A Quasi-Newton-based Approach to Load Balancing in Coordinated MultiPoint (CoMP) Green HetNets

Mohamad Khattar Awad, Ali A. M. R. Behiry,
Department of Computer Engineering, College of Engineering and Petroleum, Kuwait University
E-mail: mohamad@ieee.org, alirady@ieee.org

***Abstract*—The number of mobile connected devices have increased exponentially worldwide. The services provided across mobile networks have also become increasingly demanding for higher network capacity. This has motivated the use of the heterogeneous network (HetNet) architecture. However, despite the low power consumption of small-scale base stations (BSs), their collective power consumption in dense HetNets is significant. The use of green energy to mitigate the power consumption of mobile networks is a trend on the rise. However, traditional association schemes under utilize green energy. Furthermore, in dense HetNets, there is an increased chance of users being on cell-edges and having degraded perceived service. Coordinated Multipoint (CoMP) association can aid in improving the service perceived by cell edge users. In this work, we propose a load balancing scheme that optimizes user latency and green energy utilization. The scheme allows for a fractional solution to the user association problem, enabling CoMP transmissions for cell-edge users. The proposed algorithm is a Quasi-Newton-based approach, which applies the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method of approximating the inverse of Hessian matrices. Performance evaluation shows a reduction in latency of 79% and a reduction of power consumption by 99% in comparison to conventional schemes.**

***Index Terms*—HetNet, Green, CoMP, L-BFGS, Quasi-Newton,**

I. Introduction

Technology has become a primary component in any individual's personal and professional life. To this end, there has been a tremendous increase in the number of network connected devices; and hence the traffic they produce. Cisco predicts that total monthly mobile data traffic will nearly triple from 2018 to 2021, up to 49 exabytes [1]. Furthermore, network services are increasingly demanding in their required (QoS). Thus, mobile operators are obligated to extend the coverage and capacity of their Radio Access Network (RAN), while mitigating OPEX and CAPEX for their profitability.

HetNets emerged as a promising cost effective architecture to support the rising mobile traffic loads. In HetNets, micro BSs (MiBSs) are densely deployed in high user density areas to alleviate the load off Macro BSs (MaBSs) [2]. MiBSs consume less energy, have less transmission power, antenna height and hence a smaller coverage areas than MaBs. However, the low CAPEX for deploying them in high quantities allows the network capacity and coverage area to be extended at a minimal cost. Although indvidually MiBSs have low power consumption, their collective power consumption in a dense HetNet is massive. The energy cost ranges from 10% to 40% of the overall network OPEX [3], [4]. In order to remain competitive and profitable, mobile operators require cost reduction, particularly in long running OPEX. To this end, mobile operators are increasingly turning to the use of hybrid energy sources consisting of both conventional and renewable energy sources. Using renewable energy sources alternatives not only improves the operational efficiency of the network, but also its environmental impact. However, load balancing schemes can sometimes lead to a deterioration of individual user rates due to the association of users with BSs with a lower perceived data rate. In dense HetNet architectures, the probability of users being located on cell-edges increases dramatically from traditional networks. Coordinated Multipoint (CoMP) was introduced in LTE-A [5] in order to mitigate inter-cell interference and improve perceived rates for cell-edge users. CoMP provides an architecture where two BSs can be coordinated to serve a user in order to improve their perceived data-rate [5]. The coordination required to implement a CoMP system is facilitated by the software-defined radio access networks (SoftRAN) architecture.

In this work, we are particularly interested in balancing the loads on BSs of SoftRAN via an optimized fractional user association. The fractional user association is considered as the probability of a user being associated with BSs. We propose a Quasi-Newton-based and centralized load balancing user association scheme for C-RAN. The SoftRAN architecture provides global information about average traffic loads and available green energy from users and BSs, respectively. This information is fed to the optimization algorithm that optimizes latency and power consumption. A tradeoff coefficient is implemented in order to balance between latency and power consumption when the two objectives are at odds. The fractional solution to the problem is utilized to implement the CoMP architecture in order to improve data rates of cell-edge users.

The rest of this paper is organized as follows. In Section II, we define the system model's and assumptions. In Section III we discuss the proposed user association algorithm, its required inputs, parameters and outputs. Section IV displays results of simulating the proposed scheme as well as comparing

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

it with similar schemes, and finally, Section V follows with concluding remarks.

II. System Model

In this paper, we consider a HetNet system with two tiers of BSs: macro base stations (MaBSs) and micro base stations (MiBSs). All base stations are equipped with solar panels of appropriate sizes to harvest green energy and are capable of complementing their energy needs with on-grid power. Our system aims to load balance down-link traffic loads among the BSs in order to optimize latency across BSs, utilize green energy in order to reduce on-grid power consumption and improve cell edge user rates. The system assumes an SDN architecture where information about average traffic loads and harvested green energy are available to the RANC during any given timeslot.

A. Traffic Model

Let $\mathcal{B}$ denote the set of all BSs that serve a predefined geographical area denoted as $\mathcal{A}$, and $x \in \mathcal{A}$ references any specific location following the general model presented in [6], [7]. Then, the achievable signal-to-interference-plus-noise ratio (SINR) by a user present at location x at if it is to be associated with the jth BS 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 is the transmission power of the jth BS, $g_j(x)$ is the channel gain between BS j and a given location x. This channel gain reflects the effects of path loss and shadowing, but not fast fading. The noise power level is denoted by Y_∞^2 and the set $\mathcal{I}_j(x)$ is the subset of BSs that interfere with the transmission of BS j at a given location x, while $I_k(x)$ is the average power caused by this interference from BS k such that $k \in \mathcal{I}_j$ and is considered static due to the frequency reuse plan of the network [8].

Hence, a user's downlink rate $r_j(x)$ achievable at a given location x will be expressed by a logarithmic function of their SINR, according to the Shannon-Hartley capacity:

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

where W_j is the total bandwidth in the jth BS.

The actual achieved traffic load at a location x in the jth BS at a given timeslot t is given by

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

such that $\lambda^x)$ is the poisson distributed number of traffic arrivals at a given location x, $\nu(x)$ is the exponentially distributed size of a single traffic arrival, $r_j(x)$ is the rate available to location x by the jth BS and $\eta_j(x)$ is a binary indicator (association) variable denoted by

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

Therefore the traffic load on the jth BS at timeslot t can be expressed as:

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

The value of ρ_j also refers to the fraction of time the base station is busy; known in queuing theory as utilization. Let us define the following constant $\gamma = \frac{\nu(x)}{r_j(x)}$, where γ can be considered constant during a single user association period, such that

$$\vartheta_j = \frac{E(\gamma^2)}{2}. \tag{6}$$

Now, one can express the average latency experienced by a traffic arrival associated to a given BS as a function of that base station's traffic density

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

Since ϑ_j can be considered a constant, minimizing the total latency in the network can be equivalent to minimizing the following

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

B. Energy Model

It is assumed that both tiers of BSs are equipped with solar panels of appropriate sizes. Let the power consumption of a base station be defined by:

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

where β_j is a constant that translates a traffic load to dynamic power consumption and p_j^s is the static power consumption of an online BS, regardless of its load. Furthermore, let p_j^o denote the on-grid power consumption of a base station

$$p_j^o = \max(p_j - e_j, 0), \tag{10}$$

where e_j is the green energy available at the jth BS. This is an important metric as on-grid power consumption incurs an environmental cost to the planet and a monetary cost to network operators. Then ρ_j^g can be defined as the amount of traffic load that can be supported by only consuming green energy.

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

and ϵ is a small positive constant to guarantee $0 < \rho_j^g < 1$. Therefore in order to minimize on-grid energy consumption, a BS's load must be less than or equal ρ_j^g, the load supportable by green energy for that BS. This is the same as minimizing the following

$$\phi_j(\rho_j) = \max(0, \rho_j - \rho_j^g)^2. \tag{12}$$

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

C. Optimization Problem

In order to minimize the latency indicator function in Equation (8), users must be offloaded from a heavily loaded BS to a less utilized one. However, it might occur that the BS being offloaded to, also does not have sufficient green energy, which gives rise to the cost in Equation (12), thereby consuming more on-grid power. Therefore, the latency reduction goal and the green energy utilization goal often conflict, leading to the need for a tradeoff. Let us define a combined cost function f such that

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

where k_j sets the tradeoff between the latency indicator and the on-grid power consumption. Therefore, the user association problem can be modeled as the following optimization problem,

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

$$\text{subject to} \quad \rho_j = \int_{x \in A} \frac{\lambda(x)\nu(x)\eta_j(x)}{r_j(x)} dx \tag{15}$$

$$\epsilon \leq \rho_j \leq 1 - \epsilon \qquad \forall j \tag{16}$$

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

$$\sum_{\forall j} \eta_j(x) = 1 \qquad \forall x \tag{18}$$

This problem is a mixed-integer-nonlinear-programming (MINLP) problem which is an $\mathcal{NP}$-hard problem [9]. In order to be able to solve this problem, the binary association variable $\eta_j(x)$ must be relaxed. Therefore, we introduce the new association variable $\hat{\eta_j(x)}$ where $0 \leq \hat{\eta_j(x)} \leq 1$, and the reformulated optimization problem becomes

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

$$\text{subject to} \quad \rho_j = \int_{x \in A} \frac{\lambda(x)\nu(x)\hat{\eta}_j(x)}{r_j(x)} dx \tag{20}$$

$$\epsilon \leq \rho_j \leq 1 - \epsilon \qquad \forall j \tag{21}$$

$$0 \leq \hat{\eta}_j(x) \leq 1 \qquad \forall j, x \tag{22}$$

$$\sum_{\forall j} \hat{\eta}_j(x) = 1 \qquad \forall x \tag{23}$$

This relaxed problem is a convex optimization problem where $\hat{\eta}_j(x)$ represents the probability of BS j serving requests to a user at location x. Solving this convex optimization problem provides a minimizing user association that optimizes the cost function in (19), while maintaining constraints (20), (21), (22) and (23).

III. Proposed Scheme

The centralized RANC collects information about average user traffic profiles, channel conditions and the availability of green energy at each BS. These information are fed to our proposed algorithm in order to determine solve the relaxed convex optimization problem and provide a fractional user association. The generated fractional user association is then used to associate users to single or multiple BSs based on the rounding algorithm.

In order to transform the constrained relaxed UA problem into an unconstrained relaxed UA problem, the constraints must be included within the objective function. Using an interior barrier function allows this for inequalities by exponentially increasing the cost when a constraint is nearly violated [10]. We use a logarithmic barrier function B_1 to represent the cost of approaching constraint (21) such that,

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

and another logarithmic barrier function B_2 to represent the cost of approaching constraint (22) such that,

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

Similarly, a quadratic penalty function [11] is used to represent the cost of deviating from equality constraint (23) where,

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

Given these function, the new unconstrained UA problem becomes

$$\underset{\hat{\eta}_j \forall j}{\text{minimize}} \quad f_{un} = f - \mu_1 \sum_{\forall j} B_1(\rho_j) - \mu_2 \sum_{\forall j,x} B_2(\eta_j(x)) + \frac{\omega}{2} \sum_{\forall x} Q(x), \tag{27}$$

where μ_1, μ_2 and ω are multipliers to dictate the severity of approaching or violating the conditions.

The information from second order methods allow optimization algorithms to enjoy faster convergence than their first order counterparts [12]. However, the processing time and memory required to calculate the Hessian of multivariate functions, when the network is reasonably dense, challenge practicality of the proposed algorithm. The proposed scheme uses a limited memory BFGS (L-BFGS) approximation to the Hessian, requiring less processing time to compute, and less memory to store. The algorithm uses the L-BFGS two loop method from [12] to estimate the inverse Hessian H.

For convenience, we assume all variables $\hat{\eta}_j(x)$ to be stacked in a single vector $\hat{\boldsymbol{\eta}}$ and can then be decomposed into separate association values. The proposed algorithm is an iterative algorithm performing the main step

$$\hat{\boldsymbol{\eta}}^+ = \hat{\boldsymbol{\eta}} + \alpha \mathbf{d} \tag{28}$$

where $\mathbf{d}$ is a search direction and α is the step size. The search direction is required to be in a descent direction relative to Equation (27), the objective function. There are a number of different feasible search directions. However, the most used is

Algorithm 1 L-BFGS User Association Algorithm

Require: Initial values for $\hat{\eta}_j(x)$, tradeoff coefficients k_j.
Ensure: new user association $\hat{\eta}_j^*(x)$

1: Set initial μ_1, μ_2, ω and t.
2: Stack association variables $\hat{\eta}_j(x)$ into vector $\hat{\boldsymbol{\eta}}$
3: **while** termination condition of barrier method **do**
4: **while** termination condition of LBFGS method **do**
5: // Compute inverse Hessian approximation H
6: $\mathbf{q} := \nabla \mathbf{f_{un}}$
7: **for** $i = 0$ to $m - 1$ **do**
8: $a_i := \frac{1}{\mathbf{y_i}'\mathbf{s_i}} \mathbf{s_i}'\mathbf{q}$
9: $\mathbf{q} := \mathbf{q_i} - a\mathbf{y_i}$
10: **end for**
11: $\mathbf{r} := \frac{\mathbf{y_0}'\mathbf{s_0}}{\mathbf{y_0}'\mathbf{y_0}} \mathbf{q}$
12: **for** $i = m - 1$ to 0 **do**
13: $b = \frac{1}{\mathbf{y_i}'\mathbf{s_i}} \mathbf{y_i}'\mathbf{r}$
14: $\mathbf{r} := \mathbf{r} + \mathbf{s_i}(a_i - b)$
15: **end for**
16: // $\mathbf{r}$ contains $\mathbf{H}\nabla \mathbf{f_{un}}$
17: Compute $\mathbf{d} := -\mathbf{H}\nabla \mathbf{f_{un}}$ using $\mathbf{r}$
18: $\alpha :=$ LINE SEARCH($\hat{\boldsymbol{\eta}}$,$\mathbf{d}$)
19: $\hat{\boldsymbol{\eta}}^+ = \hat{\boldsymbol{\eta}} + \alpha \mathbf{d}$
20: Remove $\mathbf{s_{m-1}}$ and $\mathbf{y_{m-1}}$
21: $\mathbf{s_{i+1}} := \mathbf{s_i}, \forall i \in [1, m-2]$
22: $\mathbf{y_{i+1}} := \mathbf{y_i}, \forall i \in [1, m-2]$
23: $\mathbf{s_0} := \hat{\boldsymbol{\eta}}^+ - \hat{\boldsymbol{\eta}}$
24: $\mathbf{y_0} := \nabla \mathbf{f_{un}^+} - \nabla \mathbf{f_{un}}$
25: **end while**
26: $\omega := t\omega$
27: $\mu_1 := \mu_1 / t$
28: $\mu_2 := \mu_2 / t$
29: **end while**
30: $\hat{\eta}_j^*(x) = \hat{\eta}_j(x)$

the Newton's direction, defined by the negative inverse Hessian of the objective function multiplied by its gradient [12].

Since the proposed scheme in Algorithm 1 solves iterative approximated version of the original constrained problem and hence speed is highly preferred over accuracy. The algorithm uses an approximation to the inverse Hessian of the objective function Equation (27) by storing $2m$ vectors of size n each, where m is an algorithm constant, n is the number of optimization variables and $m < n$. Practically, m does not need to be large. This actively decreases the need of storing and computing n^2 operations every time the Hessian is required.

Let there be two categories of vectors denoted by $\mathbf{s_i}$ and $\mathbf{y_i}$, where $i \in [0, m)$. The vectors are stored in a fixed length queue structure. In other words, new calculations of $\mathbf{s_0}$ and $\mathbf{y_0}$ are made each iteration, and the queue is pushed forward with the last two vectors being popped out of the queue. This can be seen in lines 20 to 24 of Algorithm 1.

The vector $\mathbf{s_i}$ measures the change in the optimization variable, while the vector $\mathbf{y_i}$ measures the change in the gradient of the objective function in Equation (27). Let the gradient of the objective function in Equation (27) relative to a single association variable be

$$\begin{aligned}\frac{\partial f_{un}}{\partial \hat{\eta}_j(x)} =& \frac{c_j(x)}{(1-\rho_j)^2} + 2c_j(x)k_j(\max(0, \rho_j - \rho_j^g)) \\ & - \mu_1 \left(\frac{c_j(x)}{\rho_j - \epsilon} - \frac{c_j(x)}{((1-\epsilon) - \rho_j)} \right) \\ & - \mu_2 \left(\frac{1}{\hat{\eta}_j(x)} - \frac{1}{1 - \hat{\eta}_j(x)} \right) + \omega \ \sum_{\forall j} \hat{\eta}_j(x) - 1 \Bigg],\end{aligned} \tag{29}$$

where $c_j(x) = \frac{\lambda(x)\nu(x)}{r_j(x)}$. The gradient vector $\nabla \mathbf{f_{un}}$ is obtained from stacking the derivatives relative to all association variables $\hat{\eta}_j(x)$ for all j and x into a single vector. The line search used satisfies the Wolfe conditions in order to ensure sufficient step/progress taken in each search direction [13]. During the line search phase of the algorithm, the association values $\hat{\boldsymbol{\eta}}$ and the search direction $\mathbf{d}$ are fixed. Therefore, for notational convenience, we can define

$$\psi(\alpha) = f_{un}(\hat{\boldsymbol{\eta}} + \alpha \mathbf{d}). \tag{30}$$

Furthermore, the derivative of Equation (30) is known as the directional derivative of f_{un} in $\mathbf{d}$ and is given by

$$\nabla \psi(\alpha) = \nabla \mathbf{f_{un}}(\hat{\boldsymbol{\eta}}' + \alpha \mathbf{d})\mathbf{d} \tag{31}$$

Algorithm 2 Line Search Algorithm

1: **function** LINE SEARCH($\hat{\boldsymbol{\eta}}$,$\mathbf{d}$)
2: Initialize $\alpha_0 := 0$, $\alpha_{\max} = \frac{\psi_{\min} - \psi(0)}{\beta_1 \nabla \psi(0)}$, $\alpha_1 \in (0, \alpha_{\max})$
3: $i := 1$
4: **loop**
5: **if** $[\psi(\alpha_i) > \psi(0) + \beta_1 \alpha_i \nabla \psi(0)]$ **then**
6: **return** SECTION(α_{i-1}, α_i)
7: **end if**
8: **if** $[\psi(\alpha_i) > \psi(\alpha_{i-1})]$ and $i > 1$ **then**
9: **return** SECTION(α_{i-1}, α_i)
10: **end if**
11: **if** $|\nabla \psi(\alpha_i)| \leq -\beta_2 \nabla \psi(0)$ or $\psi(\alpha_i) < \psi_{\min}$ **then**
12: **return** α_i
13: **end if**
14: **if** $\nabla \psi(\alpha_i) \geq 0$ **then**
15: **return** SECTION(α_i, α_{i-1})
16: **end if**
17: **if** $2\alpha_i - \alpha_{i-1} < \alpha_{\max}$ **then**
18: Interpolate for $\alpha_{i+1} \in [\alpha_{\text{low}}, \alpha_{\text{high}}]$
19: **end if**
20: $\alpha_{lb} := 2\alpha_i - \alpha_{i-1}$
21: $\alpha_{ub} := \min(\alpha_{\max}, \alpha_i + \tau_1(\alpha_i - \alpha_{i-1})$
22: Interpolate for $\alpha_{i+1} \in [\alpha_{lb}, \alpha_{ub}]$
23: $i := i + 1$
24: **end loop**
25: **end function**

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

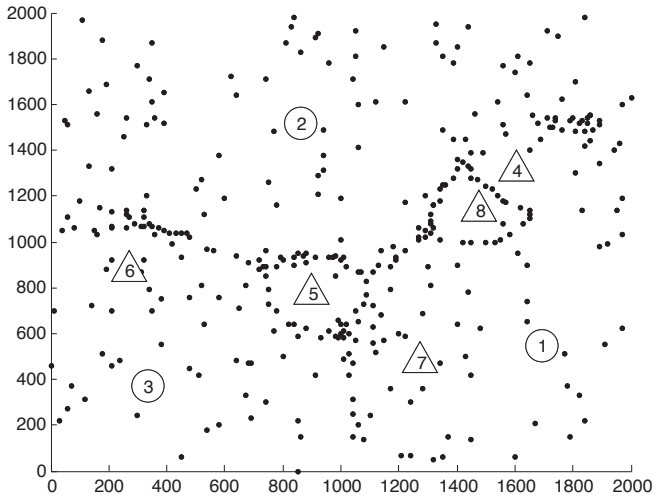


Fig. 1. Base station deployment scenario with increased user density on cell edges

Algorithm 3 Sectioning Phase Algorithm

function SECTION(a, b)
  loop
    $\alpha_{lb} := \alpha_{\text{low}} + \tau_2(\alpha_{\text{high}} - \alpha_{\text{low}})$
    $\alpha_{ub} := \alpha_{\text{high}} - \tau_3(\alpha_{\text{high}} - \alpha_{\text{low}})$
    Interpolate for $\alpha \in [\alpha_{lb}, \alpha_{ub}]$
    if $\psi(\alpha) > \psi(0) + \beta_1\alpha\nabla\psi(0)$ or $\psi(\alpha) \geq \psi(\alpha_{\text{low}})$ **then**
      $\alpha_{\text{high}} := \alpha$
    else
      if $|\nabla\psi(\alpha)| \leq -\beta_2\psi'(0)$ **then**
        return α
      end if
      if $\nabla\psi(\alpha)(\alpha_{\text{high}} - \alpha_{\text{low}}) \geq 0$ **then**
        $\alpha_{\text{high}} := \alpha_{\text{low}}$
      end if
      $\alpha_{\text{low}} := \alpha$
    end if
  end loop
end function

The line search operates in two nested phases. The first phase in Algorithm 2 attempts to find a viable step length or a bracket that is guaranteed to contain a range of viable step lengths. The second phase in Algorithm 3 shrinks down the bracket iteratively until a viable step length is found [13]. Here, we define the viability of a step length by meeting two conditions known as the strong Wolfe conditions. The first being

$$\psi(\alpha) \leq \psi(0) + \beta_1\alpha\nabla\psi(0), \tag{32}$$

and is known as the sufficient decrease condition. The second is

$$|\nabla\psi(\alpha)| \leq -\beta_2|\psi'(0)| \tag{33}$$

and is known as the curvature condition [12]. These conditions ensure that the step length taken achieves sufficient progress in minimizing the objective function in Equation (27). This is essential as it avoids making unnecessary steps of small length.

Here, τ_1, τ_2, τ_3, β_1 and β_2 are all algorithm related constant values with a flexible range of use. In other words, there is no need to tune the parameters for different network scenarios. The sectioning algorithm in Algorithm 3 selects a value for α among the found bracket and then shrinks that bracket down. This is done iteratively until a viable step length is found. The interpolation step mentioned in Algorithms 2 and 3 is used to minimize the function $\psi(\alpha)$ within the provided bracket. The interpolation method used was cubic interpolation [13]. Therefore, the α found by Algorithms 2 and 3 then satisfies the strong Wolfe conditions in Equations (32) and (33).

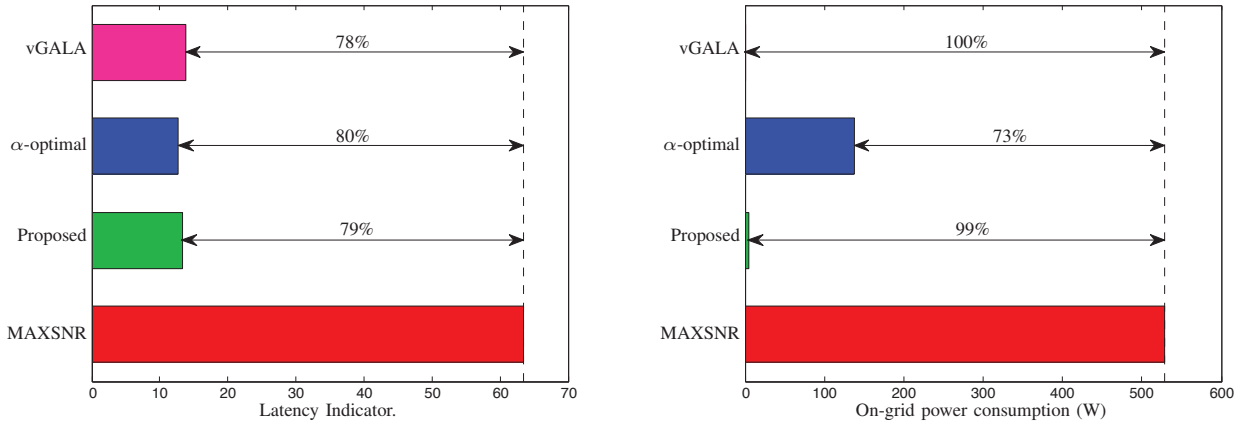
IV. Simulation Results

In this section we evaluate the evaluate the performance of the proposed scheme and evaluate significance of improvement it brings to cell edge users in comparison to existing schemes. These schemes are MAXSNR, α-optimal and vGALA. The MAXSNR associates users merely based on the strength received from a BS. The other two schemes α-optimal and vGALA are presented in [6] and [7], respectively. Figure 1 shows the simulation scenario, where 250 users are uniformly distributed. In addition, 100 users are placed along the edges of the BSs' coverage areas, where signals received from different BSs are comparable. This is not an uncommon scenario as HetNets are often densely deployed without thorough planning of coverage areas. This increases the possibility of any user being an edge user as the density of the HetNet rises.

Figure 2 shows the performance of each of the tested schemes in terms of latency and power consumption. Clearly, the MAXSNR scheme had the worse latency response out of all schemes. The other tested schemes had approximately the same latency response. The α-optimal scheme had the largest reduction, decreasing the latency indicator over MAXSNR by 80%, followed by our proposed scheme, decreasing latency by 79% and lastly vGALA, decreasing it by 78%. As seen in Figure 2(b), the MAXSNR scheme also had the highest power consumption. The α-optimal scheme consumed 73% less power than the MAXSNR scheme. Our proposed scheme and the vGALA scheme consumed approximately no power, saving 99% and 100% respectively. It can be seen that our scheme performed well compared to the other two schemes based on both latency and energy consumption.

Figure 3 shows the summation of data rates for all cell edge users across the tested schemes. It can be seen that our proposed scheme achieved sum rate improvement for cell edge users over all other schemes. Our proposed algorithm increased rates over MAXSNR by 24%, over α-optimal by 32% and by vGALA by 40%. These improvements can be attributed to the implemented CoMP association resulting from

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE



(a) Average Latency Indicator

(b) On-grid power consumption

Fig. 2. Comparison between the different schemes in a network scenario with large number of cell edge users.

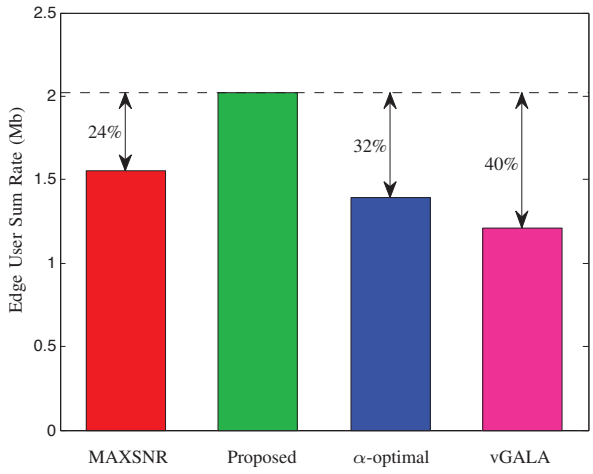


Fig. 3. Sum rate performance of system edge users.

our fractional association.

V. Conclusions

In this paper, we proposed a centralized load balancing scheme for hybrid and CoMP-enabled HetNets. The algorithm used the global information provided by the SoftRAN architecture to optimize user associations. The scheme was designed to be scalable to dense networks by using the L-BFGS method, not requiring explicit storage or computation of the Hessian. The algorithm considered average latency and on-grid power consumption jointly using a tradeoff coefficient. Furthermore, the algorithm allowed for a fractional result which facilitated CoMP association for cell-edge users in order to improve their data rates. Simulations showed that the algorithm accomplished significant improvements in reducing latency, power consumption and increasing cell-edge user rates over existing association schemes.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016 - 2021," Cisco Systems Inc, White Paper, Feb 2017.

[2] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, June 2011.

[3] S. Kumar, "Tower power africa: Energy challenges and opportunities for the mobile industry in africa," GSMA, https://www.gsma.com, Tech. Rep., 2014.

[4] D. Lister, "An operator's view on green radio," 2009, presented at the IEEE International Workshop on Green Communications.

[5] 3GPP TR 36.814, "3rd generation partnership project technical specification group radio access network evolved universal terrestrial radio access (e-utra)further advancements for e-utra physical layer aspects (release 9)," China Mobile Research, Technical Report, Oct 2010.

[6] 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, April 2016.

[7] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 177–190, Feb 2012.

[8] ——, "alpha-optimal user association and cell load balancing in wireless networks," in *2010 Proceedings IEEE INFOCOM*, March 2010, pp. 1–5.

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

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

[11] E. Chong and S. Zak, *An Introduction to Optimization*, ser. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2013. [Online]. Available: https://books.google.com.kw/books?id=iD5s0iKXHP8C

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

[13] R. Fletcher, *Practical Methods of Optimization; (2Nd Ed.)*. New York, NY, USA: Wiley-Interscience, 1987.

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE