Distributed Inter-BS Cooperation Aided Energy Efficient Load Balancing for Cellular Networks

[1]Md. Farhad Hossain, [2]Kumudu S. Munasinghe and [3]Abbas Jamalipour

[1,3]School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

[2]Discipline of Information Technology and Engineering, University of Canberra, ACT 2601, Australia

[1]md.hossain@sydney.edu.au, [2]kumudu.munasinghe@canberra.edu.au, [3]a.jamalipour@ieee.org

Abstract

We propose a distributed inter-base station (BS) cooperation assisted load balancing framework for improving energy efficiency of OFDMA-based cellular access networks. Proposed cooperation is formulated following the principle of ecological self-organization. Based on the network traffic, BSs mutually cooperate for distributing traffic among themselves and thus, the number of active BSs is dynamically adjusted for energy savings. For reducing the number of communications among BSs, a three-step measure is taken by using estimated load factor (LF), initializing the algorithm with only the active BSs and differentiating neighboring BSs according to their operating modes for distributing traffic. An exponentially weighted moving average (EWMA)-based technique is proposed for estimating the LF in advance based on the historical data. Various selection schemes for finding the best BSs to distribute traffic are also explored. Furthermore, we present an analytical formulation for modeling the dynamic switching of BSs. A thorough investigation under a wide range of network settings is carried out in the context of an LTE system. Results demonstrate a significant enhancement in network energy efficiency yielding up to 30% higher savings than the compared schemes. Moreover, frequency of correspondence among BSs can be reduced up to 80%.

Index Terms

Energy efficiency, inter-BS cooperation, load balancing, radio access network.

*This work is supported by the Australian Research Council under the Discovery Project (DP 1096276)

I. INTRODUCTION

Due to the rapidly growing number of subscribers and diverse types of applications, energy utilization in cellular networks is increasing at an incredible rate. This ever increasing energy consumption has economical as well as environmental implications resulting in higher network operating cost and rising global warming respectively. To this end, energy expenditure by the radio access network (RAN) equipment of a typical cellular network is around 60%-80% of the total amount [1] - [3]. Therefore, improving the energy efficiency of RANs has become the centre of focus to the researchers of green cellular networks.

As a consequence of random call generation and mobility patterns of users, and uneven user distribution, cellular network traffic exhibits significant temporal and spatial diversity [4] - [6]. Traditionally, load balancing schemes are proposed to use this load imbalances for redistributing traffic among base stations (BSs). By doing this, a cellular network can benefits in many ways, such as, efficient use of frequency bands, coverage enhancement for cell edge users, increment of the overall network throughput, and so on [7] - [8]. In this paper, we apply the concept of load balancing for improving the energy efficiency of cellular access networks.

In our previous work [9], an energy saving scheme using predefined switching patterns for deactivating redundant BSs was outlined. However, traffic distribution is solely dependent on the instantaneous traffic levels of BSs making it signaling intensive, whereas the impacts of locations and user data rates are completely ignored. In addition, the system is applicable only for regular cell layouts and the load-dependent power usage in BSs is not considered. All of these issues including other improvements are addressed in this paper. The main contributions of this paper can be summarized as below:

• We propose an energy efficient load balancing framework for orthogonal frequency division multiple access (OFDMA)-based cellular access networks. Being inspired by ecological self-organization, a distributed inter-BS cooperation mechanism is proposed for load balancing. As per our proposal, based on the network traffic and other system settings, BSs are engaged into mutual cooperation for redistributing traffic, adjusting their transmit power, and switching between high power active mode and low power sleep mode. Thus, RANs are adaptively reconfigured with time using a reduced number of active BSs for achieving energy savings. Quality of service (QoS), more specifically, session blocking, user data rate and network coverage are also maintained.

• By employing a three-step measure, namely, use of estimated load factor (LF) for triggering the load balancing procedure, initialization of the proposed algorithm with only the active BSs and prioritization of the active neighbors over the sleeping ones for distributing traffic, we limit the number of communications among BSs. We also propose an exponentially weighted moving average (EWMA)-based estimator for predicting the LF envelope of a day in advance from the historical LF data.

• A range of selection schemes is outlined, which can be used by a BS to decide on the best neighboring BSs for distributing its traffic.

• An analytical model in terms of system parameters is also formulated for predicting the switching dynamics of BSs.

• We thoroughly investigate the system performance in the context of long term evolution (LTE) systems. Impact of various traffic scenarios, BS power profiles and their capacities, user data rates, BS selection schemes and other design parameters on the degree of energy savings and other system parameters is analyzed. Results demonstrate that the proposed system is capable to significantly improve the energy efficiency (subject to network settings). Comparisons with the existing works are also provided for further validating the proposal.

The rest of the paper is organized as follows. Section II presents a thorough study on the related works. System model with other system parameters is outlined in section III followed by the algorithm in Section IV. An analytical model is formulated in section V. Section VI presents the simulation results followed by the conclusions in section VII.

II. Related Works

Cellular network operators and vendors are highly concerned about the ever increasing cost of energy [10] - [11]. Consequently, during the recent years, various proposals have emerged for minimizing the energy consumption at RAN level by switching off BSs [3], [9], [12] - [23]. Aiming to leverage the natural temporal-spatial traffic diversity, LTE proposes turning off evolved node Bs (eNBs) at low-traffic times for saving energy [12]. However, the standard has left the issue of implementation schemes open for further research.

In light of this, a comprehensive framework for evaluating the overall global metrics for assessing energy efficiency of cellular networks has been developed under the EARTH project [6].

In line with this framework, we proposed an energy saving cellular access network by employing predefined switching patterns for BSs [9]. Authors in [13] and [14] also used deterministic patterns for switching BSs through mutual cooperation among BSs. However, these schemes in [9], [13] - [14] are applicable only for regular cell layouts. In contrast, [15] - [17] proposed centralized algorithms for dynamically shutting down BSs, but provided no mechanism for dynamically switching them back ON. Concept of cell zooming along with the switching of BSs was introduced in [18], whereas a theoretical optimization approach for evaluating the potential energy savings from turning off BSs was outlined in [19]. Along with the switching off BSs, scope of heterogeneous cell sizes for energy savings was also investigated in [20]. However, the schemes in [9], [13], [15] - [20] failed to capture the load-dependent power utilization in BSs resulting in overestimations. Moreover, many of them presented either very basic algorithms ignoring the actual locations of users [9], [15] - [16] or no algorithm at all [13], [17], [19].

For the networks with non-deterministic traffic patterns, an actor-critic based centralized learning framework for switching BSs into sleep mode was proposed in [21]. Besides, analysis of trade-off between energy savings and delay in the form of cost minimization problems was formulated in [22]. However, the systems in [21] - [22] do not guarantee user data rate leading to the potential degradation of service quality. On the other hand, a grid based traffic profiling scheme for selecting the best BSs to turn off was proposed in [23]. This system is designed under the framework of wideband code division multiple access (WCDMA)-based systems making it incompatible for next generation OFDMA-based systems. Moreover, [15], [16], [20] - [22] initialize their algorithms assuming all BSs in active mode leading to higher computational load, correspondences among the network entities and switching in BSs.

Our proposed system is free from many of the above issues. A wide range of power consumption models of BSs, instantaneous operating modes of BSs, actual location of users and QoS constraints are taken into account. Measures are taken to reduce computations. Moreover, proposed algorithm is of distributed type and applicable for any cell layout.

III. System Model

In this section, we present the proposed system model and other system components. The system model is presented in the context of OFDMA-based LTE systems, which may also be

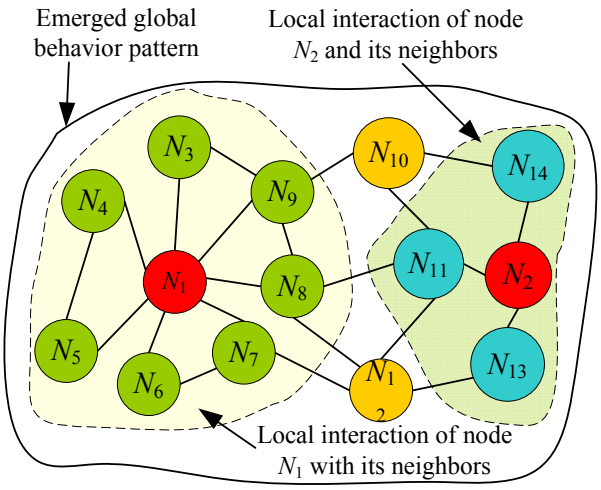


Fig. 1: Concept of ecological self-organization.

adopted to worldwide interoperability for microwave access (WiMAX) systems. Although in LTE, BSs are named as eNB, both the terms are used interchangeably throughout the paper.

A. Ecological Self-Organization

Proposed cellular network model is developed based on the principle of ecological self-organization. A self-organizing system in ecology has many small components, which follows a set of basic local rules. The interactions among components are executed using only local information, without being aware of the global pattern. This global pattern is an emergent property, which emerges spontaneously in a well-organized structured system starting from an initial random state and without any guidance from an external body [24] - [25]. The concept of ecological self-organization is presented in Fig. 1.

Local interaction domain of node N_1 and N_2 are illustrated by the dotted lines. The solid lines between two nodes represent the interaction between them. For instance, as shown in the figure, node N_1 and its neighboring nodes N_3-N_9 are interacting using their own local rules. Here, N_1 has no knowledge on the behavior of other nodes beyond its own domain (i.e., N_2 and N_{10}-N_{14}). However, neighboring nodes of N_1 are also interacting with their respective neighboring nodes, and thus, N_1 is indirectly influencing the behavior of all the nodes in the system. In turn, local behaviors of all the nodes generate the global behavioral pattern of the system. This type of

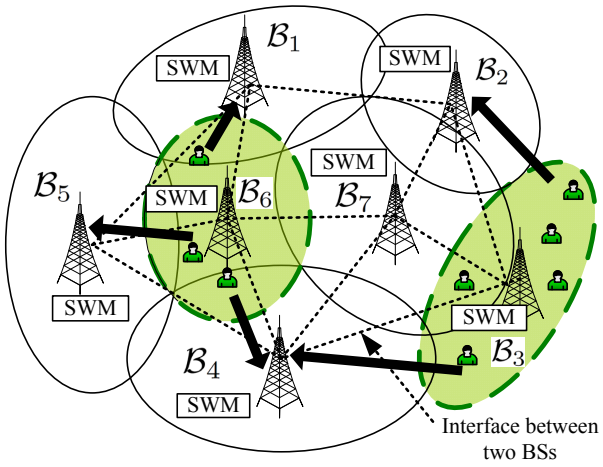


Fig. 2: Proposed network model ($\mathcal{B}_3$, $\mathcal{B}_6$ in sleep mode).

interaction is commonly observed among different species, such as, a flock of birds, a school of fishes, etc.

B. Proposed Energy Efficient Cellular Access Network

We consider the downlink of a multi-cell cellular network serving by a set of N BSs $\mathbf{B} = \{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_N\}$. We assume that the sectors in a BS are assigned orthogonal frequency bands, while the same frequency bands are reused among BSs. It is also considered that unused frequency bands and transmit power can be shared among the sectors.

Similar to the ecological system explained in Fig. 1, network wide distributed cooperation among BSs are employed in the proposed system. Concept of the network operation is demonstrated in Fig. 2. Proposed cooperation between a BS and its neighbors is governed by the traffic of itself and its neighbors, QoS requirements and other design parameters. In this paper, adjacent BSs capable to reach each other (e.g., over X2 interface in LTE) are considered as neighbors. For instance, neighboring BSs $\mathcal{B}_1$, $\mathcal{B}_4$ and $\mathcal{B}_5$ are serving as acceptors for $\mathcal{B}_6$ by cooperatively sharing its traffic (as shown by the arrows) allowing $\mathcal{B}_6$ to switch into sleep mode for saving energy. At the same time, for supporting the extended coverage zones, transmit power of these acceptors $\mathcal{B}_1$, $\mathcal{B}_4$ and $\mathcal{B}_5$ are also adjusted. BS $\mathcal{B}_6$ is totally unaware about the traffic environment of the other BSs except $\mathcal{B}_1$, $\mathcal{B}_4$, $\mathcal{B}_5$ and $\mathcal{B}_7$. On the other hand, during the high traffic time, upon receiving wake-up request from active BSs, sleep mode BSs switch to active mode for reducing

the traffic load on others. Through this dynamic switching of BSs, number of active BSs is adaptively adjusted and thus, energy savings is achieved.

Each BS is assumed to have a sleep-wake module (SWM), which enables it to switch between active and sleep modes. An active mode BS is fully functional as of conventional BSs. In contrast, a sleep mode BS neither carries user traffic nor performs control signaling; rather keeps a sleep-wake module (SWM) on for intercepting wake-up requests from other BSs.

In this paper, energy requirement for a SWM compared to that of an entire BS set is assumed insignificant. Furthermore, adaptive transmit power allocation in the downlink is assumed for maintaining the required signal strength at the user equipments (UEs). Hence, user data rates are not affected due to handover from an imminent sleeping BS to a neighbor.

C. Schemes for Selecting the Best BSs for Traffic Distribution

Candidate space of a BS is the collection of all the possible combinations of its neighbors. For distributing its traffic, a BS has to select the best combination from the candidate space, which can support its traffic as well as maintain QoS. We propose different selection schemes, which can be used by a BS for deciding on the best combination of BSs. Let us denote a selection scheme by S_n. This implies that the candidate space for a BS contains all the candidate combinations of up to n BSs taken from its neighbors. If K is the number of neighbors and the selection scheme is $S_n(n \leq K)$, then the total number of candidate combinations for a BS is equal to $\sum_{c=1}^{n} \binom{K}{c}$. For example, in Fig. 2, $\mathcal{B}_3$ has three neighbors $\mathcal{B}_2$, $\mathcal{B}_4$ and $\mathcal{B}_7$. If the selection scheme is S_2, then the candidate space for $\mathcal{B}_3$ contains the following six combinations of BSs: $\{\mathcal{B}_2\}, \{\mathcal{B}_4\}, \{\mathcal{B}_7\}, \{\mathcal{B}_2, \mathcal{B}_4\}, \{\mathcal{B}_2, \mathcal{B}_7\}$ and $\{\mathcal{B}_4, \mathcal{B}_7\}$.

D. EWMA-Based LF Estimation

In this paper, we use the aggregate LF of the network as an indicator for triggering the load distribution of BSs. Either the instantaneous LF or an estimated LF can be used for triggering this procedure. Let N_S be the number of instances per day at which traffic distribution is planned to be attempted. Use of instantaneous LF implies the necessity of measuring and gathering network parameters, and attempting to distribute traffic at all these N_S instances. This imposes a lot of signalling and computational cost. In contrast, from the estimated envelope, we can determine

the subset of N_S instances *a priori* at which load balancing procedure should start resulting a significant reduction in the number of computations and communications among BSs. Therefore, we here propose an EWMA-based technique for estimating the LF envelope of an entire day in advance from the historical data. Since the traffic level differs much from weekdays to weekends, we estimate the weekdays (weekends) LF from the weekdays (weekends) data.

Let the target is to estimate the LF for $(D+1)^{th}$ day from the given $D \times N_S$ LF matrix $\boldsymbol{\rho} = [\boldsymbol{\rho}_1; \boldsymbol{\rho}_2; ...; \boldsymbol{\rho}_D]$, where $\boldsymbol{\rho}_d, d = 1, 2, ..., D$ is a row vector containing N_S samples of LFs taken over d^{th} day. For achieving the smooth version of LF data by reducing the abruptness, we then apply a robust version of local regression technique using weighted linear least squares on each day [26], which is given as $\widehat{\boldsymbol{\rho}} = [max(\boldsymbol{\rho}_1, Sm(\boldsymbol{\rho}_1)); max(\boldsymbol{\rho}_2, Sm(\boldsymbol{\rho}_2)); ...; max(\boldsymbol{\rho}_D, Sm(\boldsymbol{\rho}_D))]$ = $[\widehat{\boldsymbol{\rho}}_1; \widehat{\boldsymbol{\rho}}_2; ...; \widehat{\boldsymbol{\rho}}_D]$. Here, $Sm(.)$ does the smoothing operation and $max(\boldsymbol{\rho}_d, Sm(\boldsymbol{\rho}_d))$ takes the sample-wise maximum between $\boldsymbol{\rho}_d$ and $Sm(\boldsymbol{\rho}_d)$, which increases the reliability that we are not underestimating the envelope. Overestimation can decrease the potential energy savings. However, it is a conservative approach for suppressing the potential false triggering of load distribution procedure at infeasible instances. Then, the moving average $\overline{\boldsymbol{\rho}}_{D+1}$, the standard deviation $\overline{\boldsymbol{\sigma}}_{D+1}$ and the estimated LF $\widetilde{\boldsymbol{\rho}}_{D+1}$ for $(D+1)^{th}$ day are given by [23], [27]

$$\overline{\boldsymbol{\rho}}_{D+1} = (1-\zeta)\overline{\boldsymbol{\rho}}_D + \zeta\widehat{\boldsymbol{\rho}}_D \tag{1}$$

$$\overline{\boldsymbol{\sigma}}_{D+1} = (1-\eta)\overline{\boldsymbol{\sigma}}_D + \eta|\overline{\boldsymbol{\rho}}_D - \widehat{\boldsymbol{\rho}}_D| \tag{2}$$

$$\widetilde{\boldsymbol{\rho}}_{D+1} = \overline{\boldsymbol{\rho}}_{D+1} + \varrho\overline{\boldsymbol{\sigma}}_{D+1} \tag{3}$$

where $0.2 \leq \zeta, \eta \leq 0.3$ are smoothing constants [27]. In this paper, $\zeta = \eta = 0.2$ are used. Here, ϱ is another constant usually taken equal to 3 [27]. At the end of each day, LF database used for estimation is updated by appending the actual LF data of the day and deleting that of the oldest day.

E. System Parameters

i. Power Consumption Profiles of BSs

Energy savings may largely depend on the power consumption profiles of BSs. A wide range of power consumption profiles of BSs having a single transceiver chain can be modeled as below

[21] - [22]

$$p_{op}(t) = \begin{cases} (1-\delta)\rho_b(t)P_{OP} + \delta P_{OP} & \text{(active mode)} \\ p_{op}^s & \text{(sleep mode)} \end{cases} \quad (4)$$

Here, $0 \leq \rho_b(t) \leq 1$ is the instantaneous actual LF of a BS. For LTE, we adopt the definition of LF as the ratio of the number of resource blocks (RBs) in use at time t to the total number [28]. Whereas, $p_{op}(t)$ is the instantaneous operating power of a BS, while $P_{OP} = aP_{Tx} + b$ is its maximum required at the full utilization of a BS. Here, P_{Tx} is the maximum transmit power, and a and b are constants. On the other hand, sleep mode power $p_{op}^s \geq 0$ may vary from vendor to vendor. Again, by varying the constant $0 \leq \delta \leq 1$, we can define three different models of BSs: (a) $\delta = 0$: fully energy proportional (FEP) model, (b) $\delta = 1$: constant energy consumption (CEC) model, and (c) $0 < \delta < 1$: non-energy proportional (NEP) model. The ultimate target of BS manufacturers is the FEP BS, which consumes power changing linearly with the LF, becoming zero at no traffic. Power consumption in NEP BSs also changes linearly with a fixed amount of consumption even at no traffic [6], [14], [21] - [23], [29]. However, most of the contemporary macro BSs are close to CEC type drawing nearly equal power irrespective of traffic level [13], [15] - [17], [19] - [20], [30] - [31].

ii. Traffic Generation Model

In real cellular networks, session generation process in BSs is time-inhomogeneous, where the traffic generation intensity varies with time. In this paper, we model a time-inhomogeneous process by multiplying a time-homogeneous process with a time-varying rate function. Thus, we can write

$$\lambda_{i,j}(t) = \alpha_{i,j} f_{i,j}(t - \theta_i)\lambda_j, \forall i = 1, 2, ..., N; \forall j = 1, 2, ..., Q \quad (5)$$

Here, Q is the number of classes of services, λ_j is the time-homogeneous session generation rate for class j, $f_{i,j}(t)$ is the rate function for class j in BS $\mathcal{B}_i$, and $\alpha_{i,j} \geq 0$ and $\theta_i \in [0, 24]$ hour are constant. By varying $\alpha_{i,j}$, $f_{i,j}(t)$ and θ_i, both temporal and spatial traffic variation among BSs can be captured.

iii. RB Allocation for UEs

Received signal-to-interference-plus-noise-ratio (SINR) at u^{th} UE located in BS $\mathcal{B}_i$ denoted

by $\gamma_{i,u}$ is given by

$$\gamma_{i,u} = \frac{P_{i,u}^{Rx}}{\mathcal{I}_{i,u}^{intra} + \mathcal{I}_{i,u}^{inter} + \mathcal{P}_N} \tag{6}$$

where $P_{i,u}^{Rx}$, $\mathcal{I}_{i,u}^{intra}$, $\mathcal{I}_{i,u}^{inter}$ and $\mathcal{P}_N$ are the received power, intra-cell interference, inter-cell interference and the additive white Gaussian noise power respectively. Considering adaptive modulation and coding (AMC), $\gamma_{i,u}$ can then be mapped to an achievable spectral efficiency given in bps/Hz [32]

$$\psi_{i,u} = \begin{cases} 0 & \textit{if } \gamma_{i,u} < \gamma_{min} \\ \xi \log_2(1 + \gamma_{i,u}) & \textit{if } \gamma_{min} \leq \gamma_{i,u} < \gamma_{max} \\ \psi_{max} & \textit{if } \gamma_{i,u} \geq \gamma_{max} \end{cases} \tag{7}$$

where $0 \leq \xi \leq 1$, γ_{min}, ψ_{max} and γ_{max} are the attenuation factor accounting the implementation loss, minimum SINR, maximum spectral efficiency and the SINR at which ψ_{max} is achieved. Then the number of required RBs for UE u of class j can be estimated by

$$\beta_{i,u}^{(j)} = \left\lceil \frac{R_{i,u}^{(j)}}{W_{RB}\psi_{i,u}} \right\rceil \tag{8}$$

where $R_{i,u}^{(j)}$ is the required data rate in bps, W_{RB} is the bandwidth per RB in Hz (e.g., 180 kHz in LTE), and $\lceil x \rceil$ is the nearest integer equal to or larger than x. On the other hand, if the number of RBs per UE is set fixed, (7)-(8) can be used for estimating the required SINR for any data rate.

iv. Interference Estimation

Use of orthogonal frequency bands in the sectors results in no intra-cell interference. On the other hand, because of dynamic switching of BSs, inter-cell interference can alter throughout the network, which is extremely challenging to keep track. Therefore, for the computational tractability, we combine the following strategies in dealing with the inter-cell interference. *Firstly*, we ignore the dynamic inter-cell interference and consider it as static Gaussian-like noise, which can be regarded as a worst case consideration [33]. Opportunity of adopting efficient fractional and soft frequency reuse, and interference randomization techniques in OFDMA-based system makes this assumption more feasible [7], [21] - [22], [34]. *Secondly*, under the above

approximations, we calculate the average interference a user can experience during the worst-case peak-traffic time in the original network. This is then used as an interference margin for all the users over the entire day. Due to the adoption of these conservative strategies, system performance evaluated in this paper can be considered as a lower bound. This type of static and Gaussian interference model has also been adopted in other energy efficiency and load balancing works [7], [20] - [22], [35].

v. Session Admission Control (SAC)

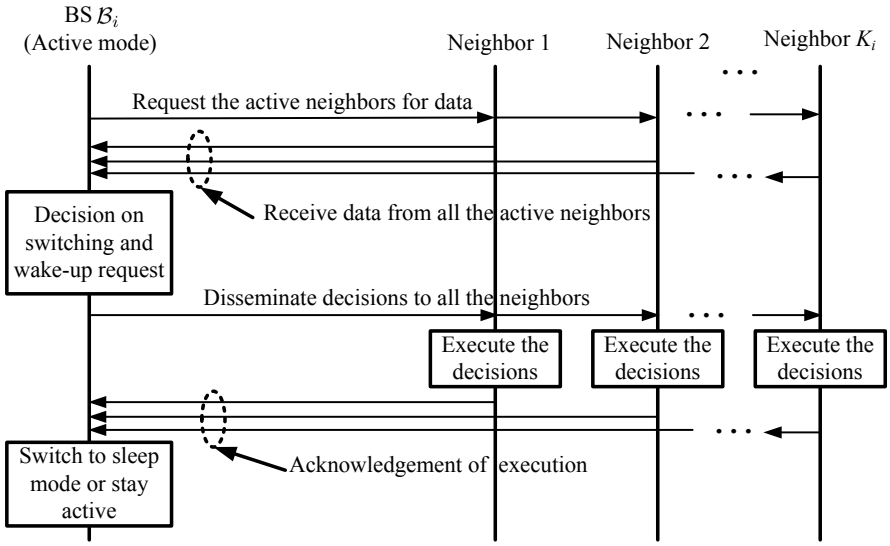
Since SAC is not our main focus, for the convenience of analysis, a simple first-come/first-served based mechanism is adopted in this paper. We assume that all UEs from the same class are allocated equal number of RBs, i.e., $\beta_{i,u}^{(j)} = \beta_j, \forall i, \forall u$. Thus, when a session request arrives, based on the requested data rate $R_{i,u}^{(j)}$, required RB(s) β_j and the location of UE, (6)-(8) along with a path loss model are used to estimate the required SINR $\gamma_{i,u}$, transmit power $P_{i,u}^{Tx}$, and the modulation and coding scheme (MCS). If β_j RBs and transmit power equal to $P_{i,u}^{Tx}$ are available in $\mathcal{B}_i$, the session is admitted. Otherwise, the session request is rejected. It is also assumed that the RBs remain dedicated for the entire session.

IV. Proposed Algorithm

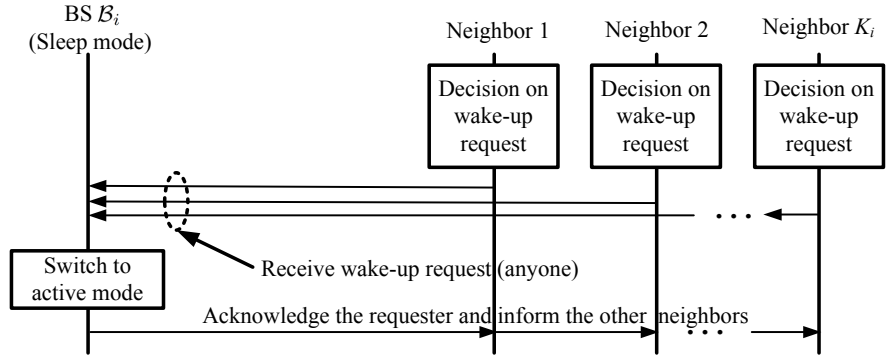
An algorithm, implemented in each BS, for energy saving load balancing by swapping traffic and switching BSs through distributed inter-BS cooperation is presented here. In this paper, we assume that the proposed cooperation is carried out periodically in every T_S time units, where T_S is an adjustable parameter. For avoiding any possibility of load distribution by two neighbors at the same time, BSs one after another distribute traffic. We assume that the sequence of BSs at which this distribution is carried out is preset by the operator. Without losing the generality, the sequence of BSs is taken as $\mathcal{B}_1$, $\mathcal{B}_2$, ..., $\mathcal{B}_N$[1].

For realizing the distributed cooperation, three LF thresholds normalized to unity are defined - lower threshold L_f, upper threshold H_f and acceptance threshold A_f. They are related as $A_f > H_f \geq L_f > 0$. Here, L_f and H_f are equally applicable for each BS and the total network

[1]This sequence can also be dynamically evaluated based on different strategies, such as, the instantaneous traffic levels, power profiles of BSs or even randomly. It is worth noting that our algorithm is independent of the underlying sequencing strategies and can work well with any of them.



(a) $\mathcal{B}_i$ is in active mode.



(b) $\mathcal{B}_i$ is in sleep mode.

Fig. 3: Coordination between BS $\mathcal{B}_i$ and its neighbors. K_i is the number of neighbors of BS $\mathcal{B}_i$.

as well, while A_f is only defined for BSs[2].

Now, let $\widetilde{\rho}(t)$ be the estimated LF for the aggregate traffic of the network at time t. Then, if $\widetilde{\rho}(t) < L_f$ or $\widetilde{\rho}(t) \geq H_f$, load distribution process is triggered by the network controller. Once the process is initiated by the controller, each BS in its turn uses the thresholds for distributing its traffic. Thus, at the beginning of $\mathcal{B}_i$'s turn, if it is in active mode, it checks its LF with the

[2]In this paper, we thoroughly analyze the system performance under various settings of these thresholds without specifying the methodology for selecting and updating them for a network, which is left for future works. Selecting the suitable values can be decided by the network operator, which mainly depends on the traffic generation characteristics, QoS requirements, cell deployment layout, etc. For instance, one strategy can be to use an offline optimization technique for estimating their values corresponding to the maximum energy savings for a given historical traffic database. Once the optimal values of L_f, H_f and A_f are found, each BS has to be instructed accordingly.

thresholds as $\rho_i(t) < L_f$ or $\rho_i(t) \geq H_f$ for starting its load distribution, where $\rho_i(t)$ is the actual LF of $\mathcal{B}_i$. Since the actual traffic condition in each BS is taken into account for distribution, thresholds are checked against the actual LF of each BS. On the other hand, BS $\mathcal{B}_i$ can accept traffic from other BSs as long as its new LF (i.e., own traffic plus shared traffic) $\rho_i^*(t) < A_f$ and session blocking is within target limit.

If either of the criteria is true for $\mathcal{B}_i$, it sends request to the active mode neighbors for the information on available RBs and the remaining transmit power. It is to be noted that exchange of information among BSs is supported in both 3GPP LTE [12] and WiMAX [36] systems. On the other hand, information on itself, namely, RBs and transmit power usage data, locations of own UEs, operating modes of its neighbors, list of BSs whose traffic is sharing $\mathbf{S}_i = \{S_1, S_2, ..., S_{z_i}\}$, and the candidate space $\mathbf{C}_i = \{\mathbf{C}_{i,1}, \mathbf{C}_{i,2}, ..., \mathbf{C}_{i,M_i}\}$ are required. Here, $\mathbf{C}_{i,n}(n = 1, 2, ..., M_i)$ is the n^{th} candidate combination. Feedback from UEs is assumed for receiving the location data of UEs. Using the gathered information, $\mathcal{B}_i$ decides its best neighbors for distributing its traffic as well as the necessary transmission range adjustments for itself and the neighbors for maintaining coverage. The formulated decision message is then propagated among its neighbors. The neighbors send back acknowledgements to $\mathcal{B}_i$ after their execution of the decisions followed by $\mathcal{B}_i$'s own execution. Thereafter, the algorithm proceeds to the next BS and so on. In sleep mode, BS $\mathcal{B}_i$ monitors for any wake-up request from its neighbors and if there is any, it switches to active mode. Fig. 3 presents a summary of the coordination between $\mathcal{B}_i$ and its neighbors during its both the operating modes. Use of the estimated LF as explained above and checking the status of BSs during their turns, reduces the computations compared to the many other schemes (e.g., [15] - [16], [20] - [22]), which start their algorithms assuming all BSs are in active mode.

At one instance, BS $\mathcal{B}_i$ can switch to sleep mode, while it can immediately be switched back to active mode by the next BS $\mathcal{B}_{i+1}$ or so on resulting in Ping-Pong effect. For reducing this switching, by utilizing the knowledge of operating modes of the neighbors, $\mathcal{B}_i$ divides its candidate space $\mathbf{C}_i$ into two mutually exclusive subsets: i) $\mathbf{C}_{i,a} = \{\mathbf{X}_{i,1}, \mathbf{X}_{i,2}, ..., \mathbf{X}_{i,q}\} \subseteq \mathbf{C}_i$, where each BS in $\mathbf{X}_{i,k} = \{x_1, x_2, ..., x_{P_k}\}, \forall i, \forall k$, is in active mode, and ii) $\mathbf{C}_{i,s} = \{\mathbf{Y}_{i,1}, \mathbf{Y}_{i,2}, ..., \mathbf{Y}_{i,r}\} \subseteq \mathbf{C}_i$, where at least one BS in $\mathbf{Y}_{i,m} = \{y_1, y_2, ..., y_{P_m}\}, \forall i, \forall m$, is in sleep mode. Here, q and r are the number of combinations in $\mathbf{C}_{i,a}$ and $\mathbf{C}_{i,s}$ respectively. BS $\mathcal{B}_i$ has to choose the best combination of neighboring BSs denoted by $\mathbf{B}_i^*$ from either of these two subsets for distributing its traffic.

In our algorithm, we impose higher priority on $\mathbf{C}_{i,a}$ over $\mathbf{C}_{i,s}$ for choosing $\mathbf{B}_i^*$, which ensures that the sleep mode BSs are not switched to active mode unless it is absolutely necessary for maintaining QoS. This in turn limits the Ping-Pong effect as well as the computational load. Proposed algorithm treats the low-traffic and the high-traffic periods in different ways, which are presented below in detail. Complete flow diagram of the algorithm is presented in Fig. 4.

A. Low-Traffic Period of $\mathcal{B}_i$ (i.e., $\rho_i(t) < L_f$)

In this case, $\mathcal{B}_i$ searches for the best combination $\mathbf{B}_i^*$ only in $\mathbf{C}_{i,a}$. This policy of not choosing $\mathbf{B}_i^*$ from $\mathbf{C}_{i,s}$ avoids the unnecessary switching of sleep mode BSs.

i) If $\mathcal{B}_i$ is active and $\mathbf{S}_i = \emptyset$, $\mathcal{B}_i$ looks for $\mathbf{B}_i^* \in \mathbf{C}_{i,a}$ (as explained in Section IV-C) for distributing its full traffic such that it can switch to sleep mode. If $\mathbf{C}_{i,a} = \emptyset$, then $\mathcal{B}_i$ stops searching.

ii) Else, if $\mathcal{B}_i$ is active and $z_i > 0$ (i.e., $\mathbf{S}_i \neq \emptyset$), $\mathcal{B}_i$ switches to sleep mode only when it can redistribute its traffic to all the z_i BSs in $\mathbf{S}_i$. Hence, $\mathcal{B}_i$ searches for $\mathbf{B}_i^*$ among those $\mathbf{X}_{i,k} \in \mathbf{C}_{i,a}, (0 \leq k \leq q)$ such that $\mathbf{S}_i \subseteq \mathbf{X}_{i,k}$.

If $\mathbf{B}_i^*$ is found, $\mathcal{B}_i$ distributes its full traffic to BSs in $\mathbf{B}_i^*$ and then, switches to sleep mode. Otherwise, it remains active.

B. High-Traffic Period of $\mathcal{B}_i$ (i.e., $\rho_i(t) \geq H_f$)

In this case, four alternatives are attempted one-after-another by BS $\mathcal{B}_i$ for distributing its traffic. This priority order allows the sleep mode BSs to stay in sleep mode for longer durations. If none of these four alternatives are met, $\mathcal{B}_i$ continues to operate. These alternatives are presented as below:

i) Total traffic distribution among BSs in $\mathbf{B}_i^* \in \mathbf{C}_{i,a}$: The procedure of traffic distribution in this option is same as Section IV-A except the condition that this procedure starts if $\rho_i(t) \geq H_f$. If this alternative is met, $\mathcal{B}_i$ switches to sleep mode after distributing its traffic to the BSs in $\mathbf{B}_i^*$.

ii) Partial traffic distribution among BSs in $\mathbf{B}_i^* \in \mathbf{C}_{i,a}$: The procedure is similar to Section IV-A with two exceptions. Along with the condition of $\rho_i(t) \geq H_f$, the other exception is that the best combination has to be capable to share the excess traffic equal to $(\rho_i(t) - H_f\ )$. If $\mathbf{B}_i^*$ is found, the excess traffic is distributed to BSs in $\mathbf{B}_i^*$, while $\mathcal{B}_i$ stays in active mode with its new lower traffic. This partial load distribution results in better load balancing.

iii) Total traffic distribution among BSs in $\mathbf{B}_i^* \in \mathbf{C}_{i,s}$: In this option, $\mathcal{B}_i$ looks for the best combination $\mathbf{B}_i^*$ in $\mathbf{C}_{i,s}$. Except this change and the condition of $\rho_i(t) \geq H_f$, the procedure is same as Section IV-A. However, the result is quite different. If $\mathbf{B}_i^*$ is found, $\mathcal{B}_i$ switches to sleep mode, while sleeping BSs in $\mathbf{B}_i^*$ switch to active mode for supporting the traffic of $\mathcal{B}_i$.

iv) Partial traffic distribution among BSs in $\mathbf{B}_i^* \in \mathbf{C}_{i,s}$: This step is same as IV-B(ii) except that the best combination $\mathbf{B}_i^*$ is determined from $\mathbf{C}_{i,s}$. Thus, $\mathcal{B}_i$ continues to operate with the new lower traffic and at the same time, other sleep mode BSs in $\mathbf{B}_i^*$ wake-up to share its excess traffic.

Necessary adjustments in the transmission range of all BSs in $\mathbf{B}_i^*$ are made to provide full coverage for $\mathcal{B}_i$. For this purpose, we propose to increase the transmit power in the required sectors of BSs as demonstrated in our previous work [9]. Advanced beam-forming techniques [14], [37], or any suitable technique can also be adopted in the system.

C. Selecting the Best Combination for Traffic Distribution

Here, we explain the procedure of selecting the best combination $\mathbf{B}_i^*$ for low traffic period. For selecting $\mathbf{B}_i^*$, expected LF (ELF) and the required additional transmit power for all BSs in all the candidate combinations $\mathbf{X}_{i,k} \in \mathbf{C}_{i,a} (\forall k = 1, 2, ..., q)$ are first calculated. To calculate the ELFs of P_k BSs in $\mathbf{X}_{i,k}, \forall k$, we first evaluate the number of available RBs in these P_k BSs. Then, based on the resource allocation policy, number of RBs required in each of the P_k BSs for supporting all the users of $\mathcal{B}_i$ and their own users is calculated, which gives the ELFs. In calculating the ELFs, it is assumed that if $\mathcal{B}_i$ could switch to sleep mode, a UE in $\mathcal{B}_i$ would hand off to the nearest BS in $\mathbf{X}_{i,k}$. Similarly, required additional transmit power for supporting the UEs of $\mathcal{B}_i$ is calculated.

Then, the *minimax* value among these ELFs is determined, which is equal to the minimum of the maximum values taken from each combination in $\mathbf{C}_{i,a}$. The combination corresponding to this *minimax* value is the probable $\mathbf{B}_i^*$ and let us denote it as $\widetilde{\mathbf{B}}_i^*$. Next, the expected session blocking probability in all BSs in $\widetilde{\mathbf{B}}_i^*$ is evaluated. Now, if *minimax* $< A_f$, session blocking is within the target and all BSs of $\widetilde{\mathbf{B}}_i^*$ can provide the respective required additional transmit power, then $\mathbf{B}_i^* = \widetilde{\mathbf{B}}_i^*$ for $\mathcal{B}_i$. Use of *minimax* value ensures traffic distribution to the least loaded BS combination and helps to reduce session blocking. Similar concept is used to find $\mathbf{B}_i^*$ for

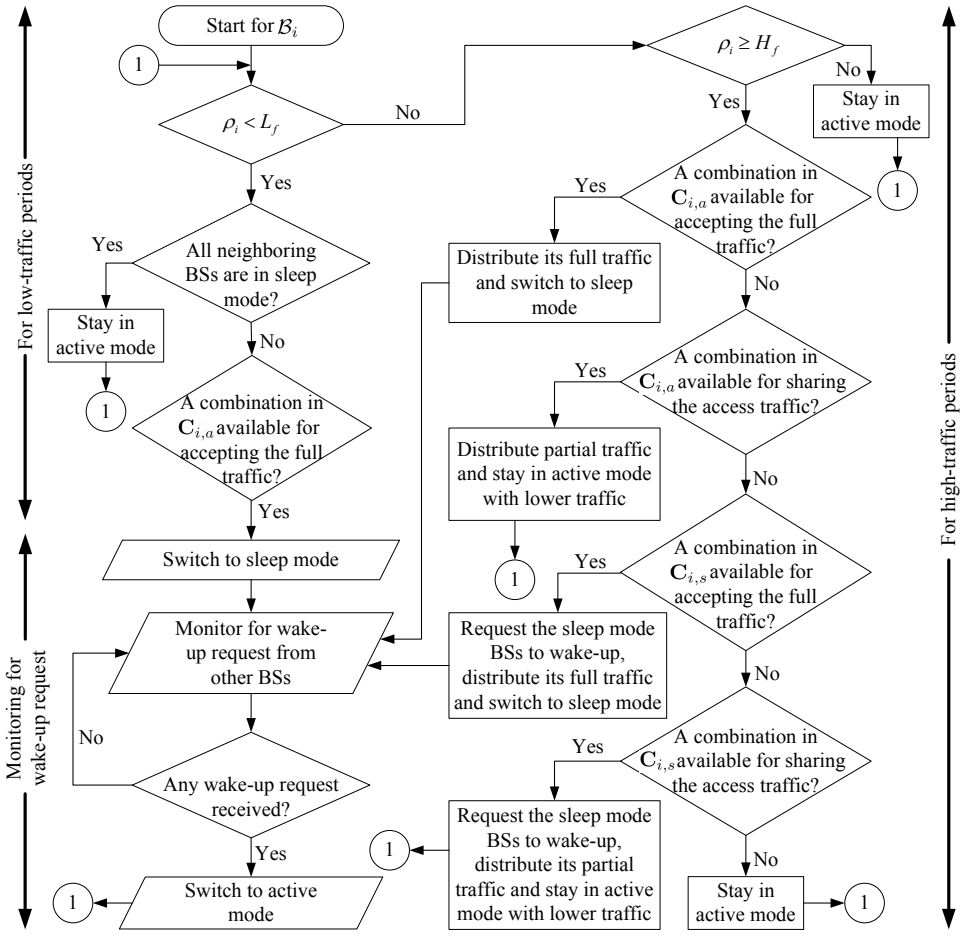


Fig. 4: Flow diagram of the algorithm.

the high traffic period, which is not presented here for the sake of brevity.

V. Analytical Model Formulation

Here, we present an analytical model for evaluating the probabilities of BSs to switch into sleep mode. Changing operating modes of BSs depends on its and neighbors' current traffic, and the operating modes of BSs at the previous instance, which can be modeled as a Markov process. A state of the network at time t is one of the possible combinations of the operating modes of all BSs. Thus the total number of states is 2^N growing exponentially with N, which makes it challenging for solving using Markov chains. To reduce the complexity, following the proposed algorithm, a heuristically guided formulation is presented here. For the convenience of

presentation, we have omitted the time index from some of the following equations and brought back later.

Let S_{P_i}, $\mathbf{N}_i = \{\mathcal{B}_{i,1}, \mathcal{B}_{i,2}, ..., \mathcal{B}_{i,K_i}\} \subset \mathbf{B}$ and $\mathbf{C}_{i,n} = \{C_{i,n}^{(1)}, C_{i,n}^{(2)}, ..., C_{i,n}^{(P_n)}\} \in \mathbf{C}_i$ be the selection scheme, the set of neighbors and the n^{th} candidate combination for $\mathcal{B}_i$ respectively. Here, $C_{i,n}^{(k)} \in \mathbf{N}_i, \forall i, \forall n, \forall k$, and $P_n \leq P_i, \forall i, \forall n$.

Let us define two different events, $E_{i,n}$ = {$\mathcal{B}_i$ distributes traffic to $\mathbf{C}_{i,n}$ and switch to sleep mode}, $\forall i, \forall n$; and $A_{i,n}^{(k)} = \{C_{i,n}^{(k)}$ is in active mode$\}$, $\forall i, \forall n, \forall k$.

Assuming that $\mathcal{B}_i$ is not sharing the traffic of any of its neighbors, we can write the probability of occurring event $E_{i,n}(n = 1, , M_i)$ at time t as below

$$P\{E_{i,n}\} = \left[P\left\{\rho_i < L_f\right\} \prod_{k=1}^{P_n} P^*\left\{A_{i,n}^{(k)}\right\} + P(\rho_i \geq H_f)\right] \times \prod_{k=1}^{P_n} \left[P\left\{A_f - \rho_{i,n}^{(k)} > \phi_{i,n}^{(k)} \rho_i\right\} P\left\{P_{i,n}^{(k)} \leq P_b^{th}\right\}\right] \tag{9}$$

where $\rho_{i,n}^{(k)}$ and $P_{i,n}^{(k)}$ are the actual LF and session blocking in $C_{i,n}^{(k)}$ respectively; P_b^{th} is the target session blocking; $\phi_{i,n}^{(k)}$ is the fraction of ρ_i to be distributed to $C_{i,n}^{(k)}$ and thus, $\sum_{k=1}^{P_n} \phi_{i,n}^{(k)} = 1$; and $P^*\left\{A_{i,n}^{(k)}\right\}$ is the probability that $C_{i,n}^{(k)}$ was active at the last instance, while $P^*\left\{A_{i,n}^{(k)}\right\} = 1, \forall t \leq 0, \forall i, \forall n, \forall k$. For no candidate combination of $\mathcal{B}_i$, $P\left\{E_{i,n}\right\} = 0$. Derivation of the probabilities in (9) is presented in Appendix.

However, if $\mathcal{B}_i$ has been supporting the traffic of any of its neighboring BSs other than those in $\mathbf{C}_{i,n}$, according to the algorithm, $\mathcal{B}_i$ is not allowed to switch into sleep mode by distributing its current traffic to this n^{th} combination. For accounting this, we calculate the probability that $\mathcal{B}_i$ is not the acceptor for any of the remaining $(K_i - P_n)$ BSs as below

$$P\left\{X_{i,n}\right\} = \prod_{\substack{k=1 \\ \mathcal{B}_{i,k} \notin \mathbf{C}_{i,n}}}^{K_i} \left(1 - F_{i,k} P^*\left\{s_{i,k}\right\}\right) \tag{10}$$

where $P^*\left\{s_{i,k}\right\}$ is the probability that $\mathcal{B}_{i,k}$ was in sleep mode at the last instance, and

$$F_{i,k} = \begin{cases} \frac{G_{i,k}}{M_{i,k}}, & \text{if } M_{i,k} \neq 0 \\ 0, & \text{if } M_{i,k} = 0 \end{cases} \tag{11}$$

where $M_{i,k}$ is the number of candidate combinations for $\mathcal{B}_{i,k}$, and $G_{i,k} \leq M_{i,k}$ is the number of

these $M_{i,k}$ combinations containing $\mathcal{B}_{i,k}$. Modified $P_i\{E_{i,n}\}$ can then be given as

$$\widehat{P}\{E_{i,n}\} = P\{E_{i,n}\} \times P\{X_{i,n}\} \tag{12}$$

Since the events of distributing traffic of $\mathcal{B}_i$ to one of the M_i candidate combinations are independent, probability of switching of $\mathcal{B}_i$ to sleep mode at time t can be written as

$$P_i(t) = \widehat{P}\left\{\bigcup_{n=1}^{M_i} E_{i,n}\right\} = \sum_{d=1}^{M_i}(-1)^{d+1} \sum_{\substack{n_1,n_2,\ldots,n_d: \\ 1\le n_1\le n_2\ldots\le n_d\le M_i}} \widehat{P}\{E_{i,n_1}\}\,\widehat{P}\{E_{i,n_2}\}\ldots\widehat{P}\{E_{i,n_d}\} \tag{13}$$

Average sleeping probability P_S per BS over any duration T can then be written as

$$P_S = \left[\frac{1}{NT}\sum_{i=1}^{N}\int_{t_0}^{t_0+T} P_i(t)dt\right] \times 100\% \tag{14}$$

VI. Results and Discussions

A. Simulation Setup

We evaluate the proposed scheme through extensive simulations. Although the scheme is applicable for any cell layout, for the convenience of comparison with other works and creating a benchmark, hexagonal layout is chosen for simulations. We consider a cellular access network serving a geographical area covered by three-sector 50 macro cells having an inter-site distance equal to $\sqrt{3}\times$ 500 m and uniformly distributed UEs. Carrier frequency = 2 GHz, channel bandwidth = 5 MHz per sector (i.e., 25 RBs) and BS transmit power = 43 dBm per sector are assumed. AMC code set parameters $\{\gamma_{min}$ = -6.5 dB, γ_{max} = 19 dB, ψ_{max} = 4.8 bps/Hz, ξ = 0.75$\}$, noise figure = 9 dB (5 dB) for UE (BS), BS antenna gain including feeder loss = 15 dBi, shadow fading standard deviation = 8 dB, penetration loss = 10 dB and Gaussian noise power density = -174 dBm/Hz are chosen in reference to the 3GPP LTE suggestions [32]. WINNER+ non-line-of-sight urban macro-cell channel model with BS antenna height = 25 m and UE antenna height = 1.5 m is used [38].

We consider three classes of real-time services requiring constant bit rate (CBR) equal to 64 kbps, 384 kbps and 512 kbps, which include packet headers and payloads. New sessions arrive following a Poisson process. We assume that only one RB can be allocated to a UE from any class. Thus, using (7)-(8), we can calculate the required SINR for the three classes equal to -4.1

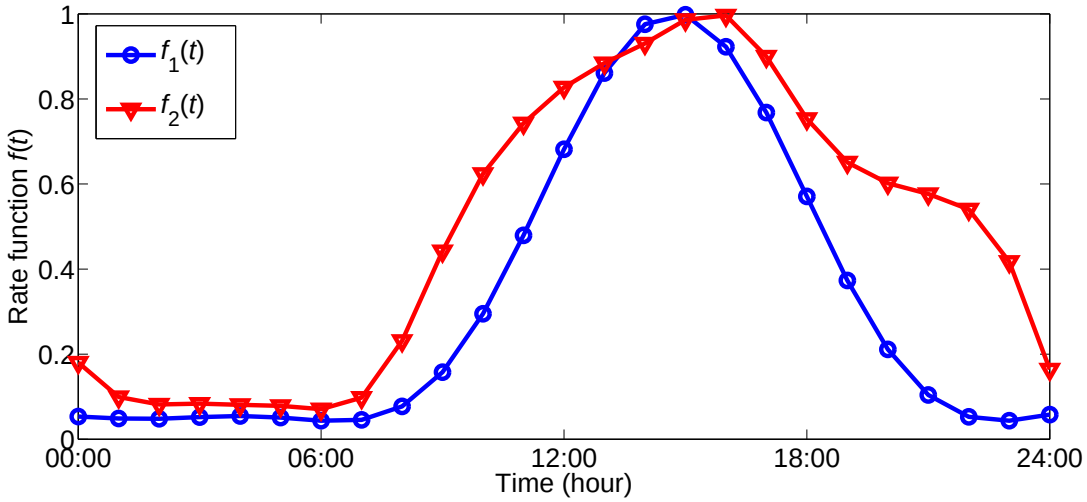


Fig. 5: Rate functions.

dB, 7.9 dB and 11.1 dB respectively. LTE Frequency division duplexing (FDD) frame structure is considered with the assumption that one transmission time interval (TTI) of 1 ms carries exactly one packet. For the convenience of simulations, we assume that the data volume per session of all UEs from the same class is equal. Also, for fair comparison, constant session duration equal to 3 minutes is considered for all the classes. For generating the time-inhomogeneous traffic, normalized rate functions shown in Fig. 5 are used [17], [23], [39]. Average session arrival rate λ_j is chosen such that the peak-time session blocking in the original network for $\alpha_{i,j} = 1, \forall i, \forall j$, becomes equal to 1%. Target session blocking in the proposed network is also set to 1%. Unless otherwise specified, we set $\alpha_{i,j} = \alpha = 1, \forall i, \forall j$, and $\theta_i = 0, \forall i$, for the simulations. On the other hand, two settings for BS power profile parameter are considered. They are: Set1: a = 21.45, b = 354.44 [30], and Set2: a = 7.8, b = 605 [31]. We assume optical backhaul link among BSs having energy requirement in the order of 1 pJ/bit/m [40]. Thus, energy cost for signaling among BSs is assumed insignificant. Below, we only present the results for weekdays. Evaluation for the weekends can be done in the same way.

B. EWMA Estimation and the Number of Communications

Performance of our estimation technique is evaluated by using the historical LF data of the last one month generated for a network having only 64 kbps UEs and demonstrated in Fig. 6.

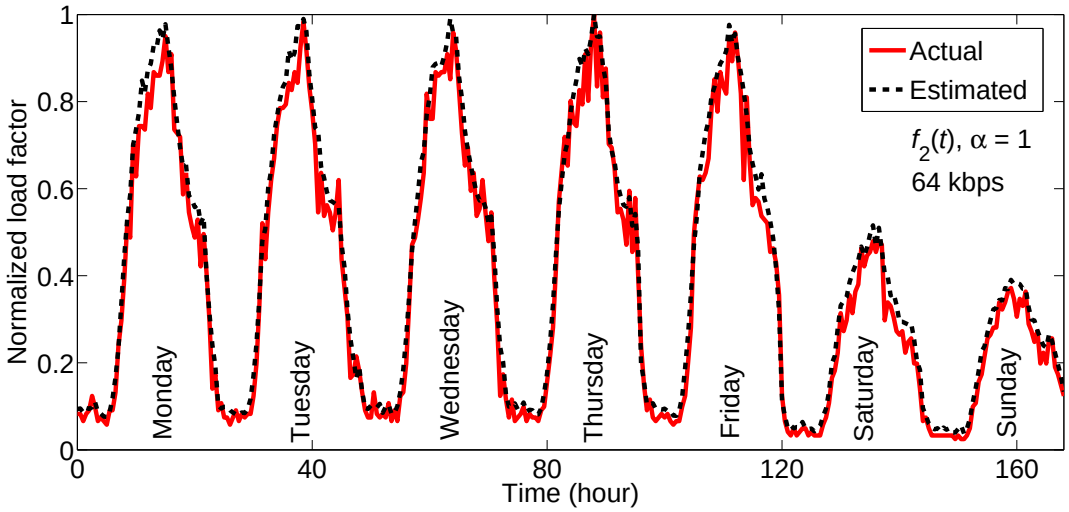


Fig. 6: Performance of EWMA estimator.

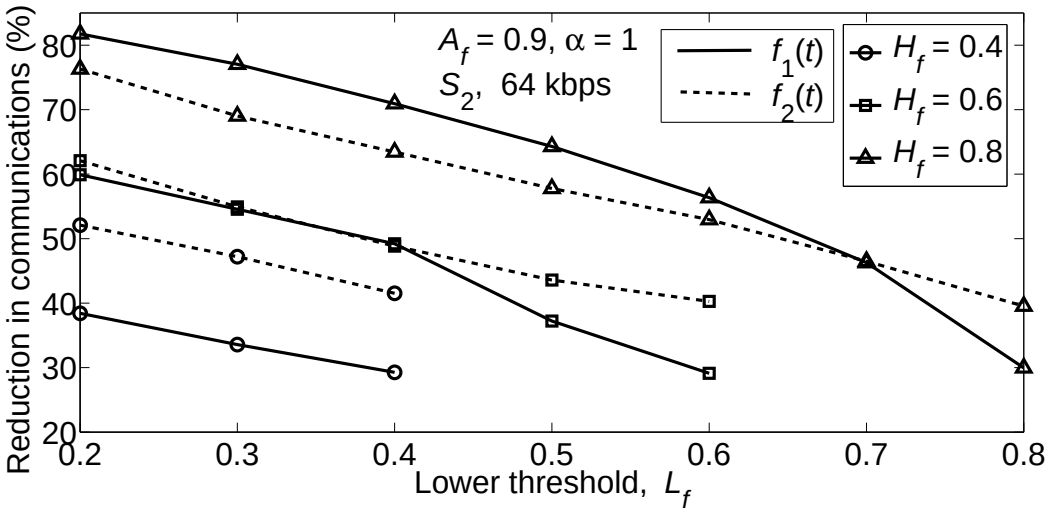


Fig. 7: Reduction in the number of communications among BSs.

Weekday and weekend data are separated for estimating the envelope for the corresponding days. From the figure, it is clear that the estimated values are very close and almost always higher than the actual data.

Percentage of reduction in the number of instances at which a BS communicates with its neighbors is presented in Fig. 7. Sampling interval 30 minutes is taken. This reduction of communications results from the combined effect of the use of estimated LF, allowing only

the active mode BSs for deciding on traffic distributions and considering the operating mode of the neighbors for distributing traffic. As seen, depending on the network parameters, as high as 80% reduction is achieved. Also, the higher the gap between L_f and H_f, the higher is the reduction. Reduction corresponding to $L_f = H_f$ solely results from the checking of the operating status of each BS amounting around 30% and 40% for $f_1(t)$ and $f_2(t)$ respectively.

C. Percentage of Sleep Mode BSs and RB Utilization

Figs. 8 and 9 present the average percentage of sleep mode BSs per day P_S (i.e., probability of sleeping of BSs) with the LF thresholds L_f and A_f respectively. Simulation parameters are also shown in the figures. As seen, P_S has an increasing trend with the increase of both L_f and A_f. This is because, for a higher value of L_f, higher number of BSs has the probability to have LF less than L_f. Consequently, higher number of BSs has the chance to distribute their traffic to other BSs and switching to sleep mode. On the other hand, higher value of A_f implies that BSs are capable to accept more traffic from the neighbors. Therefore, higher number of BSs can switch to sleep mode leading to higher P_S.

Impact of the rate functions is also evident from both the figures. Rate function $f_1(t)$ corresponds to lower traffic generation than that with $f_2(t)$. Therefore, in a network with $f_1(t)$, higher percentage of BSs can switch to sleep mode than in a network with $f_2(t)$. In addition, impact of traffic parameter α is also included in Fig. 8. Values of $\alpha < 1$ refer to a lower level of loading of BSs than their available capacities. Therefore, higher number of BSs switches to sleep mode by distributing their traffic to the neighboring BSs. Moreover, for all the cases, analytical results are reasonably close to the simulation results, which validate our simulation model.

Impact of the candidate selection schemes for distributing traffic is illustrated in Fig. 10. It is evident that as we move from S_1 to S_6, higher percentage of BSs can switch into sleep mode. For example, at $A_f = 1$, on an average 41.8% BSs switch into sleep mode under S_1 scheme, while the figure is much higher in S_6 scheme amounting to 53.2%. For S_1, only 1-BS combinations from the neighbors are selected for distributing traffic. On the other hand, all the possible combinations of neighbors are considered in S_6 scheme causing a higher P_S. Fig. 10 can also be explained as a trade-off between the number of computations and P_S. For example, due to the inclusion of all possible combinations, S_6 scheme involves the highest number of

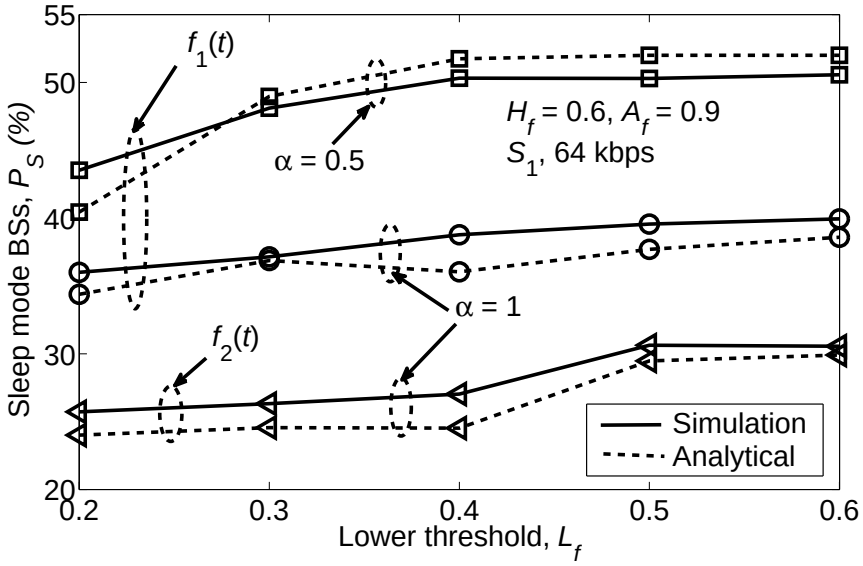


Fig. 8: Sleep mode BSs per day with L_f.

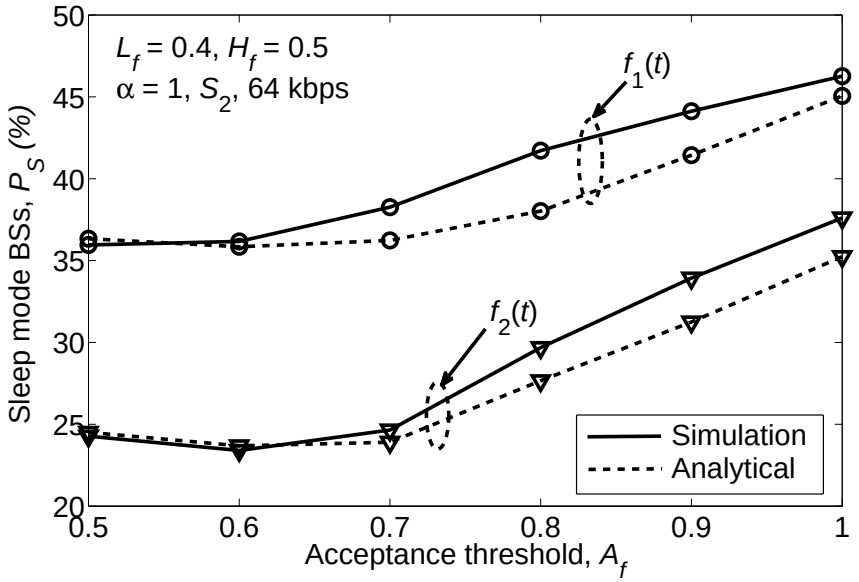


Fig. 9: Sleep mode BSs per day with A_f.

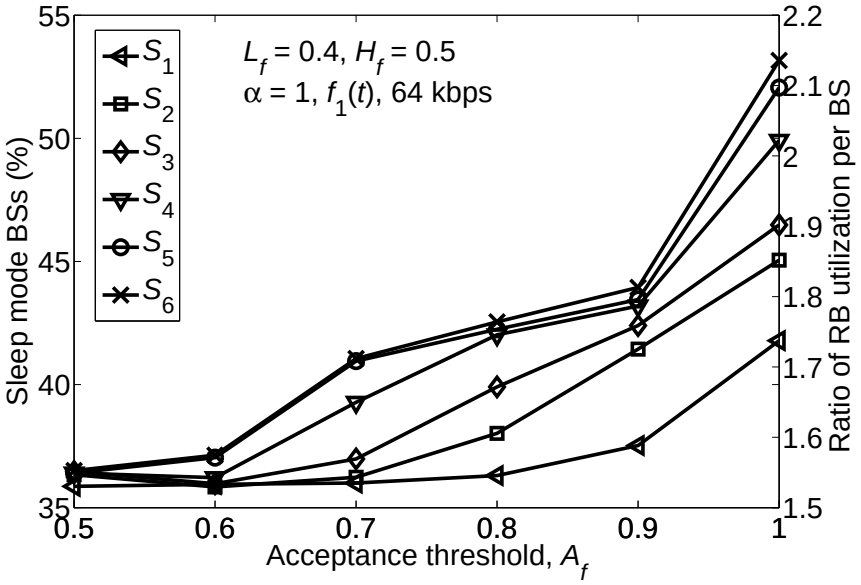


Fig. 10: Sleep mode BSs and RB utilization per day with the selection schemes.

computations. At the same time, under S_6, highest number of BSs can stay in sleep mode.

Ratio of the utilized RBs per BS in the proposed network to that of the original network also follows the same trends of P_S and its scale is shown on the right hand side of Fig. 10. In the proposed network, fewer BSs serve the same number of UEs in the original network, and hence, RB utilization per BS has increased by a factor over two as seen in the figure.

D. Energy Savings, Data Rate and BS Power Profile

Average daily energy savings of a network corresponding to $f_1(t)$ and S_6 scheme with various data rates is presented in Fig. 11. Similar to P_S, energy savings per day is also increasing with A_f. Furthermore, higher energy savings is found for lower data rate scenarios. Since higher data rate requires higher SINR, i.e., higher transmit power, additional power requirement increases with the increase of UE data rate resulting in reduced energy savings. A case where 60% UEs from 384 kbps and the other 40% requiring 512 kbps denoted as 'Mixed' is also considered. Energy savings line for the 'Mixed' case lies between the ones of 384 kbps and 512 kbps.

Fig. 12 demonstrates the dependency of energy savings on the BS power profile parameters. Since $\delta = 1$ corresponds to CEC BSs requiring constant power irrespective of traffic level, it results in the highest amount of energy wastage. Consequently, our proposed system can achieve

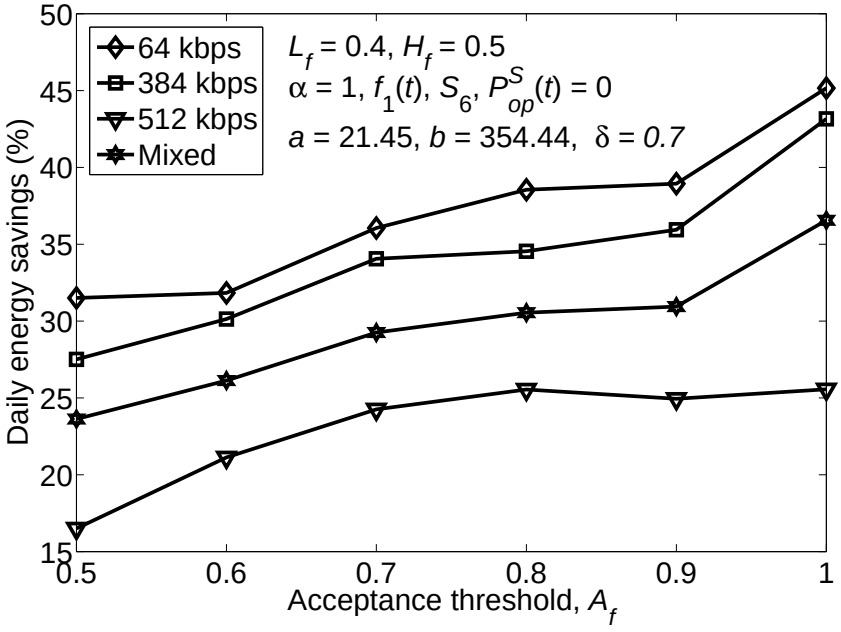


Fig. 11: Daily energy savings for various data rates.

the highest energy savings as illustrated. Then, as δ moves toward zero, BS hardware tends to be more and more energy proportional and hence, energy wastage as well as savings decreases. Finally, when $\delta = 0$, BSs are of FEP type consuming power proportional to LF and hence, no additional energy savings is possible from switching BSs. Instead, the system may consume extra power because of additional transmit power. In addition, the impact of sleep mode power p_{op}^s, and the parameters a and b is also evident from the figure. As we can see, Set1 and Set2 achieve approximately equal savings for the case of $p_{op}^s = 0$, while the savings significantly differs for $p_{op}^s(t) = \delta b$. Since, in Set2, b = 605 is much higher than that of Set1 (b = 354.44), BSs of Set1 consumes much less power in sleep mode leading to higher savings than that of Set2.

E. Comparison with the Related Works

A detail qualitative comparison of the proposed system with the other works has been provided in Section II (Related Works). Here, a quantitative comparison is presented in Fig. 13. From our thorough survey of the publications on BS on/off based energy efficient cellular networks, we broadly divide them into three switching categories: static type, network wide dynamic type and

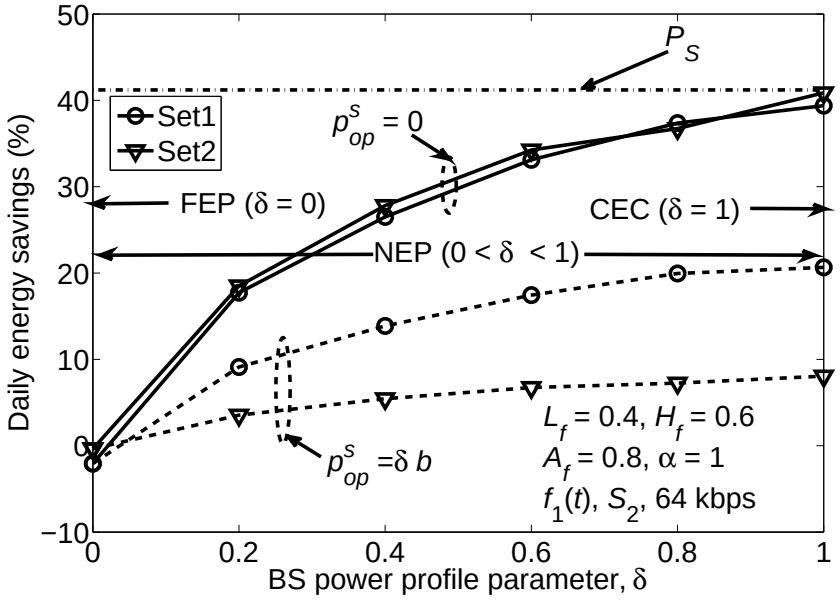


Fig. 12: Impact of BS power profile on daily savings.

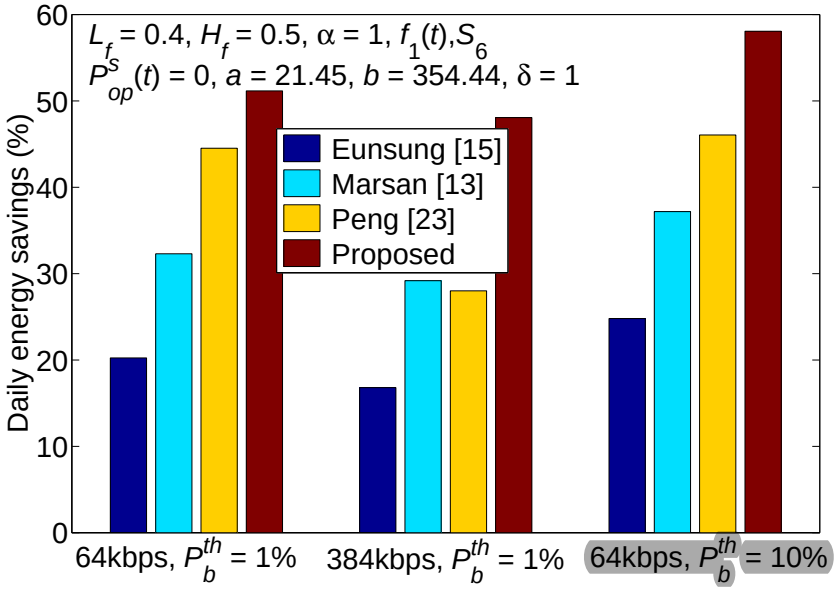


Fig. 13: Performance comparison with the other proposals.

sub-region based dynamic type. Following three works, one from each group, are then chosen for the comparison: a deterministic predefined switching pattern based static type scheme by Marsan *et al.* [13], a network wide dynamic switching based scheme by Eunsung *et al.* [15]

and a grid-based dynamic scheme by Peng *et al.* [23]. We have implemented all these schemes for the hexagonal deployment scenario. For [13], 3/4 hexagonal scheme is simulated, while a switching threshold equal to 0.5 is set for simulating the scheme in [15]. Also, following the grid-forming criteria in [23], we divide the network into grids and perform the traffic profiling. Energy savings of our system presented in the figure corresponds to scheme S_6. Comparison is presented for data rates $\in \{64, 384\}$ kbps and target session blocking $P_b^{th} \in \{1\%, 10\%\}$. As seen in the figure, our system can achieve significantly higher savings than all these schemes, which is more evident in the higher data rate scenarios.

VII. Conclusion

An ecological self-organization inspired distributed inter-BS cooperation assisted energy efficient load balancing framework for OFDMA-based cellular access networks has been presented in this paper. BSs in the proposed system mutually cooperate for dynamically switching redundant BSs into sleep mode for saving energy. Various schemes for selecting the best BSs for distributing traffic are also exploited. Use of the estimated LF derived by our proposed EWMA-based estimator, accounting the operating modes of BSs for initialization of algorithm and the prioritization of the active neighbors over sleeping BSs are also explored for reducing the number computations. Moreover, an analytical model for evaluating the probabilities of BSs switching into sleep mode is also presented. System performance over a wide range of BS selection schemes, BS power models, switching thresholds, traffic scenarios and data rates has been investigated. Higher savings is identified for the networks with lower data rate users. Furthermore, energy savings increases with the increase of non-proportionality in power consumption of BSs. Effectiveness of the framework is also exhibited by realizing as much as 30% higher savings than the compared works. In addition, depending on the network settings, an improvement in RB utilization is observed by a factor over two and communications among BSs is reduced up to 80%.

We will next focus on developing the techniques for evaluating the optimal thresholds as well as the potential faster algorithms for highly dense networks. Analysis of energy savings-delay-throughput trade-off under multi-tier networks using generalized stochastic geometry-based model is also under our consideration.

APPENDIX

Expressions for the probabilities in (9) are derived here for an especial case of having constant session durations for all UEs. Number of active users of class j in $\mathcal{B}_i$ at time t can then be written as, $u_{i,j}(t) = \alpha_{i,j} f_{i,j}(t) \omega_j h_j, \forall i, \forall j$, where h_j is the session duration and ω_j is a Poisson random variable (RV) with rate parameter λ_j. Using the approximation that a Poisson RV with large λ is equivalent to a Normal RV with both mean and variance equal to λ, we can write

$$P\left\{\rho_i(t) < L_f\right\} = P\left\{\frac{1}{\beta_{i,T}} \sum_{j=1}^{Q} \sum_{u=1}^{u_{i,j}(t)} \beta_j < L_f\right\} = P\left\{\Theta_i < \frac{\beta_{i,T} L_f - \mu_{X_i}}{\sqrt{\sigma_{X_i}^2}}\right\}$$
$$= 1 - Q\left(\frac{\beta_{i,T} L_f - \mu_{X_i}}{\sqrt{\sigma_{X_i}^2}}\right) \tag{A.1}$$

where $\beta_{i,T}$ is the total RBs in BS $\mathcal{B}_i$ and $\Theta_i \sim \mathcal{N}(0,1)$. Here, μ_{X_i} and $\sigma_{X_i}^2$ are given by

$$\mu_{X_i} = \sum_{j=1}^{Q} \sum_{u=1}^{\alpha_{i,j} f_{i,j}(t) \lambda_j h_j} \beta_j \tag{A.2}$$

$$\sigma_{X_i}^2 = \sum_{j=1}^{Q} \sum_{u=1}^{(\alpha_{i,j} f_{i,j}(t) h_j)^2 \lambda_j} \beta_j^2 \tag{A.3}$$

Similarly, we can derive

$$P\left\{\rho_i(t) \geq H_f\right\} = Q\left(\frac{\beta_{i,T} H_f - \mu_{X_i}}{\sqrt{\sigma_{X_i}^2}}\right) \tag{A.4}$$

$$P\left\{A_f - \rho_{i,n}^{(k)}(t) > \phi_{i,n}^{(k)} \rho_i(t)\right\} = 1 - Q\left(\frac{A_f - \mu_{Y_i}}{\sqrt{\sigma_{Y_i}^2}}\right) \tag{A.5}$$

where

$$\rho_{i,n}^{(k)}(t) = \frac{1}{\beta_{i,T}^{n,k}} \sum_{j=1}^{Q} \sum_{u=1}^{\alpha_{i,j}^{n,k} f_{i,j}^{n,k}(t) \omega_j h_j} \beta_j \tag{A.6}$$

$$\mu_{Y_i} = \frac{1}{\beta_{i,T}^{n,k}} \sum_{j=1}^{Q} \sum_{u=1}^{\alpha_{i,j}^{n,k} f_{i,j}^{n,k}(t) \lambda_j h_j} \beta_j + \frac{\phi_{i,n}^{(k)}}{\beta_{i,T}} \sum_{j=1}^{Q} \sum_{u=1}^{\alpha_{i,j} f_{i,j}(t) \lambda_j h_j} \beta_j \tag{A.7}$$

$$\sigma_{Y_i}^2 = \frac{1}{\left(\beta_{i,T}^{n,k}\right)^2} \sum_{j=1}^{Q} \sum_{u=1}^{\left(\alpha_{i,j}^{n,k} f_{i,j}^{n,k}(t) h_j\right)^2 \lambda_j} \beta_j^2 + \left(\frac{\phi_{i,n}^{(k)}}{\beta_{i,T}}\right)^2 \sum_{j=1}^{Q} \sum_{u=1}^{(\alpha_{i,j} f_{i,j}(t) h_j)^2 \lambda_j} \beta_j^2 \tag{A.8}$$

where $\alpha_{i,j}^{n,k}$, $f_{i,j}^{n,k}$ and $\beta_{i,T}^{n,k}$ are the parameter α, rate function and the total number of RBs of $C_{i,n}^{(k)}$ respectively.

REFERENCES

[1] A. Fehske, G. Fettweis, J. Malmodin and G. Biczok, "The Global Footprint of Mobile Communications: The Ecological and Economic Perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55-62, Aug 2011.

[2] R. Bolla, R. Bruschi, F. Davoli and F. Cucchietti, "Energy Efficiency in the Future Internet: a Survey of Existing Approaches and Trends in Energy-Aware Fixed Network Infrastructure," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 2, pp. 223-244, 2011.

[3] Z. Hasan, H. Boostanimehr and V. K. Bhargava, "Green Cellular Networks: A Survey, Some Research Issues and Challenges," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 4, pp. 524-540, 2011.

[4] U. Paul, A. P. Subramanian, M. M. Buddhikot and S. R. Das, "Understanding Traffic Dynamics in Cellular Data Networks," *Proc. IEEE INFOCOM*, pp. 882-890, China, Apr 2011.

[5] M. Z. Shafiq, L. Ji, A. X. Liu and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," *Proc. ACM SIGMETRICS*, pp. 305-316, Jun 2011.

[6] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, A. Fehske, "How Much Energy is Needed to Run a Wireless Network?," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40-49, Oct 2011.

[7] K. Son, S. Chong and G. Veciana, "Dynamic Association for Load Balancing and Interference Avoidance in Multi-Cell Networks," *IEEE Transactions on Wireless Comm.*, vol. 8, no. 7, pp. 3566-3576, July 2009.

[8] W. Song, W. Zhuang and Y. Cheng, "Load Balancing for Cellular/WLAN Integrated Networks," *IEEE Net.*, vol. 21, no. 1, pp. 27-33, Jan-Feb 2007.

[9] M. F. Hossain, K. S. Munasinghe and A. Jamalipour, "On the Energy Efficiency of Self-Organizing LTE Cellular Access Networks," *Proc. IEEE GLOBECOM*, pp. 5536-5541, Anaheim, CA, USA, Dec 2012.

[10] T. Edler and S. Lundberg, "Energy Efficiency Enhancements in Radio Access Networks," *Ericsson Review*, 2004.

[11] Huawei White Paper, "Improving Energy Efficiency, Lower CO_2 Emission and TCO," pp. 1-16, 2011.

[12] 3GPP TR 36.902 ver. 9.3.1 Rel. 9, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-Configuring and Self-Optimizing Network (SON): Use Cases and Solutions," *Technical Report*, 2011.

[13] M. A. Marsan, L. Chiaraviglio, D. Ciullo and M. Meo, "Optimal Energy Savings in Cellular Access Networks," *Proc. IEEE ICC*, Dresden, Germany, pp. 1-5, June 2009.

[14] F. Han, Z. Safar, W. S. Lin, Y. Chen and K. J. R. Liu, "Energy-Efficient Cellular Network Operation via Base Station Cooperation," *Proc. IEEE ICC*, pp. 5885-5889, Ottawa, Canada, June 2012.

[15] E. Oh and B. Krishnamachari, "Energy Savings Through Dynamic Base Station Switching in Cellular Wireless Access Networks," *Proc. IEEE GLOBECOM*, pp. 1-5, Dec. 2010.

[16] S. Zhou, J. Gong, Z. Yang, Z. Niu and P. Yang, "Green Mobile Access Network with Dynamic Base Station Energy Saving," *Proc. ACM MobiCom*, Beijing, China, pp. 1-3, Sep 2009.

[17] E. Oh, B. Krishnamachari, X. Liu and Z. Niu, "Toward Dynamic Energy-Efficient Operation of Cellular Network Infrastructure," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 56-61, June 2011.

[18] Z. Niu, Y. Wu, J. Gong and Z. Yang, "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 74-79, Nov 2010.

[19] R. Litjens and L. Jorguseski, "Potential of Energy-Oriented Network Optimisation: Switching OFF Over-Capacity in Off-Peak Hours," *Proc. IEEE PIMRC*, Turkey, pp.1660-1664, Sep 2010.

[20] K. Son, E. Oh and B. Krishnamachari, "Energy-Aware Hierarchical Cell Configuration: From Deployment to Operation," *Proc. IEEE INFOCOM*, pp. 289-294, China, Apr 2011.

[21] R. Li, Z. Zhao, X. Chen and H. Zhang, "Energy Saving through a Learning Framework in Greener Cellular Radio Access Networks," *Proc. IEEE GLOBECOM*, pp. 1574-1579, Anaheim, USA, Dec 2012.

[22] S. Kyuho, K. Hongseok, Y. Yung and B. Krishnamachari, "Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1525-1536, Sep 2011.

[23] C. Peng, S. Lee, S. Lu, H. Luo and H. Li, "Traffic-Driven Power Saving in Operational 3G Cellular Networks," *Proc. ACM MobiCom*, Nevada, USA, pp. 121-132, Sep 2011.

[24] S. Camazine, J. L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz and E. Bonabeau, *Self-Organization in Biological Systems*. Princeton Press, 2002.

[25] C. Prehofer and C. Bettstetter, "Self-Organization in Communication Networks: Principles and Design Paradigms," *IEEE Communications Magazine*, vol. 43, no. 7, pp. 78-85, July 2005.

[26] W. S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829-836, 1979.

[27] J. S. Hunter, "The Exponentially Weighted Moving Average," *Journal of Quality Technology*, vol. 18, no. 4, pp. 203-207, 1986.

[28] A. Lobinger, S. Stefanski, T. Jansen and I. Balan, "Load Balancing in Downlink LTE Self-Optimizing Networks," *Proc. IEEE VTC*, Munchen, Germany, pp.1-5, May 2010.

[29] L. Xiang, X. Ge, C. Wang, F. Li and F. Reichert, "Energy Efficiency Evaluation of Cellular Networks Based on Spatial Distributions of Traffic Load and Power Consumption," *IEEE Transactions on Wireless Communications*, vol. 12, no.3, pp. 961-973, Mar 2013.

[30] F. Richter, A. J. Fehske, and G. P. Fettweis, "Energy Efficiency Aspects of Base Station Deployment Strategies for Cellular Networks," *Proc. IEEE VTC*, pp. 1-5, USA, Sep 2009.

[31] M. Deruyck, W. Tanghe, W. Joseph and L. Martens, "Modelling and Optimization of Power Consumption in Wireless Access Networks," *Elsevier J. of Comp. Comm.*, vol. 34, no. 17, pp. 2036-2046, Nov 2011.

[32] 3GPP TR 36.942 Ver. 11.0.0 Rel. 11, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios," *Tech. Rep.*, Sep 2012.

[33] C. Seol and K. Cheun, "A statistical Inter-Cell Interference Model for Downlink Cellular OFDMA Networks Under Log-Normal Shadowing and Multipath Rayleigh Fading," *IEEE Trans. on Commun.*, vol. 57, no. 10, pp. 3069-3077, Oct 2009.

[34] T. D. Novlan, R. K. Ganti, A. Ghosh and J. G. Andrews, "Analytical Evaluation of Fractional Frequency Reuse for OFDMA Cellular Networks," *IEEE Trans. on Wireless Commun.*, vol. 10, no. 12, pp. 4294-4305, Dec 2011.

[35] D. Cao, S. Zhou, C. Zhang and Z. Niu, "Energy Saving Performance Comparison of Coordinated Multi-Point Transmission and Wireless Relaying," *Proc. IEEE GLOBECOM*, Beijing, China, pp.1-5, Dec 2010.

[36] WMF-T32-001-R021v02, "Architecture Tenets, Reference Model and Reference Points: Base Specifications," *WiMAX Forum Network Architecture*, Mar 2013.

[37] T. Han and N. Ansari, "On Greening Cellular Networks via Multicell Cooperation," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 82-89, Feb 2013.

[38] Wireless World Initiative New Radio WINNER+, "D5.3: WINNER+ Final Channel Models," Jun 2010.

[39] M. A. Marsan and M. Meo, "Energy Efficient Management of Two Cellular Access Networks," *Proc. ACM SIGMETRICS*, Washington, USA, pp. 1-5, June 2009.

[40] A. V. Krishnamoorthy, *et al.*, "Progress in Low-Power Switched Optical Interconnects," *IEEE J. of Sel. Topics in Quantum Elect.*, vol. 17, no. 2, pp. 357-376, Mar-Apr 2011.