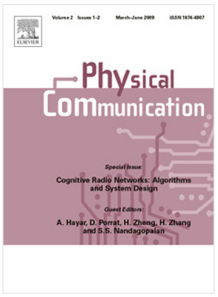
Journal Pre-proof



User association for load balancing in coordinated multipoint green HetNets: A Quasi-Newton-based approach

Mohamad Khattar Awad, Ali A.M.R. Behiry, Mohammed W. Baidas

PII: S1874-4907(21)00201-9
DOI: https://doi.org/10.1016/j.phycom.2021.101464
Reference: PHYCOM 101464

To appear in: *Physical Communication*

Received date : 25 February 2021
Revised date : 1 August 2021
Accepted date : 11 September 2021

Please cite this article as: M.K. Awad, A.A.M.R. Behiry and M.W. Baidas, User association for load balancing in coordinated multipoint green HetNets: A Quasi-Newton-based approach, *Physical Communication* (2021), doi: https://doi.org/10.1016/j.phycom.2021.101464.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Elsevier B.V. All rights reserved.

Journal Pre-proof

User Association for Load Balancing in Coordinated Multipoint Green HetNets: A Quasi-Newton-based Approach

Mohamad Khattar Awad[†], Ali A. M. R. Behiry[⋆], Mohammed W. Baidas[‡]

[†] Department of Computer Engineering, College of Engineering and Petroleum, Kuwait University

[⋆] College of Engineering and Applied Sciences, American University of Kuwait

[‡] Department of Electrical Engineering, College of Engineering and Petroleum, Kuwait University

(E-mail: mohamad@ieee.org, abehiry@auk.edu.kw, baidas@ieee.org)

September 20, 2021

Abstract

The demand for high capacity network services with stringent quality of service requirements is at a rapidly accelerating rate due to the exponential rise in the numbers of mobile-connected devices. This demand has motivated the use of the heterogeneous network (HetNet) architectures. However, even though small-cell base-stations have relatively low power consumption, the overall aggregate power consumption of a dense HetNet is significant. Due to high inter-cell-interference and imbalanced loads in dense HetNets with conventional user association techniques, cell-edge users perceive dramatically less quality of service than their cell-center counterparts. The use of a Coordinated Multipoint (CoMP) association can augment the service perceived by cell-edge users by allowing a single user to be jointly served by two base-stations. In this work, we propose a load balancing scheme for CoMP-enabled HetNets with hybrid energy supplies that jointly optimizes user latency and green energy utilization. The proposed scheme employs a fractional solution to the user association problem to decide CoMP transmission for cell-edge users, ultimately improving their data rates. Performance evaluations of the proposed scheme show a reduction in latency of 79% and on-grid power consumption by 99% compared to conventional user association schemes that associate users based on the maximum received signal strength. Furthermore, an improvement in the network sum-rate for cell-edge users by 24% has been achieved compared to the traditional association scheme and as much as 40% over other existing schemes.

Index Terms

Heterogeneous networks (HetNets), green energy, coordinated multipoint (CoMP) transmission, Broyden–Fletcher–Goldfarb–Shanno (BFGS), load balancing, user association

I. Introduction

5G is envisioned to introduce new use cases, such as the Internet-of-Things (IoT), mission-critical services, ultra-reliable communications, and real-time control applications [1]. Such use cases impose extreme quality-of-

Journal Pre-proof

Journal Pre-proof

service (QoS) and seamless connectivity requirements, and hence entail massive network capacity and improved coverage, which are already on the rise due to the exponential increase in usage of smart mobile devices [2,3]. The deployment of heterogeneous networks (HetNets) fulfills these demands, where small-cell base-stations (SBSs) are densely deployed underlaying the macro base-station (MBS) [4]. This deployment not only improves the network capacity and coverage, but also offloads the traffic from the MBS to SBSs [5]. Moreover, due to the close proximity of the users and their respective SBSs, their received signal strength is significantly improved [3]. However, due to the dense deployment of base-stations (BSs) in the network, co- and cross-tier[1] interferences increase, leading to a drastic degradation in the users' throughput [3,6]. This degradation in throughput is more substantial for cell-edge users.

Coordinated Multipoint (CoMP) transmission is a technique where users are served by more than one base-station simultaneously. It is considered as a promising solution for mitigating inter-cell interference (ICI) for cell-edge users [7], thereby improving their signal quality. Particularly, multiple neighboring BSs can coordinate and jointly transmit signals to users in order to mitigate such interference. Therefore, in HetNets, coordination between the MBS and SBSs and/or among the SBSs could help reduce the cross- and co-tier interferences, respectively, for cell-edge users and consequently improve their data rates [6,8]. An illustration of the CoMP HetNets powered by grid (in gray) and green energy sources is shown in Fig. 1.

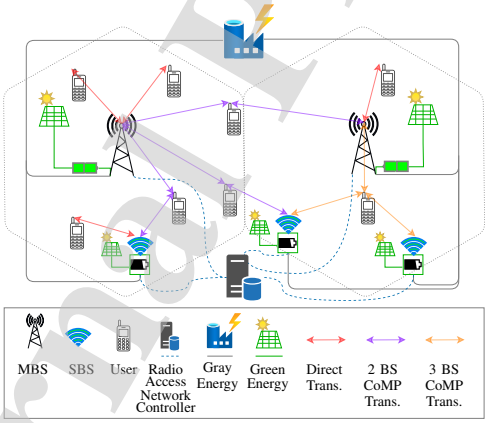


Fig. 1. Illustration of a CoMP-enabled HetNets powered by grid and green energy sources.

Two significant challenges emerge due to the deployment of CoMP-enabled dense HetNets, namely increased power consumption [9,10] and imbalanced traffic load distribution among cells [4,5,8,10]. The utilization of renewable energy sources could aid in reducing the on-grid power consumption, and thus minimizing CO_2 emissions [11].

[1]Co-tier interference is the interference among BSs that belong to the same tier, e.g., among SBSs. On the other hand, cross-tier interference is the interference caused between entities of different tiers, i.e., between MBSs and SBSs.

Journal Pre-proof

To support the utilization of renewable energy sources and flexible control of the network traffic load distribution, softwarization of radio access networks (RAN) becomes essential [12]. The software-defined RAN (SoftRAN) is an emerging centralized architecture, which abstracts BSs in a geographical area as one virtual BS and considers physical BSs as remote radio heads (RRHs) [13,14]. SoftRAN facilitates unprecedented programmability of network control functions, (e.g., power management and load balancing), based on a broader view of the network status, i.e., BSs' stored renewable energy, users distribution, and traffic demand; hence, overcoming challenges posed by enabling CoMP transmissions in HetNets.

For renewable energy-powered SoftRAN-based and CoMP-enabled HetNets, optimization of user association, i.e., assigning users to BSs, is critical to exploit network capacity and energy resources. However, associating users with MBSs or SBSs storing significant renewable energy overloads these BSs with network traffic, resulting in a substantial increase in traffic latency. On the other hand, associating users with BSs in such a way traffic latency is minimized may shift the network traffic load to BSs with limited access to green energy, consequently, increasing the on-grid power consumption. Hence, this trade-off between renewable energy utilization and network capacity exploitation must be warily managed through user association-driven traffic load balancing. However, conventional user association schemes, (e.g. [15]–[18] reviewed in Section II), aim to maximize perceived user signal quality. Such schemes are optimized for homogeneous and grid-powered traditional networks and therefore, may fail in dense HetNets, or at best, yield performance that is severely far from optimal [15]. In particular, such schemes may result in extremely imbalanced cells' traffic loads [8], and thus the degradation of user rates. The latter problem is particularly severe at cell-edges, where users suffer from poor reception. Furthermore, conventional schemes overlook the impact of balancing traffic load distribution on the network power efficiency. Moreover, the previous works reviewed in Section II employ CoMP-enabled techniques to enhance network throughput and data rates of users located at the edge of the cell. However, the question of how to load-balance traffic while providing QoS guarantees in CoMP-enabled hybrid energy powered HetNets remains unanswered.

In this work, we tackle the problem of balancing the loads among BSs of CoMP-enabled and SoftRAN-based HetNets powered by both on-grid and renewable-energy supplies. To this end, we propose a centralized load balancing user association scheme, which jointly optimizes latency and power consumption, given the network-wide information on average traffic loads and available green energy. This joint optimization is modeled as a cost function and minimized using the approximated Newton direction, which is evaluated using the method jointly proposed by Broyden, Fletcher, Goldfarb and Shanno (BFGS) [19]. A trade-off coefficient is implemented to balance power consumption and latency, while leveraging CoMP transmissions in order to improve QoS perceived by cell-edge users. Specifically, the objectives of our proposed scheme are as follows: (1) balance the distribution of downlink traffic loads among the BSs in such a way that the latency across BSs is optimized, (2) utilize the available green energy sources to minimize the on-grid power consumption, and (3) improve users' rates located at the cell-edge via CoMP-enabled transmissions. The main contributions of this work are summarized as follows:

- We develop a user association optimization model for software-defined networks. The model handles the trade-off between latency minimization and on-grid power consumption conservation, while accounting for CoMP

Journal Pre-proof

transmissions.

- Accordingly, we develop a centralized user association scheme that optimizes this trade-off, while enabling CoMP transmissions for cell-edge users.
- We adopt a Quasi-Newton method leveraging the BFGS method of approximating the inverse Hessian of an objective function. This allows second-order information to be employed without any storage or computational constraints.

The rest of this paper is organized as follows. We review related works in Section II. In Section III, we present the system model and related network assumptions. In Section IV, we devise the proposed user association scheme. Section V presents the simulation results of the proposed scheme as well as other benchmark schemes. In Section VI, conclusions are drawn from simulation results.

II. Literature Review

There are numerous efforts dedicated to traffic load balancing and distribution for cellular HetNets. One approach is completely turning OFF a BS after offloading its traffic to neighboring BSs. For instance, the authors of [20] proposed a low-complexity ON/OFF switching and radio resource management algorithm to optimize the energy efficiency in dense HetNets. Likewise, the authors of [21] proposed an energy-saving algorithm employing joint user association, clustering, and ON/OFF strategies. However, switching BSs ON/OFF merely based on traffic demands without consideration of available green energy may lead to under-utilization of renewable energy resources. The authors of [22] switch BSs ON/OFF, and optimize subcarrier assignment and energy allocation to minimize the average network power consumption. The authors of [23] implement ON/OFF BS switching as well as power control to satisfy QoS requirements, while utilizing green energy. However, under heavy traffic conditions—where all BSs are ON to support traffic loads—such ON/OFF switching-based algorithms lose their advantage and struggle with imbalanced traffic load distributions on BSs.

An alternative approach to ON/OFF switching is redistributing traffic to active BSs in order to maintain a specific performance measure while optimizing resource allocation. In [24], a joint BS activation and user association scheme is proposed to load-balance backhaul in dense HetNets. The authors of [25] proposed a distributed load-balancing scheme for renewable energy-powered HetNets, which redistributes traffic loads to optimize users' throughput. In [26], traffic loads were redistributed to achieve proportional-fairness of energy and load-based logarithmic utility functions. The authors of [16] formulated the users association problem as a control problem, where the controller aims to guarantee the required QoS given the available renewable energy.

The aforementioned works have mainly focused on conserving the on-grid consumed energy while maintaining proportional fairness of traffic loads on BSs, or optimizing user rates. However, the impact of high traffic loads on traffic delivery latency is not captured by users' achievable throughput. In [17], latency-aware load balancing is considered. A distributed latency and energy-aware user association scheme for 3-tier HetNets is proposed in [27], while a software-defined centralized latency and energy-aware load balancing scheme is devised for HetNets in

Journal Pre-proof

[28]. Despite the significant energy savings and overall network QoS enhancements achieved by the aforementioned schemes, they provide no guarantees on the QoS performance of cell-edge users.

In HetNets, CoMP transmissions can be used to improve the QoS preceived by cell-edge users. The authors of [18] proposed a greedy algorithm combined with a neural network-based algorithm to balance network throughput and traffic loads for CoMP-enabled HetNets. Results demonstrate significant improvements in the QoS perceived by cell-edge users. A load-aware CoMP-enabled HetNet with an arbitrary number of BS tiers is considered in [3]. The approach aims to decrease the probability of void BSs, while improving downlink data rates and coverage area of the network. The authors of [29] proposed a resource allocation algorithm to maximize energy efficiency for both dense green HetNets and network backhaul. Cell range expansion (CRE) is a popular technique to load-balance users in HetNets. However, biasing different BSs to achieve specific performance measures can be a challenging problem. The authors of [30] propose a Q-learning based selection strategy (QSS) to bias user rates and achieve load-balancing. In [31], particle swarm optimization is used to optimize the biases for the CRE approach and achieve load-balancing. In [32], the authors evaluate the effectiveness of load-balancing using CRE in CoMP-enabled HetNets under different biases.

III. Network Model and Problem Formulation

A downlink HetNet consisting of two types of BSs, MBSs and SBSs, is considered. Both types of BSs are connected to local renewable energy sources as well as the power grid. Particularly, the grid complements the network energy needs when renewable energy is insufficient. In the following subsections we present our adopted network architectural model, traffic model, and energy model, followed by problem.

A. Network Architectural Model

Enabling CoMP transmissions in HetNets requires a high level of coordination among MBSs and SBSs, which the SoftRAN architecture can naturally facilitate [33]. SoftRAN first appeared in the pioneering work of Gudipati et al. [14]; when they proposed softwarization of the RAN via abstraction of the BSs as a single programmable virtual-BS consisting of a radio access network controller (RANC) and RRHs representing the physical BSs. Therefore, the control plane is decoupled from the user plane, allowing for an intuitive transition to programmable networks and innovative development of novel services [33]. The RANC receives periodic updates from the physical BSs on the interference map, traffic flows data, and user's channel state information, based on which control decisions are made. These decisions are reported back to physical BSs as commands and configurations for implementation.

Several recent efforts have been devoted to the realization of the SoftRAN architecture in evolving wireless networks. For example, in [33], a flexible and programmable platform for SoftRAN consisting of custom-tailored southbound application-programming-interfaces (APIs) was developed. Similarly, the authors in [34] developed a flexible and programmable platform specifically for heterogeneous 5G RANs. Furthermore, recently the authors in [12] presented a benchmarking tool for the SoftRAN architecture and its controllers. In addition, in recent 3rd

Journal Pre-proof

Generation Partnership Project (3GPP) standards on 5G and next-generation RAN (NG-RAN) architectures [35], the separation between the control plane and user plane is strongly adopted.

In this paper, we adopt a SoftRAN architecture under which the physical MBSs and SBSs provide the RANC with a full view of the network status; e.g. estimates of harvested green energy and average traffic loads. Based on this status, the proposed scheme associates users with BSs to balance the traffic-load, such that a well-balanced trade-off between traffic latency and on-grid power consumption is maintained. Then, the RANC reports the user association decisions to the physical BSs, which routes flows and transmits signals accordingly.

B. Network Traffic Model

Let the set of BSs serving a geographical area $\mathcal{A}$ be denoted by $\mathcal{B}$. Furthermore, let $x \in \mathcal{A}$ refer to a specific location on a two-dimensional grid in the same geographical area[2] Therefore, a user is uniquely identifiable by x and individual users can be referred to by x. Then, the average signal-to-interference-plus-noise ratio (SINR) measured at location x from BS j can be expressed as

$$\mathrm{SINR}_j(x) = \frac{P_j g_j(x)}{\sigma^2 + \sum_{k \in \mathcal{I}_j(x)} I_k(x)}, \tag{1}$$

where P_j denotes the transmission power of BS j to all users. Since the focus of this work is on user association rather than power allocation and resource allocation, it is common to assume a fixed equal transmission power to all users, as in [17,28,36,37]. Then, users channel gains are mainly differentiated by the large-scale fading and shadowing they experience. In addition, user association is performed at a large time-scale, whereas power allocation is performed at a smaller time-scale; therefore, by optimizing one, the other can be considered constant. Moreover, $g_j(x)$ is the average channel gain at location x measured from BS j, which captures the quasi-static effects of path-loss and log-normal shadowing. The channel gain is assumed to be measured at a large time scale [28]. The variable σ^2 models the power level of the noise. The set $\mathcal{I}_j(x)$ denotes the subset of BSs that introduce interference to user's reception at location x from BS j; whereas, $I_k(x)$ is the average interference power introduced by BS $k \in \mathcal{I}_j$. Due to the fractional frequency reuse plan of the network, the interference is modeled as static noise [17]. It is important to mention that our proposed scheme is applicable in CoMP-enabled HetNets adopting either fractional frequency reuse or interference randomization, in which interference can be considered as static noise [17]. A user's downlink rate $r_j(x)$ at x is given by

$$r_j(x) = W_j \log_2 \left(1 + \mathrm{SINR}_j(x)\right), \tag{2}$$

where W_j is the total bandwidth of the j^{th} BS. The traffic load at the j^{th} BS that is designated for a user located at x is expressed as

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

[2]The traffic modeling in this section follows the model used in similar works in the literature, such as in [17,28].

Journal Pre-proof

Journal Pre-proof

where $\lambda(x)$ denotes the number of traffic arrivals (i.e. number of packets) at x. The traffic arrivals follow a Poisson distribution, while the size of traffic arrivals, denoted by $\nu(x)$, is exponentially distributed. The variable $\eta_j(x)$ denotes a binary association variable, defined as

$$\eta_j(x) = \begin{cases} 1, & \text{if user at } x \text{ is associated with BS } j, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Hence, the traffic load at BS j can be expressed as

$$\rho_j = \int_{x \in \mathcal{A}} \varrho_j(x) dx. \tag{5}$$

The traffic load $0 \leq \rho_j < 1$, known in queuing theory as utilization, refers to the fraction of time the BS is busy serving users. Let us define the following parameter $\gamma = \frac{\nu(x)}{r_j(x)}$, where γ can be considered constant during a single user association period, such that [28]

$$\vartheta_j = \frac{\mathrm{E}[\gamma^2]}{2}, \tag{6}$$

where $\mathrm{E}[\cdot]$ is the expectation operator. Now, the average latency experienced by a traffic arrival can be expressed as [28]

$$L_j(\rho_j) = \frac{\vartheta_j \rho_j}{1 - \rho_j}. \tag{7}$$

Note that $L_j(\rho_j)$ is a function of the j^{th} BS traffic density, and since ϑ_j is a constant, minimizing the latency at the j^{th} BS is equivalent to minimizing the approximation of $L_j(\rho_j)$, given by the following latency indicator function

$$\hat{L}_j(\rho_j) = \frac{1}{1 - \rho_j}. \tag{8}$$

C. Energy Model

Both types of BSs are powered by renewable energy sources (e.g. solar panels and/or wind turbines), and are also connected to the grid. Although the priority is always given to utilization of green energy, whenever green energy is fully depleted, a BS complements its excess power requirements from the grid[3]. Let the power consumption of BS j be modeled by [28]

$$p_j = \beta_j \rho_j + p_j^s, \tag{9}$$

where β_j is a constant variable that translates the traffic load to its corresponding load-dependent power consumption, and p_j^s is the static load-independent power consumption. Furthermore, let p_j^o denote the on-grid power consumption of the j^{th} BS, which is defined as

$$p_j^o = [p_j - e_j]^+, \tag{10}$$

where e_j is the amount of green energy stored at the j^{th} BS, and $[\cdot]^+ = \max(0, \cdot)$. Then, ρ_j^g is defined as the amount of traffic load that can be served only by utilizing the available green energy, as [28]

$$\rho_j^g = \max\left[\epsilon, \min\left[\frac{e_j - p_j^s}{\beta_j}, 1 - \epsilon\right]\right]. \tag{11}$$

[3]Battery management and energy redistribution are beyond the scope of this work. For reference, a battery management system for HetNets powered by hybrid energy sources can be found in [38].

Journal Pre-proof

Here, ϵ is a positive small constant that ensure $0 < \rho_j^g < 1$. Therefore, minimization of on-grid power consumption, requires the traffic load assigned to a given BS be less than or equal to ρ_j^g. It should be noted that renewable energy does not incur any cost to our system, and therefore it should not be penalized in the objective function. To reflect this, we use the Courant-Beltrami function, given by [39]

$$\phi_j(\rho_j) = \left[[\rho_j - \rho_j^g]^+\right]^2, \tag{12}$$

which penalizes only on-grid power consumption. The function is differentiable, continuous and smooth, making it suitable for iterative optimization [39]. Minimizing the objective function in (12) achieves best green energy utilization, and therefore minimizes the on-grid energy consumption. Notice that $\hat{L}_j(\rho_j)$ in (8) and $\phi_j(\rho_j)$ in (12) are functions of the BSs load $\rho_j \ \forall j \in \mathcal{B}$, which is determined by the network parameters and association variables $\eta_j(x) \ \forall j \in \mathcal{B}, \forall x \in \mathcal{A}$ in (4). In the following subsection, we present our formulation of the users association problem, in which we optimize $\eta_j(x)$ to minimize the network traffic latency and on-grid power consumption.

D. User Association Problem

The minimization of the latency indicator function in (8) entails offloading users from a heavily loaded BS to a lightly loaded BS. However, it may be the case that a lightly loaded BS does not have sufficient green energy, ultimately raising the cost in (12) and consuming more on-grid power. In addition, the lightly loaded BS may be far from a user, thereby lowering its data rate and increasing traffic delivery latency. Thus, latency minimization and power conservation often conflict with each other. In turn, define a cost function f that combines both latency minimization and power conservation, such that

$$f = \sum_{\forall j} \hat{L}_j(\rho_j) + \sum_{\forall j} k_j \phi_j(\rho_j), \tag{13}$$

where k_j sets the trade-off between the on-grid power consumption and traffic delivery latency. Then, the user association problem can be posed as

$$\underset{\eta_j(x) \forall j, \forall x}{\text{minimize}} \quad f = \sum_{\forall j} \hat{L}_j(\rho_j) + \sum_{\forall j} k_j \phi_j(\rho_j) \tag{14}$$

$$\text{s.t.} \quad \epsilon \leq \rho_j \leq 1 - \epsilon \qquad \forall j \in \mathcal{B}, \tag{15}$$

$$\sum_{\forall j} \eta_j(x) \geq 1, \qquad \forall x \in \mathcal{A}, \tag{16}$$

$$\eta_j(x) \in \{0, 1\}, \qquad \forall j \in \mathcal{B}, \forall x \in \mathcal{A}, \tag{17}$$

$$\rho_j = \int_{x \in A} \frac{\lambda(x)\nu(x)\eta_j(x)}{r_j(x)} dx. \tag{18}$$

For a sufficiently small ϵ, Constraint (15) ensures BS queue stability, while Constraint (16) guarantees that all users are associated with at least one BS, and enables CoMP transmission by allowing users to be associated with more than one BS. It is worth noting that our problem formulation does not differentiate between cell-center users and cell-edge users. All users are considered CoMP candidates; however naturally, the users most benefiting from

Journal Pre-proof

such a CoMP system are those with comparable perceived signal strength from multiple BSs (i.e. cell-edge users). The binary association variable $\eta_j(x)$ is defined in Constraint (17). The equation (18) defines the traffic load of a BS in terms users traffic arrivals, traffic size, data rates and binary association variable.

The association problem is a mixed-integer nonlinear programming (MINLP) problem, which is known to be NP-hard [40]. Classical MINLP solving techniques, such as branch-and-bound, branch-and-cut or even heuristic solutions are too computationally expensive to the point where their solution would be redundant in the scale of a single user association period. Therefore, solving this problem efficiently necessitates relaxation of the binary association variable $\eta_j(x)$ into a new association variable $\hat{\eta}_j(x)$, where $0 \leq \hat{\eta}_j(x) \leq 1$. Furthermore, in order to guarantee that all users are served and associated with at least one BS, we replace Constraint (16) with

$$\sum_{\forall j} \hat{\eta}_j(x) = 1, \qquad \forall x \in \mathcal{A}.$$

Thus, a sub-space of the original problem feasible space is defined. In Section IV, this constraint is relaxed in the proposed algorithm for users that can be served by more than one BS, and thus allowing for CoMP transmissions. Hence, the user association (UA) problem is reformulated as

$$\underset{\hat{\eta}_j(x)\forall j,\forall x}{\text{minimize}} \quad f = \sum_{\forall j} \hat{L}_j(\rho_j) + \sum_{\forall j} k_j \phi_j(\rho_j) \tag{19}$$

$$\text{s.t.} \quad \epsilon \leq \rho_j \leq 1 - \epsilon, \quad \forall j \in \mathcal{B}, \tag{20}$$

$$0 \leq \hat{\eta}_j(x) \leq 1, \qquad \forall j \in \mathcal{B},\ x \in \mathcal{A}, \tag{21}$$

$$\sum_{\forall j} \hat{\eta}_j(x) = 1, \qquad \forall x \in \mathcal{A}, \tag{22}$$

$$\rho_j = \int_{x \in A} \frac{\lambda(x)\nu(x)\hat{\eta}_j(x)}{r_j(x)} dx. \tag{23}$$

Unlike the original problem, the relaxed problem is convex. Here $\hat{\eta}_j(x)$ denotes the probability that the j^{th} BS serve a user located at x. This sub-space is later mapped back to the original feasibility space using the fractional association solution; therefore, achieving CoMP associations.

IV. Proposed Scheme

The RANC gathers information on traffic profiles of users, channel measurements, and green energy availability at BSs. This information is employed by the proposed scheme in order to solve the relaxed problem and compute a fractional user association. A local search algorithm is developed to project the generated fractional user association solution onto the feasible space of the original problem; thereby, associating users to a single or multiple BSs. The CoMP transmissions are enabled for users associated with multiple BSs.

The proposed UA scheme uses augmented Lagrangian and barrier function methods in transforming the relaxed constrained UA problem into a series of relaxed unconstrained UA problems that can be solved iteratively using a second order method. Moreover, the constrained relaxed UA problem is transformed into an unconstrained relaxed UA problem, via the use of an interior barrier function, which increases exponentially when inequality constraints

Journal Pre-proof

are nearly violated [41]. A logarithmic barrier function B_1 is used to model the cost of closely satisfying Constraint (20) as

$$B_1(\rho_j) = \log\left((1-\epsilon) - \rho_j\right) + \log(\rho_j - \epsilon), \tag{24}$$

and a logarithmic barrier function B_2 to model the cost of closely satisfying Constraint (21), as

$$B_2(\hat{\eta}_j(x)) = \log\left(\hat{\eta}_j(x)\right) + \log\left(1 - \hat{\eta}_j(x)\right). \tag{25}$$

Negative logarithmic functions are used for inequalities, as they heavily penalize approaching the constraint from feasibility, reaching up to infinity at the constraint. Similarly, the cost of deviating from the equality Constraint (22) is represented by a quadratic penalty function given by [39]

$$Q(x) = \left[\sum_{\forall j} \hat{\eta}_j(x) - 1\right]^2. \tag{26}$$

A quadratic function is used to penalize any deviation (positive or negative) from the equality line, while having a penalty of zero when the equality constraint is met. Given these functions, the unconstrained UA problem can be written as

$$\underset{\hat{\eta}_j(x)\forall j\in\mathcal{B},\forall x\in A}{\text{minimize}} f_{uc} = f - \mu_1 \sum_{\forall j} B_1(\rho_j) - \mu_2 \sum_{\forall j,x} B_2\left(\hat{\eta}_j(x)\right) + \frac{\omega}{2} \sum_{\forall x} Q(x), \tag{27}$$

where μ_1, μ_2 and ω are scaling multipliers used to model the severity of violating or approaching the constraints. Although μ_1, μ_2 and ω can be set to arbitrarily large numbers, this would introduce poor convergence properties [41]. Therefore, the problem is solved iteratively, while decreasing the values of μ_1, μ_2 and increasing the values of ω in each iteration. The values of μ_1 and μ_2 are initially set high and decrease in order to iteratively decrease the numerical significance of the barrier functions on the objective function, and speed up the convergence. On the other hand, the value of ω is initially set to a small value and increases in order to prevent the violation of the equality constraint. However, the cost of meeting the equality constraint approaches zero, resulting in the original constrained objective function when the constraints are met.

Unlike first order methods, second order methods—mainly based on deriving the Hessian—allow optimization algorithms to enjoy faster convergence. On the other hand, the computational time and memory required to evaluate the Hessian of multivariate functions may hinder the applicability of second order methods in dense HetNets. The proposed scheme applies a limited–memory–Broyden–Fletcher–Goldfarb–Shanno BFGS (L-BFGS) algorithm to approximate the Hessian, requiring shorter time to compute and less memory to store. The algorithm uses the L-BFGS two loop method from [42] to estimate the inverse of the Hessian $\mathbf{H}$.

For notational brevity, we define $\hat{\boldsymbol{\eta}}$ to be a matrix of $\hat{\eta}_j(x)$, $\forall j \in \mathcal{B}$, $\forall x \in \mathcal{A}$. Then, the proposed algorithm performs the following iterative update, $\hat{\boldsymbol{\eta}}^+ = \hat{\boldsymbol{\eta}} + \alpha\mathbf{d}$, where $\mathbf{d}$ denotes the search direction, and α denotes the search step size. The search direction is computed based on Newton's direction, which is defined by the negative Hessian inverse of the objective function multiplied by its gradient [42]. This search direction is in a descent direction of the objective function in (27). However, in order to avoid the overhead associated with using the fully

Journal Pre-proof

formed Hessian, we use a quasi-newton method that implicitly approximates it. The objective function in (27) has a gradient with respect to a the relaxed association variable, given by

$$\frac{\partial f_{uc}}{\partial \hat{\eta}_j(x)} = \frac{c_j(x)}{(1-\rho_j)^2} + 2c_j(x)k_j[\rho_j - \rho_j^g]^+ - \mu_1\left[\frac{c_j(x)}{\rho_j - \epsilon} - \frac{c_j(x)}{((1-\epsilon)-\rho_j)}\right] - \mu_2\left[\frac{1}{\hat{\eta}_j(x)} - \frac{1}{1-\hat{\eta}_j(x)}\right] + \omega\left[\sum_{\forall j}\hat{\eta}_j(x) - 1\right], \tag{28}$$

where $c_j(x) = \frac{\lambda(x)\nu(x)}{r_j(x)}$. The gradient vector $\nabla \mathbf{f_{uc}}$ is computed based on stacking the derivatives with respect to all association variables $\hat{\eta}_j(x)$ for all j and x into a single matrix.

The proposed scheme iteratively solves an approximated version of the original constrained problem as outlined in **Algorithm 1**. The scheme approximates the Hessian inverse of the objective function in (27) by storing $2 \times m$ vectors. The size of each vector is n, where n is the number of optimization variables and $m << n$. Practically speaking, m is in the order of tens, whereas n is in the problem is in the order of thousands. This actively alleviates the need for computing and storing n^2 operations every time the Hessian is computed [42].

Journal Pre-proof

Journal Pre-proof

Algorithm 1 L-BFGS User Association

Input: Initial values of $\hat{\eta}_j(x)$ and trade-off coefficients k_j.

Output: User association $\hat{\eta}_j^*(x) \forall j \in \mathcal{B},\ \forall x \in \mathcal{A}$

Set initial μ_1, μ_2, ω and t.
while termination condition of barrier method not satisfied **do**
  while termination condition of L-BFGS method not satisfied **do**
    $\mathbf{q} := \nabla \mathbf{f_{uc}}$ ▷ Find a search direction $\mathbf{d}$
    for $i = 0$ to $m - 1$ **do** ▷ L-BFGS two-loop recursion
      $a_i := \frac{1}{\mathbf{y_i' s_i}} \mathbf{s_i}' \mathbf{q}$
      $\mathbf{q} := \mathbf{q} - a_i \mathbf{y}_i$
    end for
    $\mathbf{r} := \frac{\mathbf{y_0}' \mathbf{s_0}}{\mathbf{y_0}' \mathbf{y_0}} \mathbf{q}$
    for $i = m - 1$ to 0 **do**
      $b = \frac{1}{\mathbf{y}_i' \mathbf{s_i}} \mathbf{y}_i' \mathbf{r}$
      $\mathbf{r} := \mathbf{r} + \mathbf{s_i}(a_i - b)$
    end for
    Compute search direction $\mathbf{d} := -\mathbf{H} \nabla \mathbf{f_{uc}} = -\mathbf{r}$
    $\alpha :=$ LINE SEARCH($\hat{\boldsymbol{\eta}}$,$\mathbf{d}$) ▷ Find a step length α that satisfies the strong Wolfe conditions
    $\hat{\boldsymbol{\eta}}^+ = \hat{\boldsymbol{\eta}} + \alpha \mathbf{d}$ ▷ Update association variables $\hat{\eta}$
    Discard $\mathbf{s_{m-1}}$ and $\mathbf{y_{m-1}}$ ▷ Update vector lists with new local information
    $\mathbf{s_{i+1}} := \mathbf{s_i}, \forall i \in [1, m-2]$
    $\mathbf{y_{i+1}} := \mathbf{y_i}, \forall i \in [1, m-2]$
    $\mathbf{s_0} := \hat{\boldsymbol{\eta}}^+ - \hat{\boldsymbol{\eta}}$
    Generate the updated gradient vector $\nabla \mathbf{f}_{\mathbf{uc}}^+$
    $\mathbf{y_0} := \nabla \mathbf{f}_{\mathbf{uc}}^+ - \nabla \mathbf{f_{uc}}$
  end while
  $\omega := t\omega$, $\mu_1 := \mu_1 / t$, $\mu_2 := \mu_2 / t$ ▷ Update multipliers of penalty functions
end while
for all fractional association variables $\hat{\eta}_j(x), \forall x, \forall j$ **do** ▷ Associate users with BSs having fractional association higher than threshold ϵ_η
  if $\hat{\eta}_j(x) \leq \epsilon_\eta$ **then**
    $\eta_j(x) := 0$
  else
    $\eta_j(x) := 1$
  end if
end for

Journal Pre-proof

Journal Pre-proof

The solution to the approximated unconstrained problem is obtained using the L-BFGS method. Firstly, in lines 4 to 14, the two-loop recursion method uses two types of vectors in order to implicitly approximate $-\mathbf{H}\nabla\mathbf{f_{uc}}$. These vectors are denoted by $\mathbf{s}_i$ and $\mathbf{y}_i$, where $i \in [0, m)$. The vectors $\mathbf{s}_i$ measure local changes in the optimization variables, while the vectors $\mathbf{y}_i$ measure local changes in the gradients of the objective function in (27). Because of the implicit use of the approximation, the need to store or arithmetically use the n^2 sized matrix $\mathbf{H}$ is eliminated. Instead, the two-loop recursion method requires $4 \times m \times n$ iterations in order to calculate the search direction $\mathbf{d}$ [42].

In Line 15, a line search is performed to determine the step length to move the association variables $\hat{\boldsymbol{\eta}}$ along the previously calculated search direction $\mathbf{d}$. Firstly, we note that the association values $\hat{\eta}$ and the search direction $\mathbf{d}$ do not change during the line search phase of the algorithm. Therefore, for convenience, we define

$$\psi(\alpha) = f_{uc}(\hat{\boldsymbol{\eta}} + \alpha\mathbf{d}). \tag{29}$$

Furthermore, the directional derivative of f_{uc} in $\mathbf{d}$ (i.e. the derivative of (29)) is given by

$$\nabla\psi(\alpha) = \nabla\mathbf{f_{uc}}(\hat{\boldsymbol{\eta}} + \alpha\mathbf{d})\mathrm{d}', \tag{30}$$

where ∇ is the gradient operator. In order to ensure the change to the association variables achieves a significant improvement in minimization of the objective function, the step length is required—by the line search—to meet a pair of inequalities, known as the strong Wolfe conditions [19,42]. The first condition is

$$\psi(\alpha) \leq \psi(0) + \beta_1\alpha\nabla\psi(0), \tag{31}$$

which is referred to as "the sufficient decrease condition". The second condition is

$$|\nabla\psi(\alpha)| \leq \beta_2|\nabla\psi(0)|, \tag{32}$$

which is known as "the curvature condition" [42]. Both conditions guarantee that the step length taken achieves sufficient progress in minimizing the objective function in (27). This is necessary because it avoids making unnecessary steps of small length with negligible impact on the minimization of the objective function. Once a step length that satisfies the strong Wolfe conditions is found, the association variables $\hat{\boldsymbol{\eta}}$ are updated, as per Line 16.

The vectors $\mathbf{s}_i$ and $\mathbf{y}_i$ are stored in a queue structure of fixed length. The updates of $\mathbf{s_0}$ and $\mathbf{y_0}$ are made in each iteration using local information, and the queue is pushed forward with the least recent two vectors being discarded, as can be seen in Lines 17 to 22.

After the L-BFGS method converges to a solution of the approximate unconstrained problem, the multipliers ω, μ_1 and μ_2 are adjusted to better model the original constrained problem, as in Lines 24 to 26. Adjusting the multipliers produces a new unconstrained problem and the process is repeated until a solution for the original constrained problem is found.

The fractional association values $\hat{\eta}_j(x)$ $\forall x \in \mathcal{A}$ $\forall j, \in \mathcal{B}$ obtained from the barrier method in Lines 2 to 25 represent the probability of a user located at x receiving traffic from the j^{th} BS. In order to obtain a CoMP-enabled

Journal Pre-proof

association, the algorithm associates users with BSs having a fractional association value larger than a threshold ϵ_η, as seen in Lines 27 to 32.

The line search algorithm in **Algorithm 2** iteratively finds a step length α_i that meets the strong Wolfe conditions (Line 12) or a range of values, i.e., bracket, that contains such a step length (Lines 6, 9 and 15). If neither is found during an iteration of the line search algorithm, a new trial step length α_{i+1} is interpolated, as in Lines 18 and 22. Here, interpolation refers to modeling the objective function as a cubic function that is a function of the step length α and solving for that model's minimum. The function references "SECTION$(\cdot,\cdot)$" in Lines 6, 9, and 15 call a sectioning method implemented in **Algorithm 3** to find a viable α within the provided bracket. The pseudo-code of **Algorithm 3** is presented in Appendix A. The interpolation steps mentioned in **Algorithms 2** and **3** (given in Lines 22 and 5, respectively) are used to minimize $\psi(\alpha)$ within the determined bracket, where a cubic interpolation method is used [19]. Therefore, the step length α found by **Algorithms 2** and **3** fulfills the strong Wolfe conditions in (31) and (32).

Journal Pre-proof

Journal Pre-proof

Algorithm 2 Line Search Algorithm

function LINE SEARCH($\hat{\boldsymbol{\eta}}$,$\mathbf{d}$)
  Initialize $\alpha_0 := 0$, $\alpha_{\max} = \frac{\psi_{\min} - \psi(0)}{\beta_1 \nabla \psi(0)}$, $\alpha_1 \in (0, \alpha_{\max})$
  $i := 1$
  loop
    if $[\psi(\alpha_i) > \psi(0) + \beta_1 \alpha_i \nabla \psi(0)]$ **then**
      return SECTION(α_{i-1}, α_i) ▷ Bracket containing viable α found
    end if
    if $[\psi(\alpha_i) > \psi(\alpha_{i-1})]$ and $i > 1$ **then**
      return SECTION(α_{i-1}, α_i) ▷ Bracket containing viable α found
    end if
    if $|\nabla \psi(\alpha_i)| \leq -\beta_2 \nabla \psi(0)$ or $\psi(\alpha_i) < \psi_{\min}$ **then**
      return α_i ▷ Viable step length α found
    end if
    if $\nabla \psi(\alpha_i) \geq 0$ **then**
      return SECTION(α_i, α_{i-1}) ▷ Bracket containing viable α found
    end if
    if $2\alpha_i - \alpha_{i-1} < \alpha_{\max}$ **then**
      Interpolate for $\alpha_{i+1} \in [\alpha_{\text{low}}, \alpha_{\text{high}}]$ ▷ Determine next trial step length α
    end if
    $\alpha_{lb} := 2\alpha_i - \alpha_{i-1}$
    $\alpha_{ub} := \min(\alpha_{\max}, \alpha_i + \tau_1(\alpha_i - \alpha_{i-1})$
    Interpolate for $\alpha_{i+1} \in [\alpha_i, \alpha_{\max}]$ ▷ Determine next trial step length α
    $i := i + 1$
  end loop
end function

The line search guaranteeing the strong Wolfe conditions does not only improve the algorithm's performance, but also provides useful theoretical properties on convergence. The L-BFGS method is proven to converge at a linear rate, as well as a local super-linear convergence rate near the solution. This convergence behavior is similar to the one observed in Newton's method and is guaranteed, given that the strong Wolfe conditions are maintained [19]. The complexity of the proposed L-BFGS User Association in **Algorithm 1** is analyzed in Appendix B. Furthermore, a proof of convergence of **Algorithm 1** is given in Appendix C.

In summary, given the network parameters reported to the RANC and the trade-off coefficients, the proposed scheme computes the cost of an initial association, which can be set as the last used association or according to maximum received signal strength. The proposed scheme then minimizes the objective function in (19) by iteratively

Journal Pre-proof

Journal Pre-proof

solving unconstrained versions of the problem in (27). The resulting solution is a fractional user association that minimizes the objective function in (19). The association values are either 0 or 1 for most cell-center users, and hence are associated with a single BS. However, some users, particularly cell-edge users, experience a split association value. The algorithm simultaneously associates cell-edge users with all BSs that have non-trivial fractional association values, i.e., above a predefined threshold ϵ_η.

V. Simulation Results

A. *Uniformly Distributed User Locations*

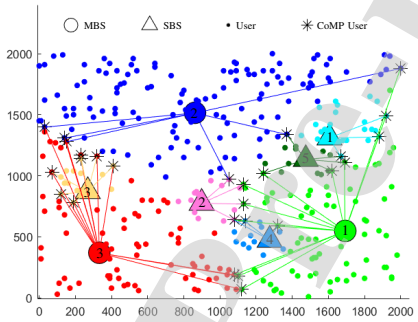


Fig. 2. Base-station deployment scenario with uniformly distributed users.

In order to evaluate the proposed scheme, we consider a deployment scenario in a 4 km^2 square geographical region. The area is covered by 3 MBSs and 5 SBSs. These BSs serve 350 users, with uniform and independently distributed locations. Figure 2 shows the positioning of BSs and users, where circles represent MBSs, triangles represent SBSs and dots represent users. Assigned CoMP users are indicated by a star, and are connected with a line to each BS they are associated with. Users associated with only a single BS are colored with the respective BS's color. On average, there are 5 traffic arrivals per user with each request averaging 250 Kbs in size. In order to estimate user data rates, the radio propagation modeling follows baseline test conditions from [43], where the path-loss for an MBS is given by

$$\mathrm{PL}_{\mathrm{MBS}} = 130.19 + 37.6 \log(r_p), \tag{33}$$

while for a SBS, it is

$$\mathrm{PL}_{\mathrm{SBS}} = 37 + 30 \log(r_p). \tag{34}$$

The variable r_p denotes the distance in kilometers between the serving BS and the user location. Furthermore, the radio channel is affected by log-normal shadowing with standard deviation of 8 dB. The noise power level is -174 dBm/Hz [43]. The antenna used has a receiver sensitivity of -123 dBm and an antenna gain of 15 dB [43]. The system bandwidth is 20 MHz. Each BS is equipped with a solar panel of sizes 4 and 1 m^2, for MBSs

Journal Pre-proof

TABLE I
PROPOSED SCHEME IMPLEMENTATION PARAMETERS USED

Parameter	k_j	β_1	β_2	τ_1	τ_2	τ_3	ϵ_η
Value	700	0.01	0.9	3	0.1	0.5	0.01

and SBSs, respectively. We assume standard test conditions such that the temperature is 25°C, an air mass of 1.5 spectra and a solar irradiation incidence angle of 45 degrees. Under these conditions, the highest achievable direct irradiance is 1 $\mathrm{kW/m^2}$ [44]. The available solar irradiance is bound by this maximum and is assumed to follow a Beta distribution [45,46]. MBSs and SBSs transmit at powers of 43 and 33 dBm, and statically consume 750 and 40 W of power, respectively. The load to power relation coefficient β is 500 for a MBS and 4 for a SBS [47]. The specific parameters used for simulations are summarized in Table I. The proposed scheme is compared with existing association schemes. The scheme named MAXSINR associates each user with the BS from which the highest maximum SINR is measured. Two other load balancing user association schemes are considered as well. The α-optimal scheme aims to minimize latency by minimizing a latency indicator function, with $\alpha = 2$ [17]. Also, the vGALA scheme—presented in [28]—aims to associate users in order to minimize both latency and on-grid power consumption. The vGALA scheme is configured with the default parameters used in [28]. All results are obtained by averaging over 500 independent network instances.

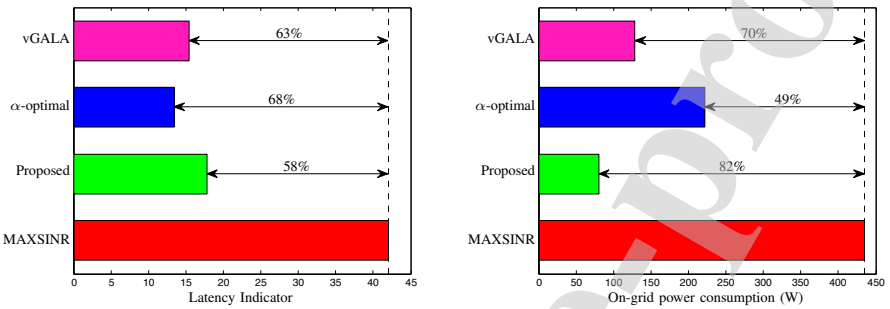
TABLE II
DISTRIBUTION OF USER-TO-BS ASSOCIATION OBSERVED IN THE COMPARED SCHEMES WITH VARIOUS STORAGE OF GREEN ENERGY (G.E.).

BS	G.E. (W)	MAXSINR	α-optimal	vGALA	Proposed
MBS 1	150	41%	35.4%	32.5%	33%
MBS 2	200	23%	22.5%	21%	23%
MBS 3	335	21.7%	21.1%	21%	25%
SBS 1	5	1.4%	2.5%	2.5%	2.5%
SBS 2	5	1.4%	3.7%	3.7%	4%
SBS 3	4	1.7%	4%	4%	4%
SBS 4	3	4.5%	8%	8%	8%
SBS 5	4	4.8%	7.4%	7.4%	8%

Table II illustrates the distribution of users among the available MBSs and SBSs as percentages of the total number of users in the system. It is also clear from Table II that users in the MAXSINR scheme are more heavily biased towards the MBSs. The α-optimal scheme considers latency as its major and sole performance measure, and thus offloads users towards SBSs to minimize latency. Furthermore, because in the randomly generated scenario, MBS 1 and MBS 2 have less available green energy than MBS 3; it can be seen that the proposed scheme offloads more users towards MBS 3 to reduce the amount of power supplied from the grid. Our proposed scheme distributes as much load as possible over the SBS and decreasing the most load off MBS 2, which was heavily loaded. Our

Journal Pre-proof

proposed scheme's percentages add up to over 100% due to the CoMP associations given to users located at the cell-edge to improve their assigned rates. This is a part of our scheme's trade-off, where some extra load is carefully imposed on some BSs to improve cell-edge user rates.



(a) Average Latency Indicator

(b) On-grid power consumption

Fig. 3. Performance comparison of the considered schemes in minimizing: (a) latency and (b) power consumption.

Figure 3 illustrates the performance of all considered schemes in comparison to the conventional MAXSINR scheme. Figure 3(a) shows that the MAXSINR scheme demonstrates the highest latency, because it heavily associates users with MBSs, as can be seen from Table II. The other schemes have comparable performance in terms of latency optimization. The proposed scheme reduces the latency indicator in comparison to MAXSINR by 58%. The vGALA scheme reduces the latency by 63% compared to MAXSINR, while the α-optimal scheme reduces the average latency by 68%, which is the greatest among all compared schemes. This is expected, as the sole objective of the α-optimal scheme is to optimize latency.

In comparison, the proposed scheme has the least on-grid power consumption, consuming 82% less on-grid power than MAXSINR, as demonstrated in Fig. 3(b). The vGALA scheme consumes 70% less power than the MAXSINR. The α-optimal scheme consumes 49% less power than the MAXSINR scheme. Although, the α-optimal scheme does not actively optimize power consumption, the load balancing aspect of the system leads to the observed power saving results. It can be observed that the proposed scheme yields the best trade-off, as it only has 10% and 5% increase in latency relative to α-optimal and vGALA, respectively, while consuming 33% and 12% less power compared to these schemes. Therefore, the performance of the proposed scheme and vGALA are generally comparable in this scenario, where users are uniformly distributed.

B. Increased User Density at Cell Edges

We simulate a scenario with higher user density at cell edge than the scenario considered in the previous sub-section. Figure 4 depicts a simulation scenario similar to the previous scenario, in which 250 users are uniformly distributed; however, 100 users out of the 250 are placed along the edges of coverage areas. Along the edges of

Journal Pre-proof

cells, signals received from multiple BSs are comparable in strength. Users that have been associated with a single BS by our scheme are represented by a black dot, users associated with two BSs are represented by a red star, and those associated with three BSs are represented by a slightly larger blue star. Except for user distribution, all other simulation parameters remain the same as the previous scenario. This leads to further degradation of rates received by cell-edge users.

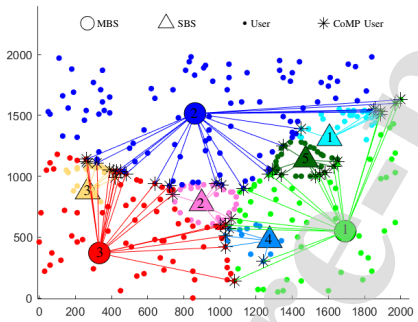


Fig. 4. A scenario of BSs deployment with larger density of users at the edges of cells.

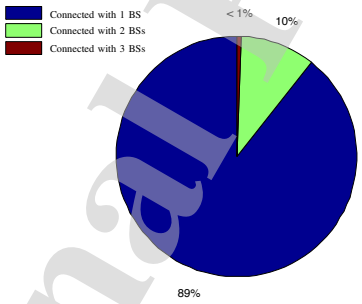


Fig. 5. Percentages of users associated to a single BS or employing CoMP transmission in the proposed scheme.

Figure 5 displays the distribution of users associated with a single BS, two or three BSs in the scenario shown in Fig. 4. It can be seen that the proposed scheme allows 10% of users to be associated with two BSs in order to improve their data rates. A very small percentage of users were associated with three BSs, as it is unlikely for a user to have comparable signal strengths from 3 or more BSs.

Table III shows the distribution of users in the proposed scheme under the increased cell-edge density scenario. It can be noted that cell-center users are more heavily associated with MBSs than SBSs, as the user rates provided to them by MBSs are higher than those provided by SBSs. Therefore, offloading them to SBSs incurs a much higher penalty to the total network latency indicator, in comparison to cell-edge users. On the other hand, cell-edge users

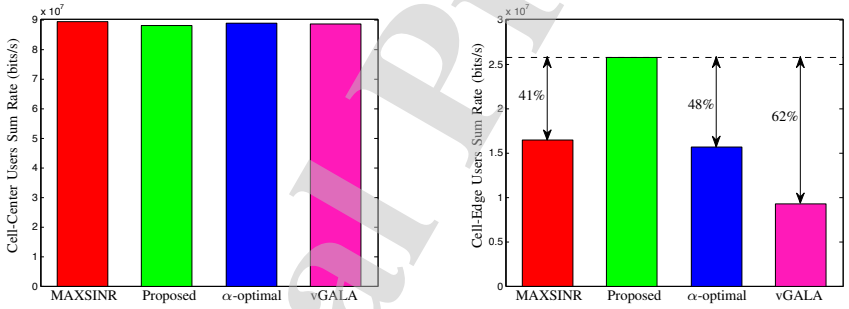
Journal Pre-proof

are more likely to be distributed among SBSs, as they have similar data rates from several BSs. It is worth noting that the proposed scheme does not overload the SBSs with CoMP users, even if they are more likely to associate with cell-edge users. In this way, the CoMP association does not severely impact the other performance measures.

TABLE III

DISTRIBUTION OF USERS (US.) BY BS WITH VARIOUS STORAGE OF GREEN ENERGY (G.E.) UNDER INCREASED NUMBER OF CELL-EDGE USERS.

BS	G.E. (W)	Total Us.	Cell-edge Us.	Cell-center Us.	$\frac{\text{CoMP Us.}}{\text{Cell-edge Us.}}$ (%)
MBS 1	150	67	34%	64%	74%
MBS 2	200	110	24%	76%	77%
MBS 3	335	81	43%	57%	51%
SBS 1	4	32	84%	16%	22%
SBS 2	5	32	81%	19%	19%
SBS 3	4	24	75%	25%	33%
SBS 4	4	14	50%	50%	57%
SBS 5	5	33	79%	21%	31%



(a) Cell-Center Users

(b) Cell-Edge Users

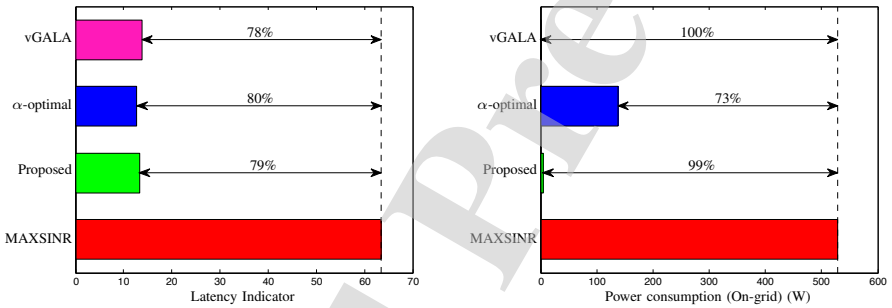
Fig. 6. Sum-rate performance of: (a) cell-center and (b) cell-edge users.

Figure 6 shows the sum-rate for users located at the center of the cell and the edge of the cell, under the compared schemes. Figure 6(a) illustrates that the sum-rate of cell-center users remains comparable across the different schemes, with negligible differences. However, it can be seen in Fig. 6(b) that the proposed scheme achieves a significant sum-rate enhancement for cell-edge users in comparison to other schemes. Specifically, the proposed scheme enhances the sum-rate of cell-edge users by 41%, 48% and 62% over MAXSINR, α-optimal and vGALA, respectively. Such enhancements are due to the CoMP-enabled transmission by associating cell-edge users to more than one BS. However, there is a minimal trade-off in terms of slightly decreased overall latency and power consumption performance in return for the enhancement cell-edge users' rates.

Journal Pre-proof

Journal Pre-proof

Figure 7 shows a performance comparison of the considered schemes in terms of power consumption and latency. From Fig. 7(a), the MAXSINR scheme demonstrates the worst latency response among all considered schemes. The other compared schemes demonstrate comparable latency responses. The α-optimal scheme showed the most significant reduction in latency equivalent to 80% in comparison to MAXSINR, followed by the proposed scheme which demonstrates a decrease in latency equivalent to 79% and lastly vGALA shows a drop that amounts to 78%. It is clear from Fig. 7(b) that the MAXSINR scheme also entails the largest power consumption. The α-optimal scheme consumes 73% less power than the MAXSINR scheme. The proposed scheme and the vGALA scheme consume approximately no power, saving 99% and 100%, respectively. A 100% savings of energy consumption implies that the proposed scheme did not consume any on-grid energy and satisfied its energy needs only from green energy. It can clearly be observed that the scheme has comparable performance to the other schemes when considering both power consumption and latency, while improving the cell-edge user rates as seen in Fig. 6(b).



(a) Average Latency Indicator

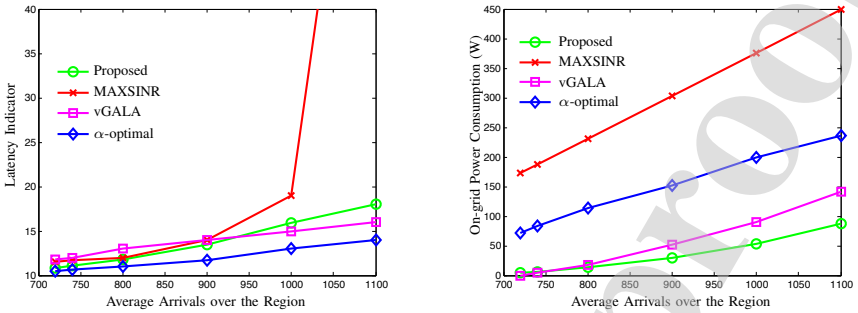
(b) On-grid power consumption

Fig. 7. Comparison between different schemes with increased density of cell-edge users in terms of: (a) average latency Indicator, and (b) on-grid power consumption.

Figure 8 displays the impact of increasing arrival rate on latency and power consumption. It can be clearly seen from Fig. 8(a) that MAXSINR's traffic latency increases exponentially as the arrival rate increases. The other schemes enjoy an almost linear increase in latency in response to the increase in traffic arrival. Fig. 8(b) shows that the proposed scheme consumes the least on-grid power as the arrival rate increases. It is note-worthy that these effects are achieved while maintaining the sum-rate improvements displayed in Fig. 6.

Figure 9 shows the trade-off between latency and power consumption for all compared schemes as total arrival rate over the region $\sum_{\forall x} \lambda(x)$ changes from 700 to 1100. The higher the slope of a curve in Fig. 9, the more latency-aware the scheme is, while the lower the slope, the more power-efficient the scheme is. It can be seen that the MAXSINR scheme scales poorly with increased arrivals, rising rapidly to high latency and high power consumption. It is worth-mentioning that the α-optimal and the vGALA schemes at their given parameters are more latency inclined, while the proposed scheme enjoys a fair balance between latency and power consumption.

Journal Pre-proof



(a) Average latency indicator (b) On-grid power consumption

Fig. 8. Effect of increased total arrival rates on the performance of considered scheme in the uniform user distribution scenario.

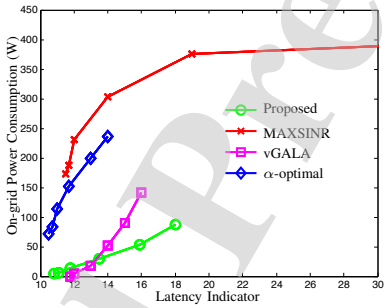


Fig. 9. Trade-off between on-grid power consumption and latency for the proposed schemes, by varying total arrival rate from 700 to 1100.

In Fig. 10, we explore the effect of changing the trade-off coefficient k_j on both on-grid power consumption and average latency. For simplicity, we assume that the trade-off for all BSs is the same. Practically, the trade-off coefficients can be managed by network operators independently. It can be noted from Fig. 10 that when $k_j = 0$, the problem becomes focused solely on latency, and thus achieving the best latency results and the highest power consumption. As k_j increases, the scheme becomes more sensitive to power consumption; therefore, the on-grid power consumption decreases; whereas, the traffic delivery latency increases, as can be seen observed Fig. 10. Eventually, as $k \to \infty$, the system reduces power consumption as feasibly possible without violating the problem's constraints.

Figure 11 illustrates the convergence behavior of the barrier method in the proposed scheme. The larger the value of t and the higher the initial values of the multipliers, the larger the progress made by each unconstrained sub-problem in the barrier method. However, as the number of sub-problems decreases, their difficulty increases, leading the number of L-BFGS iterations required to solve them to increase. For the simulated scenario, it was

Journal Pre-proof

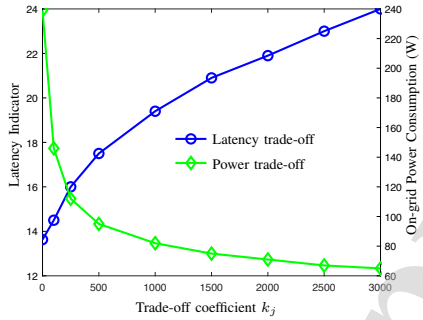


Fig. 10. Effect of increasing the trade-off coefficient k_j on the system performance in the uniform user distribution scenario.

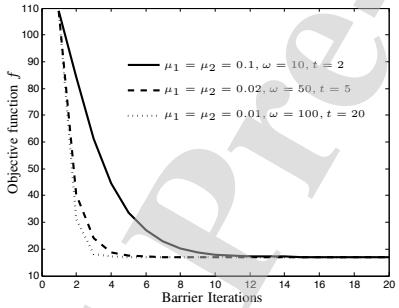


Fig. 11. Convergence behavior under different initial values of barrier method parameters.

noted that the values of $\mu_1 = \mu_2 = 0.02, \omega = 50, t = 5$ achieves the best performance among all simulation runs, as it balances the number of barrier sub-problems and their difficulty.

VI. Conclusions

In this paper, we investigated the user association problem in CoMP-enabled and SoftRAN-based HetNets with hybrid energy supplies. The user association problem is modeled as MINLP, which is known to be NP hard. Therefore, a sub-optimal scheme is developed to optimize the trade-off between the on-grid power consumption and traffic delivery latency, while ensuring BS queue stability and users connectivity. The scheme has been designed to be scalable to dense networks by using the L-BFGS method, a limited memory version of the well-studied BFGS algorithm that does not require explicit storage, and formation and computation of the Hessian. The proposed scheme jointly reduces the average latency and on-grid power consumption using a per-BS trade-off coefficient to emphasize importance of the two objectives conflict. Furthermore, the proposed scheme enabled CoMP transmissions by using fractional user association to select users eligible to be jointly serviced by multiple BSs to improve cell-edge users'

Journal Pre-proof

data rates. Simulations showed that the proposed scheme achieves significant improvements in reducing latency, power consumption and improving cell-edge user rates over existing schemes.

The centralized control enabled by the SoftRAN architecture in the considered network facilitates application of machine learning (ML) techniques. In our future work, we plan to explore applicability of machine learning techniques towards load balancing. Particularly, we plan to compare and contrast the performance of several reinforcement learning techniques in handling the trade-off between latency and energy efficiency in HetNets powered by hybrid energy sources.

APPENDIX A

The Sectioning phase in **Algorithm 3** iteratively reduces the width of the bracket that was found in Lines 6, 9 and 15 until a viable step length is found [19]. The value α_{lb} is updated, such that it is guaranteed to meet the first Wolfe condition. The bracketed interval is iteratively reduced (Lines 7, 13 and 15) in order to converge on a step length α that meets the second Wolfe condition (Line 10). Here, τ_1, τ_2, τ_3, β_1 and β_2 are constant parameters. The algorithm is insensitive to small changes in these parameters; thus, they do not have to be re-tuned for different network scenarios. **Algorithm 3** sets a value for α within the identified bracket, and then iteratively reduces that bracket width until a viable step length is identified.

Algorithm 3 Sectioning Phase

1: **function** SECTION(α_{low}, α_{high})
2: **loop**
3: $\alpha_{lb} := \alpha_{\text{low}} + \tau_2(\alpha_{\text{high}} - \alpha_{\text{low}})$
4: $\alpha_{ub} := \alpha_{\text{high}} - \tau_3(\alpha_{\text{high}} - \alpha_{\text{low}})$
5: Interpolate for $\alpha \in [\alpha_{lb}, \alpha_{ub}]$ ▷ Generate trial α
6: **if** $\psi(\alpha) > \psi(0) + \beta_1\alpha\nabla\psi(0)$ or $\psi(\alpha) \geq \psi(\alpha_{\text{low}})$ **then**
7: $\alpha_{\text{high}} := \alpha$ ▷ Shrink bracket
8: **else**
9: **if** $|\nabla\psi(\alpha)| \leq -\beta_2\nabla\psi(0)$ **then**
10: **return** α ▷ Found step length α satisfying strong Wolfe conditions
11: **end if**
12: **if** $\nabla\psi(\alpha)(\alpha_{\text{high}} - \alpha_{\text{low}}) \geq 0$ **then**
13: $\alpha_{\text{high}} := \alpha_{\text{low}}$ ▷ Shrink bracket
14: **end if**
15: $\alpha_{\text{low}} := \alpha$ ▷ Shrink bracket
16: **end if**
17: **end loop**
18: **end function**

Journal Pre-proof

Journal Pre-proof

APPENDIX B

The total number of iterations performed by **Algorithm 1** in the outermost loop is κ, where κ is a constant, determined by the update factor t. Let the total number of all decision variables $\hat{\eta}_j(x)$ in the system be n. Also, the maximum number of saved vector pairs used to approximate the Hessian is m, where m is usually between 10 and 15 [42]. The value of m is comparatively small to the problem size n, and is not scenario specific. The two-loop recursion method used to approximate the Hessian has a worst-case complexity of $\mathcal{O}(mn)$ [42]. Experimentally, the line search algorithm does not contribute significantly to the computational load in comparison to the two-loop recursion. Therefore, for brevity, we assume that line search has worst-case complexity $\mathcal{O}(ln)$. Therefore, the worst-case complexity of the algorithm is $\mathcal{O}(\kappa mn + \kappa ln)$. It can be shown that the convergence rate of the algorithm is linear and locally superlinear near the solution of each sub-problem, when meeting the strong Wolfe conditions. This is similar to the convergence behavior achieved when using Newton's direction [42].

APPENDIX C

The proof in this section follows the approach and employs properties from [42]. Throughout this section, we refer to the exact Hessian of the unconstrainted objective function $\nabla^2 f_{uc}$ as $\mathbf{G}$. The L-BFGS approach used closely approximates the BFGS update. In this section, we prove the global convergence of our unconstrained optimization subproblem using a BFGS update. The unconstrainted optimization subproblem depends on the following iterative step

$$\hat{\boldsymbol{\eta}}^{+} = \hat{\eta} + \alpha \mathbf{d}, \tag{35}$$

where the step size α is determined by a line search that satisfies the Wolfe conditions in (31) and (32), and the search direction is determined by

$$\mathbf{d} = -\mathrm{B}^{-1} \boldsymbol{\nabla} \mathbf{f_{uc}}, \tag{36}$$

where $\mathbf{B}$ is the BFGS approximation to the Hessian updated iteratively such that

$$\mathbf{B}^{+} = \mathbf{B} - \frac{\mathbf{B}\mathbf{s}\mathbf{s}'\mathbf{B}}{\mathbf{s}'\mathbf{B}\mathbf{s}} + \frac{\mathbf{y}\mathbf{y}'}{\mathbf{y}'\mathbf{s}}. \tag{37}$$

Proving that the algorithm converges to the optimal value $\hat{\eta}^*$ is equivalent to proving that

$$\liminf_{k \to \infty} \|\boldsymbol{\nabla} \mathbf{f_{uc}}\| = 0. \tag{38}$$

The following proof relies on properties of the objective function, which are as follows:

(i) The objective function f_{uc} is twice differentiable.

(ii) The objective function f_{uc} is continuous on the open set $\mathcal{N}$ containing the level set $\mathcal{L} := \{\hat{\eta} : f_{uc}(\hat{\eta}) \leq f_{uc}(\hat{\eta}_{\mathbf{0}})\}$.

(iii) The gradient $\nabla \mathbf{f_{uc}}$ is Lipschitz continuous on $\mathcal{N}$.

(iv) The level set $\mathcal{L}$ is convex and there exists positive constants such that

$$m \|z\|^2 \leq z'Gz \leq M \|z\|^2 \tag{39}$$

Journal Pre-proof

Journal Pre-proof

for all $z \in \mathbb{R}^n$. This assumption is preserved by the L-BFGS two-loop method from [42]. The set convexity, along with the lower bound on the objective function implies the existence of a unique minimizer $\hat{\eta}^*$ to the objective function.

We define the average Hessian $\bar{G}$ as

$$\bar{G} = \int_0^1 \nabla^2 f_{uc}(\hat{\eta} + \tau\alpha\mathbf{d})d\tau \tag{40}$$

and the property following from the Taylor series representation of the BFGS update $y = \bar{G}\alpha\mathbf{d} = \bar{G}s$. Using this property and (40), we arrive at

$$\frac{y's}{s's} = \frac{s'\bar{G}s}{s's} \geq m. \tag{41}$$

Let $z = \bar{G}^{1/2}s$, then

$$\frac{y'y}{y's} = \frac{s'\bar{G}^2s}{s'\bar{G}s} = \frac{z'\bar{G}z}{z'z} \leq M. \tag{42}$$

The subsequent proof relies on the following theorem, called Zoutendijk's theorem [42].

Theorem C.1. *Consider an algorithm driven by the iterative update in (36), where* $\mathbf{d}$ *is a descent direction and* α *satisfies the Wolfe conditions in (31) and (32). Furthermore, the function* f_{uc} *is bounded below and is continuously differentiable on an open set* $\mathcal{N}$ *containing the level set* $\mathcal{L} := \{\hat{\eta} : f_{uc}(\hat{\eta}) \leq f_{uc}(\hat{\eta}_\mathbf{0})\}$*. Also, the gradient* $\nabla\mathbf{f_{uc}}$ *is Lipschitz continuous on* $\mathcal{N}$*. Then*

$$\sum_{\forall k} \left[\cos^2(\theta_k) \left\|\nabla\mathbf{f_{uc}}\right\|^2\right] < \infty \tag{43}$$

where θ_k *is the angle between the search direction* $\mathbf{d}$ *and the steepest descent search direction* $-\nabla f_{uc}$*.*

Theorem C.2. *Let* B_0 *be any symmetric positive definite initial Hessian approximation. Then the sequence generated by (36) converges to* $\hat{\eta}^*$ *of the objective function* f_{uc}*.*

Proof. We define the following

$$m_k = \frac{y's}{s's}, \quad M_k = \frac{y'y}{y's}. \tag{44}$$

Note that from (41) and (42) that

$$m_k \geq m, \quad M_k \leq M. \tag{45}$$

It can be shown that the trace of the BFGS update in (37) is

$$\text{trace}(B_{k+1}) = \text{trace}(B_k) - \frac{\|B_k s\|^2}{s'Bs} + \frac{\|y\|^2}{y's}. \tag{46}$$

It can also be shown that the determinant of the BFGS update is

$$\det(B_{k+1}) = \det(B_k)\frac{y's}{s'Bs}. \tag{47}$$

We can now define the angle θ_k as

$$\theta_k = \frac{s'Bs}{\|s\|\,\|Bs\|}, \tag{48}$$

Journal Pre-proof

and

$$q = \frac{s'Bs}{s's}. \tag{49}$$

We can then rewrite

$$\frac{\|Bs\|^2}{s'Bs} = \frac{\|Bs\|^2}{(s'Bs)^2}\frac{s'Bs}{\|s\|^2} = \frac{q}{\cos^2(\theta_k)}, \tag{50}$$

as well as rewrite

$$\det(B_{k+1}) = \det(B_k)\frac{y's}{s's}\frac{s's}{s'Bs} = \det(B_k)\frac{m_k}{q}. \tag{51}$$

Let us now define the following function of the approximate Hessian

$$\psi(B) = \text{trace}(B) - \log(\det(B)), \tag{52}$$

where $\psi(B)$ is always positive. We can now write

$$\psi(B_{k+1}) = \text{trace}(B_k) + M_k - \frac{q}{\cos^2(\theta_k)} - \log(\det(B_k)) - \log(m_k) + \log(q_k), \tag{53}$$

which in turn can be rewritten as

$$\psi(B_{k+1}) = \psi(B_k) + u_k + \left[1 - \frac{q}{\cos^2(\theta_k)} + \log\left(\frac{q}{\cos^2(\theta_k)}\right)\right] + \log(\cos^2(\theta_k)), \tag{54}$$

where $u_k = M_k - \log(m_k) - 1$. The term inside the square bracket is non-positive, and therefore from (45) and (54), we can state the following on the initial Hessian approximation,

$$0 < \psi(B_{k+1}) \leq \psi(B_0) + u(k+1) + \sum_{i=0}^{k}\log(\cos^2(\theta_i)), \tag{55}$$

where $0 < u = M - \log(m) - 1$. From (43), we know that the gradient does not converge to zero if and only if $\cos\theta_i$ goes to zero. Let us assume that $\cos\theta_i$ does go to zero. Therefore, there will exist an iteration k_1 such that for all $i > k$, we have

$$\log(\cos^2(\theta_i)) < -2c. \tag{56}$$

Then we can rewrite (55) as

$$0 < \psi(B_0) + u(k+1) + \sum_{i=0}^{k_1}\log(\cos^2(\theta_i)) + \sum_{i=k_1+1}^{k}(-2c)), \tag{57}$$

or equivalently,

$$0 < \psi(B_0) + u(k+1) + \sum_{i=0}^{k_1}\log(\cos^2(\theta_i)) + 2ck_1 + c - ck, \tag{58}$$

The right hand side of the previous equation is negative for large k, therefore $\cos(\theta_i)$ does not go to zero, and therefore due to Zoutendijk's theorem, $\|\nabla f_{uc}\| \to 0$ and the problem converges. Due to the convexity of the problem, the converged solution is also the optimal solution $\hat{\eta}^*$. □

Journal Pre-proof

Journal Pre-proof

ACKNOWLEDGMENTS

The authors would like to thank Dr. Tao Han at the University of North Carolina, Charlotte, USA for the insight given on the methods in [28]. This work was supported and funded by Kuwait University Research Grant No. EO-08/18. This work was also partially supported by the Kuwait Foundation for the Advancement of Sciences (KFAS), under project code PN17-15EE-02.

REFERENCES

[1] European Comission, "5G vision: The 5G infrastructure public private partnership: the next generation of communication networks and services," European Comission, European Union, Tech. Rep., Feb. 2015. [Online]. Available: https://espas.secure.europarl.europa.eu/orbis/document/5g-vision-5g-infrastructure-public-private-partnership-next-generation-communication

[2] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 65–75, Nov. 2014.

[3] C. Liu and P. Chen, "Load-aware coordinated multipoint joint transmission in dense heterogeneous networks: Downlink coverage and throughput limits," in *Proc. of IEEE International Conference on Communications (ICC)*. IEEE International Conference on Communications, May 2017, pp. 1–7.

[4] I. Siomina and D. Yuan, "Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization," in *Proc. of IEEE International Conference on Communications (ICC)*. IEEE International Conference on Communications, Jun. 2012, pp. 1357–1361.

[5] L. You, L. Lei, and D. Yuan, "Load balancing via joint transmission in heterogeneous LTE: Modeling and computation." Proc. of IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Aug. 2015, pp. 1173–1177.

[6] A. H. Sakr and E. Hossain, "Location-aware cross-tier coordinated multipoint transmission in two-tier cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 6311–6325, Nov. 2014.

[7] G. T. . V11.0.0, "Coordinated multi-point operation for LTE," *3GPP TSG RAN WG1*, Sept. 2011.

[8] L. Liu, Y. Zhou, V. Garcia, L. Tian, and J. Shi, "Load aware joint CoMP clustering and inter-cell resource scheduling in heterogeneous ultra dense cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2741–2755, Mar. 2018.

[9] G. Cili, H. Yanikomeroglu, and F. R. Yu, "Cell switch off technique combined with coordinated multi-point (CoMP) transmission for energy efficiency in beyond-LTE cellular networks." Proc of. IEEE International Conference on Communications (ICC), Jun. 2012, pp. 5931–5935.

[10] N. Saxena, A. Roy, and H. Kim, "Traffic-aware cloud RAN: A key for green 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 1010–1021, Apr.. 2016.

[11] Vodafone, "Sustainable Business Report 2018," Vodafone Group Plc., Tech. Rep., 2018.

[12] A. Papa, R. Durner, E. Goshi, L. Gorattiy, T. Rasheedy, A. Blenk, and W. Kellerer, "MARC: On modeling and analysis of software-defined radio access network controllers," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2021.

[13] A. Papa, R. Durner, L. Goratti, T. Rasheed, and W. Kellerer, "Controlling next-generation software-defined RANs," *IEEE Communications Magazine*, vol. 58, no. 7, pp. 58–64, 2020.

[14] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network." Proc. of the Second ACM SIGCOMM Workshop on Hot Topics in Networking (HotSDN), 2013, pp. 25–30.

[15] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, 2014.

[16] L. A. Fletscher, J. Barreiro-Gomez, C. Ocampo-Martinez, C. V. Peroni, and J. M. Maestre, "Atomicity and non-anonymity in population-like games for the energy efficiency of hybrid-power HetNets," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1600–1614, Dec. 2018.

[17] 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, 2012.

Journal Pre-proof

Journal Pre-proof

[18] J. Yu, Y. Wang, Q. Zhang, Q. Jiang, and S. Yue, "A greedy algorithm-based BP neural network for user association in HetNets." Proc. of 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), May 2018, pp. 340–345.

[19] R. Fletcher, *Practical Methods of Optimization; 2nd Ed.* New York, NY, USA: Wiley-Interscience, 1987.

[20] H. Ghazzai, M. J. Farooq, A. Alsharoa, E. Yaacoub, A. Kadri, and M. Alouini, "Green networking in cellular HetNets: A unified radio resource management framework with base station on/off switching," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 5879–5893, Jul. 2017.

[21] L. Tang, W. Wang, Y. Wang, and Q. Chen, "An energy-saving algorithm with joint user association, clustering, and on/off strategies in dense heterogeneous networks," *IEEE Access*, vol. 5, pp. 12 988–13 000, 2017.

[22] Y. Li, M. Sheng, Y. Sun, and Y. Shi, "Joint optimization of BS operation, user association, subcarrier assignment, and power allocation for energy-efficient HetNets," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3339–3353, Dec. 2016.

[23] S. Zhang, N. Zhang, S. Zhou, J. Gong, Z. Niu, and X. Shen, "Energy-aware traffic offloading for green heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1116–1129, May 2016.

[24] Y. L. Lee, W. L. Tan, S. B. Y. Lau, T. C. Chuah, A. A. El-Saleh, and D. Qin, "Joint cell activation and user association for backhaul load balancing in green HetNets," *IEEE Wireless Communications Letters*, vol. 9, no. 9, pp. 1486–1490, 2020.

[25] Q. Han, B. Yang, C. Chen, and X. Guan, "Energy-aware and QoS-aware load balancing for HetNets powered by renewable energy," *Computer Networks*, vol. 94, pp. 250 – 262, Jan. 2016.

[26] T. Zhang, H. Xu, and Y. Chen, "User association for energy balancing in HetNets with hybrid energy sources," in *Proc. of IEEE 85th Vehicular Technology Conference (VTC Spring)*. IEEE Vehicular Technology Conference, Jun. 2017, pp. 1–5.

[27] D. Liu, Y. Chen, K. K. Chai, and T. Zhang, "Distributed delay-energy aware user association in 3-tier HetNets with hybrid energy sources." 2014 IEEE Globecom Workshops (GC Wkshps), Dec. 2014, pp. 1109–1114.

[28] T. Han and N. Ansari, "A 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.

[29] K. M. S. Huq, S. Mumtaz, J. Bachmatiuk, J. Rodriguez, X. Wang, and R. L. Aguiar, "Green HetNet CoMP: Energy efficiency analysis and optimization," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4670–4683, Oct. 2015.

[30] C. Jialing, Y. Mingxi, D. Xiaohui, and J. Bingli, "Q -learning based selection strategies for load balance and energy balance in heterogeneous networks," in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, 2020, pp. 728–732.

[31] H. P. Kuribayashi, M. A. De Souza, D. De Azevedo Gomes, K. Da Costa Silva, M. S. Da Silva, J. C. Weyl Albuquerque Costa, and C. R. Lisbôa Francês, "Particle swarm-based cell range expansion for heterogeneous mobile networks," *IEEE Access*, vol. 8, pp. 37 021–37 034, 2020.

[32] Y. Cao, H. Xia, and C. Feng, "Evaluation of diverse cell range expansion strategies applying CoMP in heterogeneous network." Proc of. IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Sept. 2013, pp. 1962–1966.

[33] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A flexible and programmable platform for software-defined radio access networks," in *Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies*, 2016, pp. 427–441.

[34] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: A software-defined networking platform for 5G radio access networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 715–728, 2019.

[35] 3GPP, "TS 138 401 V15.9.0 (2020-11) 5G and NG-RAN architecture description (3GPP TS 38.401 version 15.9.0 Release 15)," Standard, 11 2020.

[36] T. Han and N. Ansari, "Green-energy aware and latency aware user associations in heterogeneous cellular networks," in *2013 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2013, pp. 4946–4951.

[37] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Two-dimensional optimization on user association and green energy allocation for HetNets with hybrid energy sources," *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 4111–4124, Nov. 2015.

[38] X. Liu, T. Han, and N. Ansari, "Intelligent battery management for cellular networks with hybrid energy supplies." Proc. of IEEE Wireless Communications and Networking Conference, Apr. 2016, pp. 1–6.

[39] E. Chong and S. Zak, *An Introduction to Optimization*, ser. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2013.

Journal Pre-proof

Journal Pre-proof

[40] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan, "Mixed-integer nonlinear optimization," *Acta Numerica*, vol. 22, 05 2013.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[42] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

[43] IEEE, "IEEE 802.16m evaluation methodology document (EMD)," IEEE, Tech. Rep., 2009.

[44] IEEE, "IEEE recommended practice for testing the performance of stand-alone photovoltaic systems," *IEEE STD 1526-2003*, pp. 1–18, 2004.

[45] S. H. Karaki, R. B. Chedid, and R. Ramadan, "Probabilistic performance assessment of autonomous solar-wind energy conversion systems," *IEEE Transactions on Energy Conversion*, vol. 14, no. 3, pp. 766–772, Sept. 1999.

[46] Y. M. Atwa, E. F. El-Saadany, M. M. A. Salama, and R. Seethapathy, "Optimal renewable resources mix for distribution system energy loss minimization," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 360–370, Feb. 2010.

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

Mohamad Khattar Awad (S'02, M'09, SM'17), earned the B.A.Sc. in electrical and computer engineering (communications option) from the University of Windsor, Ontario, Canada, in 2004 and the M.A.Sc. and Ph.D. in electrical and computer engineering from the University of Waterloo, Ontario, Canada, in 2006 and 2009, respectively.

From 2004 to 2009 he was a research assistant in the Broadband Communications Research Group (BBCR), University of Waterloo. From 2009 to 2012, he was an Assistant Professor of Electrical and Computer Engineering at the American University of Kuwait. Since 2012, he has been with Kuwait University, where currently he is an Associate Professor of Computer Engineering.

Dr. Awad's research interest includes wireless and wired communications, software-defined networks resource allocation, wireless networks resource allocation, and acoustic vector-sensor signal processing. He is a frequent reviewer for several journals and conferences. Dr. Awad served on the editorial board of the IEEE Transactions on Green Communications and Networking (TGCN) between October 2016 and May 2021.

He received the Ontario Research & Development Challenge Fund Bell Scholarship in 2008 and 2009, the University of Waterloo Graduate Scholarship in 2009, and a fellowship award from the Dartmouth College, Hanover, NH in 2011. In 2015 and 2017, he received the Kuwait University Teaching Excellence Award and Best Young Researcher Award, respectively.

Ali A. M. R. Behiry Ali A. M. R. Behiry (S'18) earned his B.Eng. in computer engineering from the American University of Kuwait, Kuwait, in 2015 and the M.Sc. in computer engineering from Kuwait University, Kuwait, in 2019.

He has served as a part-time lab assistant and tutor during his undergraduate years from 2012 to 2015 at the American University of Kuwait, as well as a lab coordinator and instructor at Kuwait University from 2016 to 2019, during his graduate studies. Currently, Ali is an undergraduate lab instructor in the Department of Engineering at the American University of Kuwait.

His recent research interests include wireless mobile networks, wireless sensor networks, optimization and machine learning. Ali received the highest academic merit scholarship from the American University of Kuwait in 2011, as well as the full Excellence Scholarship from Kuwait University in 2016.

Journal Pre-proof

Mohammed W. Baidas received the B.Eng. (Hons.) degree in communication systems engineering from the University of Manchester, Manchester, U.K., in 2005, the M.Sc. degree (with distinction) in wireless communications engineering from the University of Leeds, Leeds, U.K., in 2006, the M.S. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2009, and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2012. He was a Visiting Researcher with the University of Manchester in the Academic Years of 2015/2016 and 2018/2019. He is currently an Associate Professor with the Department of Electrical Engineering, Kuwait University, Kuwait, where he has been on the faculty since May 2012. He is also a frequent reviewer for several IEEE journals and international journals and conferences, with over 80 publications. His research interests include resource allocation and management in cognitive radio systems, game theory, cooperative communications and networking, and green and energy-harvesting networks. He also serves as a technical program committee member for various IEEE and international conferences. He was a recipient of the Outstanding Teaching Award of Kuwait University for the academic year of 2017/2018.

Journal Pre-proof