Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks

Kyuho Son, *Member, IEEE*, Hongseok Kim, *Member, IEEE*, Yung Yi, *Member, IEEE* and Bhaskar Krishnamachari, *Member, IEEE*

***Abstract*—Energy-efficiency, one of the major design goals in wireless cellular networks, has received much attention lately, due to increased awareness of environmental and economic issues for network operators. In this paper, we develop a theoretical framework for BS energy saving that encompasses dynamic BS operation and the related problem of user association together. Specifically, we formulate a total cost minimization that allows for a flexible tradeoff between flow-level performance and energy consumption. For the user association problem, we propose an optimal energy-efficient user association policy and further present a distributed implementation with provable convergence. For the BS operation problem (i.e., BS switching on/off), which is a challenging combinatorial problem, we propose simple greedy-on and greedy-off algorithms that are inspired by the mathematical background of submodularity maximization problem. Moreover, we propose other heuristic algorithms based on the distances between BSs or the utilizations of BSs that do not impose any additional signaling overhead and thus are easy to implement in practice. Extensive simulations under various practical configurations demonstrate that the proposed user association and BS operation algorithms can significantly reduce energy consumption.**

***Index Terms*—Energy-efficient, cellular system, user association, base station operation, flow-level dynamics, greedy algorithms;**

I. Introduction

ENERGY-efficient design of wireless networks has received significant attention for decades with emphasis on prolonging battery life-time for sensor nodes and mobile terminals [1]. More recently, there has been a renewed focus on energy-efficiency in wireless networks from the different perspectives of reducing the potential harms to the environment caused by CO_2 emissions (e.g., global warming) and the depletion of non-renewable energy resources. Reducing energy consumption has also economic impact on revenue, e.g., the wireless network operators are estimated to spend more than 10 billion dollars for electricity [2], a significant portion of their operational expenditure (OPEX). Energy consumed by ICT (Information and Communication Technology) industry is rising at 15-20% per year, doubling every five years [3].

Pushed by such needs of energy reduction, the operators have been seeking ways to improve energy-efficiency in all available dimensions and all components across base stations (BSs), mobile terminals (MTs) [4]–[6], and backhaul networks [7]. The focus of this paper is on reducing the energy consumption in BSs, since BSs use a significant portion of energy used in cellular networks, reported to amount to about 60-80% [8]. Energy reduction in BSs is achieved in various ways: novel hardware designs (e.g., adopting energy-efficient power amplifiers [9] and fanless cooler, or even cooler based on natural resources [10]), resource management scheme (e.g., power control [11]), smart topological designs [12]–[14] from deployment to operation (e.g. using relay, cite optimization or dynamic switching), etc.

In this paper, we study dynamic load-aware on/off operation of BSs. BSs are typically deployed and operated on the basis of peak traffic volume and also stayed turned-on irrespective of traffic load. Recent research based on the real temporal traffic trace over one week [2] reports that BSs are largely underutilized; the time portion when the traffic is below 10% of peak during the day is about 30% and 45% in weekdays and weekends. However, even BSs with small or no activity consume more than 90% of its peak energy (e.g., a typical UMTS BS consumes 800-1500W and has an RF output power of 20-40W). Therefore, instead of turning off just radio transceivers, *dynamic BS operations*, which allow the system to entirely turn off some underutilized BSs and transfer the imposed loads to neighboring BSs during low traffic periods such as nighttime, has substantial potential to reduce the energy waste.

Turning on/off of BSs must be coupled with *user association*: when a set of active BSs changes, a MT may need to be associated with a new BS. This coupling makes solving the problem more challenging. To decouple user association and dynamic BS on/off operation and purely focus on each of problems, we make a reasonable assumption of *time-scale separation* such that user association is determined at a much faster time-scale than that of dynamic BS operation. Under this assumption, we decompose our problem into two sub-

Manuscript received 1 October 2010; revised 15 February 2011. This research was supported in part by National Science Foundation NetSE grant 1017881 and the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA-2011-09913-04005).

K. Son and B. Krishnamachari are with the Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 (e-mail: kyuhoson@usc.edu; bkrishna@usc.edu).

H. Kim is with the Department of Electronic Engineering, Sogang University, Seoul, Korea (e-mail: hongseok@ieee.org).

Y. Yi is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea (e-mail: yiyung@kaist.edu).

Digital Object Identifier 10.1109/JSAC.2011.110903.

problems. This assumption can be confirmed by measurement data in real networks, and also follows our intuition that the operator will apply dynamic BS on/offs depending on the traffic load, e.g., every a couple of hours or even less frequently, whereas user association is decided over a much finer time granularity.

There have been many studies [15]–[22] on user association policies under different models and assumptions. They basically considered two metrics for selecting the serving BS, (i) received signal quality (pilot signal strength, SINR or corresponding achievable rate) and (ii) cell traffic load. It should be noted that the user association policy proposed in this paper, which is both energy and load aware, is a generalized version of [22], which was focused on load balancing only. Specifically, we further consider the operational power consumption of BSs as the user association metric so that MTs are likely to be associated with energy-efficient BSs. Recently, several papers studied energy conservation effect through BS on/off [8], [23]–[25]. Luca *et al*. [23] showed the possibility of energy saving by simulations. Marsan *et al*. [8] presented a predefined BS sleep pattern for a deterministic traffic profile over one day. Dynamic BS switching algorithms [24], [25] were also proposed under the setting where traffic patterns are not predictable from day to day. In addition, the concept of BS sharing was introduced in [26], where different operators pool their BSs together to further conserve energy.

We aim at developing algorithmic solutions based on *flow-level dynamics* [27] where new file transfers are initiated at random and leave the system after being served over time. Despite analytical challenges of this dynamic model, this model seems to be more practical than a static model with a fixed set of backlogged users which was studied in many previous papers [18]–[20], [24]. Moreover, our model considers spatially inhomogeneous traffic distribution and also capture signal degradation due to being served by further BSs when a previously associated BS is switched off. Such a dynamic model also enables us to investigate the tradeoff between flow-level performance and energy consumption.

The main contributions of this paper are summarized as follows:

1) First, we develop a theoretical framework for BS energy saving that considers dynamic BS operation and the related problem of user association jointly. Specifically, we formulate a total cost minimization problem that allows for a flexible tradeoff between flow-level performance and energy consumption by adjusting a single parameter η that will be defined later. In addition, our cost function for energy consumption are general enough to encompass several types of BSs, from energy-proportional BSs to non-energy-proportional BSs and, in the extreme, constant energy consumption BSs as well.
2) Second, we decompose our general problem into two subproblems, (i) energy-efficient user association and (ii) energy-efficient BS operation, and tackle these problems one by one. For the user association problem, we propose an optimal energy-efficient user association policy and further present a distributed implementation with provable convergence irrespective of the initial condition. For the BS operation problem (i.e., BS switching on/off), finding the optimal set of active BSs, is very difficult because it is a challenging combinatorial problem where the number of possible cases exponentially increases as the total number of BSs increases. In order to overcome the prohibitive complexity, we propose simple greedy-on (GON) and greedy-off (GOFF) algorithms that are inspired by the mathematical background of submodularity maximization problem. We further propose other operator-friendly heuristic algorithms (GON-Dist, GOFF-Dist and GOFF-Util) based on the distances between BSs or the utilizations of BSs that do not impose any additional signaling overhead on MTs and thus are easy to implement in practice. We find that GOFF-Util offers comparable performance to GON and GOFF.
3) Third, through extensive simulations under various practical configurations, we have obtained many interesting results: (i) The proposed energy-efficient user association and BS operation algorithms can significantly reduce the energy consumption by up to 70-80%, in which the amount of energy saving depends on the arrival rate of traffic and its spatial distribution as well as the density of BS deployment. For example, we can obtain large energy savings in the case of low traffic loads and/or in the urban and suburban environments, but almost no or low energy saving in the rural environment. (ii) When the portion of fixed power consumption for BSs is very small (i.e., for more energy-proportional BSs), the energy saving mostly comes from the proposed user association algorithm that allows MTs to select energy-efficient BSs. However, when the portion is large, which is a reasonable assumption for BSs in practice today, much larger energy savings can be obtained by turning off some underutilized BSs.

The remainder of this paper is organized as follows. In Section II, we formally describe our system model and general problem. In Section III, we propose an energy-efficient user association policy and present its distributed implementation. In Section IV, we propose several greedy BS operation algorithms. In Section V, we provide extensive simulation results followed by conclusion in Section VI.

II. System Model and Problem Formulation

A. System Model

We consider an infrastructure-based wireless network with multiple BSs, where our focus is on downlink communication, i.e., from BSs to MTs. We consider a region $\mathcal{L} \subset \mathbb{R}^2$ served by a set of BSs $\mathcal{B}$. Let $x \in \mathcal{L}$ denote a location and we use $i \in \mathcal{B}$ to index a typical i-th BS. We assume that file transfer requests arrive following an inhomogeneous Poisson point process with arrival rate per unit area $\lambda(x)$ and file sizes which are independently distributed with mean $1/\mu(x)$ at location $x \in \mathcal{L}$, so the traffic load density is defined as $\gamma(x) \doteq \frac{\lambda(x)}{\mu(x)} < \infty$. This captures spatial traffic variability. For example, a hot spot can be characterized by a high arrival rate and/or possibly large file sizes.

When the set of BSs $\mathcal{B}_{\text{on}}$ is turned on, the transmission rate of a user located at x and served by BS $i \in \mathcal{B}_{\text{on}}$ is denoted by $c_i(x, \mathcal{B}_{\text{on}})$. For analytical tractability, we assume

that $c_i(x, \mathcal{B}_{\text{on}})$ does not change over time, i.e., we do not consider fast fading or dynamic inter-cell interferences. Instead, $c_i(x, \mathcal{B}_{\text{on}})$ can be considered as an *time-averaged* transmission rate. This assumption is reasonable in the sense that the time scale of user association is much larger than the time scale of fast fading or dynamic inter-cell interferences. Hence, the inter-cell interference is considered static Gaussian-like noise, which is feasible under interference randomization or fractional frequency reuse [18], [19], [28]. This static inter-cell interference model has also been adopted in previous load-balancing work as well [18], [19], [28]. It should be noted, however, that $c_i(x, \mathcal{B}_{\text{on}})$ is *location-dependent* but not necessarily determined by the distance from the BS i. Hence, $c_i(x, \mathcal{B}_{\text{on}})$ can capture shadowing effect as well.

The *system-load density* $\varrho_i(x, \mathcal{B}_{\text{on}})$ is then defined as $\varrho_i(x, \mathcal{B}_{\text{on}}) \doteq \frac{\gamma(x)}{c_i(x, \mathcal{B}_{\text{on}})}$, which denotes the fraction of time required to deliver traffic load $\gamma(x)$ from BS $i \in \mathcal{B}_{\text{on}}$ to location x. We further introduce a routing function $p_i(x)$, which specifies the *probability* that a flow at location x is associated with BS i. Intuitively, $p_i(x)$ can be interpreted as the time fraction that a flow arrived at location x is routed to BS i. We will see that, however, the optimal $p_i(x)$ will be either 1 or 0, i.e., deterministic routing (or user association) is the solution of our optimization problem, which is defined in (17).

Definition 2.1 (Feasibility): The set $\mathcal{F}(\mathcal{B}_{\text{on}})$ of *feasible* BS loads (or utilization) $\rho = (\rho_1, \cdots, \rho_{|\mathcal{B}|})$ when the set of BSs $\mathcal{B}_{\text{on}} \subseteq \mathcal{B}$ is turned on, is given by

$$\mathcal{F}(\mathcal{B}_{\text{on}}) = \Big\{ \rho \mid \rho_i = \int_{\mathcal{L}} \varrho_i(x, \mathcal{B}_{\text{on}}) p_i(x) dx, \forall i \in \mathcal{B}, \quad (1)$$

$$0 \le \rho_i \le 1 - \epsilon, \forall i \in \mathcal{B}, \quad (2)$$

$$\sum_{i \in \mathcal{B}} p_i(x) = 1, \quad (3)$$

$$0 \le p_i(x) \le 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L}, \quad (4)$$

$$p_i(x) = 0, \ \forall i \in \mathcal{B} \backslash \mathcal{B}_{\text{on}}, \forall x \in \mathcal{L} \Big\}, \quad (5)$$

where ϵ is an arbitrarily small positive constant[1]. Hence, the feasible BS loads ρ has the associated routing probability vector $p(x) = (p_1(x), \cdots, p_{|\mathcal{B}|}(x))$ for all $x \in \mathcal{L}$.

Lemma 2.1: The feasible set $\mathcal{F}(\mathcal{B}_{\text{on}})$ is convex.

Proof: The proof can be done similarly as in [22] with the additional constraint (5). ■

B. Problem formulation

In this paper, we consider both the cost of flow-level performance such as file transfer delay and the cost of energy. Our problem is to find an optimal set of active BSs ($\mathcal{B}_{\text{on}}$) and BS loads ρ (i.e., user association) that minimize the total system cost function, given by

$$\min_{\mathcal{B}_{\text{on}}, \rho} \Big\{ \phi_\alpha(\rho, \mathcal{B}_{\text{on}}) + \eta \psi(\rho, \mathcal{B}_{\text{on}}) \ \Big| \rho \in \mathcal{F}(\mathcal{B}_{\text{on}}), \mathcal{B}_{\text{on}} \subseteq \mathcal{B} \Big\}, \quad (6)$$

where $\eta \ge 0$ is the parameter that balances the *tradeoff between the flow-level performance and the energy consumption*. When η is zero, we only focus on the flow-level performance, however, as η grows, more emphasis is given to energy conservation.

(i) The cost function of flow-level performance: A generalized α-delay performance function in [22] is adopted and modified such that it becomes zero at $\rho = 0$.

$$\phi_\alpha(\rho, \mathcal{B}_{\text{on}}) = \begin{cases} \displaystyle\sum_{i \in \mathcal{B}_{\text{on}}} \frac{(1-\rho_i)^{1-\alpha} - 1}{\alpha - 1}, & \alpha \ne 1 \\ \displaystyle\sum_{i \in \mathcal{B}_{\text{on}}} \log\left(\frac{1}{1-\rho_i}\right), & \alpha = 1 \end{cases} \quad (7)$$

where $\alpha \ge 0$ is a parameter specifying the desired *degree of load balancing*. When $\alpha = 0$ (rate-optimal), $\phi_0(\cdot)$ becomes $\sum_i \rho_i$. Thus, each MT prefers the BS offering the highest transmission rate to minimize the total utilization regardless of traffic load distribution. When $\alpha = 2$ (delay-optimal), $\phi_2(\cdot)$ becomes to $\sum_i \frac{\rho_i}{1-\rho_i}$, which is equal to the total number of flows in the system if we consider the system as M/GI/1 multi-class processor sharing system [29]. From Little's law, minimizing $\phi_2(\cdot)$ is equivalent to minimizing the average delay experienced by a typical flow. As α further increases, $\phi_\alpha(\cdot)$ places more emphasis on the traffic loads rather than the transmission rate. Especially, when α goes to infinity (load-equalizing), we try to minimizes the maximum utilization, which equalizes the utilization of all the BSs.

(ii) The cost function of energy: We newly introduce a general model for BSs consisting of two types of power consumptions: fixed power consumption and adaptive power consumption that is proportional to BS's utilization.

$$\psi(\rho, \mathcal{B}_{\text{on}}) = \sum_{i \in \mathcal{B}_{\text{on}}} \Big[(1 - q_i) \rho_i P_i + q_i P_i \Big], \quad (8)$$

where $q_i \in [0, 1]$ is the portion of the fixed power consumption for BS i, and P_i is the maximum operational power of BS i when it is fully utilized, i.e., $\rho_i = 1$, which includes power consumptions for transmit antennas as well as power amplifiers, cooling equipment and so on. When $q_i = 0$, BSs are assumed to consist of only energy-proportional devices. Such BSs would ideally consume no power when idle, and gradually consume more power as the activity level increases. This type of BSs will be referred to as *energy-proportional* BS.

However, this energy-proportional BS is still far from reality because several devices in the BSs dissipate standby power while inactive. As an example, a class-A amplifier, which is a typical power amplifier for macro BSs and one of the most power consuming devices in BSs, has the maximum theoretical efficiency of 50% [30]. This type of BSs, which consume the fixed power irrespective of its activity unless they are totally turned off, i.e., $q_i > 0$, will be referred to as *non-energy-proportional* BS. It is worthwhile mentioning that small BSs such as micro, pico and femto BSs may have smaller values of q_i than that of macro BSs because they do not usually have either a big power amplifier or a cooling equipment. Note that when $q_i = 1$, our model can also capture a constant energy consumption model, which is widely used in many literatures [2], [8], [23]–[25].

[1]It should be noted that ϵ is required to make $\mathcal{F}(\mathcal{B}_{on})$ a compact set on which our objective function (6) is defined. This is essential for the existence of the fixed point, which is proven to be the solution of our optimization problem.

C. Our Approach

Solving the general problem in (6) is very challenging due to highly complex coupling of BS operation and user association. For analytical tractability, we shall make an assumption on time-scale separation that flow arrival and departure process and the corresponding user association process are much faster than the period on which the set of active BSs are determined. From many measurement data in real networks [2], [8], [24], the traffic pattern clearly varies over time (as well as space), but could be assumed almost constant during a certain period of time, e.g., one hour. Since the time-scale for determining the set of active BSs is similar to the order of traffic-pattern changing, it is definitely much larger than that of flow arrival and departure process, e.g., typically less than several minutes.

Under this assumption, our general problem given in (6) can be decomposed into two subproblems, in which BS operation problem is solved at a slower time scale than user association problem. We will discuss each of problems in the consequent Sections III and IV.

1) User association problem: For any given set of active BSs $\mathcal{B}_{\text{on}}$, the problem in (6) reduces to the following load balancing problem by ignoring the constant fixed power consumption term $\sum_{i\in\mathcal{B}_{\text{on}}} q_i P_i$. This problem can be also interpreted as user association problem because it finds the optimal BS utilization vector ρ by determining which BS each MT should associated with.

User association problem **[P-UA]:**

$$\min_{\rho\in\mathcal{F}(\mathcal{B}_{\text{on}})} \quad \phi_\alpha(\rho,\mathcal{B}_{\text{on}}) + \eta \sum_{i\in\mathcal{B}_{\text{on}}} (1-q_i)\rho_i P_i. \tag{9}$$

2) BS operation problem: Now our focus moves to the following BS operation problem that finds the optimal set of active BSs $\mathcal{B}_{\text{on}}$ on the longer time-scale.

BS operation problem **[P-BO]:**

$$\min_{\mathcal{B}_{\text{on}}\subseteq\mathcal{B}} \quad G(\mathcal{B}_{\text{on}}) + \eta \sum_{i\in\mathcal{B}_{\text{on}}} q_i P_i, \tag{10}$$

where the function $G(\mathcal{B}_{\text{on}})$ is defined as the obtaining optimal value from the underlying user association in (9), i.e., $G(\mathcal{B}_{\text{on}}) \doteq \min_{\rho\in\mathcal{F}(\mathcal{B}_{\text{on}})} \phi_\alpha(\rho,\mathcal{B}_{\text{on}}) + \eta\sum_{i\in\mathcal{B}_{\text{on}}}(1-q_i)\rho_i P_i$.

Remark 2.1: Note that **[P-UA]** and **[P-BO]** may have conflicting interest. **[P-UA]** tries to distribute traffic loads to improve the flow-level performance $\phi_\alpha(\rho,\mathcal{B}_{\text{on}})$. On the other hand, to minimize $\sum_{i\in\mathcal{B}_{\text{on}}} q_i P_i$, **[P-BO]** tries to concentrate traffic loads to a subset of BSs $\mathcal{B}_{\text{on}}$ and turn off the other BSs.

III. Energy-efficient User Association

In this section, given the set of active BSs $\mathcal{B}_{\text{on}}$, we focus on solving **[P-UA]** in (9), i.e., associating users with BSs in an energy-efficient manner, considering load-balancing. We present an optimal user association policy and further present its online distributed implementation with provable convergence.

A. Example: A Linear Two-Cell Network

We start with a simple case to give insight into the structure of the optimal solution. Consider a linear two-cell topology consisting of one macro BS and one micro BS a distance D apart. We use subscripts M and m for the macro and micro BSs, respectively. The macro BS is assumed to have higher fixed operational power consumption than the micro BS, i.e., $(1-q_M)P_M > (1-q_m)P_m$. We further assume a simple channel model only depending on distance and a feasible spatially homogeneous traffic load with $\gamma(x)=\gamma$ for all $x\in[0,D]$. Then, the user association problem **[P-UA]** can be rewritten as the following optimal coverage division problem (i.e., $[0,R]$ and $[R,D]$ are the coverages of the macro and micro BSs):

$$\min_{R\in[0,D]} \quad \phi_\alpha(\rho,\mathcal{B}_{\text{on}}) + \eta\big[(1-q_M)\rho_M P_M + (1-q_m)\rho_m P_m\big], \tag{11}$$

where the BS utilizations, the functions of R, are given by $\rho_M = \int_0^R \frac{\gamma}{c_M(x,\mathcal{B})}dx$ and $\rho_m = \int_R^D \frac{\gamma}{c_m(x,\mathcal{B})}dx$.

We now state a lemma that as the tradeoff parameter η increases, the energy-efficient micro BS will have larger coverage.

Lemma 3.1: When $\alpha>0$, the optimal solution R^* for the problem in (11) is a decreasing function of η.

Proof: We prove the lemma by contradiction. First, we obtain the following optimality condition by taking the derivative of the objective function in (11) with respect to R:

$$\begin{aligned}&\left[\frac{1}{(1-\rho_M(R))^\alpha} + \eta(1-q_M)P_M\right]\frac{\partial\rho_M(R)}{\partial R}\\ &\quad+\left[\frac{1}{(1-\rho_m(R))^\alpha} + \eta(1-q_m)P_m\right]\frac{\partial\rho_m(R)}{\partial R} = 0,\end{aligned} \tag{12}$$

We put $\frac{\partial\rho_M(R)}{\partial R} = \frac{\gamma}{c_M(R)}$ and $\frac{\partial\rho_m(R)}{\partial R} = -\frac{\gamma}{c_m(R)}$ into (12). Also, denote by $K(R) \doteq \frac{c_m(R)}{c_M(R)}$ the ratio between transmission rates. It is clear that $K(R)$ is increasing in R. Then, we have:

$$\begin{aligned}&\eta\left[K(R)(1-q_M)P_M - (1-q_m)P_m\right]\\ &\qquad+\frac{1}{(1-\rho_M(R))^\alpha} = \frac{1}{(1-\rho_m(R))^\alpha}.\end{aligned} \tag{13}$$

Now, suppose that the optimal solution R^* is increasing in η. While the first and second terms in the left side of eq. (13) increase as R increases, the right side of eq. (13) decreases, which is the contradiction. This completes the proof. ■

B. Optimal user association policy

Let us denote the optimal BS load vector by $\rho^* = (\rho_1^*,\cdots,\rho_{|\mathcal{B}|}^*)$, i.e., solution to the problem **[P-UA]**, and further denote the optimal user association at location x by $i^*(x)$. We now present the optimality condition of the problem that describes an optimal user association policy.

Theorem 1: If the problem **[P-UA]** is feasible, then the optimal user association made by the MT located at x to join BS $i^*(x)$ is given by

$$i^*(x) = \underset{j\in\mathcal{B}_{\text{on}}}{\operatorname{argmax}} \frac{c_j(x,\mathcal{B}_{\text{on}})}{(1-\rho_j^*)^{-\alpha} + \eta(1-q_j)P_j}, \quad \forall x\in\mathcal{L}. \tag{14}$$

Its implication is as follows. When $\eta=0$, the user association is determined by the flow-level performance. However, as η grows, the decision metric is shifted to power consumption

(or energy). Note that as η goes to infinity, the decision is purely made by

$$i^*(x) = \underset{j\in\mathcal{B}_{\text{on}}}{\operatorname{argmax}} \frac{c_j(x,\mathcal{B}_{\text{on}})}{(1-q_j)P_j}, \quad \forall x \in \mathcal{L}, \tag{15}$$

which implies that the MT joins the BS that maximizes bits per joule. Considering the fact that typically $c_i(x, B_{\text{on}})$ is a logarithm function[2] of P_i, the impact of P_i in the denominator is expected to be very high in the BS selection.[3]

Proof: Now we prove that (14) is the optimal user association rule of **[P-UA]**. Let ρ^* be the optimal solution of **[P-UA]**. We first assume that the optimal load vector ρ^* is known to the MT. This assumption will be relaxed in Section III-C where a distributed user association algorithm converging to the optimal load vector ρ^* is described. **[P-UA]** is a convex optimization because its feasible set $\mathcal{F}(\mathcal{B}_{\text{on}})$ has been proved to be convex in Lemma 2.1 and the objective function is also convex (due to the summation of convex function $\phi_\alpha(\cdot)$ [22] and linear function $\eta\sum_{i\in\mathcal{B}_{\text{on}}}(1-q_i)\rho_i P_i$). Hence, it is sufficient to check the following inequality condition for optimality:

$$\left\langle \nabla\left(\phi_\alpha(\rho^*) + \eta\textstyle\sum_{i\in\mathcal{B}_{\text{on}}}(1-q_i)\rho_i^* P_i\right), \Delta\rho^*\right\rangle \geq 0, \tag{16}$$

for all $\rho \in \mathcal{F}(\mathcal{B}_{\text{on}})$ where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Then, (14) generates the *deterministic* cell coverage, and thus the association rule is given by

$$p_i^*(x) = \mathbf{1}\left\{ i = \underset{j\in\mathcal{B}_{\text{on}}}{\operatorname{argmax}} \frac{c_j(x,\mathcal{B}_{\text{on}})}{\left(1-\rho_j^*\right)^{-\alpha} + \eta(1-q_j)P_j} \right\}, \tag{17}$$

and then the inner product can be computed such as

$$\begin{aligned}
&\left\langle \nabla\left(\phi_\alpha(\rho^*) + \eta\textstyle\sum_{i\in\mathcal{B}_{\text{on}}}(1-q_i)\rho_i^* P_i\right), \Delta\rho^*\right\rangle \\
&= \sum_{i\in\mathcal{B}_{\text{on}}} \left((1-\rho_i^*)^{-\alpha} + \eta(1-q_i)P_i\right)(\rho_i - \rho_i^*) \\
&= \sum_{i\in\mathcal{B}_{\text{on}}} \left((1-\rho_i^*)^{-\alpha} + \eta(1-q_i)P_i\right)\int_{\mathcal{L}} \varrho_i(x)(p_i(x) - p_i^*(x))dx \\
&= \int_{\mathcal{L}} \gamma(x) \sum_{i\in\mathcal{B}_{\text{on}}} \frac{(1-\rho_i^*)^{-\alpha} + \eta(1-q_i)P_i}{c_i(x,\mathcal{B}_{\text{on}})}(p_i(x) - p_i^*(x))dx.
\end{aligned}$$

From (17), the following inequality,

$$\begin{aligned}
&\sum_{i\in\mathcal{B}_{\text{on}}} \frac{(1-\rho_i^*)^{-\alpha} + \eta(1-q_i)P_i}{c_i(x,\mathcal{B}_{\text{on}})} p_i(x) \\
&\geq \sum_{i\in\mathcal{B}_{\text{on}}} \frac{(1-\rho_i^*)^{-\alpha} + \eta(1-q_i)P_i}{c_i(x,\mathcal{B}_{\text{on}})} p_i^*(x)
\end{aligned} \tag{18}$$

holds because $p_i^*(x)$ in (17) is an indicator for the minimizer of $\frac{(1-\rho_i^*)^{-\alpha}+\eta(1-q_i)P_i}{c_i(x,\mathcal{B}_{\text{on}}))}$. Hence, (16) holds. ■

[2]More precisely, $c_i(x,\mathcal{B}_{\text{on}})$ is a logarithm function of BS i's transmission power based on Shannon's formula that is strongly related to operational power P_i, e.g., the higher transmission power implies the higher operational power.

[3]Note that (15) may not be able to stabilize the system even if it can be stabilized. This is because, as can be seen in (15), user association is performed without any knowledge of BS loads ρ.

C. Distributed Implementation Achieving Optimality

Now we propose an online distributed algorithm that achieves the global optimum of **[P-UA]** in an iterative manner. The distributed algorithm involves two parts.

Mobile terminal: At the start of the k-th iteration period, MTs receive BS loads $\rho^{(k)}$, e.g., through broadcast control messages from BSs.[4] Then, a new flow request for a MT located at x simply selects the BS $i(x)$ using the deterministic rule given by

$$i^{(k)}(x) = \underset{j\in\mathcal{B}_{\text{on}}}{\operatorname{argmax}} \frac{c_j(x,\mathcal{B}_{\text{on}})}{\left(1-\rho_j^{(k)}\right)^{-\alpha} + \eta(1-q_j)P_j}, \ \forall x \in \mathcal{L} \tag{19}$$

Base station: During the k-th period BSs *measure* their average utilizations after some period of time, i.e., when the system exhibits stationary performance. Then, BSs broadcast the average utilization vector $\rho^{(k+1)}$ for the next iteration.

This simple iteration provably converges to the global optimal point with a simple modification of the proof in [33].

IV. Energy-efficient BS Operation

BSs typically consume large amounts of energy in power amplifier circuit, air conditioning unit, etc, irrespective of offered loads. As a simple intuition, the ratio of the overhead power to the total power ($= q_i P_i / [q_i P_i + (1-q_i)\rho_i P_i]$) is close to 100% for small loads ρ_i. Thus, it would be definitely beneficial to turn off BSs with low activity, in conjunction with energy-efficient user association. In this section, we propose algorithms that reduce the energy consumption by solving the BS operation problem **[P-BO]** in (10) determining the set of BSs that can be switched off.

Recall that the function $G(\mathcal{B}_{\text{on}}) = \min_{\rho\in\mathcal{F}(\mathcal{B}_{\text{on}})} \phi_\alpha(\rho,\mathcal{B}_{\text{on}}) + \eta\sum_{i\in\mathcal{B}_{\text{on}}}(1-q_i)\rho_i P_i$ is obtained by the optimal user association policy in (14). The objective function in (9) is convex in ρ given $\mathcal{B}_{\text{on}}$, but becomes a nonconvex and also discontinuous function when $\mathcal{B}_{\text{on}}$ is considered as a variable. Thus, this BS operation problem is a challenging combinatorial problem with $O(2^{|\mathcal{B}|})$ possible cases, which makes it very difficult to find an optimal solution through exhaustive search, especially, when the number of BSs is large. Thus, we propose greedy-style heuristic algorithms, each of which has slightly different design rationale.

A. Greedy Turning On Algorithm for BS operation

1) Algorithm description: We first describe a greedy turning on algorithm, called GON, that iteratively finds BSs that have some benefit of delay reduction per their power usages.

Greedy on algorithm (GON)

1: Initialize $\mathcal{B}_{\text{on}} = \mathcal{B}_{\text{init}}$
2: **while** $\mathcal{B}_{\text{on}} \neq \mathcal{B}$
3: Calculate $M_{\text{GON}}(i) = \frac{G(\mathcal{B}_{\text{on}}) - G(\mathcal{B}_{\text{on}}\cup i)}{q_i P_i}, \forall i \in \mathcal{B}\setminus\mathcal{B}_{\text{on}}$
4: Find the BS $i^* = \operatorname{argmax}_{i\in\mathcal{B}\setminus\mathcal{B}_{\text{on}}} M_{\text{GON}}(i)$,
5: **if** $M_{\text{GON}}(i^*) > \eta$, **then** $\mathcal{B}_{\text{on}} \leftarrow \mathcal{B}_{\text{on}} \cup \{i^*\}$,
6: **else**, stop the algorithm.
7: **end while**

[4]IEEE 802.16m facilitates this type of message structure [31], [32].

We introduce a metric $M_{\text{GON}}(i)$ for BS i that represents the *turn-on benefit per fixed power consumption* for BS i. The GON starts with a initial set of BSs $\mathcal{B}_{\text{init}}$ and iteratively finds the best BS as a candidate among the set of inactive BSs $\mathcal{B} \setminus \mathcal{B}_{\text{on}}$ that has the highest M_{GON} (step 4). Then, the algorithm finally adds the selected BS to the list of BSs to turn on, only if its metric is greater than η (step 5), and stops otherwise. Note that the criterion $M_{\text{GON}}(i^*) > \eta$ is directly obtained from the condition that additionally turning on BS i^* is beneficial (i.e., minimizing the total system cost), given by:

$$G(\mathcal{B}_{\text{on}})+\eta \sum_{i\in\mathcal{B}_{\text{on}}} q_i P_i > G(\mathcal{B}_{\text{on}}\cup\{i^*\})+\eta\Big(\sum_{i\in\mathcal{B}_{\text{on}}} q_i P_i+q_{i^*} P_{i^*}\Big).$$

2) Design rationale: Consider the following problem that is closely related to (10):

$$\min_{\mathcal{B}_{\text{on}}\subseteq\mathcal{B}} \; G(\mathcal{B}_{\text{on}}) \;\; \text{subject to} \;\; \sum_{i\in\mathcal{B}_{\text{on}}} q_i P_i \le C, \qquad (20)$$

where we essentially move the power consumption cost in the objective function into the constraint of power consumption with some nonnegative budget C. For a given η, we can find $C = C(\eta)$[5], such that the same optimal solutions are achieved for (10) and (20), in which η is interpreted as a Lagrange multiplier of the dual formulation of (20).

We transform (20) into:

$$\max_{\mathcal{A}\subseteq\mathcal{B}\setminus\mathcal{B}_{\text{init}}} H(\mathcal{A}) \;\; \text{subject to} \;\; c(\mathcal{A}) = \sum_{i\in\mathcal{A}} c(i) \le \tilde{C}, \qquad (21)$$

where $\mathcal{A} = \mathcal{B}_{\text{on}} \setminus \mathcal{B}_{\text{init}}$, $H(\mathcal{A}) = G(\mathcal{B}_{\text{init}}) - G(\mathcal{B}_{\text{init}} \cup \mathcal{A}) = G(\mathcal{B}_{\text{init}}) - G(\mathcal{B}_{\text{on}})$, $c(i) = q_i P_i$ and $\tilde{C} = C - \sum_{i\in\mathcal{B}_{\text{init}}} c(i)$. If it can be shown that H is a non-decreasing submodular set function, then a variant greedy algorithm of GON, where the only difference lies in the stopping condition (step 5), can be shown to achieve a constant factor $(1 - 1/e)$ approximation[6] of the optimal value of (21).

Submodularity, informally, is an intuitive notion of *diminishing returns*, which states that adding an element to a small set helps more than adding that same element to a larger set. Formally, it is defined as follows.

Definition 4.1: A real-valued set function H, defined on subsets of a finite set $\mathcal{S}$ is called *submodular* if for all $\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \mathcal{S}$ and for all $s \in \mathcal{S} \setminus \mathcal{A}_2$, if it satisfies that

$$H(\mathcal{A}_1 \cup s) - H(\mathcal{A}_1) \ge H(\mathcal{A}_2 \cup s) - H(\mathcal{A}_2). \qquad (22)$$

In the appendix, we provide an affirmative simulation result although $-G$ is not exactly submodular in general, it empirically satisfies the condition of submodularity most of time. This implies that the proposed greedy algorithms work well and the constant-factor approximation is likely to hold in practice.

The choice of the initial set of active BSs $\mathcal{B}_{\text{init}}$ should be made carefully, such that the system is stable for $\mathcal{B}_{\text{init}}$. The $\mathcal{B}_{\text{init}}$ with a small number of BSs may not support the system, i.e., $\phi_\alpha(\rho, \mathcal{B}_{\text{on}})$ (and thus G as well) goes to infinity. In constructing $\mathcal{B}_{\text{init}}$, we randomly choose the first BS and sequentially add the BS that has the maximum distance to the previous set of BSs until the system can be stabilized.

[5] $C(\eta)$ is a non-increasing function of η.

[6] The submodular maximization problem (SMP) in (21) is in general a NP-hard problem. It has been proved that the greedy algorithm of SMP can achieve a constant factor $(1 - 1/e)$ approximation and its ratio is an optimal in the sense that no other polynomial algorithms with better constant approximation ratio exist. We refer the readers to [34]–[36] for the details.

B. Greedy Turning Off Algorithm for BS operation

We propose another greedy algorithm, called GOFF (Greedy Off), which can be interpreted as the opposite of GON. The GOFF, unlike GON, starts from the entire BSs $\mathcal{B}$ and finds a solution by iteratively removing the BS with the lowest *turn-off detriment per fixed power consumption*. Note that GOFF does not have the issue of choosing $\mathcal{B}_{\text{init}}$.

Greedy off algorithm (GOFF)

1: Initialize $\mathcal{B}_{\text{on}} = \mathcal{B}$
2: **while** $\mathcal{B}_{\text{on}} \neq \emptyset$
3: Calculate $M_{\text{GOFF}}(i) = \frac{G(\mathcal{B}_{\text{on}}\setminus\{i\})-G(\mathcal{B}_{\text{on}})}{q_i P_i}, \forall i \in \mathcal{B}_{\text{on}}$
4: $i^* = \arg\min_{i\in\mathcal{B}_{\text{on}}} M_{\text{GON}}(i)$
5: **if** $M_{\text{GON}}(i^*) < \eta$, **then** $\mathcal{B}_{\text{on}} \leftarrow \mathcal{B}_{\text{on}} - \{i^*\}$,
6: **else**, stop the algorithm.
7: **end while**

C. Discussion: GON and GOFF

1) Convergence: Let $i^*_{(k)}$ be the k-th turning on BS according to GON and let $\mathcal{B}^{(k)}_{\text{on}} = \mathcal{B}_{\text{init}} \cup_{j=1}^{k} \{i^{(j)}\}$. By the definition of GON, we have the following:

$$\begin{aligned} M_{\text{GON}}\big(i^{(k)}\big) &= \frac{G\big(\mathcal{B}^{(k-1)}_{\text{on}}\big) - G\big(\mathcal{B}^{(k-1)}_{\text{on}} \cup \{i^{(k)}\}\big)}{q_{i^{(k)}} P_{i^{(k)}}} \\ &> \frac{G\big(\mathcal{B}^{(k-1)}_{\text{on}}\big) - G\big(\mathcal{B}^{(k-1)}_{\text{on}} \cup \{i^{(k+1)}\}\big)}{q_{i^{(k+1)}} P_{i^{(k+1)}}}. \end{aligned} \qquad (23)$$

Due to the submodularity of $-G$ and $\mathcal{B}^{(k-1)}_{\text{on}} \subset \mathcal{B}^{(k)}_{\text{on}}$,

$$\begin{aligned} &G\big(\mathcal{B}^{(k-1)}_{\text{on}}\big) - G\big(\mathcal{B}^{(k-1)}_{\text{on}} \cup \{i^{(k+1)}\}\big) \\ \ge\;& G\big(\mathcal{B}^{(k)}_{\text{on}}\big) - G\big(\mathcal{B}^{(k)}_{\text{on}} \cup \{i^{(k+1)}\}\big) \end{aligned} \qquad (24)$$

holds. By substituting (24) into (23), it can be obtained that $M_{\text{GON}}(\cdot)$ is a monotone decreasing function, i.e.,

$$M_{\text{GOFF}}\big(i^{(k)}\big) > M_{\text{GON}}\big(i^{(k+1)}\big). \qquad (25)$$

Since we have the monotone behavior of $M_{\text{GON}}(\cdot)$ and moreover, the cardinality of $\mathcal{B}_{\text{on}}$ increases by one at each stage, GON converges to the unique solution by meeting the fixed criteria ($M_{\text{GON}}(i^*) < \eta$) or stopping condition (while $\mathcal{B}_{\text{on}} \neq \mathcal{B}$). In a similar way, we can prove the convergence of GOFF.

Even with the absence of submodularity, the proposed greedy heuristic algorithms will converge anyway because they have the fixed criteria and stopping condition. For example, in the case of GON, it stops adding a BS when the benefit becomes minus (i.e., the total system cost increases), and stops at the end (i.e., $\mathcal{B}_{\text{on}} = \mathcal{B}$) otherwise. However, the absence of submodularity might degrade the quality of solution. Nevertheless, in our case, the function $-G$ is close to a submodularity function as shown in the appendix. Therefore, the proposed algorithms achieve a good solution.

2) Interpretation: We now discuss the implication of the metric $M_{\text{GOFF}}(i)$ used in GOFF (or $M_{\text{GON}}(i)$ used in GON). Note that this metric can be interpreted as the *network-wide impact* per unit power cost. GOFF tends to choose and remove the BS that will bring the small impact on the network when turned off, whereas GON tends to choose and add the BS that will bring the large impact on the network when turned on. From the BS i's perspective, the following internal and external factors, coupled in a complex manner each other, affect the metric $M_{\text{GOFF}}(i)$[7] and the choice of the final set B_{on}.

- *Internal factors of BS i: Traffic loads* imposed on BS i is one of the dominant factors. Turning-off BS i with high utilization will cause high impact on neighboring BSs because the large amount of traffic loads needs to be transferred (or handed over) to its neighboring BSs with potentially low signal strengths.
- *External factors around BS i:* When turning-off the BS i, its network-wide impact also depends on the neighboring environment, e.g, the *number* of, the *distance* to, and the *utilization* of the neighboring BSs. As the number is small, the distance is far, and/or the utilization is high, we can expect high network-wide impact.

It should be noted that GON and GOFF require information about spatial system-load density ϱ in order to to compute their metrics. This is because G inside the metrics depends on ρ, where ρ is defined as the integral (or summation) of ϱ over the space in (1). We can obtain this information by exchanging signaling messages or use the predetermined traffic profile over a period (e.g., one day) as in [8]. In the next subsection, we propose other purely heuristic algorithms that are more *operator-friendly* in the sense that no signaling or measurement overhead is necessary, yet with possibly slight performance degradation compared to GON and GOFF.

D. Other Heuristic Algorithms

We first exploit the distances between BSs, such that BSs distant from (resp. close to) each other are turned on (resp. off). Motivated by this fact, we propose distance-based greedy heuristics based on GON and GOFF, called GON-DIST and GOFF-DIST, by simply modifying the metrics in the step 3 of GON and GOFF as follows:

$$M_{\text{GON-DIST}}(i) = \left[\prod_{j\in\mathcal{B}_{\text{on}}} d(i,j)\right]^{1/|\mathcal{B}_{\text{on}}|}, \quad \forall i \in \mathcal{B}\setminus\mathcal{B}_{\text{on}}, \tag{26}$$

$$M_{\text{GOFF-DIST}}(i) = \left[\prod_{j\in\mathcal{B}_{\text{on}}, j\neq i} d(i,j)\right]^{1/|\mathcal{B}_{\text{on}}|}, \quad \forall i \in \mathcal{B}_{\text{on}}, \tag{27}$$

where the geometric mean of the distances to the other BSs are used for the distance metric.

We also propose a greedy algorithm, called GOFF-UTIL[8], that chooses the most underutilized BS by modifying the metric in the step 3 of GOFF as follows:

$$M_{\text{GOFF-UTIL}}(i) = \rho_i, \quad \forall i \in \mathcal{B}_{\text{on}}, \tag{28}$$

These metrics are easy to calculate because the distances between BSs are given and BSs' utilizations are easy to measure. In Section V, we compare the performance of the proposed algorithms GON, GOFF, GON-DIST, GOFF-DIST and GOFF-UTIL with that of exhaustive search under various practical scenarios.

[7]The case of $M_{\text{GON}}(i)$ can be understood similarly.

[8]Note that it does not make sense to have GON-UTIL policy since BSs that are turned off cannot have utilizations by the definition.

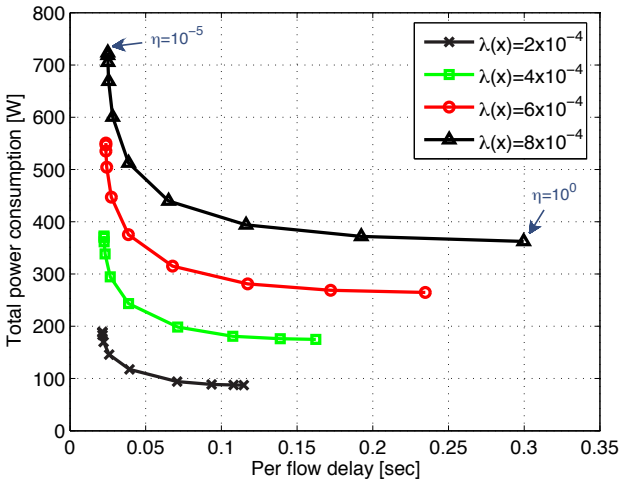


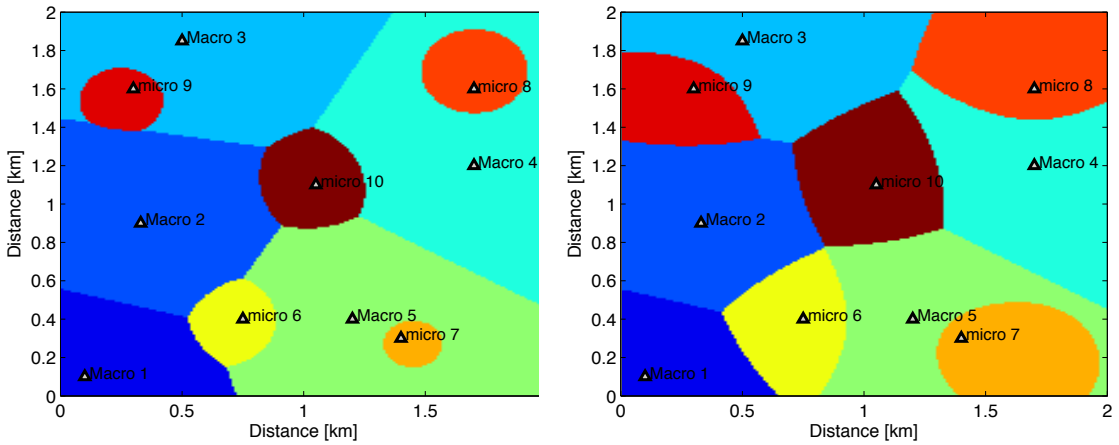
Fig. 1. Energy-delay tradeoff for the case of energy-proportional BSs ($q_i = 0$) by varying the energy-delay tradeoff parameter $\eta = 10^{-5} \sim 10^{0}$. As η increases, energy saving can be obtained at the cost of delay increase.

V. Numerical Results

We verify the proposed energy-efficient user association and BS operation algorithms through extensive simulations under various practical configurations. A network topology composed of five macro BSs and five micro BSs in 2×2 km^2 as shown in Fig. 2 is considered for our simulations. A real 3G BS deployment topology consisting of heterogeneous environments (urban, suburban and rural areas) is also considered in subsections V-B and V-C in order to provide more realistic simulation results. Among several typical levels of maximum transmission powers for BSs given in [37], we used the intermediate values for our simulations, i.e., 43dBm and 30dBm for macro and micro BSs, respectively. Based on the linear relationship[9] between transmission and operational power consumptions, we could calculate the maximum operational powers for BSs, i.e., 865W and 38W for macro and micro BSs, respectively.

For the traffic model, we assume that file transfer requests follow a Poisson point process with arrival rate $\lambda(x)$. Each request has exactly one file that is log normally distributed with mean $1/\mu(x) = 100$ kbyte. In modeling propagation environment, the modified COST 231 path loss model with macro BS height $h = 32$m and micro BS height $h = 12.5$m is used. Other parameters for the simulations follow the suggestions in the IEEE 802.16m evaluation methodology document [39]. We consider the average delay experienced by a typical flow as our system performance metric, i.e., setting the degree of load balancing parameter as $\alpha = 2$ for the cost function of level performance. For the cost function of energy,

[9]The linear relationship we adopted is as follows: $P^{\text{op}} = a\cdot p^{\text{tx}} + b$, where $a_{\text{macro}} = 22.6$, $b_{\text{macro}} = 412.4W$, $a_{\text{micro}} = 5.5$ and $b_{\text{micro}} = 32W$. The model and the parameters are derived in [37], [38] based on the data sheets of several existing GSM and UMTS BSs are analyzed at the component level, e.g., power amplifier, antenna, cooling equipment, etc



(a) Low $\eta = 10^{-5}$ (b) High $\eta = 10^{0}$

Fig. 2. Snapshots of cell coverage by the proposed energy-efficient user association algorithm. As η increases, the energy-efficient micro BS (indexed by 6 to 10) will have larger coverage.

we vary the portion of fixed power consumption q_i between 0 and 1 so that we can cover several types of BSs from energy-proportional BSs to non-energy-proportional BSs.

A. Energy-delay Tradeoff for Energy-proportional BSs

We first consider energy-proportional BSs ($q_i = 0$) and investigate the performance obtained purely by the proposed energy-efficient user association algorithm. Fig. 1 shows the energy-delay tradeoff curves by varying the energy-delay tradeoff parameter from $\eta = 10^{-5}$ to 10^{0} for the different values of arrival rate $\lambda(x)$. As expected, we can obtain energy saving at the cost of delay increase when we increase η. The percentage of maximum energy saving (moving from $\eta = 10^{-5}$ to $\eta = 10^{0}$) is about 50%. This result is obtained under homogeneous traffic distribution, i.e., $\lambda(x) = \lambda$ for all $x \in \mathcal{L}$. Similar trends can be also observed in inhomogeneous traffic distribution, although we do not provide it due to space limitation.

In order to examine the details of where these energy savings come from, we plot Figs. 2 (a) and (b) that illustrate the snapshots of cell coverage for the cases of low $\eta = 10^{-5}$ and high $\eta = 10^{0}$. By comparing two figures, we can clearly see that the micro BS, which is more energy-efficient than the macro BSs, will have large coverage for the case of high η (i.e., giving more emphasis on conserving energy). In other words, more MTs are associated with and served by the energy-efficient micro BSs that are indexed by 6 to 10 in the figures. This result coincides with Lemma 3.1. However, as the traffic loads are concentrated in the micro BSs, the large utilizations at the micro BSs will results in the increase of per-flow delay.

B. Energy-Delay Tradeoff for Non-energy-proportional BSs

We now consider non-energy-proportional BSs ($q_i = 0.5$) and investigate the performance obtained by both the proposed energy-efficient user association and BS operation algorithms. We consider GON, GOFF, GON-DIST, GOFF-DIST and GOFF-UTIL as the BS operation algorithm, and compare their performance with the optimal solution obtained by exhaustive search. Fig. 3 depicts the map of BS layout [40] that we use for more realistic simulations. It is a part of real 3G network operated by one of the major mobile network operators in Korea (the source has to be anonymized to preserve confidentiality). There are totally 30 BSs within 20×10 km^2 rectangular area. Particularly, we choose this partial map to include a scenario that three environments (urban, suburban and rural) coexit together.

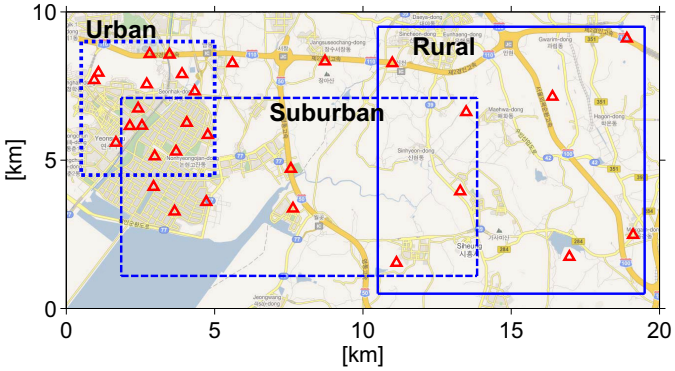
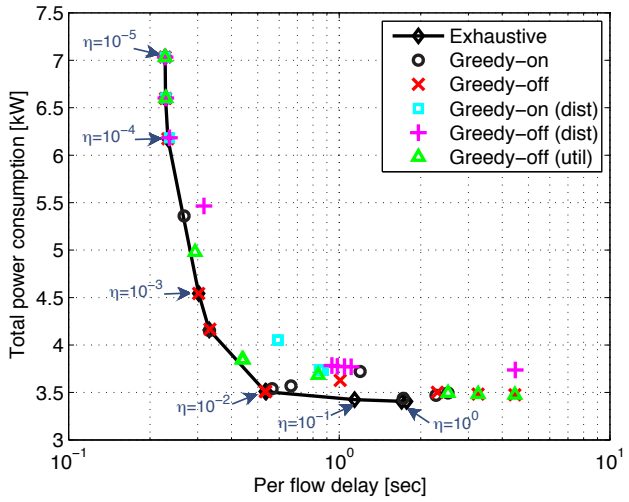
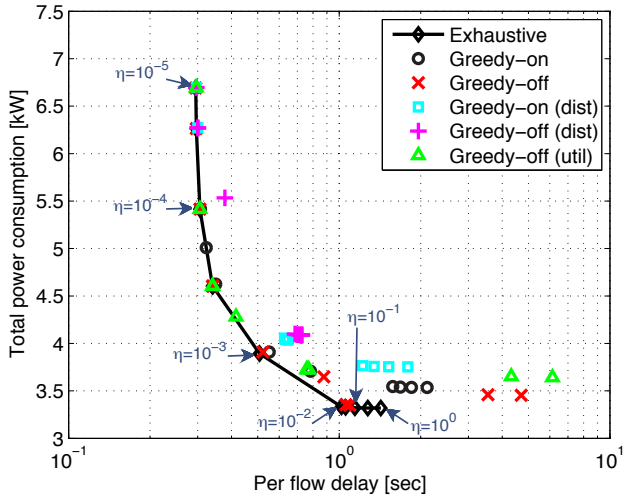


Fig. 3. Real 3G BS deployment map (30 BSs in 20 x 10 km^2).

Fig. 4(a) shows the energy-delay tradeoff curves of different algorithms by varying the energy-delay tradeoff parameter $\eta = 10^{-5} \sim 10^{0}$ in urban area with the homogeneous traffic distribution of $\lambda(x) = 10^{-4}$ for all $x \in \mathcal{L}$. This offered load corresponds to about 10% of BSs utilizations when all BSs are turned on. Recall the real traffic measurement report [2] showing that the time fraction when the traffic is below 10% of peak during the day is about 30% in weekdays and about 45% in weekends. As can be seen from Fig. 4(a), energy is saved at the cost of per-flow delay increase. The proposed greedy algorithms perform close to the optimal solution when η is small. For example, from $\eta = 10^{-5}$, up to $\eta = 10^{-2}$ for GON and GOFF, $\eta = 10^{-3}$ for GOFF-UTIL, and $\eta = 10^{-4}$ for GON-DIST and GOFF-DIST, respectively, we can have almost the same solution to the optimal (i.e., Exhaustive $\geq$ GON = GOFF $\geq$ GOFF-UTIL $\geq$ GON-DIST = GOFF-DIST). However, there is a performance gap when η becomes large.



(a) Homogeneous traffic distribution



(b) Inhomogeneous traffic distribution

Fig. 4. Energy-delay tradeoff of different algorithms for the case of non-energy-proportional BSs ($q_i = 0.5$) by varying the energy-delay tradeoff parameter from $\eta = 10^{-5}$ to 10^{0}. While greedy algorithms perform close to the optimal solution when η is small, there is a gap when η is large. Especially, GON-DIST and GOFF-DIST have large performance gaps under the inhomogeneous traffic distribution.

Fig. 4(b) shows the energy-delay tradeoff curves under the inhomogeneous traffic distribution. As an example of inhomogeneous traffic loads, a linearly increasing load along the diagonal direction from right bottom to left top is considered. They are normalized over the space so as to have the same amount of total traffic as the homogeneous traffic loads have. Similar tradeoff curve can be observed in inhomogeneous traffic distribution as well. The proposed greedy algorithms GON, GOFF and GOFF-UTIL still perform close to the optimal solution up to $\eta = 10^{-3}$, however, GON-DIST and GOFF-DIST start to deviate much form the optimal solution after $\eta = 10^{-4.5}$. There is a reason why such GON-DIST and GOFF-DIST based on the distance do not work well under the inhomogeneous traffic distribution: turning on (resp. off) the BSs distant from (resp. close to) each other is no longer reasonable because the BSs distant from (resp. close to) each other but located in the area of low (resp. high) traffic loads

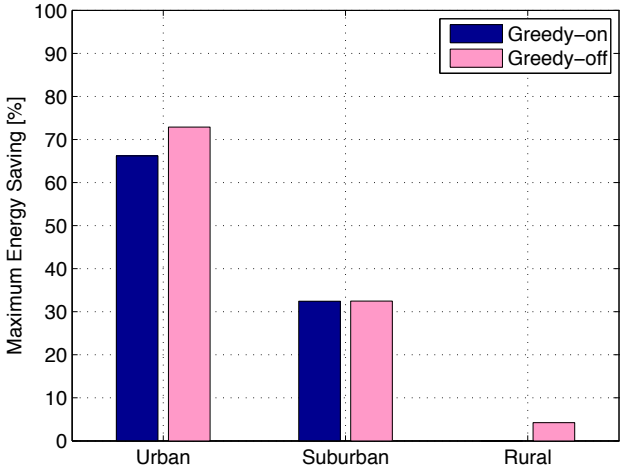


Fig. 5. Effect of BS density on energy saving. Much energy saving can be expected in the urban and suburban environments, but almost no or low energy saving in the rural environment.

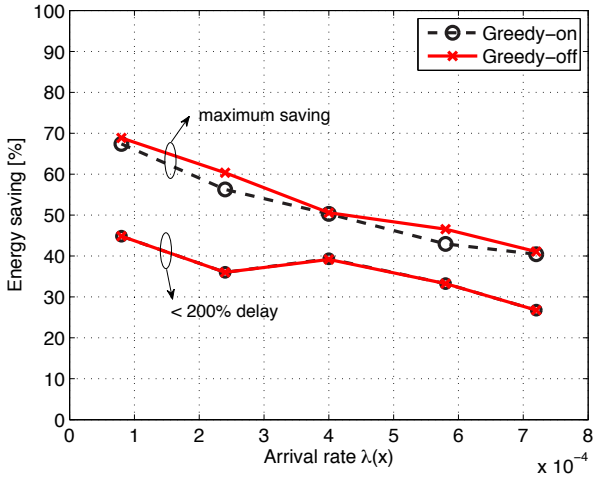
may not be beneficial to turn on (resp. off). It is noteworthy that GOFF-UTIL has the almost comparable performance to that of GON and GOFF under both homogeneous and inhomogeneous traffic distribution. This is a desirable observation for wireless network operators who do want to implement a simple yet efficient BS operation algorithm.

C. Effect of BS Density and Traffic Load on Energy Saving

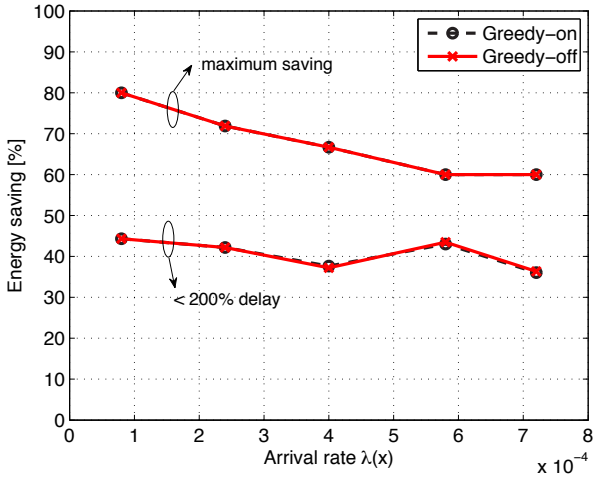
We also examine how much energy saving can be achieved according to the density of BS deployment and the traffic load. Fig. 5 shows the effect of BS deployment density on energy saving. We change the BS density by adopting the BS topology from three different environments: urban (15 BSs in 4.5×4.5 km^2), suburban (15 BSs in 12×6 km^2) and rural (8 BSs in 9×9 km^2) areas. As can be seen, we can obtain much energy saving in the urban and suburban environments, but almost no or low energy saving in the rural environment. This is because the degradation of signal strength is significant in the rural environment when traffic loads are transferred from the switched-off BS to neighboring BSs.

Fig. 6 shows the effect of traffic load on energy savings for the different values of the portion of fixed power consumption $q_i = 0.5$ and 1.0. Here, the maximum energy saving is defined as follows: $1-$[the ratio between energy consumptions when all BSs are turned on (as η goes 10^{-5}) and when the maximum number of BSs are turned off (as η goes 10^{0}) in Fig. 4]. Energy saving within 200% delay is defined as the percentage of energy saving while maintaining the delay lower than twice that of the minimum delay when all BSs are turned on.

As expected, the significant amounts of maximum energy saving and energy saving within 200% delay can be achieved when the traffic load is small, e.g., up to 70-80% energy reduction maximally and up to 45% energy reduction within 200% delay. By comparing Figs. 6 (a) with (b), one can easily notice that the larger energy saving can be obtained from the larger value of the portion of fixed power consumption q_i. In other words, the farther BSs are from energy-proportional BSs, the larger energy conservation can be expected. This is because the benefit by turning off the BSs mainly comes from

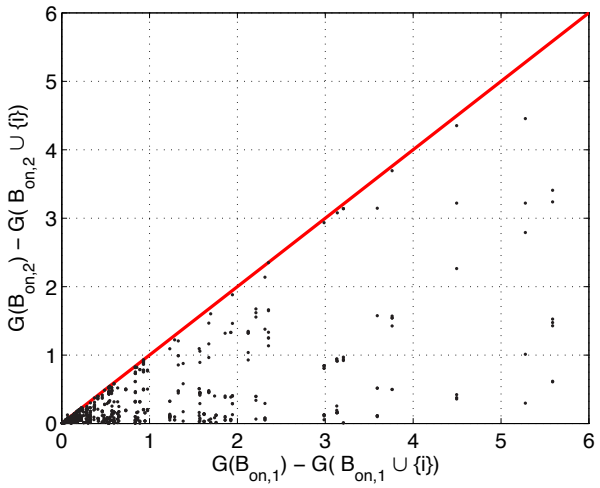


(a) $q_i = 0.5$

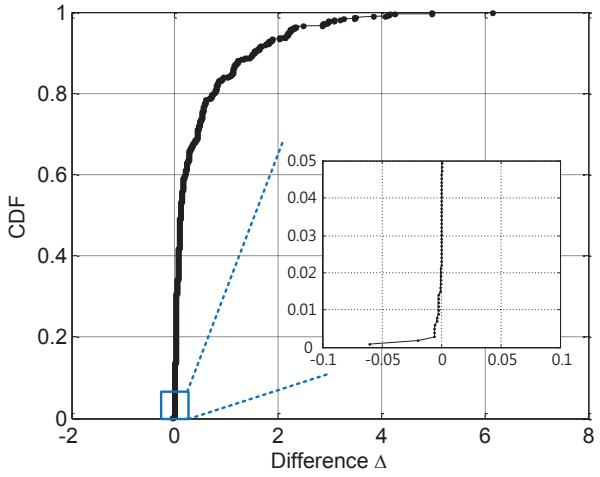


(b) $q_i = 1.0$

Fig. 6. Effect of traffic loads on energy saving for the different values of the portion of fixed power consumption $q_i = 0.5$ and 1.0. As traffic load increases, less energy saving can be expected because the number of BSs that can be turned off is reduced.



(a) Scatter plot



(b) CDF plot

Fig. 7. Submodularity test for $\alpha = 2$.

reducing the fixed power consumption term. Since the BSs dissipate the large portion of its peak power while inactive in practice as of now, the proposed energy-efficient algorithms enable wireless network operators to conserve energy, which in turn will boost the efficiency of overall wireless network operators by reducing their operational expenditures.

VI. Conclusion

In this paper, we developed a theoretical (and also practical) framework for BS energy saving that encompasses both dynamic BS operation and user association. We specifically formulated a total cost minimization problem that allows for a flexible tradeoff between flow-level performance and energy consumption. For the user association problem, we proposed an optimal user association policy and further presented a distributed implementation with provable convergence. For the BS operation problem, we proposed simple greedy turning on and off algorithms that perform close to the optimal solution. Moreover, we proposed other heuristic algorithms based on the distances between BSs or the utilizations of BSs that do not impose any additional signaling overhead and thus are easy to implement in practice. Numerical results based on the acquired real BS topologies under practical configurations showed that the proposed energy-efficient user association and BS operation algorithms can dramatically reduce the total energy consumption by up to 70-80%, depending on the arrival rate of traffic and its spatial distribution as well as the density of BS deployment.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments that greatly improved the quality of this paper. The authors also would like to thank Dr. Daniel Golovin at Caltech for his helpful discussion of the submodularity.

Appendix: Submodularity Test

In order to test the submodularity of $-G$, it is equivalent to check that the following condition holds for all $\mathcal{B}_{\text{on},1} \subseteq \mathcal{B}_{\text{on},2} \subseteq \mathcal{B}$ and $i \in \mathcal{B} \setminus \mathcal{B}_{\text{on},2}$,

$$G(\mathcal{B}_{\text{on},1}) - G(\mathcal{B}_{\text{on},1} \cup \{i\}) \geq G(\mathcal{B}_{\text{on},2}) - G(\mathcal{B}_{\text{on},2} \cup \{i\}). \quad (29)$$

This submodularity condition (29) is not exactly satisfied for some cases. Fig. 7 shows the scatter plot of the

left and right hand sides in (29) and the cumulative distribution function (CDF) of the their difference, i.e., $\Delta = [G(\mathcal{B}_{\text{on},1}) - G(\mathcal{B}_{\text{on},1} \cup \{i\})] - [G(\mathcal{B}_{\text{on},2}) - G(\mathcal{B}_{\text{on},2} \cup \{i\})]$. These are the results for the case of $\alpha = 2$ (i.e., delay objective), which is our special interest in this paper. We also tested for other cases and obtained similar trends, although we do not provide them due to space limitation. As can be seen in the figures, it is not exactly submodular because about 4% of points violate the inequality (i.e., above the red line in the scatter plot of Fig. 7(a) and below $\Delta = 0$ in the CDF of Fig. 7(b)). However, the absolute values of Δ are relatively small for such points and most of points ($> 96\%$) still satisfy the condition. The implication of our finding is that the constant-factor approximation is likely to hold in practice. This is further verified by the simulation results in Fig. 4.

REFERENCES

[1] C. E. Jones, K. M. Sivalingam, P. Agrawal, and J. C. Chen, "A survey of energy efficient network protocols for wireless networks," *Wireless Networks*, vol. 7, no. 4, pp. 343–358, Aug. 2001.

[2] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Towards dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, June 2011.

[3] G. Fettweis and E. Zimmermann, "ICT energy consumption–trends and challenges," in *Proc. IEEE WPMC*, Lapland, Finland, Sept. 2008.

[4] H. Kim, C.-B. Chae, G. de Veciana, and R. W. Heath Jr., "A cross-layer approach to energy efficiency for adaptive mimo systems exploiting spare capacity," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 2745–53, Aug. 2009.

[5] H. Kim and G. de Veciana, "Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals energy," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 802–815, June 2010.

[6] G.-W. Miao, N. Himayat, G. Y. Li, and A. Swami, "Cross-layer optimization for energy-efficient wireless communications: A survey (invited paper)," *Wiley Journal Wireless Communications and Mobile Computing*, vol. 9, no. 4, pp. 529–542, Apr. 2009.

[7] L. Chiaraviglio, M. Mellia, and F. Neri, "Energy-aware backbone networks: a case study," in *Proc. first International Workshop on Green Communications (GreenComm)*, Dresden, Germany, June 2009.

[8] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. first International Workshop on Green Communications (GreenComm)*, Dresden, Germany, June 2009.

[9] "Optimizing performance and efficiency of PAs in wireless base stations," White Paper, Texas Instruments, Feb. 2009.

[10] "Sustainable energy use in mobile communications," White Paper, Ericsson, Aug. 2007.

[11] J. Kwak, K. Son, Y. Yi, and S. Chong, "Impact of Spatio-Temporal Power Sharing Policies on Cellular Network Greening," in *Proc. WiOpt*, Princeton, NJ, USA, May 2011, pp. 167–174.

[12] F. Richter, A. J. Fehske, and G. P. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks," in *Proc. IEEE VTC*, Anchorage, AK, Sept. 2009.

[13] P. Rost and G. Fettweis, "Green communications in cellular networks with fixed relay nodes," *Cooperative Cellular Wireless Networks*, Sept. 2010.

[14] S. Bhaumik, G. Narlikar, S. Chattopadhyay, and S. Kanugovi, "Breathe to stay cool: adjusting cell sizes to reduce energy consumption," in *Proc. first ACM SIGCOMM workshop on Green networking*, New Delhi, India, Aug. 2010, pp. 41–46.

[15] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. IEEE INFOCOM*, San Francisco, CA, Apr. 2003, pp. 786–796.

[16] T. Bonald, S. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," in *Proc. European Wireless Conference*, Nicosia, Cyprus, 2005, pp. 566–572.

[17] K. Navaie and H. Yanikomeroglu, "Downlink joint base-station assignment and packet scheduling for cellular CDMA/TDMA networks," in *Proc. IEEE ICC*, Istanbul, Turkey, June 2006, pp. 4339–4344.

[18] A. Sang, M. Madihian, X. Wang, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," in *Proc. ACM Mobicom*, Philadelphia, PA, Sept. 2004, pp. 302–314.

[19] K. Son, S. Chong, and G. de Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3566–357, Jul. 2009.

[20] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[21] B. Rengarajan and G. de Veciana, "Network architecture and abstractions for environment and traffic aware system-level coordination of wireless networks: The downlink case," in *Proc. IEEE INFOCOM*, Apr. 2008.

[22] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "α-optimal user association and cell load balancing in wireless networks," in *Proc. IEEE INFOCOM*, Mar. 2010.

[23] L. Chiaraviglio, D. Ciullo, M. Meo, M. A. Marsan, and I. Torino, "Energy-aware UMTS access networks," in *Proc. WPMC Symposium*, Lapland, Finland, Sept. 2008.

[24] S. Zhou, J. Gong, Z. Yang, Z. Niu, and P. Yang, "Green mobile access network with dynamic base station energy saving," in *Proc. ACM MobiCom*, Beijing, China, Sept. 2009.

[25] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proc. IEEE Globecom*, Miami, FL, Dec. 2010.

[26] M. A. Marsan and M. Meo, "Energy efficient management of two cellular access networks," in *Proc. ACM GreenMetrics*, Seattle, WA, Jun 2009.

[27] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM*, Apr. 2003, pp. 321–331.

[28] P. Hande, S. Patil, and H. Myung, "Distributed load-balancing in a multi-carrier system," in *Proc. IEEE Wireless Comm. and Net. Conf.*, 2009.

[29] J. Walrand, *An introduction to queueing networks.* Prentice Hall, 1998.

[30] A. Grebennikov, *RF and Microwave Power Amplifier Design.* New York: McGraw-Hill, 2004.

[31] "IEEE P802.16m-2007 draft standards for local and metropolitan area networks part 16: Air interface for fixed broadcast wireless access systems," *IEEE Standard 802.16m*, 2007.

[32] H. Kim, X. Yang, M. Venkatachalam, Y.-S. Chen, K. Chou, I.-K. Fu, and P. Cheng, "Handover and load balancing rules for 16m," in *IEEE C802.16m-09/0136r1*, Jan. 2009, pp. 1024–1037.

[33] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," accepted to *IEEE/ACM Trans. Netw.*, 2011 (to appear).

[34] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of the approximations for maximizing submodular set functions-I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, Dec. 1978.

[35] M. Sviridenko, "A note on maximizing a submodular set function subject to knapsack constraint," *Operations Research Letters*, vol. 32, pp. 41–43, 2004.

[36] D. Golovin and A. Krause, "Adaptive submodularity: A new approach to active learning and stochastic optimization," in *Proc. 23rd Annual Conference on Learning Theory*, Haifa, Israel, June 2010, pp. 333–345.

[37] A. J. Fehske, F. Richter, and G. P. Fettweis, "Energy efficiency improvements through micro sites in cellular mobile radio networks," in *Proc. the second International Workshop on Green Communications (GreenComm)*, Honolulu, HI, Dec. 2009.

[38] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Proc. ICT MobileSummit*, Florence, Italy, June 2010.

[39] *IEEE 802.16m-08/004r5: IEEE 802.16m Evaluation Methodology Document (EMD)*, IEEE Std. 802.16m, 2009.

[40] K. Son, S. Lee, Y. Yi, and S. Chong, "REFIM: A practical interference management in heterogeneous wireless access networks," *IEEE J. Sel. Areas Commun.: Special Issue on Distributed Broadband Wireless Communications*, vol. 29, no. 6, pp. 1260–1272, June 2011.

Kyuho Son (S'03-M'10) received his B.S., M.S. and Ph.D. degrees all in the Department of Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2002, 2004 and 2010, respectively. From 2004 to 2005, he was a visiting researcher at Wireless Networking and Communication Group in the Department of Electrical and Computer Engineering at the University of Texas at Austin. He is currently a post-doctoral research associate in the Department of Electrical Engineering at the University of Southern California. His current research interests range over the various areas of resource allocation and optimization, with a focus on cross-layer design for broadband wireless networks, green networks, smart grid and network economics. He had served as the web chair of WiOpt 2009.

Hongseok Kim (S'06-M'10) received his B.S. and M.S. degree in Electrical Engineering from Seoul National University in 1998 and 2000, respectively, and Ph.D. degree in Electrical and Computer Engineering at the University of Texas at Austin in 2009. From 2000 to 2005 he was a member of technical staff at Korea Telecom Network Laboratory. He participated in FSAN and ITU-T SG16 FS-VDSL standardization from 2000 to 2003. He was a Post-Doctoral Research Associate in the Department of Electrical Engineering at Princeton University, Princeton, NJ from 2009 to 2010. He was a Member of Technical Staff with Bell Laboratories, Alcatel-Lucent, Murray Hill, NJ, from 2010 to 2011. His research interests are network resource allocation, optimization including cross layer design of wireless communication systems, MIMO, network economics, and Smart Grid. Dr. Kim is the recipient of Korea Government Overseas Scholarship in 2005.

Yung Yi (S'04-M'06) received his B.S. and the M.S. in the School of Computer Science and Engineering from Seoul National University, South Korea in 1997 and 1999, respectively, and his Ph.D. in the Department of Electrical and Computer Engineering at the University of Texas at Austin in 2006. From 2006 to 2008, he was a post-doctoral research associate in the Department of Electrical Engineering at Princeton University. Now, he is an assistant professor at the Department of Electrical Engineering at KAIST, South Korea. His current research interests include the design and analysis of computer networking and wireless communication systems, especially congestion control, scheduling, and interference management, with applications in wireless ad hoc networks, broadband access networks, economic aspects of communication networks, and green networking systems. He has been serving as a TPC member at various conferences such as ACM Mobihoc, Wicon, WiOpt, IEEE Infocom, ICC, Globecom, ACM CFI, ITC, the local arrangement chair of WiOpt 2009 and CFI 2010, and the networking area track chair of TENCON 2010.

Bhaskar Krishnamachari received his B.E. in Electrical Engineering at The Cooper Union, New York, in 1998, and his M.S. and Ph.D. degrees from Cornell University in 1999 and 2002 respectively. He is currently an Associate Professor and a Ming Hsieh Faculty Fellow in the Department of Electrical Engineering at the University of Southern California's Viterbi School of Engineering. His primary research interest is in the design and analysis of algorithms and protocols for next-generation wireless networks.