Energy Consumption Analysis and Minimization in Multi-Layer Heterogeneous Wireless Systems

Elias Chavarria-Reyes, *Member, IEEE*, Ian F. Akyildiz, *Fellow, IEEE*, and Etimad Fadel

Abstract—Cellular network technologies have traditionally evolved to meet the ever-increasing need for capacity and coverage. Particularly, there has been a significant focus on exploiting the use of small cells and heterogeneous networks (HetNets). In the latter, the economic and environmental impact of the energy consumption is a key concern. Although much research has been done to address the energy consumption in HetNets, existing approaches have failed to capture the key factors affecting it. In this paper, the energy consumption in HetNets is analyzed with a focus on their multi-layer nature and the dependence of the energy consumption on the spatio-temporal traffic demands and the internal base station hardware components. The problem of minimizing the energy consumption is then studied and characterized in terms of a 0-1 Knapsack-like problem. Due to the differences with the classical 0-1 Knapsack problem, an efficient algorithm is introduced to minimize the energy consumption by adjusting the cell-association and the base stations on-off policies. Such algorithm is shown to be applicable to the two- and m-layer HetNet cases. Performance evaluation is provided to identify the achievable energy savings of our algorithm and its effect on the energy consumption, activity, and load across multiple layers.

Index Terms—Heterogeneous networks, energy saving, cellular networks, cell-association, on-off schemes

1 Introduction

The evolution of cellular network technologies has traditionally been driven by the ever-increasing need for capacity and coverage. The current forecasts predict that by 2018, high-speed coverage will reach over 85 percent of the world's population, and global traffic in mobile networks will rise with a compound annual growth rate (CAGR) of 50 percent, reaching a 12-fold increase since 2012 [1]. Foreseeing this exponential growth, the Third Generation Partnership Project (3GPP) developed the Long Term Evolution (LTE) Advanced (LTE-A) for the GSM/UMTS standards [2]. To reach these objectives, LTE-A incorporated four core technologies in its Release 10 (Rel-10) [3], [4]: carrier aggregation, enhanced multiple input multiple output (MIMO), relay nodes, and cooperative multipoint transmission. In addition, there has been a significant focus on exploiting, more than ever before, the use of small cells with heterogeneous networks (HetNets) in LTE-A. However, in HetNets many more base stations (BSs) are installed in any given geographical area. Each BS constantly operates at its maximum power, regardless of the amount of traffic being handled, which immediately raises the concern about the environmental impact and operational expenses (OPEX) incurred by operators in such HetNets, particularly in terms of energy consumption.

From an economic point of view, the energy consumption is a key concern for operators, as the energy costs constitute 7 to 20 percent of the entire OPEX. The network represents nearly 75 percent of this amount. Inside the network, around 70 percent of the energy is used by the radio access network (RAN) [5], with base stations consuming the most. From an environmental point of view, the impact of the information and communications technology (ICT) sector is also significant. Today, the ICT carbon footprint is comparable to that of the global aviation industry [6] or the one of 50 million cars. By 2020, this footprint is projected to grow at a rate of 3.8 percent, contributing to 2.3 percent of the global green house gas (GHG) emissions [7], which represents 1.27 $\mathrm{GtCO_2e}$.[1]

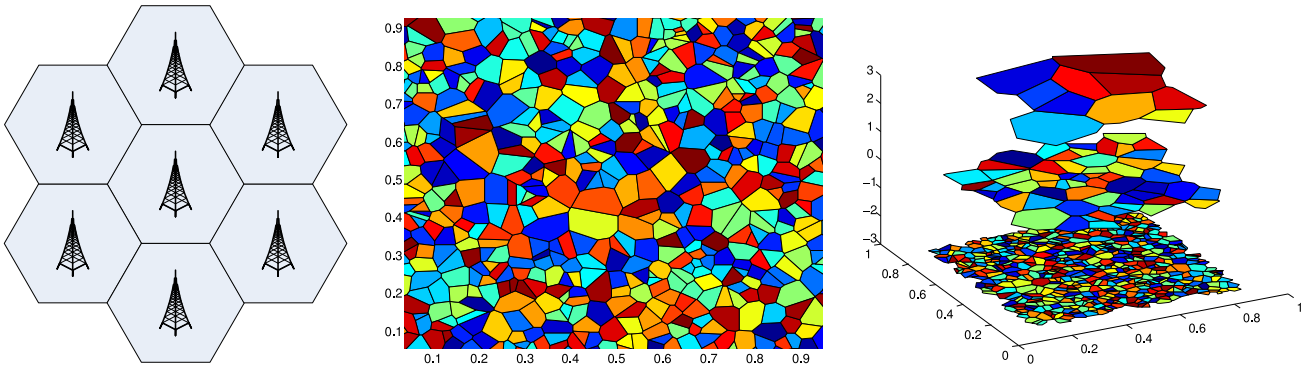
Due to the aforementioned reasons, both industry and academia have engaged in addressing the energy efficiency in cellular networks, also called "green cellular networks". Since the major energy inefficiencies and wastage occur at the RAN, it has received the most attention. Beyond improving the hardware components of the BS [8], [9], one of the more effective ways to reduce the energy consumption at the RAN is by turning BSs off. In HetNets, such action is possible since more than one BS may provide coverage in a particular area. To address the energy consumption at the RAN in HetNets, appropriate models are required for the (i) network deployment strategies, (ii) energy consumption at the BSs, and (iii) traffic demands (TDs).

In terms of network deployment strategies, initial work [10], [11], [12], [13], [14], [15], [16] focused on analyzing the energy consumption in a deployment following a hexagonal grid layout [17], as shown in Fig. 1a. Even though

- *E. Chavarria-Reyes is with the Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: elias.chavarria@gatech.edu.*
- *I.F. Akyildiz is with the Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, and with the Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia. E-mail: ian@ece.gatech.edu.*
- *E. Fadel is with the Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia. E-mail: eafadel@kau.edu.sa.*

Manuscript received 14 Feb. 2014; revised 3 Dec. 2014; accepted 8 Jan. 2015. Date of publication 26 Jan. 2015; date of current version 30 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2015.2393352

1. Giga tonnes of $\mathrm{CO_2}$ equivalent.



(a) Hexagonal Layout (b) Homogenous Poisson Point Process Layout (c) Multi-Layer Deployment

Fig. 1. Layouts for cellular networks.

it provides useful and tractable insights regarding the general network behavior, real deployments are far from following this layout. Recently, a layout where BSs are deployed according to some homogeneous Poisson Point Process distribution, as depicted in Fig. 1b, was shown to be tractable and to provide a better approximation to real network deployments [18]. Utilizing this second layout, the optimal BS density has been analyzed [19], as well as methods to adapt the active BSs and the cell-association policies [20], [21]. Nevertheless, these two layouts assume a single layer of BSs. This is far from truth in current systems and even more so in future ones. Operators owning multiple frequency bands use them for multi-layer HetNet deployments, as shown in Fig. 1c. The lower frequency bands are generally preferred for the BSs meant for coverage (i.e., the layer with largest cells), while the higher frequency bands are preferred for the BSs meant for capacity (i.e., the layer with smaller cells). In contrast to the existing work, the focus of our work is on multi-layer HetNets.

The results in the existing literature are also hindered by the BS energy consumption model used. Such model has typically been a long-term function of the output power [11], [12], [13], [14], [15], [16], [22] or of the overall BS traffic [20]. And even though it provides good insights, it is not readily usable to characterize the expected amount of energy that would be consumed according to the specific set of users with their traffic demands that the BS is expected to serve. This is a serious drawback, given that the traffic demands across cellular networks can fluctuate significantly during the day and across BSs [23], [24]. Furthermore, since a user equipment (UE) in a multi-layer HetNet may be under the coverage of more than one BS [23], it is necessary to quantify the energy consumption effects of each user to determine the BS that should serve a particular UE. This process is called *cell-association* [25]. Its role in the energy consumption in HetNets has been previously studied [20], [25], but only through simple energy models and deployment strategies. Another limitation of existing approaches is that they generally focus on satisfying simple performance thresholds, such as service outage probability, peak traffic, and minimum SINR, among others. However, such metrics are not adequate for the analysis of the energy efficiency of a network when spatially—and temporally—varying traffic demands are considered.

In this paper, we address the key limitations of the previous work in modeling and analyzing the energy consumption in infrastructure-based heterogeneous wireless system (IbHWS), where cellular systems are a particular case, and we propose a joint cell-association and on-off scheme to minimize such energy consumption. Our contributions can be summarized as follows:

- We model IbHWSs accounting for their multi-layer nature and the spatio-temporal variations of the traffic demands. According to the properties of the coverage area of the different BSs, we characterize all IbHWSs as general or regular IbHWSs.
- We demonstrate that, for a regular IbHWS, the solution for the energy minimization problem can be reduced to solving for an IbHWS whose last layer has a single BS.
- We characterize the energy minimization problem for the two-layer and the m-layer regular IbHWSs and demonstrate its mapping to a 0-1 Knapsack-like problem. We determine the conditions required for a direct mapping to the 0-1 Knapsack problem.
- We develop an efficient algorithm to perform a joint cell-association and on-off scheme to minimize the energy consumption in a two-layer regular IbHWS. We extend this algorithm to address the m-layer case.
- From the performance results of our algorithm, we observe energy savings of 35 percent across multiple scenarios, an inverse relation between the network load and the energy savings, and the significant role of small cells in handling the traffic in energy-efficient deployments.
- More importantly, we show that the cell-association and on-off policies need to be jointly adjusted according to the actual network deployment, the energy efficiency of the BSs, and the traffic dynamics, in order to achieve an energy-efficient network operation.

The rest of the paper is organized as follows. In Section 2, we describe the network architecture and traffic model. In Section 3, we describe and characterize the BS energy consumption model. In Section 4, we present the energy minimization problem, its analysis for regular IbHWSs, and an energy-efficient algorithm to address the two-layer and

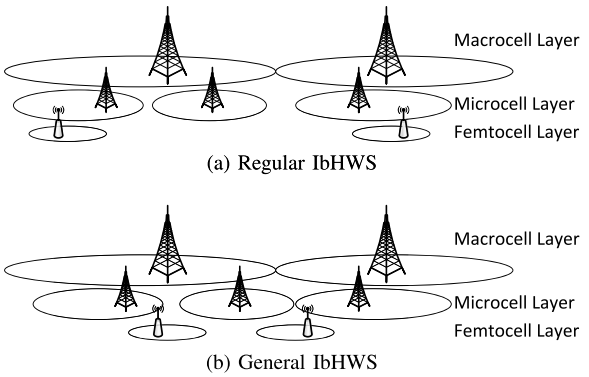


Fig. 2. Infrastructure-based heterogeneous wireless system.

m-layer cases. The performance of our algorithm is evaluated in Section 5. Finally, we conclude the paper in Section 6.

2 Network Architecture and Traffic Model

In this section, we describe the models utilized for an IbHWS, as well as its energy consumption.

2.1 Network Architecture

Typically, an IbHWS is composed of multiple layers of BSs of increasing coverage area,[2] as shown in Fig. 2. Each layer operates in a different frequency band. In Fig. 2a, the coverage area of each femtocell is a subset of the coverage area of a BS that belongs to the microcell layer. Similarly, the coverage area of each microcell is a subset of the coverage area of a BS that belongs to the macrocell layer. On the other hand, in Fig. 2b, the coverage area of the femtocell on the left is a subset not of the coverage area of any single microcell BS, but of the coverage area of a macrocell BS; the coverage area of the femtocell BS on the right is not a subset of any single BS. Formally, we define these two scenarios as follows.

Definition 1. *A general IbHWS is a set $\mathcal{A}$ of layers. Each layer $a \in \mathcal{A}$ contains a set $\mathcal{B}_a$ of BSs. Each BS $b \in \mathcal{B}_a$ is deployed in a coordinate $x_b \in \mathbb{R}^3$. Furthermore, there is no overlap between the effective coverage areas $\nu_{b^{(1)}} \subset \mathbb{R}^3$ and $\nu_{b^{(2)}} \subset \mathbb{R}^3$ of any two different BSs $b^{(1)}$ and $b^{(2)}$ that belong to the same layer a, i.e.,*

$$\nu_{b^{(1)}} \cap \nu_{b^{(2)}} = \varnothing, \forall a \in \mathcal{A}, b^{(1)} \in \mathcal{B}_a, b^{(2)} \in \mathcal{B}_a, b^{(1)} \neq b^{(2)}. \quad (1)$$

The effective coverage area represents the area in which users will prefer to connect to BS b rather than to any other BS belonging to the same layer as b.

In general, the BSs within the same layer share certain general characteristics, but may differ in their hardware and configurations. For example, the coverage area of the BSs in the femtocell layer is typically just enough for residential or small office environments. However, different femtocells may vary in their number of antennas, power amplifier energy efficiency, and other hardware characteristics. Moreover, even two femtocells with the exact same hardware may be configured to operate differently, e.g., with different transmission power.

Definition 2. *A regular IbHWS is a general IbHWS for which there exists an ordering of layers $a_0, a_1, \ldots$ so that*

$$\text{if } \nu_{b^{(i)}} \cap \nu_{b^{(j)}} \neq \varnothing \qquad \forall b^{(i)} \in a_i, \forall b^{(j)} \in a_j, j > i \\ \nu_{b^{(i)}} \subseteq \nu_{b^{(j)}}. \quad (2a)$$

Intuitively, in a regular IbHWS, the coverage area of each BS is either a subset of the coverage area of a larger BS or it is the BS with largest coverage area in a particular area.

It is important to highlight that an operator requires more resources, particularly spectrum, to deploy the multi-layer IbHWSs previously described. So far, the sub-3 GHz spectrum is the preferred and most commonly used by operators and is becoming increasingly crowded [26]. Gaining access to more sub-3 GHz spectrum can be achieved by repurposing, i.e., refarming, of the existing spectrum or by applying spectrum-sharing techniques [27]. However, the best case scenario expected from exploiting these two techniques is doubling the current cellular bandwidth, which may not be enough to allow a significant number operators to migrate to a multi-layer IbHWS. Another option is to exploit the mmWave frequencies, which range from 3 to 300 GHz [28]. This approach is expected to provide several tens of gigahertz for 5 G and future cellular systems [27], [29]. Thus, the use of mmWave frequency bands will enable the widespread adoption of multi-layer IbHWSs.

2.2 Traffic Model

In the existing literature, a traffic demand is typically characterized by the session arrival rate and the average file size [20]. With these two values, the overall load of a BS is then estimated in terms of (a) the total number of bits served by the BS, or (b) the average bitrate, i.e., total number of bits divided by the total time. Then, based on the value of overall load, the energy consumption of the BS is predicted. However, this approach is highly inaccurate, mainly because it assumes a linear relationship between a particular bitrate and the corresponding power needed to provide it.

If we were to transmit the same number of bits ρ at two different bitrates R_1 and R_0, where $R_1 = mR_0$ and $m > 1$, then the amounts of time t_1 and t_0 that would be required at each rate, respectively, would be related by

$$t_0 = mt_1. \quad (3)$$

Therefore, for the energy consumption at both rates to be equal it would be needed that

$$P_{RX_1} = mP_{RX_0}, \quad (4)$$

where P_{RX_1} and P_{RX_0} represent the received power required to achieve each bitrate, respectively. However, utilizing Shannon's Channel Capacity theorem, it can be shown that the relationship between P_{RX_1} and P_{RX_0} is exponential rather than linear:

2. For simplicity, we will utilize the term "coverage area" to denote the 3D coverage space. Similarly, diagrams will depict 2D versions of the coverage area for ease of understanding. Nevertheless, the analysis is done considering a 3D coverage space.

$$P_{RX_1} = \left[\left(1 + \frac{P_{RX_0}}{\eta}\right)^m - 1\right]\eta, \tag{5}$$

where η represents the noise plus interference. Even if the analysis is restricted to a unique bitrate, i.e., $m = 1$, the value of P_{RX} will vary with the distance between the BS and the intended receiver. In conclusion, energy consumption estimation based on the classical traffic demand modeling is inaccurate due to (a) the non-linear relationship between power and bitrate and (b) the distance-dependent power required to satisfy a particular bitrate.

To accurately characterize the energy consumption, we follow a different approach. We consider the network to be able to serve sessions that have different quality of service (QoS) requirements. These are defined in terms of bitrates, to capture the non-linear relationship between bitrate and power. Here, $\mathcal{Q}$ denotes the set of QoS requirements that the network supports.

We consider the total service area $\mathcal{U}$ to be divided into locations $u \subset \mathbb{R}^3$, satisfying the following:

1) Each location u is a connected set.
2) Locations are non overlapping: $u_1 \cap u_2 = \varnothing$.
3) Each location can generate sessions, i.e., traffic demands, of any QoS $q \in \mathcal{Q}$.
4) $\mathcal{U} = \bigcup u$.

For every BS and the locations that it serves, we use a matrix $\mathbf{N}(t)$ to characterize the number of active sessions at any time instant t corresponding to every location u and QoS q.

3 Energy Model

Using the previous network and traffic models, we developed in [30] the energy consumption model for the RAN. First, the expected value of the radio frequency (RF) energy required by a BS to serve its users during the jth time interval was shown to be

$$\mathbb{E}[E_{RF}(\Delta t_j)] = \Delta t_j \, \mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]\big), \tag{6}$$

where $\mathrm{tr}(\mathbf{X})$ denotes the trace of a matrix $\mathbf{X}$, Δt_j denotes the jth time interval of the day, and the matrix $\mathbf{P}(t)$ characterizes the amount of power required by a given BS to satisfy every location u with one session of QoS q. It was considered that both $\mathbf{N}(t)$ and $\mathbf{P}(t)$ were first-order stationary (FOS), which allowed us to drop the time dependency from the expected value.

Similarly, by modeling a BS as a set of interconnected components $c \in \mathcal{C}$ we showed that each component always consumes a minimum amount of energy, whether that component is "on" or "off":

$$E_{\mathrm{off}}(\Delta t_j) = P_{\mathrm{off}} \Delta t_j, \tag{7}$$

$$E_{\mathrm{on,min}}(\Delta t_j) = P_{\mathrm{on,min}} \Delta t_j, \tag{8}$$

where P_{off} is the standby power, and $P_{\mathrm{on,min}}$ is the power consumed with no load. For passive components

$$P_{\mathrm{on,min}} = P_{\mathrm{off}} = 0, \tag{9}$$

and for non-passive components

$$P_{\mathrm{on,min}} \geq P_{\mathrm{off}} \geq 0. \tag{10}$$

In addition to the minimum energy consumed in the "on" state, each component also has a dynamic energy consumption $\mathbb{E}[E_{\mathrm{on,dyn}}(\Delta t_j)]$ that can follow one of two models. In the first model, $\mathbb{E}[E_{\mathrm{on,dyn}}(\Delta t_j)]$ is a function of the energy that the component needs to produce at its output:

$$\mathbb{E}[E_{\mathrm{on,dyn}}(\Delta t_j)] = \frac{1}{\alpha}\mathbb{E}[E_{\mathrm{out}}(\Delta t_j)], \tag{11}$$

where α is power efficiency of the component, and $E_{\mathrm{out}}(t)$ is the energy that it needs to produce at its output. We denote the components that follow this model as Type A. This includes the power amplifier and the feeder. In the second model, $\mathbb{E}[E_{\mathrm{on,dyn}}(\Delta t_j)]$ is a function of the overall load (in terms of bitrate) that it needs to process:

$$\mathbb{E}[E_{\mathrm{on,dyn}}(\Delta t_j)] = \frac{\Delta t_j (P_{\mathrm{on,max}} - P_{\mathrm{on,min}})}{\mathrm{load}_{\max}} \mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbf{R}^T\big), \tag{12}$$

where $P_{\mathrm{on,max}}$ is the maximum power. $\mathbf{R}$ is a matrix in which (i) all rows are equal, and (ii) the values in a row represent the bitrate associated with each $q \in \mathcal{Q}$. We denote the components that follow this second model as Type B. This includes the baseband processor (BBuP) [14], [22]. Since we model the BS as a set of interconnected components of Types A and B, having the energy consumption model for each type allows us to model the total energy consumption of the entire BS.

Remark 1. The energy consumption model of the BS is an affine function of $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbf{R}^T\big)$ and $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]\big)$.

Proof. In this section, all the energy models are affine functions of $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbf{R}^T\big)$ and $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]\big)$. Since the interconnection of any two components leads to the composition of their respective energy models, their joint energy model is also an affine function of $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbf{R}^T\big)$ and $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]\big)$. By induction, it follows that the energy consumption model of the BS is also an affine function of these two variables. □

Being an affine function, the energy consumption $\hat{E}_b(\Delta t_j)$ of a BS b can be expressed in the form

$$\hat{E}_b(\Delta t_j) = \begin{cases} \hat{E}_{\mathrm{off}}(\Delta t_j) & \text{if } b \text{ is } off \\ \hat{E}_{\mathrm{on,min}}(\Delta t_j) + \hat{E}_{\mathrm{on,dyn}}(\Delta t_j) & \text{if } b \text{ is } on, \end{cases} \tag{13}$$

where $\hat{E}_{\mathrm{off}}(\Delta t_j)$ represents the total energy consumption of b when it is *off*. $\hat{E}_{\mathrm{on,min}}(\Delta t_j)$ and $\hat{E}_{\mathrm{on,dyn}}(\Delta t_j)$ represent the total minimum and dynamic energy consumption when b is *on*, respectively. $\hat{E}_{\mathrm{on,dyn}}(\Delta t_j)$ can be further expressed as the linear function

$$\hat{E}_{\mathrm{on,dyn}}(\Delta t_j) = \mathbf{A}\begin{bmatrix} \big(\mathbb{E}[\mathbf{N}]\mathbf{R}^T\big) \\ \mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]\big) \end{bmatrix}, \tag{14}$$

where $\mathbf{A}$ is a 1x2 matrix unique to every BS. The exact expressions for $\hat{E}_{\mathrm{off}}(\Delta t_j)$, $\hat{E}_{\mathrm{on,min}}(\Delta t_j)$, and $\mathbf{A}$ depend on the specific set $\mathcal{C}$ of components present in a BS b and how they are interconnected. For example, regarding $\mathcal{C}$, a macrocell is

typically equipped with an air conditioning (A/C) system; however, such A/C is not required in a femtocell. Regarding the interconnection of components, even if no users are connected to a BS b, the power supply (PS) is *on* (with load) since it provides the power to all components.

Our affine model for BS energy consumption fits the early measurements reported in [14] as part of the EARTH project. However, in [14], the model was approximated from the measured data rather than explicitly derived, as done here. Furthermore, the model in [14] is limited since it only considers the RF output power dependency. In addition, it is not readily extendable to predict the BS energy consumption in particular traffic distributions.

Remark 2. $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbf{R}^T\big)$ and $\mathrm{tr}\big(\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]\big)$ are linear functions along the dimension of locations u.

Proof. The ith element of the main diagonal of the matrices $\mathbb{E}[\mathbf{N}]\mathbf{R}^T$ and $\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]$ can be expressed as

$$\sum_{\tau} \mathbb{E}\big[\mathbf{N}_{i,\tau}\big]\mathbf{R}_{\tau,i} \qquad \text{and} \tag{15a}$$

$$\sum_{\tau} \mathbb{E}\big[\mathbf{N}_{i,\tau}\big]\mathbb{E}\big[\mathbf{P}_{\tau,i}\big], \tag{15b}$$

respectively, where $\mathbf{N}_{i,\tau}$ is the number of active sessions of QoS q_τ from location u_i, $\mathbf{R}_{\tau,i}$ is the rate associated with QoS q_τ for location u_i, and $\mathbf{P}_{\tau,i}$ is the RF power required to satisfy a single session of QoS q_τ for location u_i. Therefore, the ith element of the main diagonal in each matrix is uniquely defined by the values from location u_i. It follows that the summation of the elements of the main diagonal, i.e., the trace function, of the matrices $\mathbb{E}[\mathbf{N}]\mathbf{R}^T$ and $\mathbb{E}[\mathbf{N}]\mathbb{E}\big[\mathbf{P}^T\big]$ are linear functions along the dimension of locations u. □

From Eq. (14), we have that $\hat{E}_{\mathrm{on,dyn}}(\Delta t_j)$ is a linear function of $\mathrm{tr}(\mathbb{E}[\mathbf{N}]\mathbf{R}^T)$ and $\mathrm{tr}(\mathbb{E}[\mathbf{N}]\mathbb{E}[\mathbf{P}^T])$. Since these two matrices are linear functions along the dimension of locations u, as described in Remark 2, it follows that $\hat{E}_{\mathrm{on,dyn}}(\Delta t_j)$ is also a linear function along the dimension of locations u. Therefore, the energy consumption of a BS b, denoted by $\hat{E}_b(\Delta t_j)$ in Eq. (13), is itself an affine function along the dimension of locations u. This allows us, in Section 4, to analyze the effect on the energy consumption $\hat{E}_b(\Delta t_j)$ of the traffic coming from each location that a BS can potentially serve.

4 Energy Saving

For jth time interval Δt_j, finding the maximum energy savings at the RAN can be obtained by solving the following optimization problem:

$$\min \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}_a} \mathbb{E}\big[\hat{E}_b(\Delta t_j)\big], \tag{16}$$

subject to:

$$\sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}_a} \phi_b(\Delta t_j, u) = 1, \forall u \in \mathcal{U}, \tag{17a}$$

$$\sum_{u\in\Psi_b(\Delta t_j)} \sum_{q\in\mathcal{Q}} \mathbb{E}\big[\mathbf{N}_{u,q}\big]\mathbf{R}_{u,q} \le \Upsilon_b, \forall a \in \mathcal{A}, \forall b \in \mathcal{B}_a, \tag{17b}$$

where $\phi_b(\Delta t_j, u)$ is an indicator function equal to 1 when b serves u, $\Psi_b(\Delta t_j)$ is the set of users served by b, and Υ_b is the maximum load supported by b. Constraint (17a) indicates that each u is served by a single b, while (17b) indicates that the total load in each BS should be below the maximum it can support. This problem is in general non-convex and not-solvable through standard optimization methods. Nevertheless, the complexity of obtaining the solution at the jth time interval is reduced if any of the following conditions applies:

1) For every active location u that can only be served by a single BS, its serving BS must be in the *on* state. For any such BS, its energy contribution to the overall RAN energy consumption is at least $\hat{E}_{\mathrm{on,min}}(\Delta t_j)$.
2) Every BS b for which no location under its coverage requests any traffic can be switched to the *off* state. For any such BS, its energy contribution to the overall RAN energy consumption is determined by Eq. (13).

   From an analytical perspective, this condition should always be enforced. However, in a real IbHWS, turning off a cell should only be done when it is highly probable that no user will appear in that area (e.g., an office femtocell during the night) or when another active cell is serving the area of the cell being turned off.

An alternative way of looking at the energy minimization problem is to solve two related and coupled problems:

- Cell (de-)activation: Which BSs should be turned on? Which ones should be turned off?
- Cell-association: To which BS should each location connect in order to be served?

If the answer to any of these two is fixed, then the optimal solution for the second one can be found [20]. Rather than considering these two problems separately, we focus on solving them jointly for the Regular IbHWS. Our approach allows us to adjust the cell-association policy to maximize the achievable energy savings. For example, as later detailed in Section 4.1, by adjusting the cell-association policies so that no users are associated with a given BS it is then possible to turn that BS off and, thus, reduce the network energy consumption.

4.1 Regular IbHWS

Remark 3. Given a regular IbHWS $\mathcal{A}$, there exists a partition Γ of the total coverage area of interest

$$\Xi = \bigcup_{\substack{b\in\mathcal{B}_a \\ a\in\mathcal{A}}} \nu_b, \tag{18}$$

such that

$$\forall \gamma \in \Gamma : \exists a \in \mathcal{A}, b \in \mathcal{B}_a \,|\, \gamma = \nu_b. \tag{19}$$

Proof. The partition Γ is obtained through the following steps. First, made Γ equal to a set whose elements are all possible ν_b. Then, remove any ν_b that is a subset of another element of Γ due to condition (2a). Elements that

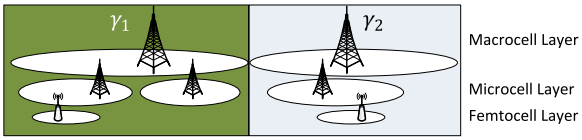


Fig. 3. Partition of regular IbHWS.

were not removed must satisfy that their intersection with any other element of Γ is the empty set. Therefore, the remaining elements form a partition of Ξ. □

From the previous proof, it also follows that:

$$\forall a \in \mathcal{A}, b \in \mathcal{B}_a : \exists \gamma \in \Gamma \mid \nu_b \subseteq \gamma. \tag{20}$$

Therefore, any IbHWS $\mathcal{A}$ can be partitioned into mutually disjoint areas such that the effective coverage area of each BS is a subset of some element of the partition. At the same time, all the BSs for which $\nu_b \subseteq \gamma$, for a particular $\gamma \in \Gamma$, also define an IbHWS where the last layer has a single BS. Fig. 3 depicts this partitioning for the regular IbHWS of Fig. 2a.

Remark 4. For a regular IbHWS, the optimal solution to the energy minimization problem is achieved by finding the optimal solution for each IbHWS defined by the elements of Γ.

Proof. From Eq. (20), it follows that the on/off configuration of any BS b only affects the energy consumption of the IbHWS defined by the γ for which $\nu_b \subseteq \gamma$. Thus, the IbHWSs defined by the rest of the elements of Γ can be independently configured from the on/off state of b. □

In conclusion, finding the optimal solution for the energy minimization problem in a regular IbHWS reduces to being able to find the optimal solution for a regular IbHWS whose last layer has a single BS. We exploit this property to develop an efficient algorithm to minimize the energy consumption for the two-layer and multi-layer regular IbHWS.

4.2 Two-Layer Regular IbHWS

The two-layer regular IbHWS is depicted in Fig. 4. While simple, this type of scenario is important since it is the most commonly deployed when operators are migrating their networks from a single-layer layout to a multi-layer one. Furthermore, it allows us to establish the concepts that become the building blocks in addressing the m-layer regular IbHWS.

The approach to tackle the energy minimization problem at the jth time interval for a two-layer regular IbHWS is now presented. For simplicity, we utilize $b_{i,k}$ to denote the ith BS of layer $a_k \in \mathcal{A}$ and we drop the Δt_j dependency.

First, we consider that layer a_1 is not present. Therefore, every location u must be served by a BS $b_{i,0}$ (i.e., that belongs to layer a_0), and we can obtain the energy consumption of the network by applying Eq. (13) to every BS $b_{i,0}$. We denote the network energy consumption for this case by E_{N-0}. For n BSs, this has complexity $O(n)$.

Second, we consider that layer a_1, which only has one BS $b_{0,1}$, is also present. At this point, we have a decision point:

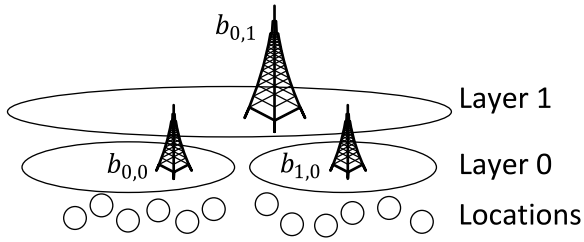


Fig. 4. Two-layer regular IbHWS.

should $b_{0,1}$ be *on* or *off*? In order to answer the question, we need to compare the minimum energy required by the network when $b_{0,1}$ is *off* to the minimum one required when $b_{0,1}$ is *on*. We now proceed to find these two values.

If $b_{0,1}$ is set to *off*, the overall network consumption E_{N-1} becomes

$$E_{N-1} = E_{N-0} + \hat{E}_{\text{off}}(b_{0,1}), \tag{21}$$

where $\hat{E}_{\text{off}}(b_{0,1})$ corresponds to the energy consumed by $b_{0,1}$ while *off*, as defined in Eq. (13).

On the other hand, if $b_{0,1}$ is set to be *on*, the network consumption must increase by at least $\hat{E}_{\text{on,min}}(b_{0,1})$, as defined in Eq. (13). Therefore, E_{N-1} becomes (at least)

$$E_{N-1} = E_{N-0} + \hat{E}_{\text{on,min}}(b_{0,1}). \tag{22}$$

When $b_{0,1}$ is *on* and all locations are served by layer a_0, there are two types of actions that can be further executed to reduce the energy consumption E_{N-1}.

1) *Action Type 1:* Use $b_{0,1}$ to a serve a particular location u instead of layer a_0. For this action to be executable, the following condition must be satisfied:

$$\text{RemCap}(b_{0,1}) \geq \text{ReqCap}(\text{AT1}_u), \tag{23}$$

where $\text{RemCap}(b_{0,1})$ denotes the remaining capacity in $b_{0,1}$, and $\text{ReqCap}(\text{AT1}_u)$ is the required capacity to serve location u. If executable, the change in the energy consumption $\Delta E_1(u)$ of the overall network is

$$\Delta E_1(u) = \hat{E}_{\text{on,dyn}}(b_{0,1}, u) - \hat{E}_{\text{on,dyn}}(b_{i,0}, u), \tag{24}$$

where $b_{i,0}$ is the BS of layer a_0 initially serving the particular location u. However, if the location u was the only one served by $b_{i,0}$, then such BS can be turned off. Therefore, there would be an additional change in the overall energy consumption:

$$\Delta E_1(u) = \Delta E_1(u) + \hat{E}_{\text{off}}(b_{i,0}) - \hat{E}_{\text{on,min}}(b_{i,0}). \tag{25}$$

From Eq. (10), it immediately follows that

$$\hat{E}_{\text{off}}(b_{i,0}) - \hat{E}_{\text{on,min}}(b_{i,0}) \leq 0. \tag{26}$$

A value $\Delta E_1 < 0$ implies a reduction in the overall energy consumption of the network.

2) *Action Type 2:* Use $b_{0,1}$ to serve all locations u originally served by a particular BS $b_{i,0}$. For this action to be executable, the following condition must be satisfied:

$$\text{RemCap}(b_{0,1}) \geq \text{ReqCap}(\text{AT2}_{b_{i,0}}), \tag{27}$$

where $\mathrm{ReqCap}(\mathrm{AT2}_{b_{i,0}})$ denotes the required capacity to serve all locations u originally served by $b_{i,0}$. If executable, the change in the overall energy consumption $\Delta E_2(b_{i,0})$ of the network is

$$\Delta E_2(b_{i,0}) = \hat{E}_{\mathrm{off}}(b_{i,0}) + \sum_{u \in \Psi_{b_{i,0}}} E_{\mathrm{on,dyn}}(b_{0,1}, u) - \hat{E}_{\mathrm{on,min}}(b_{i,0}) - \sum_{u \in \Psi_{b_{i,0}}} E_{\mathrm{on,dyn}}(b_{i,0}, u), \tag{28}$$

where Ψ_b is the set of locations served by b. A value $\Delta E_2 < 0$ implies a reduction in the overall energy consumption of the network.

When $b_{i,0}$ serves a single location u, Eq. (28) reduces to Eq. (25). This can be generalized as follows: the application of Action Type 2 can be interpreted as the execution of Action Type 1 to all u originally served by $b_{i,0}$, followed by turning $b_{i,0}$ off.

For n BSs in layer a_0, p locations, $p \geq n$, calculating the effect ΔE_1 of Action Type 1 for all locations u has an overall complexity of $O(p)$. Similarly, calculating the effect ΔE_2 of Action Type 2 for all $b_{i,0}$ has an overall complexity of $O(p)$.

Utilizing the above formulation, we now describe how the energy saving problem is mapped to a Knapsack-like problem. Consider the following concept mapping:

- $-\Delta E_1(u)$: "Profit" of executing Action Type 1 on u.
- $-\Delta E_2(b_{i,0})$: "Profit" of executing Action Type 2 on $b_{i,0}$.
- $\mathrm{ReqCap}(\mathrm{AT1}_u)$: "Weight" of executing Action Type 1 on u.
- $\mathrm{ReqCap}(\mathrm{AT2}_{b_{i,0}})$: "Weight" of executing Action Type 2 on $b_{i,0}$.
- $\mathrm{Cap}(b_{0,1})$: Maximum "weight" supported by BS $b_{0,1}$.

Then, finding the minimum energy consumed by the network when $b_{0,1}$ is *on* can be represented as a 0-1 Knapsack-like problem where each action has a "weight" and provides a "profit". The major differences with the traditional 0-1 Knapsack problem are:

- The "profit" of an action might be negative. This occurs if the action causes an increase in the overall energy consumption of the network. No such action should be executed.
- The "profit" of an action may change depending on previously executed actions. For example, if Action Type 1 is executed on a location u, then the additional achievable "profit" of executing Action Type 2 on the serving BS $b_{i,0}$ of u will be less than the original value of $-\Delta E_2(b_{i,0})$.

For these reasons, we introduce Algorithm 1 to tackle the Knapsack-like problem of finding the minimum energy consumption when $b_{0,1}$ is *on*. Without loss of generality, we assume that every location u could be served by some $b_{i,0}$. If this were not the case, the algorithm would be modified to make $b_{0,1}$ serve such locations before executing any other action.

The first step in Algorithm 1 consists of initializing a list of all possible actions whose execution would reduce the network energy consumption. Such initiation is addressed by Algorithm 2. There, steps 2-9 and 10-17 calculate all possible actions of Type 1 and Type 2, respectively. Actions whose execution would increase the network energy consumption, i.e., whose $\Delta E > 0$, are discarded in steps 5 and 13. For the non-discarded actions, the efficiency is computed in steps 7 and 15 by the function Eff(), defined as

$$\mathrm{Eff}() = \frac{\Delta E}{\mathrm{ReqCap}}. \tag{29}$$

Algorithm 1. Minimum E_{N-1} for $b_{0,1}$ *on*

1: **Call** Init-Vars ▷ Calls Algorithm 2
2: **for** $i \leftarrow 1, |H|$ **do**
3: $h \leftarrow i$th element from H.
4: $\mathrm{ATX} \leftarrow$ action associated with h
5: **if** $\mathrm{RemCap}(b_{0,1}) \geq \mathrm{ReqCap}(\mathrm{ATX})$ **then**
6: **Call** Execute-Action ▷ Calls Algorithm 3
7: **else**
8: Discard ATX
9: **end if**
10: **end for**

Algorithm 2. Init-Vars

1: $E_{N-1} = E_{N-0} + \hat{E}_{\mathrm{on,min}}(b_{0,1})$
2: **for all** u **do** ▷ Calculate all AT1
3: Calculate $\Delta E_1(u)$
4: **if** $\Delta E_1(u) > 0$ **then**
5: Discard $\mathrm{AT1}_u$
6: **else**
7: Calculate $\mathrm{Eff}(\mathrm{AT1}_u)$
8: **end if**
9: **end for**
10: **for all** $b_{i,0}$ **do** ▷ Calculate all AT2
11: Calculate $\Delta E_2(b_{i,0})$
12: **if** $\Delta E_2(b_{i,0}) > 0$ **then**
13: Discard $\mathrm{AT2}_{b_{i,0}}$
14: **else**
15: Calculate $\mathrm{Eff}(\mathrm{AT2}_{b_{i,0}})$
16: **end if**
17: **end for**
18: List $H \leftarrow$ Jointly sort all $\mathrm{Eff}(\mathrm{AT1}_u)$ and $\mathrm{Eff}(\mathrm{AT2}_{b_{i,0}})$ in ascending order

The efficiency of the non-discarded actions is always non-positive. A lower efficiency value represents a higher reduction in energy consumption per capacity required to execute an action. Finally, step 18 jointly sorts all non-discarded actions of Type 1 and Type 2 in ascending order.

Having the sorted list of actions that would reduce the energy consumption, steps 2-10 of Algorithm 1 execute each action, as long as $b_{0,1}$ has enough remaining capacity to support it. For the executable ones, Algorithm 1 relies on Algorithm 3 to perform the action execution.

In Algorithm 3, steps 1-11 and steps 12-20 deal with the execution of an AT1 and AT2, respectively. For an AT2, $\mathrm{Prev} - \mathrm{mod}(b)$ indicates whether one or more locations originally served by b have already been reallocated to $b_{0,1}$. If $\mathrm{Prev} - \mathrm{mod}(b)$ is false, the AT2 is executed normally, as shown in steps 3-6. If $\mathrm{Prev} - \mathrm{mod}(b)$ is true, the AT2 is

executed in steps 7-9 only if the energy savings provided by the AT2 are greater than those achieved by the previous reassociation of locations from b to $b_{0,1}$. In other words, $\text{Prev} - \text{mod}(b)$ allows us to account for the co-dependency of an AT2 on previously executed actions. For an AT1, the co-dependency on previously executed actions arises from the fact that an AT1 cannot be executed if the BS b has already been turned off, as shown in steps 15-20.

Algorithm 3. Execute-Action

```
1:  if Type(ATX) = Type 2 then                          ▷ Type 2
2:    b ← BS associated with ATX
3:    if Prev − Mod(b) = False then
4:      Apply ATX, and Update RemCap(b_{0,1})
5:      E_{N−1} = E_{N−1} + ΔE_2(b)
6:    else
7:      if ΔE_2(b) < ΔE_previous then
8:        Apply ATX, and Update RemCap(b_{0,1})
9:        E_{N−1} = E_{N−1} + ΔE_2(b) − ΔE_previous
10:     end if
11:   end if
12: else                                                ▷ Type 1
13:   u ← location associated with ATX.
14:   b ← servings BS of u.
15:   if b is already off then
16:     Discard ATX.
17:   else                                              ▷ b is on
18:     Apply ATX, and Update RemCap(b_{0,1})
19:     E_{N−1} = E_{N−1} + ΔE_1(u)
20:   end if
21: end if
```

Together, these three algorithms have an overall complexity of $O(p \log(p))$, which is dominated by the sorting performed in the last step of Algorithm 2. These algorithms allow us to:

1) Find the minimum amount of energy consumed when $b_{0,1}$ is *on*, which we can compare to the energy required when $b_{0,1}$ is *off* (Eq. (21)).
2) Jointly determine which BSs should be *on* or *off*, and to which BS each location should be associated.

Remark 5. The energy minimization problem reduces to the classical 0-1 Knapsack problem if

$$\hat{E}_{\text{off}}(b_{i,0}) = \hat{E}_{\text{on,min}}(b_{i,0}) \forall i. \tag{30}$$

Proof. When the previous condition applies, executing an Action Type 2 on a BS b has the same effect on the overall energy consumption as the application of Action Type 1 over all the locations served by b. Therefore, we can omit Actions Type 2 from the problem formulation and only consider Actions Type 1. □

The condition established in Remark 5 applies to most BSs nowadays since they lack the capability to enter into a dormant state for energy savings. However, in the most recent studies for 4G networks, such as LTE-Advanced, the ability of entering into a dormant state is being introduced.

4.3 Multi-Layer Regular IbHWS

In this section, we build on the scheme followed in the two-layer IbHWS to develop a solution for the m-layer regular IbHWS. The proposed solution for the m-layer regular IbHWS can be efficiently obtained from the solution of the $(m-1)$-layer regular IbHWS, making the approach scalable.

Consider a regular IbHWS $\mathcal{A}$ of m layers, i.e.,

$$\mathcal{A} = \{a_0, a_1, \ldots, a_{m-1}\}, \tag{31}$$

where the top layer a_{m-1} has a single BS. It is enough to analyze this case since it can be extended to the one where the top layer has more than one BS, as discussed at the beginning of Section 4.1. In addition, consider that we have the optimal solution for the $(m-1)$-layer IbHWS $\mathcal{A}'$ defined by

$$\mathcal{A}' = \mathcal{A} \setminus a_{m-1} = \{a_0, a_1, a_{m-2}\}. \tag{32}$$

Let E_{N-2} denote the energy consumption of the solution for $\mathcal{A}'$. If we create $\mathcal{A}$ by adding a_{m-1} to $\mathcal{A}'$, we have to decide: should $b_{0,m-1}$ be *on* or *off*? As in two-layer case, we need to compare the minimum energy required by the network when $b_{0,m-1}$ is *off* to the minimum one required when $b_{0,m-1}$ is *on*. We proceed to find these two values.

If $b_{0,m-1}$ is *off*, the overall network consumption E_{N-3} becomes

$$E_{N-3} = E_{N-2} + \hat{E}_{\text{off}}(b_{0,m-1}). \tag{33}$$

On the other hand, if $b_{0,m-1}$ is set to be *on*, the network consumption must increase by at least $\hat{E}_{\text{on,min}}(b_{0,m-1})$. Therefore, E_{N-3} becomes (at least)

$$E_{N-3} = E_{N-2} + \hat{E}_{\text{on,min}}(b_{0,m-1}). \tag{34}$$

When $E_{\min}(b_{0,m-1})$ is *on* and all locations are served by the BSs in $\mathcal{A}'$, there are four types of actions that could be executed to reduce the energy consumption E_{N-3}.

1) *Action Type 1:* Use the BS in the top layer, i.e., $b_{0,m-1}$, to serve a particular location u instead of $\mathcal{A}'$. This is equivalent to Action Type 1 for the two-layer case.
2) *Action Type 2:* Use $b_{0,m-1}$ to serve all locations u originally served by a particular BS $b_{i,k}$, where $k < m-1$ (i.e., for some $a_k \in \mathcal{A}'$). This is equivalent to Action Type 2 for the two-layer case.
3) *Action Type 3:* For a location u, switch the serving BS to another BS $b_{i,k}$, where $k < m-1$.
4) *Action Type 4:* For all locations u currently served by the same BS, switch the serving BS to another BS $b_{i,k}$, where $k < m-1$.

Remark 6. Applying an Action Type 3 or Action Type 4 to the optimal solution of $\mathcal{A}'$ does not lead to an immediate reduction in the energy consumption E_{N-3}.

Proof. Assume that an Action Type 3 or Action Type 4 can lead to a reduction in the energy consumption E_{N-3}. Thus, any of them could also be applied to the optimal solution of $\mathcal{A}'$ and achieve a reduction in the energy consumption of $\mathcal{A}'$. Since the energy consumption of $\mathcal{A}'$ is already at the minimum, then it follows that no action exists that could have been applied to further reduce the energy consumption. Therefore, an Action Type 3 or

Action Type 4 cannot lead to an immediate reduction of the energy consumption of $\mathcal{A}'$ or $\mathcal{A}$. □

As a result of the previous remark, we focus our attention on Action Type 1 and Action Type 2 from now on. Since it is equivalent to the ones for the two-layer case, we can use the following mapping to avoid repeating all the formulation previously explained for the two-layer case.

- In the two-layer case, all the active BSs initially serving users belonged to layer 0, i.e., they were of the form $b_{i,0}$. In the m-layer case, they have the form $b_{i,k}$, where $k < m-1$.
- In the two-layer case, the effect of both actions leads to $b_{0,1}$ becoming the serving BS for certain locations u. In the m-layer case, the effect of both actions leads to $b_{0,m-1}$ becoming the serving BS for certain locations u.

With this mapping, we can update all the equations developed for the two-layer case, as well as the proposed algorithm, and apply them to obtain the solution for the m-layer case if we are given the solution to the case of $m-1$ layers. Therefore, given any regular IbHWS $\mathcal{A}$ we can iteratively apply the previous formulation to find the solution for increasing number of layers till we reach m, with an overall complexity $O(mp \log(p))$. Regarding how frequently the algorithm is executed, we envision it to be rerun at regular time intervals defined by the operator, e.g., every hour. The algorithm is not reactive, it runs based on traffic predictions. As a result, no over- or under-estimation of the traffic conditions will trigger an immediate algorithm rerun; instead, those differences will be incorporated in the future traffic predictions. Acquiring the data necessary for the traffic predictions represents the largest overhead associated with the execution of our algorithm. Such data includes the UE location information and the QoS requested for a traffic session. Acquiring the UE location information at the cell level has been shown to be feasible [24]. Operators could obtain higher location accuracy by triggering the UEs to accurately report their locations [31]. The operators can also fairly easily obtain the QoS requested for each traffic session by inspecting the messages exchanged during the procedures to establish a data bearer between the UE and the packet gateway [32].

5 Performance Evaluation

In this section, we evaluate the performance of the proposed energy-saving algorithm for the regular IbHWS.

TABLE 1
Simulation Parameters

Parameter	Value
Bandwidth	3.84 MHz
BS Max. bitrate	14.4 Mbps
QoS Rates	[1,20,80,200]*15 kbps
Time intervals	24 (1 hr each)
Total Coverage Radius	2,500 m
Number of Locations	500
Altitude of Locations	1.5 m
Number of Layers	3
Type of BSs (per layer)	[pico,pico,macro]
Number of BSs (per layer)	[100,25,6]
Altitude of BSs (per layer)	[10,20,25]m

5.1 Simulation Setup

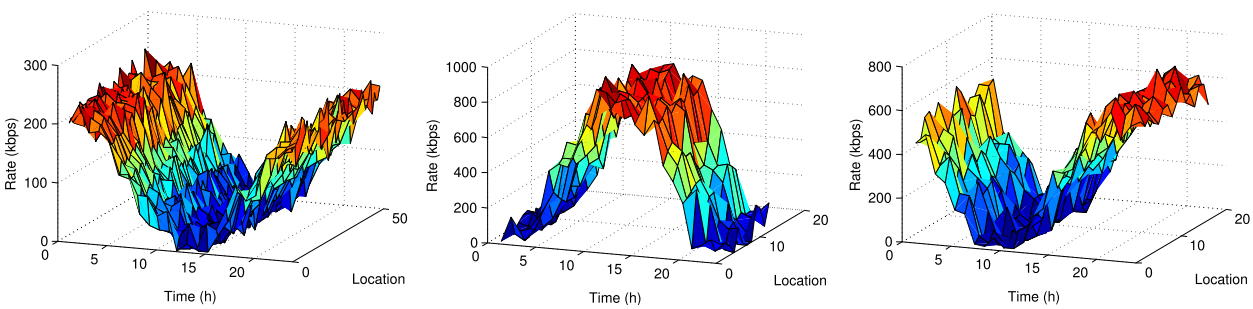
Simulation parameters for the IbHWS are shown in Table 1. BSs per layer and locations are uniformly distributed across the coverage area of interest Ξ. The IbHWS is configured to satisfy the needs of capacity and coverage, using three layers denoted by L0, L1, and L2, respectively. TDs are generated so that L1 is able to satisfy the peak TD at any time and location, providing the minimum required capacity. To capture the spatio-temporal variability, each L1 BS is setup to experience the peak TD at different times, as shown in Fig. 5 for three L1 BSs.

The previous formulation also has consequences in the other layers. L2 cannot satisfy all traffic demands, since it has less BSs than L1. Consequently, L2 is meant for coverage. On the other hand, L0 has excess capacity to satisfy all TDs. As such, it is meant to enhance capacity. Therefore, this formulation allows to capture operators' objectives of satisfying and enhancing capacity and coverage.

In Table 2, the BS components and parameters are listed. We consider the BSs in L0 require no A/C due to a low expected value of power consumption. The interconnection among these elements is shown in Fig. 6. E_{RF} is the output energy of the feeder. The energy received at the input of the feeder comes from the PA, whose power is mainly drawn from the PS. The BBuP energy is also provided by the PS. The A/C needs to compensate the energy dissipated by all the previous elements, and the energy it needs is drawn from the PS. Based on these relations and the energy models, we obtain the energy used by each BS.

For the path loss, we used the following 3GPP models for heterogeneous networks in outdoors (distance D (in km))[33]:

$$(\text{macro})PL = 128.1 + 37.6 \log(D), \tag{35}$$



(a) Average Load for BS1 in layer 1 (b) Average Load for BS2 in layer 1 (c) Average Load for BS3 in layer 1

Fig. 5. Average load for BSs in layer 1.

TABLE 2
BS Components and Parameters, per Layer

Component	P_{off}(W)	$P_{\text{on,min}}$(W)	α	$P_{\text{on,max}}$(W)
Feeder	[0,0,0]	[0,0,0]	[0,0,1]*0.63	-
PA	[1,1,1]*0.25	[0.25,1,4]	[1,1,2]*0.2	-
BBuP	[1,1,1]	[2,2,4]	-	[1,1,2]*6
A/C	[-,1,1]*0.9	[-,1,1]*0.9	[-,3,3]	-
PS	[1,1,1]*0.1	[1,1,1]*0.1	[1,1,1]*0.9	-

$$(\text{pico})PL = 140.7 + 36.7\log(D). \tag{36}$$

5.2 Simulation Results

The amount of energy savings that can be achieved compared to an always-on network depends on the specific layout of the network, locations, and traffic demands. First, we present the results obtained from a single scenario generated with the previous parameters. Then, we provide values for average achievable energy savings across multiple scenarios. To obtain an upper bound for the performance of our algorithm we relax the energy minimization problem to a fractional-Knapsack one, for which a greedy algorithm provides an optimal solution. The performance of our algorithm can be no better than that obtained from the relaxed problem. To obtain a lower bound, we consider the cases where our algorithm only applies actions of Type 1 or Type 2.

The results of applying our energy-saving scheme to a scenario generated with the parameters previously described are shown in Figs. 7 and 8. In Fig. 7, we observe that when our algorithm considers actions of Type 1 and Type 2 it achieves energy savings between 34 and 39 percent throughout a day, with a gap of less than 2 percent from the energy savings of the fractional-Knapsack upper bound. Even though the performance of our algorithm is almost the same when only actions of Type 2 are considered, it definitely outperforms the case where only actions of Type 1 are considered. In addition, we also observe that the achievable energy savings vary inversely to the overall network load. This behavior can be explained as follows. During low network load periods, the number of underutilized BSs is generally greater than during high network load periods. Therefore, during low network load periods, there is a higher probability that BSs can be turned off and, thus, the additional energy savings described by Eq. (26) can be achieved.

The effect of our energy-saving algorithm on the load, activity, and the energy consumption across the different layers is depicted in Fig. 8. We define the activity of layer i during time interval Δt_j as

$$\text{Activity}(i, \Delta t_j) = \frac{\sum_{b \in a_i} \zeta(b, \Delta t_j)}{|a_i|}, \tag{37}$$

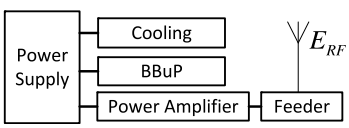


Fig. 6. Interconnections of BS components.

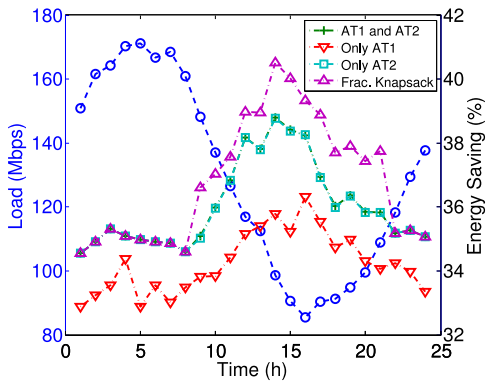


Fig. 7. Total network load and energy savings throughout a day.

where $\zeta(b, \Delta t_j)$ is an indicator function equal to 1 when BS b is *on* during time interval Δt_j, and $|a_i|$ denotes the total number of BSs in layer a_i.

L0 handled between 62 to 93 percent of the traffic throughout the day, while consuming between 64 to 80 percent of the total network energy. Even though the larger cells (i.e., the ones of BSs in L1 and L2) in general did not handle as much traffic as the ones of L0, the amount of traffic that they managed increased during the time interval 9-21 h, i.e., when the network load was low-to-medium. We also note that the level of activity for L1 and L2 tends to be fairly stable during time intervals of at least 3 hours and, even though the level of activity of L0 shows greater variations, these variations tend to be relatively low. In addition, we observe that during the time interval 10-17 h, which corresponds to the highest energy savings in Fig. 7, the activity of L0 experiences a significant drop (i.e., L0 BSs are turned off) and the activity of L2 increases (i.e., L2 BSs are turned on). This behavior suggests that the high energy savings are highly related to the de-activation of underutilized L0 BSs.

To further explore the effect of our algorithm on the cell-association policy, we obtained the locations-to-layer association, as shown in Fig. 8d. We observe that the increment of

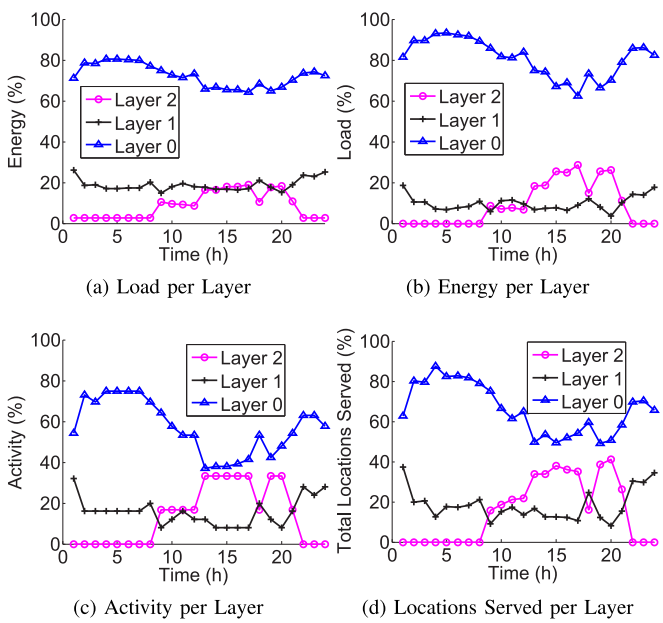


Fig. 8. Load, energy, activity, and locations served per layer.

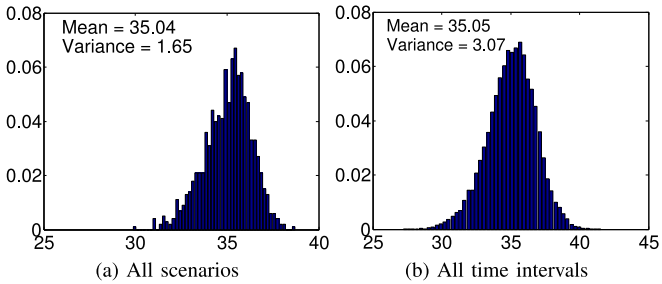


Fig. 9. Probability density functions of energy savings (percent).

the load managed by L2 in the time interval 9-21 h (as identified from Fig. 8b) results from L2 becoming the serving layer for an increased number of locations. On the other hand, the fairly stable activity of L1 in the time intervals 2-7 h and 13-17 h (as identified from Fig. 8b) still corresponds to changes in the cell-association policies during these intervals, as observed from Fig. 8d. In general, variations in the cell-association policy during the times of stable activity also occur in the other layers.

To evaluate the overall performance of our energy-saving algorithm, we applied it to 1,000 different scenarios generated using the parameters shown in Tables 1 and 2. Fig. 9 shows the probability density function (PDF) for the energy savings. Particularly, Fig. 9a shows the PDF of the daily energy savings across all scenarios. In this case, the mean is 35.04 percent and the variance is 1.65. To further analyze how the energy savings behave throughout a day for the different scenarios, we obtained the PDF for the energy savings across all time intervals of all scenarios, as shown in Fig. 9b. In this case, the mean is 35.05 percent and the variance is 3.07. Even though the variance in this case is higher by 1.42 than the one of Fig. 9a, the energy savings only differ by 0.01 percent. This result indicates that the mean energy savings across the time intervals do not experience wide fluctuations compared to the mean daily energy savings across multiple scenarios.

For the same scenarios used in Fig. 9, we obtained the PDF of the absolute optimality gap between our algorithm and the fractional-Knapsack relaxation, as shown in Fig 10. Fig. 10a shows that the optimality gap is less than 3 percent of energy savings across the scenarios. Across all time intervals, we observe that the optimality gap is less than 8 percent of energy savings, and mostly within 2 percent of energy savings.

For the same scenarios used in Fig. 9, we also obtained the PDF of the improvements in energy savings of our

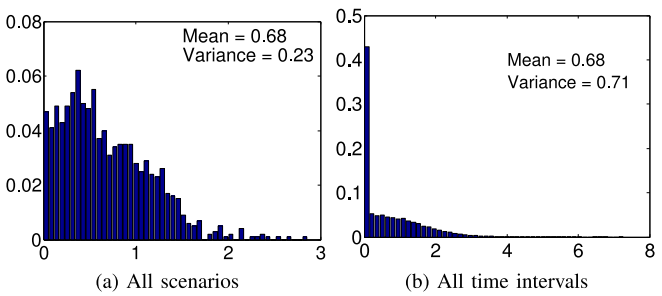


Fig. 10. Probability density function of the absolute optimality gap for energy savings (percent).

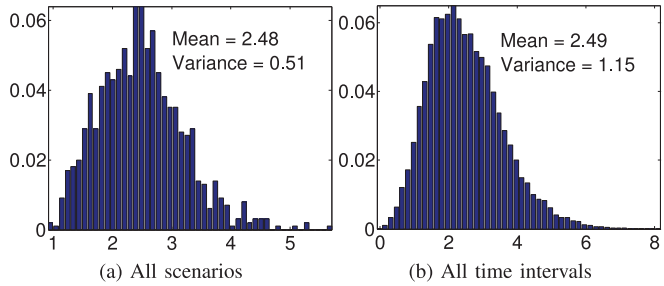


Fig. 11. Probability density function of the absolute improvements in energy savings (percent) compared to considering only actions of Type 1.

algorithm when both action types are considered compared to only considering actions of Type 1, as shown in Fig. 11. As observed from Figs. 11a and 11b, accounting for both types of actions provides improvements of up to 5 and 8 percent across all scenarios and all time intervals, respectively, compared to only considering actions of Type 1.

Across all scenarios, we also obtained the per-layer PDF for the energy, load, and activity percent, as shown in Fig. 12. We define the activity of a layer i as

$$\text{Activity}(i) = \frac{\sum_{\Delta t_j} \sum_{b \in a_i} \zeta(b, \Delta t_j)}{\sum_{\Delta t_j} |a_i|}, \tag{38}$$

where $\zeta(b, \Delta t_j)$ is an indicator function equal to 1 when BS b is *on* during time interval Δt_j, and $|a_i|$ denotes the total number of BSs in layer a_i. For these PDFs, Fig. 13 summarizes the mean and the variance. In terms of the mean values shown in Fig. 13a, we observe that L0 has an activity of around 50 percent but manages 75 percent of the network load, while consuming no more than 70 percent of the overall network energy. On the other hand, L1 shows an almost direct relation between its activity, energy consumption, and managed load. In contrast to L0, L2 had activity of around 15 percent, but only consumed 8 percent of the energy and handled 6 percent of the network load. These results highlight the relevance of small cells (i.e., L0) and the non-trivial role of larger cells (i.e., L1 and L2) in achieving energy-efficient networks.

In terms of the variance, Fig. 13b shows that the variance corresponding to the energy consumption is similar across layers and relatively low compared to the ones of the activity and the load. The high values of the latter indicate that the activity and the load per layer are highly dependent on the particular IbHWS under consideration. On the other hand, the low variance for the energy consumption indicates a low dependence of the energy consumption per layer on the particular IbHWS under consideration.

6 Conclusion

Small cells and HetNets play a key role in current and future cellular systems. Nevertheless, the energy consumption in such systems has become a key concern due to its economic and environmental impact. Even though much work has been done in recent years to develop techniques to minimize the energy consumption, existing approaches have not captured important factors that greatly affect the

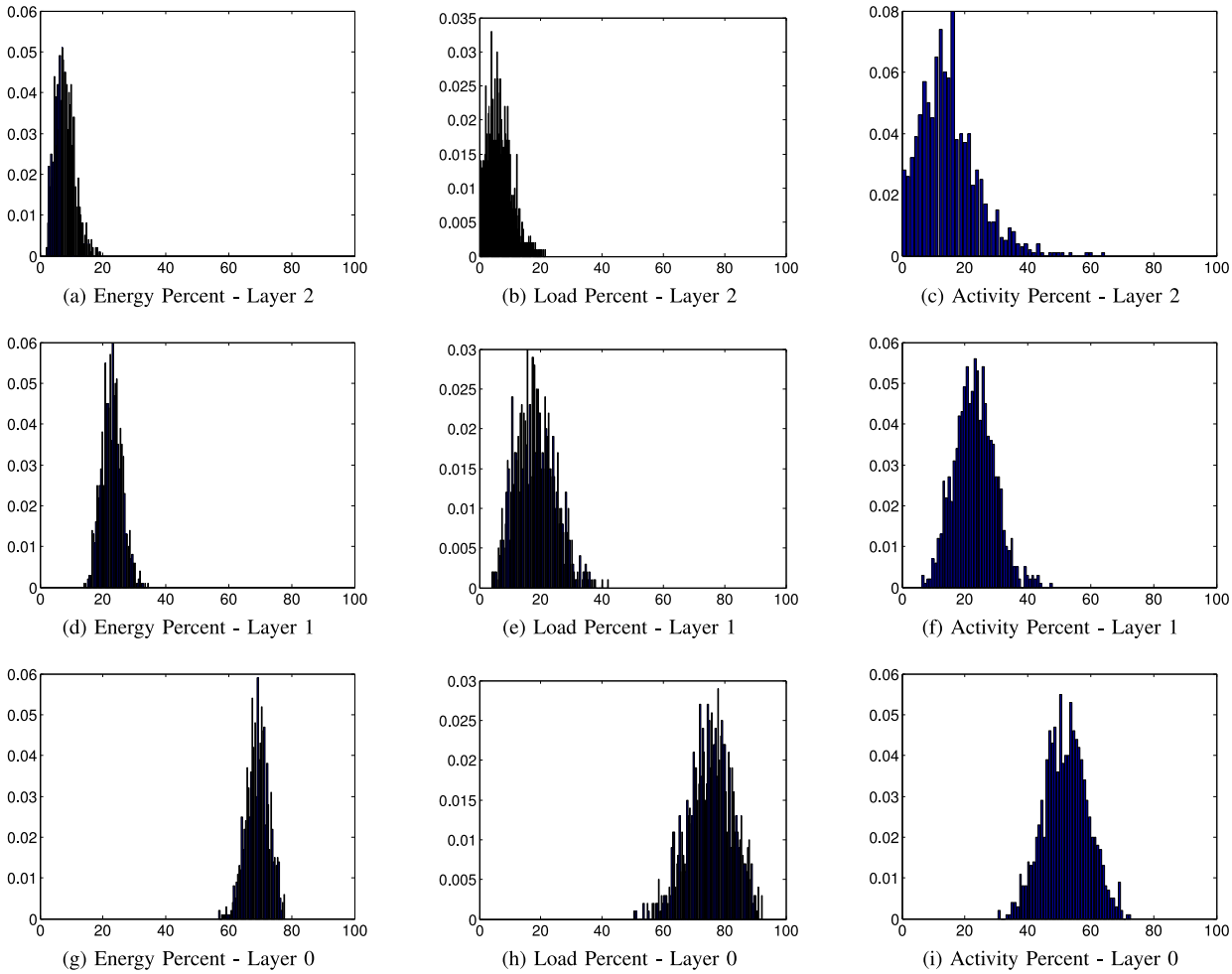
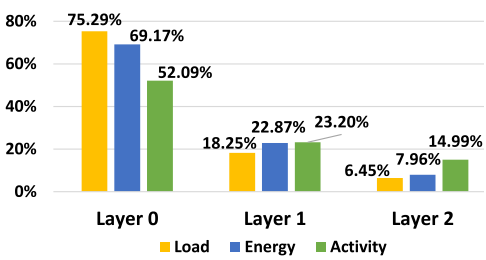
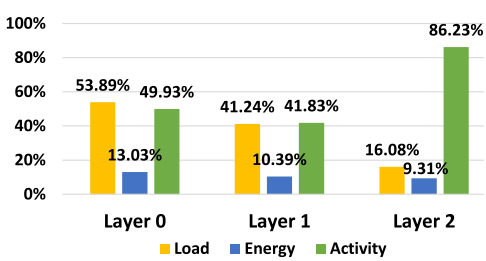


Fig. 12. Per-Layer probability density functions of energy, load, and activity.

actual energy consumption in a network. In this paper, we have analyzed the energy consumption in HetNets while accounting for such factors. Specifically, we considered heterogeneous networks composed of multiple layers, as well as the dependence of the energy consumption on the spatio-temporal variations of traffic demands and on the internal hardware components of the BSs. Considering these elements, we studied the energy minimization problem in HetNets, showed a mapping to a 0-1 Knapsack-like problem, and identified the conditions required for a direct mapping to the classical 0-1 Knapsack problem. Due to the important differences between the general energy minimization problem and the classical 0-1 Knapsack problem, we developed an efficient algorithm to minimize the energy consumption by adjusting the cell-association and BS on-off policies. We showed that our algorithm, while directly applicable to the two-layer case, is extendable to the m-layer case. We evaluated the performance of the proposed algorithm and obtained energy savings of 35 percent across a wide range of scenarios. Throughout the day, these energy savings, in general, vary inversely to the overall network load. Furthermore, we identified that small cells play a key role not just in improving the capacity, but also in increasing the energy efficiency of the network by consuming 69 percent of the energy while handling 75 percent of the traffic. However, contrary to the belief that small cells are always the most energy-efficient solution, we found that larger cells have an important role in energy-efficient deployments, particularly during low and medium network



(a) Mean Value for Activity, Energy and Load Percent



(b) Variance Value for Activity, Energy and Load Percent

Fig. 13. Per-layer statistics.

load periods. During such periods, the energy consumption of the small cells and their underutilization calls for their de-activation and the utilization of larger cells, instead. More importantly, we have shown that the cell-association and on-off policies should be jointly adjusted according to the actual network deployment, the energy efficiency of the BSs, and the traffic dynamics in order to achieve a highly energy-efficient network operation.

ACKNOWLEDGMENTS

This work was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, under grant No. (3-611-35-HiCi). The authors, therefore, acknowledge technical and financial support of KAU. In addition, this work was partly supported by the Georgia Research Alliance (GRA) and the Secretaría Nacional de Ciencia y Technología (SENACYT), Republic of Panama.

REFERENCES

[1] Ericsson, "Ericsson mobility report," Ericsson, Stockholm, Sweden, Jun. 2013, http://www.ericsson.com/res/docs/2013/ericsson-mobility-report-june-2013.pdf
[2] 3GPP, "Overview of 3GPP Release 8," 3GPP, Sep. 2013.
[3] I. F. Akyildiz, D. Gutierrez-Estevez, and E. Chavarria Reyes, "The evolution to 4G cellular systems: LTE-advanced," *Phys. Commun. J.*, vol. 3, no. 4, pp. 217–244, Dec. 2010.
[4] I. F. Akyildiz, D. M. Gutierrez-Estevez, R. Balakrishnan, and E. Chavarria-Reyes, "LTE-advanced and the evolution to beyond 4G (B4G) systems," *Phys. Commun. J.*, vol. 10, pp. 31–60, 2014.
[5] T. E. Klein, "GreenTouch consortium: Transforming ICT networks for a sustainable future," Jan. 2013.
[6] Gartner, "Gartner estimates ICT industry accounts for 2 percent of global CO2 emissions," Apr. 2007.
[7] The Boston Consulting Group, "GeSI SMARTer 2020: The role of ICT in driving a sustainable future," Global e-Sustainability Initiative, 2012, http://gesi.org/assets/js/lib/tinymce/jscripts/tiny_mce/plugins/ajaxfilemanager/uploaded/SMARTer%202020%20-%20The%20Role%20of%20ICT%20in%20Driving%20a%20Sustainable%20Future%20-%20December%202012.pdf
[8] K. Chen and D. Peroulis, "Design of adaptive highly efficient GaN power amplifier for octave-bandwidth application and dynamic load modulation," *IEEE Trans. Microw. Theory Tech.*, vol. 60, no. 6, pp. 1829–1839, Jun. 2012.
[9] H. K. Boyapati, R. Rajakumar, and S. Chakrabarti, "Quantifying the improvement in energy savings for LTE enodeb baseband subsystem with technology scaling and multi-core architectures," in *Proc. Nat. Conf. Commun.*, Feb. 2012, pp. 1–5.
[10] M. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2009, pp. 1–5.
[11] F. Richter, A. Fehske, P. Marsch, and G. Fettweis, "Traffic demand and energy efficiency in heterogeneous cellular mobile radio networks," in *Proc. IEEE Veh. Technol. Conf.*, May 2010, pp. 1–6.
[12] R. Litjens and L. Jorguseski, "Potential of energy-oriented network optimisation: Switching off over-capacity in off-peak hours," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2010, pp. 1660–1664.
[13] Y. Qi, M. Imran, and R. Tafazolli, "Energy-aware adaptive sectorisation in LTE systems," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2011, pp. 2402–2406.
[14] G. Auer, V. Giannini, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, C. Desset, and O. Blume, "Cellular Energy Efficiency Evaluation Framework," in *Proc. IEEE Veh. Technol. Conf.*, May 2011, pp. 1–6.
[15] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
[16] F. Han, Z. Safar, and K. Liu, "Energy-efficient base-station cooperative operation with guaranteed QoS," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3505–3517, Aug. 2013.
[17] D. H. Ring, "Mobile telephony-wide area coverage-case 20564," Bell Telephone Laboratories Incorporated, Murray Hill, NJ, Tech. Memoranda, Dec. 1947.
[18] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[19] D. Cao, S. Zhou, and Z. Niu, "Optimal base station density for energy-efficient heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2012, pp. 4379–4383.
[20] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1525–1536, Sep. 2011.
[21] Y. S. Soh, T. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.
[22] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. Gonzalez, H. Klessig, I. Godor, M. Olsson, M. Imran, A. Ambrosy, and O. Blume, "Flexible power modeling of LTE base stations," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2012, pp. 2858–2862.
[23] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.
[24] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Apr. 2011, pp. 882–890.
[25] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "alpha-optimal user association and cell load balancing in wireless networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Mar. 2010, pp. 1–5.
[26] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.
[27] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
[28] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5g of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, May 2014.
[29] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.
[30] E. Chavarria-Reyes and I. F. Akyildiz, "Radio access network energy minimization in multi-layer heterogeneous wireless systems," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2013, pp. 3259–3263.
[31] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); Radio measurement collection for minimization of drive tests (MDT); Overall description; Stage 2 ," Tech. Specification 37.320, Sep. 2014.
[32] 3GPP, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3," Tech. Specification 24.301, Mar. 2013.
[33] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," Tech. Rep. 36.814, Mar. 2010.

Elias Chavarria-Reyes received the BE degree in electronics and communication engineering from the Universidad de Panamá, Ciudad de Panamá, Panamá, in 2007, and the MS degree in electrical and computer engineering from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, in May 2010. He is currently working toward the PhD degree with the Broadband Wireless Networking Lab, Georgia Institute of Technology, under the supervision of Prof. Ian F. Akyildiz, with a fellowship of "SENACYT." His current research is focused on energy modeling and analysis in heterogeneous wireless systems. He is a member of the IEEE.

Ian F. Akyildiz received the BS, MS, and PhD degrees in computer engineering from the University of Erlangen-Nrnberg, Germany, in 1978, 1981, and 1984, respectively. Currently, he is the Ken Byers chair professor in telecommunications with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, the director in the Broadband Wireless Networking Laboratory and the chair in the Telecommunication Group, Georgia Tech. He is an honorary professor with the School of Electrical Engineering at Universitat Politècnica de Catalunya (UPC), Barcelona, Catalunya, Spain, and the founder of N3Cat (NaNoNetworking Center in Catalunya). Since 2011, he has been a consulting chair professor in the Department of Information Technology, King Abdulaziz University (KAU) in Jeddah, Saudi Arabia. Since January 2013, he has also been a FiDiPro professor (Finland Distinguished Professor Program (FiDiPro) supported by the Academy of Finland) at Tampere University of Technology, Department of Communications Engineering, Finland. He is the editor-in-chief of *Computer Networks (Elsevier) Journal*, and the founding editor-in-chief of the *Ad Hoc Networks (Elsevier) Journal*, the *Physical Communication (Elsevier) Journal*, and the *Nano Communication Networks (Elsevier) Journal*. He received numerous awards from IEEE and ACM. His current research interests are in nano networks, Long Term Evolution (LTE) advanced networks, cognitive radio networks and wireless sensor networks. He is a fellow of the IEEE from 1996 and a fellow of the ACM from 1997.

Etimad Fadel received the BS degree in computer science from King Abdulaziz University (KAU), Saudi Arabia, in 1994 and the Mphil/PhD degrees in computer science from De Montfort University (DMU), United Kingdom, in 2007. Currently, she is an assistant professor with the Computer Science Department, KAU, Saudi Arabia. Her research interests include wireless networks, Internet of things, Internet of nano-things, sensor networks, and LTE-A cellular systems.