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

---

Citation:
Tao Han and Nirwan Ansari, "Traffic Load Balancing Framework for Software-defined Radio Access Networks Powered by Hybrid Energy Sources",
IEEE/ACM Transactions on Networking, vol. 24, no. 2, pp. 1038-1051, Apr. 2016.

URL:
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7063976

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*

***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 considerable amount of on-grid power, e.g., 30%, at a cost of only a small increase, e.g., 8%, of the average traffic delivery latency.**

***Index Terms*—Green communications, HetNet, renewable energy, software-defined radio access networks, traffic load balancing.**

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 (SCBSs) 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 (BS) 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 loads and thus to fully exploit the capacity potential of HetNet.

In order to maximize network utilities, balancing traffic loads 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 loads. 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 usage. For instance, while

Manuscript received June 25, 2014; revised December 04, 2014; accepted January 26, 2015; approved by IEEE/ACM Transactions on Networking Editor Y. Bejerano. This work was supported in part by the National Science Foundation (NSF) under Grant CNS-1218181 and Grant CNS-1320468.

The authors are with the Advanced Networking Laboratory, Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: th36@njit.edu; nirwan.ansari@njit.edu).

balancing traffic loads, MNOs may enable BSs with sufficient green energy to serve more traffic loads while reducing the traffic loads 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 loads 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. We generate virtual users and virtual BSs (vBSs) in the radio access networks controller (RANC) to emulate a distributed user association solution that requires iterative user association adjustments between users and BSs. This scheme runs the user association optimization in the RANC, and thus significantly reduces the communication overhead over the air interface. In this scheme, users report their downlink data rates calculated based on perceived SINRs via an associating BS to the RANC where traffic loads from individual users and BSs are measured. The RANC optimizes the BS operation status that reflects the price for a user to access a BS. The user association is determined by the BS operation status and the users' downlink data rates. 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. Meanwhile, running the user association within the RANC avoids leaking energy information to users. As a result, users have no obvious incentives to counterfeit reports. Based on the above features, we name the proposed user association scheme as vGALA: **v**irtualized **G**reen energy **A**ware and **L**atency **A**ware user association.[1]

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 practicality of the vGALA scheme. Section VI shows the simulation results, and concluding remarks are presented in Section VII.

II. Related Works

Balancing traffic loads in HetNet has been extensively studied in recent years [10]. In mobile networks, traffic loads 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 [11]. 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 the handover procedure 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 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 [1]. 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 α-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 α. In addition, game theory has been exploited to model and solve the traffic load balancing problems. Aryafar *et al.* [15] modeled 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 loads to maximize the network utilities, do not consider the green energy utilization as a performance metric in balancing traffic loads. 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 loads. 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

[1]The initial idea about green energy aware and latency aware user association was presented at GLOBECOM 2013 [9].

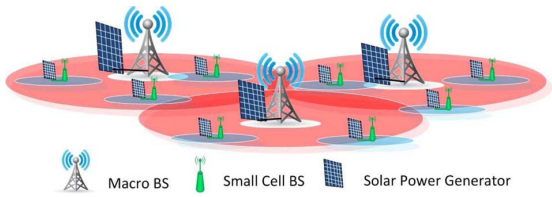

Fig. 1. A HetNet powered by hybrid energy sources: on-grid power and green energy.

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 loads 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 loads and green energy. The user association is optimized by the RANC. The specific design of the RANC is beyond the scope of this paper.

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 point process with the average arrival rate per unit area at location x equaling to $\lambda(x)$, and the traffic size (packet size) per arrival has a general distribution with the average traffic size of $\nu(x)$. Assuming a mobile user at location x is associated with the jth BS, then the user's downlink data rate $r_j(x)$ that will end up becoming available to the user can be generally expressed as a logarithmic function of the perceived SINR, $SINR_j(x)$, according to the Shannon-Hartley theorem [14],

$$r_j(x) = W_j \log_2(1 + SINR_j(x)), \tag{1}$$

where W_j is the total bandwidth in the jth BS.

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

Here, P_j is the transmission power of the jth BS, $\mathcal{I}_j$ represents the set of interfering BSs which is defined as the set of BSs whose transmission interferes the jth 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 kth BS, σ^2 denotes the noise power level and $g_j(x)$ is the channel gain between the jth 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, $\mathcal{I}_j$ contains the set of BSs who share the same spectrum with the jth 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 jth BS is

$$\varrho_j(x) = \frac{\lambda(x)\nu(x)\eta_j(x)}{r_j(x)}. \tag{3}$$

Here, $\eta_j(x)$ is an indicator function. If $\eta_j(x) = 1$, the user at location x is associated with the jth BS; otherwise, the user is not associated with the jth BS. Assuming mobile users are uniformly distributed in the area and denoting $\mathcal{A}$ as the coverage area of all the BSs, based on (3), we derive the average traffic loads in the jth BS expressed as

$$\rho_j = \int_{x \in \mathcal{A}} \varrho_j(x) dx. \tag{4}$$

The value of ρ_j indicates the fraction of time during which the jth BS is busy.

We assume that traffic arrival processes at individual locations are independent. Since the traffic arrival per unit area is a Poisson point process, the traffic arrival in the jth BS, which is the sum of the traffic arrivals in its coverage area, is a Poisson process. The required service time per traffic arrival for a user at location x in the jth BS is $\gamma_j = \frac{\nu(x)}{r_j(x)}$. Since $\nu(x)$ is the average traffic size per arrival which 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. As a result, users in different priority groups perceive different average waiting time. Since traffic arrives at 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 jth BS can be expressed as

$$\bar{L}_j = \frac{\rho_j E(\gamma_j^2)}{2(1-\rho_j)}. \tag{5}$$

This indicates that given the users' required service time in the jth BS, if the scheduling algorithm gives some users higher priority and reduces their average waiting time, it will increase the average waiting time of the other users. Therefore, $\bar{L}_j$ generally reflects the jth BS's performance in terms of users' average

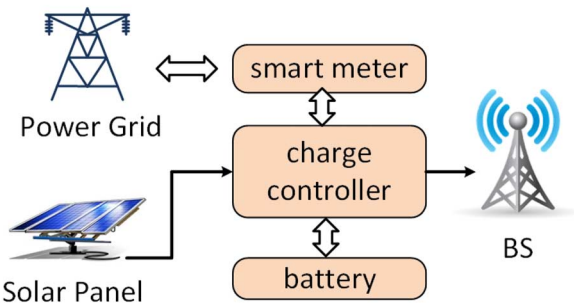

Fig. 2. A hybrid energy powered BS.

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 thus $\vartheta_j = \frac{E(\gamma_j^2)}{2}$ can be considered as a constant. Thus, we adopt

$$L(\rho_j) = \frac{\vartheta_j \rho_j}{1-\rho_j} \tag{6}$$

as a general latency indicator for the jth BS. A smaller $L(\rho_j)$ indicates that the jth BS introduces less latency to its associated users. Therefore, we use $L(\rho_j)$ to reflect the jth BS's average traffic delivery latency.

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 loads 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 an hour [21]. Denote e_j as the amount of green energy for powering the jth BS in a traffic load balancing procedure. If the power consumption of the jth 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 uploaded 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 loads in the BS, which can be well approximated by a linear function of the traffic loads [22]. Denote p_j^s as the static power consumption of the jth BS. Then, the jth BS's power consumption can be expressed as

$$p_j = \beta_j \rho_j + p_j^s. \tag{7}$$

Here, β_j is the load-power coefficient that reflects the relationship between the traffic loads and the dynamic power consumption in the jth 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 jth BS is

$$p_j^o = \max(p_j - e_j, 0). \tag{8}$$

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 a 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 green energy usage.

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 (8), 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 jth BS has to reduce its traffic loads. We define the green traffic capacity as the maximum traffic loads that can be supported by green energy. Denote $\hat{\rho}_j$ as the green traffic capacity of the jth BS. Then,

$$\hat{\rho}_j = \max\left(\epsilon, \min\left(\frac{e_j - p_j^s}{\beta_j}, 1-\epsilon\right)\right). \tag{9}$$

Here, ϵ is an arbitrary small positive constant to guarantee $0 < \hat{\rho}_j < 1$. To reduce traffic loads from ρ_j to $\hat{\rho}_j$, the jth BS has to shrink its coverage area. As a result, its traffic loads are 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 jth BS as θ_j. We further define the desired traffic loads in the jth BS after the energy-latency trade-off as

$$\tau_j = (1-\theta_j)\rho_j + \theta_j\hat{\rho}_j. \tag{10}$$

Here, $0 \leq \theta_j \leq 1$. If θ_j is set to zero, the jth BS's desired traffic loads are its actual traffic loads without considering green energy. In this case, we consider the jth BS being latency-sensitive; otherwise, if θ_j equal to one, the jth BS's desired traffic loads are dominated by its green traffic capacity and thus the BS is energy-sensitive. The selection of θ_j reflects the jth BS's energy-latency trade off that will be discussed in Section V-B. We assume θ_j remains constant within the duration of a user association process.

Since mobile devices are powered by battery, it is desirable to guarantee the energy efficiency of mobile devices while performing the traffic load balancing [23]. To ensure the energy efficiency of mobile devices, we restrict a user to only associate with the BSs to which the user's uplink pathloss is smaller than a predefined threshold. Considering all the above factors, the user association (UA) problem is formulated as

$$\min_{\boldsymbol{\rho}} \quad \sum_{j\in\mathcal{B}} w_j(\rho_j)L(\rho_j) \tag{11}$$

$$\begin{aligned} subject\ to: \quad & 0 \leq \rho_j \leq 1-\epsilon. \\ & (\alpha_j(x) - \alpha^*(x))\eta_j(x) \leq 0, \\ & \forall x \in \mathcal{A}, j \in \mathcal{B}. \end{aligned} \tag{12}$$

Here, $\alpha_j(x)$ and $\alpha^*(x)$ are the uplink pathloss from the user at location x to the jth BS and the uplink pathloss threshold for the user, respectively. $0 < \epsilon < 1$ is a small real number to ensure $\rho_j < 1$. $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_{|\mathcal{B}|})$, and

$$\begin{aligned} w_j(\rho_j) &= e^{\kappa(\rho_j - \tau_j)} \\ &= e^{\kappa(\rho_j - (1-\theta_j)\rho_j - \theta_j\hat{\rho}_j)} \\ &= e^{\kappa\theta_j(\rho_j - \hat{\rho}_j)}. \end{aligned} \tag{13}$$

In the objective function, $w_j(\rho_j)$ indicates the weight of the jth BS's latency indicator. If the jth BS has sufficient green energy ($\hat{\rho}_j \geq \rho_j$), $0 < w_j(\rho_j) \leq 1$; otherwise, $w_j(\rho_j) > 1$. This is because when the amount of available green energy in the jth BS is sufficient, the green traffic capacity, $\hat{\rho}_j$, is larger than ρ_j. Then, $\tau_j > \rho_j$ and $w_j < 1$. With a large weight, the jth BS has a high priority in reducing its latency indicator while minimizing (11) as compared with the BSs having a small weight. Therefore, as compared with $w_j(\rho_j) \leq 1$, $w_j(\rho_j) > 1$ enables the jth BS to achieve a smaller latency indicator. Since

$$\frac{dL(\rho_j)}{d\rho_j} = \frac{\vartheta_j}{(1-\rho_j)^2} > 0, \tag{14}$$

a smaller latency indicator means less traffic loads in the jth BS, which is desirable for saving on-grid power in the jth 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. κ 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. The vGALA scheme generally consists of three phases. The first phase is the initial user association and network measurement, during which the RANC collects network information, e.g., available green energy, traffic loads, and users' data rates. The second phase is the user association optimization, in which the RANC optimizes the user association and derives the corresponding BSs' operation statuses based on the information collected in the first phase. Here, a BS's operation status reflects the price for a user to access the BS. In the third phase, the user association is determined based on the optimized BSs' operation statuses and users' downlink data rates. The major optimization of the vGALA scheme is in the second phase. To be analytically tractable, we assume that (1) the RANC can successfully collect the network information from all BSs and users, and (2) the users' data rates do not change within one user association process. We will evaluate these assumptions in the next section where we discuss the practicality of the vGALA scheme.

A. The vGALA User Association Scheme

Based on the collected network information, the RANC optimizes the user association and derives the optimal BS operation status. Leveraging the SoftRAN architecture, the RANC has a global view of the traffic loads 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 calculated based on the green traffic capacity and the traffic loads. Based on the updates, the user side algorithm re-calculates the BS selection. The user association algorithm iterates until it converges. After the convergence, the optimal BS operation status is obtained and the optimal user association is subsequently determined.

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 information locally within the RANC. Here, the virtualization only virtualizes the computation resources for BSs and users rather than virtualzing all their functions.

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 kth time slot, vBSs send their operation statuses to virtual users. Let

$$\psi(\boldsymbol{\rho}) = \sum_{j\in\mathcal{B}} w_j(\rho_j)L(\rho_j). \tag{15}$$

The jth vBS's operation status in the kth time slot is defined as

$$\begin{aligned}\phi_j(\rho_j(k)) &= \frac{\partial\psi(\boldsymbol{\rho}(k))}{\partial\rho_j(k)} \\ &= \frac{\vartheta_j e^{\kappa\theta_j(\rho_j(k)-\hat{\rho}_j)}(\kappa\theta_j\rho_j(k)-\kappa\theta_j\rho_j(k)^2-1)}{(1-\rho_j(k))^2}.\end{aligned} \tag{16}$$

Here, the jth vBS is mapped to the jth BS in the mobile network.

Let $\bar{\mathcal{B}}(x) = \{j|\alpha_j(x) \le \alpha^*(x)\}$ be the set of BSs whose uplink pathloss is less than the user's pathloss threshold. Assign $r_j(x) = \zeta,\ \forall j \in \mathcal{B}\setminus\bar{\mathcal{B}}(x)$ where ζ is a very small positive number that approaches zero. This is equivalent to restricting the user from associating with the BSs outside $\bar{\mathcal{B}}(x)$. Then, the BS selection rule for a user at location x can be expressed as

$$b^k(x) = \arg\max_{j\in\mathcal{B}} \frac{r_j(x)}{\phi_j(\rho_j(k))}. \tag{17}$$

Here, $b^k(x)$ is the index of the vBS selected by the virtual user at location x in the kth time slot. The pseudo code of the user side algorithm is shown in Alg. 1. The computational complexity of the user side algorithm for an individual user is $O(|\mathcal{B}|)$.

Algorithm 1: The User Side Algorithm

Input :BSs' operation status: $\phi_j(\rho_j(k)), j \in \mathcal{B}$;

Output: The BS selection: $b^k(x)$;

1 Estimate the uplink pathloss: $\alpha_j(x)$;

2 Find $\bar{\mathcal{B}}(x) = \{j|\alpha_j(x) \le \alpha^*(x)\}$;

3 Assign $r_j(x) = \zeta, \forall j \in \mathcal{B}\setminus\bar{\mathcal{B}}(x)$;

4 Find $b^k(x) = \arg\max_{j\in\mathcal{B}} \frac{r_j(x)}{\phi_j(\rho_j(k))}$;

2) The BS Side Algorithm: Upon receiving vBSs' operation status updates, virtual users select vBSs according to the user side algorithm. Then, the coverage area of the jth vBS in the kth time slot is updated as

$$\tilde{\mathcal{A}}_j(k) = \{x|j = b^k(x), \forall x \in \mathcal{A}\}. \tag{18}$$

Then, given $\boldsymbol{\rho}(k) = (\rho_1(k), \rho_2(k), \ldots, \rho_{|\mathcal{B}|}(k))$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{|\mathcal{B}|})$, and $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_{|\mathcal{B}|})$, the jth vBS's perceived traffic loads in the kth time slot is

$$M_j(\boldsymbol{\rho}(k), \boldsymbol{\theta}, \hat{\boldsymbol{\rho}}) = \min\Big(\int_{x\in\tilde{\mathcal{A}}_j(k)} \varrho_j(x)dx, 1-\epsilon\Big). \tag{19}$$

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

After having derived the perceived traffic loads, the jth vBS updates its traffic loads as

$$\rho_j(k+1) = \delta(k)\rho_j(k) + (1-\delta(k))M_j(\boldsymbol{\rho}(k)). \tag{20}$$

Here, $0 \le \delta(k) < 1$ is a system parameter calculated by the RANC to enable

$$\begin{aligned}&\psi(\boldsymbol{\rho}(k+1)) \\ &\le \psi(\boldsymbol{\rho}(k)) + \varsigma(1-\delta(k))\sum_{j\in\mathcal{B}} \phi_j(\rho_j(k))(M_j(\boldsymbol{\rho}(k)) - \rho_j(k)).\end{aligned} \tag{21}$$

Here, $0 < \varsigma < 0.5$ is a constant. In the $(k+1)$th time slot, the jth vBS's operation status is $\phi_j(\rho_j(k+1))$. The pseudo code of the BS sid algorithm is presented in Alg. 2. The computational complexity of the BS side algorithm is determined by the "while" loop whose running time depends on the convergence of $\psi(\boldsymbol{\rho}(k))$. When $\psi(\boldsymbol{\rho}(k))$ is closer to the optimal value, it may take longer time to find $\delta(k)$. In the following, we will analyze the convergence of the vGALA scheme, which reflects the computational complexity of the BS side algorithm.

Algorithm 2: The BS Side Algorithm

Input : Users' vBS selection: $b^k(x), \forall x \in \mathcal{A}$;

Output: vBSs' operation status, $\phi_j(\rho_j(k+1)), \forall j \in \mathcal{B}$;

1 vBSs measure their perceived traffic loads, $M_j(\boldsymbol{\rho}(k))$;

2 Assign $\delta(k) = 0$;

3 **while** *(21) is not true* **do**

4 $\delta(k) = 1 - \xi(1-\delta(k))$, here, $0 < \xi < 1$ is a real number;

5 vBSs update their traffic loads: $\rho_j(k+1) = \delta(k)\rho_j(k) + (1-\delta(k))M_j(\boldsymbol{\rho}(k))$;

6 Calculate $\phi_j(\rho_j(k+1))$ based on $\rho_j(k+1), \forall j \in \mathcal{B}$;

B. 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

$$\begin{aligned}\mathcal{F} = \{\boldsymbol{\rho}|\rho_j &= \int_{x\in\mathcal{A}} \varrho_j(x)dx, \\ &0 \le \rho_j \le 1-\epsilon, \sum_{j\in\mathcal{B}} \eta_j(x) = 1, \\ &\eta_j(x) = \{0,1\}, \forall j \in \mathcal{B}, \forall x \in \mathcal{A}\}.\end{aligned} \tag{22}$$

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

$$\begin{aligned}\tilde{\mathcal{F}} = \{\boldsymbol{\rho}|\rho_j &= \int_{x\in\mathcal{A}} \varrho_j(x)dx, \\ &0 \le \rho_j \le 1-\epsilon, \sum_{j\in\mathcal{B}} \eta_j(x) = 1, \\ &0 \le \eta_j(x) \le 1,\ \forall j \in \mathcal{B}, \forall x \in \mathcal{A}\}\end{aligned} \tag{23}$$

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 $\boldsymbol{\rho}$. ■

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

Proof: The lemma is proved by showing $\bigtriangledown^2\psi(\boldsymbol{\rho}) \succeq q\boldsymbol{I}$ where $q = 4e^{-1}$ and $\boldsymbol{I}$ is an identity matrix. ■

Let $\boldsymbol{M}(\boldsymbol{\rho}) = \{M_1(\boldsymbol{\rho}), M_2(\boldsymbol{\rho}), \ldots, M_{|\mathcal{B}|}(\boldsymbol{\rho})\}$.

Lemma 3: When $\boldsymbol{M}(\boldsymbol{\rho}(k)) \neq \boldsymbol{\rho}(k)$, $\boldsymbol{M}(\boldsymbol{\rho}(k))$ provides a descent direction of $\psi(\boldsymbol{\rho})$ at $\boldsymbol{\rho}(k)$.

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

$$\langle \bigtriangledown\psi(\boldsymbol{\rho})|_{\boldsymbol{\rho}=\boldsymbol{\rho}(k)}, \boldsymbol{M}(\boldsymbol{\rho}(k)) - \boldsymbol{\rho}(k)\rangle < 0. \tag{24}$$

Let $\hat{\eta}_j(x)$ and $\eta_j(x)$ be the user association indication of the jth BS that result in the traffic load $M_j(\boldsymbol{\rho}(k))$ and $\rho_j(k)$, respectively.

$$\begin{aligned}
&\langle \bigtriangledown\psi(\boldsymbol{\rho})|_{\boldsymbol{\rho}=\boldsymbol{\rho}(k)}, \boldsymbol{M}(\boldsymbol{\rho}(k)) - \boldsymbol{\rho}(k)\rangle \\
&= \sum_{j\in\mathcal{B}} (M_j(\boldsymbol{\rho}(k)) - \rho_j(k))\phi_j(\rho_j(k)) \\
&= \sum_{j\in\mathcal{B}} \frac{\int_{x\in\mathcal{A}} \lambda(x)\nu(x)(\hat{\eta}_j(x) - \eta_j(x))dx}{r_j(x)\phi_j^{-1}(\rho_j(k))} \\
&= \int_{x\in\mathcal{A}} \lambda(x)\nu(x) \sum_{j\in\mathcal{B}} \frac{\hat{\eta}_j(x) - \eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j(k))} dx.
\end{aligned} \tag{25}$$

Since

$$\hat{\eta}_j(x) = \begin{cases} 1, & \text{for } j = b^k(x) \\ 0, & \text{for otherwise,} \end{cases} \tag{26}$$

$$\sum_{j\in\mathcal{B}} \frac{\hat{\eta}_j(x) - \eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j(k))} \leq 0. \tag{27}$$

Because $\boldsymbol{M}(\boldsymbol{\rho}(k)) \neq \boldsymbol{\rho}(k)$, there exists $j \in \mathcal{B}$ such that $\hat{\eta}_j(x) \neq \eta_j(x)$, $x \in \mathcal{A}$. Hence,

$$\sum_{j\in\mathcal{B}} \frac{\hat{\eta}_j(x) - \eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j(k))} < 0, \tag{28}$$

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

Theorem 1: The traffic load vector $\boldsymbol{\rho}$ converges to the traffic load vector $\boldsymbol{\rho}^* \in \mathcal{F}$.

Proof: Since $\sum_{j\in\mathcal{B}}(M_j(\boldsymbol{\rho}(k)) - \rho_j(k))\phi_j(\rho_j(k)) < 0$ when $\boldsymbol{M}(\boldsymbol{\rho}(k)) \neq \boldsymbol{\rho}(k)$, Alg. 2 ensures $\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k))$ in each time slot. Since $\psi(\boldsymbol{\rho}) \geq 0$, $\psi(\boldsymbol{\rho})$ will converge. Let $\psi(\boldsymbol{\rho})$ converge to $\psi(\boldsymbol{\rho}^*)$. Since

$$\begin{aligned}
\boldsymbol{\rho}(k+1) &= \delta(k)\boldsymbol{\rho}(k) + (1-\delta(k))(\boldsymbol{M}\boldsymbol{\rho}(k)) \\
&= \boldsymbol{\rho}(k) + (1-\delta(k))(\boldsymbol{M}(\boldsymbol{\rho}(k)) - \boldsymbol{\rho}(k)).
\end{aligned} \tag{29}$$

$\boldsymbol{M}(\boldsymbol{\rho})$ and $\boldsymbol{\rho}$ will converge to $\boldsymbol{\rho}^*$. Because $\boldsymbol{M}(\boldsymbol{\rho}*)$ is derived based on the user side algorithm where $\eta_j^m(x) = \{0,1\}, \forall j \in \mathcal{B}, x \in \mathcal{A}$, $\boldsymbol{\rho}^*$ is in the feasible set $\mathcal{F}$. ■

Corollary 1: The vBSs' operation status $\phi_j(\rho_j)$, $\forall j \in \mathcal{B}$, converges to $\phi_j(\rho_j^*)$.

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

Since $\psi(\boldsymbol{\rho})$ is a strong convex function, there exists $q > 0$ and $Q > 0$ such that $q\boldsymbol{I} \preceq \bigtriangledown^2\psi(\boldsymbol{\rho}) \preceq Q\boldsymbol{I}$, $\boldsymbol{\rho} \in \tilde{\mathcal{F}}$[24]. Denote $\psi(\boldsymbol{\rho}^*)$ as the optimal solution. $\psi(\boldsymbol{\rho}(k+1))$ is said to be the ϵ suboptimal solution if $\psi(\boldsymbol{\rho}(k+1)) - \psi(\boldsymbol{\rho}^*) \leq \epsilon$ where $\epsilon > 0$ is a small real number.

Lemma 4: The number of iterations required to ensure $\psi(\boldsymbol{\rho}(k+1)) - \psi(\boldsymbol{\rho}^*) \leq \epsilon$ is at most equal to

$$\frac{\log((\psi(\boldsymbol{\rho}(1)) - \psi(\boldsymbol{\rho}^*))/\epsilon)}{\log 1/z} \tag{30}$$

where $z = 1 - \min\{2q\varsigma, 2q\varsigma\xi/Q\} < 1$ and $\boldsymbol{\rho}(1)$ is the initial traffic load vector.

Proof: The lemma is proved in Appendix A. ■

Equation (30) indicates that $\psi(\boldsymbol{\rho})$ converges at least as fast as a geometric series. Such convergence is called linear convergence in the context of iterative numerical method [24]. The number of iterations required for $\psi(\boldsymbol{\rho})$ to converge depends on the gap between $\psi(\boldsymbol{\rho}(1))$ and $\psi(\boldsymbol{\rho}^*)$, ϵ, and z. Given the gap and the value of ϵ, a smaller z enables faster convergence. By properly selecting ς and ξ, we can reduce the value of z, and thus reduce the number of iterations required for the convergence. However, how to optimize the value of ς and ξ is beyond the scope of this paper.

C. The Optimality of vGALA

Since the vBSs' traffic load vector converges to $\boldsymbol{\rho}*$, we show that the corresponding user association minimizes $\psi(\boldsymbol{\rho})$.

Theorem 2: Suppose $\mathcal{F}$ is not empty and the traffic load vector converges to $\boldsymbol{\rho}*$, the user association corresponding to $\boldsymbol{\rho}*$ minimizes $\psi(\boldsymbol{\rho})$.

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

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

$$\langle \bigtriangledown\psi(\boldsymbol{\rho})|_{\boldsymbol{\rho}=\boldsymbol{\rho}^*}, \triangle\boldsymbol{\rho}^*\rangle \geq 0. \tag{31}$$

$$\begin{aligned}
&\langle \bigtriangledown\psi(\boldsymbol{\rho})|_{\boldsymbol{\rho}=\boldsymbol{\rho}^*}, \triangle\boldsymbol{\rho}^*\rangle \\
&= \sum_{j\in\mathcal{B}} (\rho_j - \rho_j^*)\phi_j(\rho_j^*) \\
&= \sum_{j\in\mathcal{B}} \frac{\int_{x\in\mathcal{A}} \lambda(x)\nu(x)(\eta_j(x) - \eta_j^*(x))dx}{r_j(x)\phi_j^{-1}(\rho_j^*)} \\
&= \int_{x\in\mathcal{A}} \lambda(x)\nu(x) \sum_{j\in\mathcal{B}} \frac{\eta_j(x) - \eta_j^*(x)}{r_j(x)\phi_j^{-1}(\rho_j^*)} dx.
\end{aligned} \tag{32}$$

According to the user side algorithm,

$$\eta_j^*(x) = \begin{cases} 1, & \text{for } j = \arg\max_{i\in\mathcal{B}} \frac{r_i(x)}{\phi_i(\rho_i^*)} \\ 0, & \text{for otherwise} \end{cases}. \tag{33}$$

Therefore,

$$\sum_{j\in\mathcal{B}} \frac{\eta_j^*(x)}{r_j(x)\phi_j^{-1}(\rho_j^*)} \leq \sum_{j\in\mathcal{B}} \frac{\eta_j(x)}{r_j(x)\phi_j^{-1}(\rho_j^*)}. \tag{34}$$

Hence, $\langle \bigtriangledown\psi(\boldsymbol{\rho})|_{\boldsymbol{\rho}=\boldsymbol{\rho}^*}, \triangle\boldsymbol{\rho}^*\rangle \geq 0$. ■

D. The Generalization of vGALA

In determining the user association, the vGALA scheme strives for a balance between the green energy utilization

and the network performance. In the problem formulation, $w_j(\rho_j)$ and $L(\rho_j)$ model the green energy utilization and the network performance, respectively. Since $w_j(\rho_j)$ and $L(\rho_j)$ are functions of the traffic load ρ_j, they are coupled by ρ_j. $L(\rho_j)$ is a general latency indicator derived under the M/G/1 processor sharing queue model. In practical networks, traffic arrivals may follow arbitrary distributions rather than a Poisson distribution. In addition, the network operators may aim to represent the network performance with other metrics instead of the average traffic delivery latency. It is desirable that the vGALA framework can be applied to a collection of network performance models. Denote $f(\rho_j)$ as a function of the traffic load ρ_j that models the jth BS's performance. Define the user association problem with a generalized network performance model, $f(\rho_j)$, as the UAG problem expressed as

$$\min_{\boldsymbol{\rho}} \quad \sum_{j\in\mathcal{B}} w_j(\rho_j) f(\rho_j) \tag{35}$$

$$subject\ to: \quad 0 \le \rho_j \le 1-\epsilon. \tag{36}$$

Lemma 5: If $f(\rho_j)$ is positive, convex and non decreasing over $\rho_j, \forall j \in \mathcal{B}$, $\tilde{\psi}(\boldsymbol{\rho}) = \sum_{j\in\mathcal{B}} w_j(\rho_j) f(\rho_j)$ is convex over $\boldsymbol{\rho} \in \tilde{\mathcal{F}}$.

Proof: Since $f(\rho_j)$ is positive, convex and non decreasing, $f(\rho_j) > 0$, $f''(\rho_j) \ge 0$, and $f'(\rho_j) \ge 0$. Because $w_j''(\rho_j) > 0$, $w_j'(\rho_j) > 0$, and $w_j(\rho_j) > 0$,

$$\begin{aligned} &\frac{\partial^2 \sum_{j\in\mathcal{B}} w_j(\rho_j) f(\rho_j)}{\partial \rho_j^2} \\ &= w_j''(\rho_j) f(\rho_j) + 2 w_j'(\rho_j) f'(\rho_j) + w_j(\rho_j) f''(\rho_j) \\ &\ge q. \end{aligned} \tag{37}$$

Here, q is a positive number. Let $\boldsymbol{I}$ be an identity matrix. Since

$$\frac{\partial^2 \sum_{j\in\mathcal{B}} w_j(\rho_j) f(\rho_j)}{\partial \rho_j \partial \rho_i} = 0, \forall i \ne j, \tag{38}$$

$\nabla^2 \tilde{\psi}(\boldsymbol{\rho}) \ge q\boldsymbol{I}$. Therefore, $\tilde{\psi}(\boldsymbol{\rho})$ is a strong convex function over $\boldsymbol{\rho}, \boldsymbol{\rho} \in \tilde{\mathcal{F}}$. ■

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

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

V. The Practicality of the vGALA Scheme

In this section, we first present how to put the vGALA framework into practice and evaluate the assumptions made for developing the scheme. Then, we discuss two related issues on applying the vGALA scheme: the energy-latency trade-off and the admission control mechanism.

A. The Practical Implementation

In practical cellular networks, the traffic load balancing among BSs is usually triggered by network-level events, e.g., some BSs are congested while others are lightly loaded, rather than by user-level events, e.g., a few users' movement and data rate changes. Since a BS's traffic loads are determined by the average traffic load density of its coverage area, without considering green energy, it is reasonable to reduce a BS's coverage area to avoid traffic congestion if the traffic load density of the BS's coverage area is increasing. Therefore, a BS's traffic load can be derived based on the location-based traffic load density that reflect the traffic load density at a location. Thus, for practical implementation, the vGALA scheme collects the location-based traffic load density and the network green energy information in the first phase as shown in Fig. 3. Given a specific location, it is realistic to assume that BSs' downlink data rates to users at the location are not changing during a traffic load balancing period. Notice that on modeling the traffic load in the UA problem, we differentiate users by their locations. Therefore, the vGALA scheme is compatible with the input of the location-based traffic load density and the location-based downlink data rates.

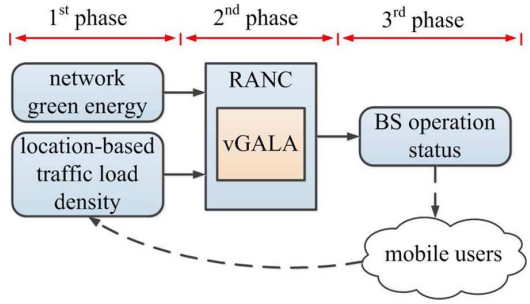

Fig. 3. The practical implementation of vGALA.

In the second phase, the vGALA scheme implemented in the RANC optimizes the user association and derives the optimal BS operation status based on the network information collected in the first phase. The optimization can be triggered either periodically or by some predefined events, e.g., a BS's traffic loads exceed a threshold or a BS's green energy utilization is lower than a threshold. What are the best strategies for triggering the traffic load balancing can be determined by network operators and is beyond the scope of this paper. The output of the second phase is the BS operation status, based on which the user association is determined in the third phase. In this phase, a user's BS association can be determined in either centralized or distributed fashion. In the first case, users send their data rate measurements to the RANC, and the RANC determines the users' BS associations based on the BS operation status and the users' date rates. In the second case, the RANC may simply let BSs broadcast their operation statuses, and based on which individual users decide their own BS associations. The users' BS selections may change the location-based traffic load density. Individual BSs translate the users' BS selections to location-based traffic load density and report it to the RANC.

In the vGALA scheme, the user association is optimized with the consideration of both the average traffic delivery latency and the green energy usage. From users' point of view (who may not care about the green energy usage), they may seek to maximize their performance and violate the BS selection rule in the vGALA scheme. However, the users, in fact, do not have any clue on maximizing their own QoS. According to (17), a

user's BS selection is based on both $r_j(x)$ and $\phi_j(\rho_j)$. Here, $\phi_j(\rho_j)$ is determined by both the jth BS's traffic loads and its available green energy. A user's average traffic delivery latency is determined by both the downlink data rate and the traffic loads of the associated BS. Since the users do not know the traffic loads of BSs, the users have no clue about which BS can provide them the best QoS. Simply selecting a BS with the largest $r_j(x)$ may lead the users to a highly congested BS and degrade the users' QoS. Thus, the users do not have obvious incentives to counterfeit their measurement reports.

B. 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 θ and κ. θ is the energy-latency coefficient of a BS. It reflects individual BSs' operation strategies. A BS with a large $\theta(\theta \rightarrow 1)$ indicates that the BS is energy-sensitive. When a BS chooses a small $\theta(\theta \rightarrow 0)$, the BS is latency-sensitive. Therefore, by choosing the value of θ, a BS adapts its sensitivity about the on-grid power consumption and the average traffic delivery latency. Hence, θ is chosen by individual BSs based on their operation strategies.

κ is chosen by the RANC based on the global view of green energy status and the mobile traffic demands. Given θ and the available green energy, $w_j(\rho_j)$ grows exponentially as the traffic demand increases. For a large κ, $w_j(\rho_j)$ grows faster than it does with a small κ. This indicates that the vGALA scheme is more energy-sensitive when κ is assigned a larger value. When κ keeps increasing, the vGALA scheme will perform similarly as 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 κ restricts the individual BSs' capability in adapting the energy-latency trade-off. The adaptation of κ can be triggered by either green energy changes or the mobile traffic demand changes. For example, when the network experiences heavy traffic loads, the RANC will focus on balancing the traffic loads to reduce the network congestion. In this case, the RANC may choose a small κ to give a high priority to the latency awareness in balancing the traffic loads. On the other hand, if the network experiences light traffic loads, the RANC may increase κ to emphasize the green energy usage.

Traffic load balancing parameters do impact the convergence of traffic load balancing algorithms such as the one reported in [25]. For the vGALA scheme, κ and θ determine the energy-latency trade-off of the network. As a result, they determine the optimal traffic loads of individual BSs. According to Lemma 4, the number of iterations required for the vGALA scheme to converge depends on $\psi(\boldsymbol{\rho}(1)) - \psi(\boldsymbol{\rho}^*)$. Here, $\boldsymbol{\rho}(1)$ and $\boldsymbol{\rho}^*$ are the initial and the optimal traffic load vectors, respectively. The optimal traffic load vector is related to the energy-latency trade-off of the network. Therefore, κ and θ affect the convergence of the vGALA scheme: when κ and θ enables the vGALA scheme to achieve a higher network performance enhancement in terms of minimizing $\psi(\boldsymbol{\rho})$, the difference between $\psi(\boldsymbol{\rho}(1))$ and $\psi(\boldsymbol{\rho}^*)$ will be larger, and as a result, the vGALA scheme requires more iterations to converge.

C. The 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 loads should be within the feasible set defined in (22). When the traffic loads are 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. Thus, the purpose of proposing a simple admission control mechanism is to ensure that the vGALA scheme works even under very heavy traffic load condition (when the UA problem is not feasible) rather than to reduce either the energy consumption or average traffic delivery latency of the network.

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. The RANC assigns $\mu(x)$ to a user at location x. $\mu(x)$ does not depend on the user's 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 of the users. The coverage area of a BS, e.g., $\tilde{\mathcal{A}}_j(k)$, is still calculated by (18). Owing to the admission control, the traffic load measurement in the jth vBS is revised as

$$M_j(\boldsymbol{\rho}(k)) = \min\Big(\int_{x \in \tilde{\mathcal{A}}_j(k)} \mu(x)\varrho_j(x)dx, 1-\epsilon\Big). \tag{39}$$

The vBS updates its traffic loads based on (20).

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

$$\begin{aligned}\hat{\mathcal{F}} = \{\boldsymbol{\rho} | \rho_j = &\int_{x \in \mathcal{A}} \mu(x)\varrho_j(x)dx, \\ &0 \leq \rho_j \leq 1-\epsilon, \sum_{j \in \mathcal{B}} \eta_j(x) = 1, \\ &0 \leq \eta_j(x) \leq 1, \forall j \in \mathcal{B}, \forall x \in \mathcal{A}\}.\end{aligned} \tag{40}$$

Since $0 \leq \mu(x) \leq 1$ is a constant, Lemma 1 still holds, which means that $\hat{\mathcal{F}}$ is a convex set. Integrating admission control does not change the objective function of the UA problem. Thus, Lemma 2 also holds. By applying the similar analysis presented in Sections IV-B and IV-C, we can prove that the vGALA scheme still enables the convergence of the traffic loads and obtains the optimal solution to the UA problem with the admission control.

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 2000 m $\times$ 2000 m area. The traffic arrival in the area follows the Poisson point process with the average arrival rate equaling to 200. The traffic size per arrival is 250 kb. The area is divided into 40,000 locations with each location representing a 10 m $\times$10 m area. The location-based traffic load density is calculated based on the traffic model. The static power consumption of the MBS and the SCBS

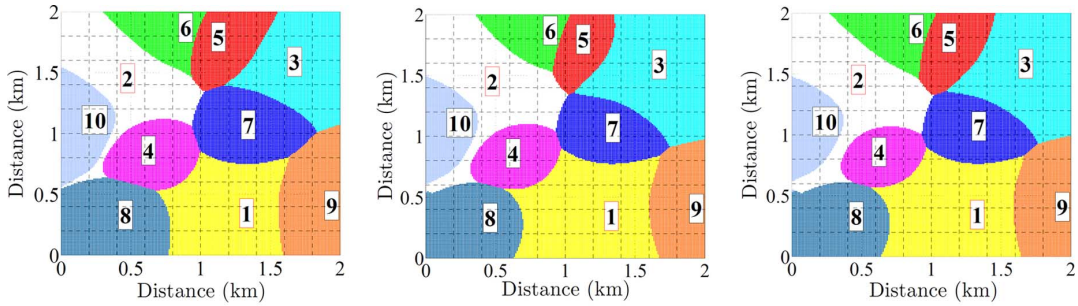

Fig. 4. The coverage areas of different user association schemes. (a) Green energy aware (GA). (b) Latency aware (LA). (c) vGALA ($\theta = 0.8$, $\kappa = 4$).

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

TABLE II
VGALA IMPLEMENTATION PARAMETERS

Parameters	θ	κ	ϵ	ς	ξ
Value	$0 < \theta < 1$	$1 \leq \kappa \leq 20$	0.001	0.4	0.7

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% [26]. We assume that the weather condition is the standard condition which specifies a temperature of 25°C, an irradiance of 1000 W/m^2, and an air mass of 1.5 spectrum. Thus, the green energy generation rate is 174 W/m^2. 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. In other words, we let $\theta_i = \theta_j = \theta, \forall i, j \in \mathcal{B}$ in the simulation. The value of θ varies for different simulation scenarios. The total bandwidth is 20 MHz in which 10 MHz is exclusively used by MBSs and the other 10 MHz is allocated to SCBSs. The frequency reuse factor for each system (MBSs and SCBSs) is one. The channel propagation model is based on COST 231 Walfisch-Ikegami [27]. 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. The values of various parameters used to implement the vGALA scheme are summarized in Table II. While the values of θ and κ vary for different simulation scenarios, ϵ, ς, and ξ are fixed.

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

$$\min_{\boldsymbol{\rho}} \quad \sum_{j \in \mathcal{B}} \max(\rho_j - e_j, 0) \tag{41}$$

$$subject\ to: \quad 0 \leq \rho_j \leq 1 - \epsilon. \tag{42}$$

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

$$\min_{\boldsymbol{\rho}} \quad \sum_{j \in \mathcal{B}} L(\rho_j) \tag{43}$$

$$subject\ to: \quad 0 \leq \rho_j \leq 1 - \epsilon. \tag{44}$$

As shown in Fig. 4, 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[2] . A larger coverage area indicates the BS serves more traffic loads. The first, second and third BSs are MBSs and the other BSs are SCBSs. Taking the coverage area of the 5th BS as an example, as compared with the GA scheme (Fig. 4(a)), the LA scheme significantly reduces the BS's coverage area as shown in Fig. 4(b). The 5th BS has sufficient green energy. Therefore, the GA scheme will redirect more traffic loads to the BS to minimize the on-grid power consumption. The LA scheme, which does not consider the energy usage, balances the traffic loads among BSs to minimize the average traffic delivery latency. As a result, the LA scheme limits the traffic loads in the BS. In terms of the power consumption and the average traffic delivery latency, as compared with the GA scheme, the vGALA scheme slightly reduces the BS's coverage area as shown in Fig. 4(c) to obtain a trade-off between the on-grid power consumption and the average traffic delivery latency.

Fig. 5 shows the trade-off achieved by the vGALA scheme between the on-grid energy consumption and the average traffic delivery latency. Fig. 5(a) shows the on-grid power consumption of the LA, vGALA, and GA schemes, respectively. As compared with the LA scheme, the vGALA scheme consumes 30% less on-grid power. Fig. 5(b) shows that the average traffic delivery latency of the vGALA scheme is only 8% more than that of the LA scheme. While the GA scheme significantly reduces the on-grid power consumption, it increases the traffic delivery latency by about 48% percent as compared with the vGALA scheme. Here, the latency indicator equals to $\sum_{j \in \mathcal{B}} L(\rho_j)$. The above observation indicates that the vGALA scheme achieves a preferable trade-off: saving 30% on-grid power at the cost of 8% increase in the average traffic delivery latency. In addition, as shown in Fig. 5, the vGALA scheme requires about 60 iterations to converge to the optimal solution. On the one hand, it proves

[2]The white color indicates the coverage area of the second BS.

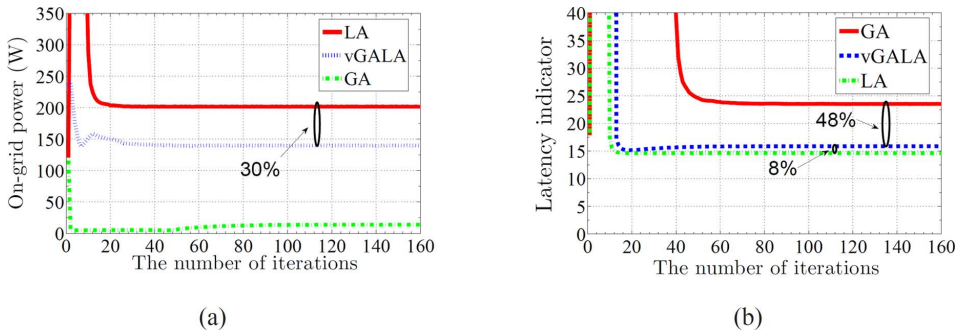

Fig. 5. The comparison of different user association scheme ($\theta = 0.8$, $\kappa = 4$). (a) The on-grid power consumption. (b) The average traffic delivery latency.

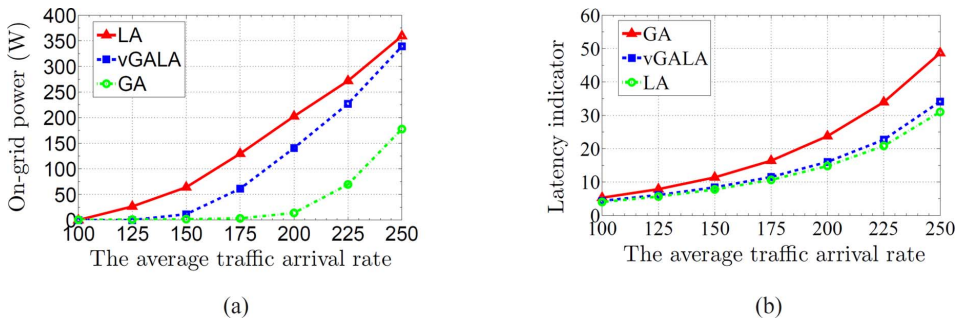

Fig. 6. The performance of user association schemes with different average traffic arrival rates ($\theta = 0.8$, $\kappa = 4$). (a) The on-grid power consumption. (b) The average traffic delivery latency.

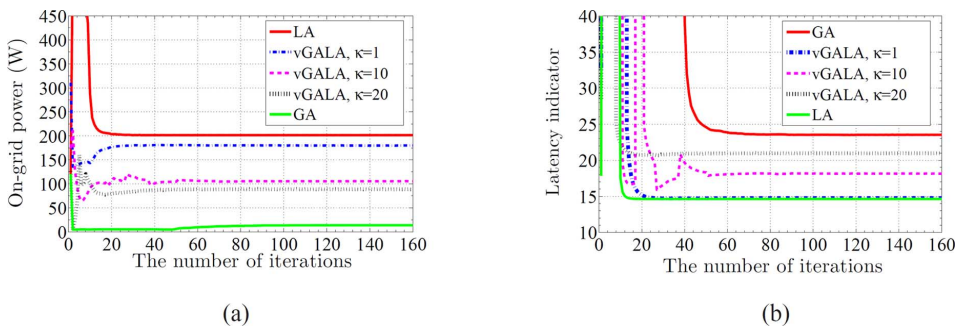

Fig. 7. The performance of vGALA with various κ ($\theta = 1$). (a) The on-grid power consumption. (b) The average traffic delivery latency.

that the vGALA scheme converges fast. On the other hand, it indicates that 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.

Fig. 6 shows the performance of the LA, vGALA and GA schemes for different average traffic arrival rates. As shown in the figure, as the average traffic arrival rate increases, the on-grid power consumption and the average traffic deliver latency of these schemes are increasing. When the average traffic arrival rate is very small, e.g., 100, these schemes exhibit similar performance because the traffic loads in individual BSs, in spite of different traffic load balancing schemes, are less than their green traffic capacity. As the average traffic arrival rate increases, the performance gap between the LA and GA schemes in terms of both on-grid power consumption and the average traffic delivery latency increases. The performance of the vGALA scheme, which is determined by θ and κ, is between that of the LA and the GA scheme. When the traffic arrival rate further increases, the traffic load balancing problem becomes infeasible.

B. *Performance Adaptation*

The trade-off between the on-grid power consumption and the average traffic delivery latency can be adapted by adjusting κ and θ in the vGALA scheme. Fig. 7 shows the performance of the vGALA scheme with different κ. By varying κ, the vGALA scheme may act as the LA scheme when $\kappa \to 0$ and performs like the GA scheme when $\kappa \to \infty$. As shown in Fig. 8, given κ, adjusting θ has a limited performance adaptation. In other words, κ defines a performance adaptation range and adjusting θ can only adapt the performance within the range. As discussed in Section V-B, the selection of θ is determined by the operation strategies of BSs while the value of κ is chosen based on network conditions, e.g., the traffic load intensity and the available green energy. In addition, the figures show how the values of κ and θ affect the number of iterations required for convergence. However, how to optimize these values is beyond the scope of this paper.

C. *Green Energy Generation Rate Evaluation*

The amount of green energy in BSs impacts the performance of the vGALA scheme. In Fig. 9, the x-axis is the solar cell

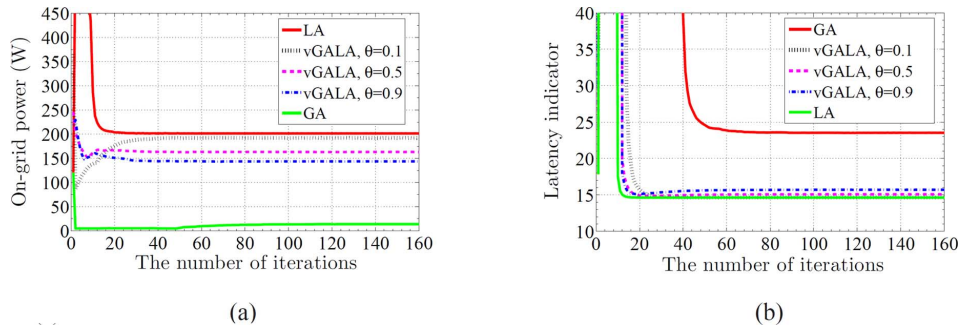

Fig. 8. The performance of vGALA with various θ ($\kappa = 4$). (a) The on-grid power consumption. (b) The average traffic delivery latency.

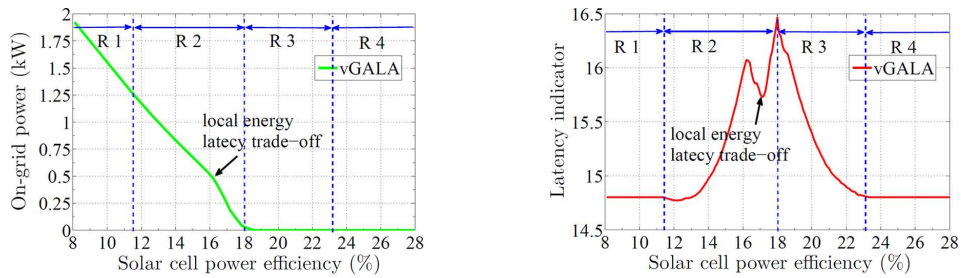

Fig. 9. The performance of vGALA versus solar cell power efficiency ($\theta = 0.8$, $\kappa = 4$). (a) The on-grid power consumption. (b) The average traffic delivery latency.

power efficiency. As the solar cell power efficiency enhances, the amount of green energy in BSs will increase. As shown in Fig. 9(a), 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. 9(b). 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 loads 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 region, green energy is not sufficient in the network. Thus, the major strategy is to trade the average traffic delivery latency for saving on-grid power. However, as the solar power efficiency increases, some BSs may have sufficient green energy and they start trading their green energy for reducing the average traffic delivery latency in the network (the solar power efficiency falls between 16% and 17%). This event reflects the local energy-latency trade-off among several BSs. 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 to further reduce 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 loads. 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. Practicality Evaluation

The cell range expansion (CRE) approach is one of the most practical traffic load balancing approach and has been proven to have similar performance as optimal traffic balancing schemes in term of maximizing network utilities [1], [8]. This simulation evaluates the traffic balancing performance of the vGALA scheme and the CRE approach. For the vGALA scheme, the simulation follows Section V-A to obtain the optimal BS operation status based on the location-based traffic load density of the coverage area and the available green energy. We adopt the two-tier data rate bias approach as the CRE approach and assume that BSs in the same tier have the data rate bias. In the simulation, MBSs are in the first tier while SCBSs are in the second tier. In the data rate bias approach, a user selects the BS to maximize the biased data rate.

$$b(x) = \arg\max_{j \in \mathcal{B}} Z_j r_j(x). \tag{45}$$

Here, $b(x)$ and Z_j are the index of the selected BS and the data rate bias of the jth BS. The data rate bias of a MBS is one. The data rate biases are selected for SCBSs to minimize 1) the average traffic delivery latency, 2) the overall on-grid power consumption, and 3) $\psi(\boldsymbol{\rho})$. Define these data rate biases as 1) CRE_LA (latency-aware), 2) CRE_GA (green energy-aware), and 3) CRE_LG (latency and green energy-aware), respectively.

In the simulation, the BS operation status and the data rate biases are calculated based on the location-based traffic load density generated in previous simulations. We randomly generate users' locations using Poisson point process[3] with average rate equalling to 200 in the area. The average traffic size per user is 250 kb. We run the simulation 10,000 times to

[3]The Poisson point process is the same as the Poisson point process used to generate the location-based traffic load density.

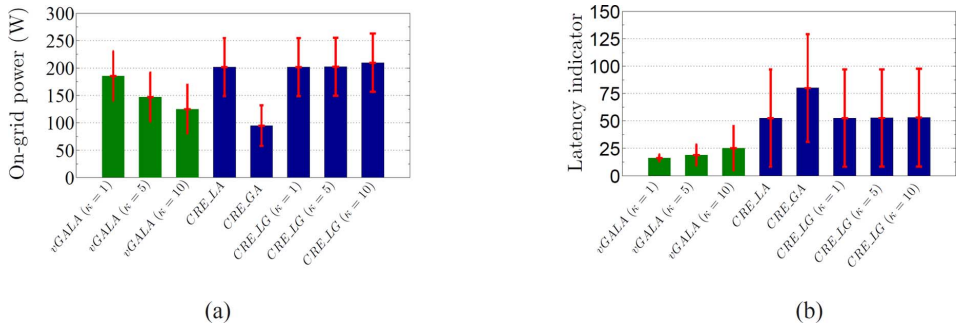

Fig. 10. The performance of vGALA versus CRE ($\theta = 0.8$). (a) The on-grid power consumption. (b) The average traffic delivery latency.

evaluate the performance of different approaches in terms of the average traffic delivery latency and the average on-grid power consumption. As shown in Fig. 10, CRE_GA achieves the minimum on-grid energy consumption among all the schemes. However, the average traffic delivery latency of CRE_GA is significantly larger than other schemes. As compared with CRE_LA and CRE_LG, the vGALA scheme not only saves the on-grid energy consumption but also reduces the average traffic delivery latency. For the vGALA scheme, when κ increases, the scheme is to gradually prioritize saving on-grid energy in balancing the traffic loads, as shown in Fig. 10(a), at the cost of a small increase of the average traffic delivery latency as shown in Fig. 10(b). For the CRE_LG scheme, increasing κ does not effectively adjust the energy-latency trade-off as the vGALA scheme does. This indicates the tier-based data rate bias approach 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, by leveraging the SoftRAN architecture, introduces virtual users and vBSs to simulate the interactions between users and BSs, thus significantly reducing the information exchanges over the air interface. The extensive simulation results have validated the performance and the practicality of the vGALA scheme.

Appendix
Proof of Lemma 4

Let $\triangle\boldsymbol{\rho}(k) = \boldsymbol{M}(\boldsymbol{\rho}(k)) - \boldsymbol{\rho}(k)$. The termination condition of the BS side algorithm (Alg. 2) can be expressed as

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) + \varsigma(1-\delta(k)) \bigtriangledown \psi(\boldsymbol{\rho})^{\top} \triangle\boldsymbol{\rho}(k). \quad (46)$$

Since $\triangle\boldsymbol{\rho}(k)$ is a descent direction of $\psi(\boldsymbol{\rho}(k))$, $\triangle\boldsymbol{\rho}(k)$ can be replaced by $-\bigtriangledown\psi(\boldsymbol{\rho})$. Thus, the termination condition of Alg. 2 can be rewritten as

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) - \varsigma(1-\delta(k))\| \bigtriangledown \psi(\boldsymbol{\rho})\|_2^2. \quad (47)$$

Next, we will prove that the termination condition is satisfied whenever $0 \leq 1-\delta(k) \leq 1/Q$. Since $\psi(\boldsymbol{\rho}) \preceq Q\boldsymbol{I}$, we can derive, according to [24],

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) + (\frac{(1-\delta(k))Q}{2} - 1)(1-\delta(k))\|\bigtriangledown\psi(\boldsymbol{\rho})\|_2^2. \quad (48)$$

When $0 \leq 1-\delta(k) \leq 1/Q$, $\frac{(1-\delta(k))Q}{2} - 1 \leq -1/2$. Therefore,

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) - \frac{(1-\delta(k))}{2}\| \bigtriangledown \psi(\boldsymbol{\rho})\|_2^2. \quad (49)$$

Since $0 < \varsigma < 0.5$, $-\frac{(1-\delta(k))}{2} \leq -(1-\delta(k))\varsigma$. Thus, we have

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) - (1-\delta(k))\varsigma\| \bigtriangledown \psi(\boldsymbol{\rho})\|_2^2, \quad (50)$$

which satisfies the termination condition of Alg. 2. Therefore, Alg. 2 terminates either with $\delta(k) = 0$ or $(1-\delta(k))$ equaling to a value that is larger than ξ/Q.

In the first case ($\delta(k) = 0$), we have

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) - \varsigma\| \bigtriangledown \psi(\boldsymbol{\rho})\|_2^2. \quad (51)$$

In the second case ($(1-\delta(k)) \geq \xi/Q$), we can derive that

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) - \varsigma\xi/Q\| \bigtriangledown \psi(\boldsymbol{\rho})\|_2^2. \quad (52)$$

Thus,

$$\psi(\boldsymbol{\rho}(k+1)) \leq \psi(\boldsymbol{\rho}(k)) - \min\{\varsigma, \varsigma\xi/Q\}\| \bigtriangledown \psi(\boldsymbol{\rho})\|_2^2. \quad (53)$$

Subtracting $\psi(\boldsymbol{\rho}^*)$ from both side, we have

$$\psi(\boldsymbol{\rho}(k+1)) - \psi(\boldsymbol{\rho}^*) \leq \psi(\boldsymbol{\rho}(k)) - \psi(\boldsymbol{\rho}^*) - \min\{\varsigma, \varsigma\xi/Q\}\| \bigtriangledown \psi(\boldsymbol{\rho})\|_2^2. \quad (54)$$

Since $q\boldsymbol{I} \preceq \bigtriangledown^2\psi(\boldsymbol{\rho})$, according to [24],

$$\| \bigtriangledown \psi(\boldsymbol{\rho}(k))\|_2^2 \geq 2q(\psi(\boldsymbol{\rho}(k)) - \psi(\boldsymbol{\rho}^*)). \quad (55)$$

Combining these together, we can derive that

$$\psi(\boldsymbol{\rho}(k+1)) - \psi(\boldsymbol{\rho}^*) \leq (1 - \min\{2q\varsigma, 2q\varsigma\xi/Q\})(\psi(\boldsymbol{\rho}(k)) - \psi(\boldsymbol{\rho}^*)). \quad (56)$$

Let $z = 1 - \min\{2q\varsigma, 2q\varsigma\xi/Q\}$ and apply the inequality recursively, we find that

$$\psi(\boldsymbol{\rho}(k+1)) - \psi(\boldsymbol{\rho}^*) \leq z^k(\psi(\boldsymbol{\rho}(1)) - \psi(\boldsymbol{\rho}^*)). \quad (57)$$

Let $z^k(\psi(\boldsymbol{\rho}(1)) - \psi(\boldsymbol{\rho}^*)) = \epsilon$; we derive that the number of iteration required to achieve ϵ optimality is

$$k = \frac{\log((\psi(\boldsymbol{\rho}(1)) - \psi(\boldsymbol{\rho}^*))/\epsilon)}{\log 1/z}. \quad (58)$$

REFERENCES

[1] J. Andrews *et al.*, "An overview of load balancing in HetNets: old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.

[2] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "Softran: Software defined radio access network," in *Proc. 2nd ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, Hong Kong, 2013, pp. 25–30.

[3] T. Han and N. Ansari, "On greening cellular networks via multicell cooperation," *IEEE Wireless Commun. Mag.*, 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 Commun. Surveys and Tutorials*, vol. 13, no. 4, pp. 524–540, 2011.

[5] T. Han and N. Ansari, "Powering mobile networks with green energy," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 90–96, Feb. 2014.

[6] "Sustainable energy use in mobile communications," Ericson Inc., White Paper, 2007.

[7] T. Han and N. Ansari, "On optimizing green energy utilization for cellular networks with hybrid energy supplies," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 3872–3882, Aug. 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," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[9] T. Han and N. Ansari, "Green-energy aware and latency aware user associations in heterogeneous cellular networks," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM'13)*, Atlanta, GA, USA, Dec. 2013.

[10] L. Wang and G.-S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks: A tutorial," *IEEE Commun. Surveys Tutorials*, vol. 15, no. 1, pp. 271–292, First Q. 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 *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2011.

[13] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[14] H. Kim, G. d. Veciana, X. Yang, and M. Venkatachalam, "Distributed $\alpha-$ optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. 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 *Proc. IEEE INFOCOM*, 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 *Proc. 2010 IEEE/ACM Int. Conf. Green Comput. Commun.*, Hangzhou, China, Dec. 2010.

[17] D. Lopez-Perez, I. Guvenc, G. D. l. Roche, M. Kountouris, T. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, 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 Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3566–3576, Jul. 2009.

[19] L. Kleinrock, *Queueing Systems: Computer Applications*. New York, NY, USA: Wiley-Interscience, 1976.

[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," *IEEE Commun. Surveys Tutorials*, 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 Trans. Mobile Comput.*, vol. 6, no. 8, pp. 960–970, Aug. 2007.

[22] G. Auer *et al.*, "How much energy is needed to run a wireless network?," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[23] M. Raj, K. Kant, and S. Das, "Energy adaptive mechanism for P2P file sharing protocols," in *Euro-Par 2012: Parallel Processing Workshops, Ser. Lecture Notes in Computer Sci Springer*, I. Caragiannis, M. Alexander, R. Badia, M. Cannataro, A. Costan, M. Danelutto, F. Desprez, B. Krammer, J. Sahuquillo, S. Scott, and J. Weidendorfer, Eds., Heidelberg, Berlin, Germany, 2013, vol. 7640, pp. 89–99.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[25] S. Imon, A. Khan, M. D. Francesco, and S. Das, "Energy-efficient randomized switching for maximizing lifetime in tree-based wireless sensor networks," *IEEE/ACM Trans. Networking*, 2014, preprint.

[26] HIT photovoltaic module, Sanyo [Online]. Available: http://us.sanyo.com/dynamic/product/

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

Tao Han (S'08) is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the New Jersey Institute of Technology (NJIT), Newark, NJ, USA. His research interests include big-data-driven communication network design, mobile and wireless networking, the Internet of Things, and green communications. He has authored 11 papers published/accepted in premium IEEE publications such as ACM/IEEE TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE *Wireless Communications*. He has also produced 13 high-quality IEEE conference papers and five U.S. non-provisional patent applications. He has been recognized with a New Jersey Inventors Hall of Fame Graduate Students Award.

Nirwan Ansari (S'78–M'83–SM'94–F'09) received the B.S.E.E. degree (*summa cum laude*) from the New Jersey Institute of Technology (NJIT), Newark, NJ, USA, the M.S.E.E. degree from the University of Michigan, Ann Arbor, MI, USA, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA.

He is Distinguished Professor of Electrical and Computer Engineering at NJIT, which he joined in 1988. He has also assumed various administrative positions at NJIT. He has been a Visiting (Chair) Professor at several universities. He coauthored *Media Access Control and Resource Allocation* (Springer, 2013) with J. Zhang and *Computational Intelligence for Optimization* (Springer, 1997) with E.S.H. Hou, and edited *Neural Networks in Telecommunications* (Springer, 1994) with B. Yuhas. He has also contributed over 450 technical papers, over one-third of which were published in widely cited refereed journals/magazines. He has guest edited a number of Special Issues, covering various emerging topics in communications and networking. His current research focuses on various aspects of broadband networks and multimedia communications.

He has served on the Editorial Board and Advisory Board of ten journals, including as a Senior Technical Editor of IEEE *Communications Magazine* (2006–2009). He was elected to serve on the IEEE Communications Society (ComSoc) Board of Governors as a Member-at-Large (2013–2015). He has chaired ComSoc Technical Committees, and has actively organized numerous IEEE international conferences/symposia/workshops, assuming various leadership roles. Some of his recognitions include several Excellence in Teaching Awards, two Best Paper Awards, the NCE Excellence in Research Award (2014), ComSoc AHSN TC Outstanding Service Recognition Award (2013), NJ Inventors Hall of Fame Inventor of the Year Award (2012), Thomas Alva Edison Patent Award (2010), and designation as an IEEE Communications Society Distinguished Lecturer (2006–2009, two terms). He has also been granted over 25 U.S. patents.