484–2497, May 2013.
[14] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, “Distributed α–optimal user association and cell load balancing in wireless networks,” IEEE/ACM Transactions on Networking, vol. 20, no. 1, pp. 177–190, Feb. 2012.
[15] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, “Rat selection games in hetnets,” in INFOCOM, 2013 Proceedings IEEE, April 2013, pp. 998–1006.
[16] J. Zhou, M. Li, L. Liu, X. She, and L. Chen, “Energy source aware target cell selection and coverage optimization for power saving in cellular networks,” in Proceedings of the 2010 IEEE/ACM Int’l Conference on Green Computing and Communications, Hangzhou, China, Dec. 2010.

[17] D. Lopez-Perez, I. Guvenc, G. De la Roche, M. Kountouris, T. Quek, and J. Zhang, “Enhanced intercell interference coordination challenges in heterogeneous networks,” Wireless Communications, IEEE, vol. 18, no. 3, pp. 22–30, 2011.

[18] K. Son, S. Chong, and G. Veciana, “Dynamic association for load balancing and interference avoidance in multi-cell networks,” IEEE Transactions on Wireless Communications, vol. 8, no. 7, pp. 3566 –3576, Jul. 2009.

[19] L. Kleinrock, Queueing Systems: Computer applications. Wiley-Interscience, 1976, ISBN: 978-0471491118.

[20] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, “Downlink packet scheduling in lte cellular networks: Key design issues and a survey,” Communications Surveys Tutorials, IEEE, vol. 15, no. 2, pp. 678–700, 2013.

[21] A. Farbod and T. D. Todd, “Resource Allocation and Outage Control for Solar-Powered WLAN Mesh Networks,” IEEE Transactions on Mobile Computing, vol. 6, no. 8, pp. 960–970, Aug. 2007.

[22] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, “How much energy is needed to run a wireless network?” Wireless Communications, IEEE, vol. 18, no. 5, pp. 40 –49, Oct. 2011.

[23] “HIT photovoltaic module.” [Online]. Available: http://us.sanyo.com/dynamic/product/Downloads/Panasonic%20HIT%20220A%20Data%20Sheet

[24] “Evolution of land mobile radio (including personal) communications: Cost 231.” [Online]. Available: http://www.awe-communications.com/Propagation/Urban/COST/