Energy Savings for Cellular Network with Evaluation of Impact on Data Traffic Performance

Kateřina Dufková*, Milan Bjelica‡, Byongkwon Moon†, Lukáš Kencl*, Jean-Yves Le Boudec†
* R&D Centre for Mobile Applications (RDC), Czech Technical University in Prague
Technicka 2, 166 27 Prague 6, Czech Republic
† Ecole Polytechnique Fédérale de Lausanne
EPFL, CH-1015 Lausanne, Switzerland
‡ Faculty of Electrical Engineering (ETF), University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia

***Abstract*—We present a concrete methodology for saving energy in future and contemporary cellular networks. It is based on re-arranging the user-cell association so as to allow shutting down under-utilized parts of the network. We consider a hypothetical static case where we have complete knowledge of stationary user locations and thus the results represent an upper bound of potential energy savings. We formulate the problem as a binary integer programming problem, thus it is NP-hard, and we present a heuristic approximation method. We simulate the methodology on an example real cellular network topology with traffic- and user distribution generated according to recently measured patterns. Further, we evaluate the energy savings, using realistic energy profiles, and the impact on the user-perceived network performance, represented by delay and throughput, at various times of day. The general findings conclude that up to 50% energy may be saved in less busy periods, while the performance effects remain limited. We conclude that practical, real-time user-cell re-allocation methodology, taking into account user mobility predictions, may thus be feasible and bring significant energy savings at acceptable performance impact.**

***Index Terms*—energy efficiency, cellular network, optimization**

I. Introduction

The energy consumption of telecommunication networks is not negligible, and the mobile communication networks alone contribute to a large fraction of it. Recent studies reveal that information and communication technology (ICT) systems cause 2% of global CO_2 emissions, which is equivalent to the emissions produced by international air traffic [1]. ICT is responsible for up to 10% of the world energy consumption [2], with telecommunication networks being one of the main consumers. The mobile communication networks alone consume approximately 60 billion kWh per year [3].

In the perspective of *energy sober* economy, it is desirable that the mobile communication sector reduces its energy consumption. Indeed, the fixed infrastructure of mobile telecommunication networks was not designed with the primary goal of saving its energy cost; there are many indications that significant gains are possible by optimizing network design, operation and traffic management.

Both the network operators and the users should be interested in increasing their energy efficiency. Besides their corporate responsibility regarding the environmental protection, the operators are also strongly motivated to lower their operational expenses. It is claimed that mobile network operators in Germany alone spend more than 200 million EUR per year for their electricity bills [3]; at the same time, the electricity prices in EU continue to rise [4]. Further investigations show that the radio network alone causes approximately 80% of total cost, making it an outstanding candidate for savings.

In [5] we showed, based on tracked user data, that it was essentially feasible to predict user to cell association in contemporary and future cellular communication networks. In this paper, we investigate how much energy can be saved if such a prediction is available, how to achieve such gains, and what may the impact be on data traffic performance. To this end, we first propose a method for adapting user to cell associations to the actual traffic as closely as possible, thus making possible to shut down some transceivers. We formulate the problem as a binary integer programming problem and propose a heuristic, the greedy algorithm, to solve it. We test our proposal on a real network topology and show that savings from 25% to up 50% can be achieved, depending on the time of the day. Then we evaluate the impact on quality of service perceived by data traffic users. The reasons for dealing with the data traffic only are: first, out of the SMS, voice and data traffic, the data traffic can be seen as the worst case because it generates the highest amount of bits that need to be transmitted and with more irregular patterns; second we believe that the growth potential of data traffic is much higher than the potential of the other types of traffic, and thus data traffic will prevail in future networks. We leave for further study the impact on quality of service for traditional voice traffic. Our results show that it is possible to formulate our optimization problem such that the impact on quality of service for data traffic is very small.

The goal of this paper is to get a first evaluation of how much can be saved with such a method, and whether quality of service degradation is involved. Our results are based on offline optimization, so they constitute an upper bound to what can be achieved in operational conditions.

The rest of the paper is organized as follows. In Section II we give the overview of the related work. Section III describes our network model and traffic assumptions. Section IV describes our proposal in detail. Section V describes simulation results and Section VI concludes the paper.

II. Related work

Energy optimization in wireless networks draws much of the research attention in the recent period. However, most of these researches heavily depend on the particular technology that is being concerned. They typically exploit some protocol properties and modify the timings or retransmissions, or force the terminal to go to the sleep/standby mode. Good examples could be found in [6] for WiFi and in [7] for multi-hop ad hoc networks. The proposed solutions are designed for non-real time and best effort applications, so it remains unclear if they could guarantee certain level of quality of service (QoS).

From the users' viewpoint, the most important factor is probably the limited battery life of their mobile terminals. It does not surprise that many authors therefore propose methods for energy savings which are aimed at reducing the consumption of the terminal. Once again, these force the standby state, like in [8], combine the exploiting of some web traffic statistical properties with switching the wireless interface off [9], or use a secondary air interface with lower energy consumption as signaling channel [10]. Our research, in the other hand, tends to be technologically independent and infrastructure oriented in terms that we are concerned about the savings in the BTS subsystem of a cellular network.

Restrepo *et al.* introduce the notion of energy profile, as the dependence of the energy consumption in function of the traffic load (or traffic throughput) of a particular network component [11]. They propose several profiles which could describe the behavior of both the existing and future devices. We adopt the idea of "on-off" and linear profiles, while modifying the latter one to account for offset (i. e. traffic-independent) consumption. We find the justification to this approach in a report by Corliano and Hufschmid [12], who monitored the consumption of some typical BTS configurations.

Louhi notices that the energy consumption of a cellular network could be reduced not only by decreasing the energy consumption of BTS sites, but by minimizing the number of BTS sites as well [13]. Marsan *et al.* further elaborate this idea in [14]. They investigate the possibility of reducing the energy consumption of the *access* part of a cellular network by switching some cells off during the periods in which they are under-utilized because the traffic is low. They observe some idealized cell configurations, like hexagonal, crossroad or Manhattan and use both trapezoidal traffic pattern and real world traces, collected in wired network. Under these assumptions, energy savings of 25–30% were possible to achieve. Chiaraviglio et al. apply the similar approach to wired networks in [15]. They propose a method to switch the network nodes and links off while still guaranteeing full connectivity and maximum link utilization. What is particularly interesting to us, they provide an Integer Linear Programming (ILP) formulation of the problem and show it to be NP-hard. They then propose several heuristics to solve this problem. In their another paper [16], the authors use the approximation of sinusoidal traffic pattern.

Unlike these, we consider an existing network, with topology being far more complex than the simple hexagonal grid. Also, we do not use any of the approximations for the traffic load, but apply the SURGE tool [17] to generate the web-like traffic traces.

Fehske *et al.* investigate the possibility of lowering the energy consumption of cellular networks by deployment of small, low power base stations, alongside the conventional sites [18]. Once again, they assume regular hexagonal grid of macro sites and assume that their power consumption virtually does not depend on traffic load. While this approach might be regarded as a large-scale network optimization (i. e. with respect to the deployment), we are more interested in run-time or short-scale optimization.

III. Network Model

To assess the energy-savings potential of the contemporary wireless networks and the trade-offs between energy efficiency, quality of service and other network parameters, a suitable network model is crucial. In this section we describe a realistic network model based on real network topology and user distribution approximating live network traces. We introduce a graph-based approach to describe relations between users and network topology. Finally, we model traffic generated by the users using the SURGE tool [17].

A. Network topology

We perform experiments on a model of real world cellular network topology, which is part of a live GSM network. We chose a rectangular district of approximate size $2\ 100$ km^2, and selected all the base transceiver stations (BTS) and cells in the district to be part of the experiments. Let $L \subset \mathbb{R}^2$ denote a set of *sites* (BTS locations[1]) inside the chosen district, and let $C \subset \mathbb{N}$ be a set of identifiers of all cells hosted by those sites. For given cell $c \in C$ we call *cell area* $s_c \subset \mathbb{R}^2$ an area served *dominantly* by directional antenna located at site $l_c \in L$ with direction azimuth $a_c \in [0°, 360°)$, and *cell coverage area* $o_c \subset \mathbb{R}^2$ a whole area where signal from the antenna is present. Obviously $s_c \subset o_c$, and the cell coverage areas overlap in practical deployments. To model the cell areas we use Voronoi tessellation [19] (see Figure 1), to model the cell coverage areas we adopt a simple uniform propagation model, where each cell c covers circular sector of fixed radius, with centre in location l_c and central angle given by azimuth a_c and azimuths of the other cells on the same BTS site [20]. The size of the topology is 47 cells on 16 BTS sites.

B. User Equivalents

Unfortunately we are not aware of any practical method to obtain information about all network users in a district. Therefore to populate the network topology we introduce a model of user called user equivalent. Each user equivalent represents a single user and its relevant properties. Let $U \subset \mathbb{N}$ be a set of identifiers of all user equivalents acting in the experiment (called user equivalent set onwards). For given user equivalent $u \in U$ we know its location $l_u \in \mathbb{R}^2$, which is

[1] All BTS locations were used in anonymous form. The coordinate system was transformed from original longitude and latitude values to distance from origin, using a distance preserving operation.

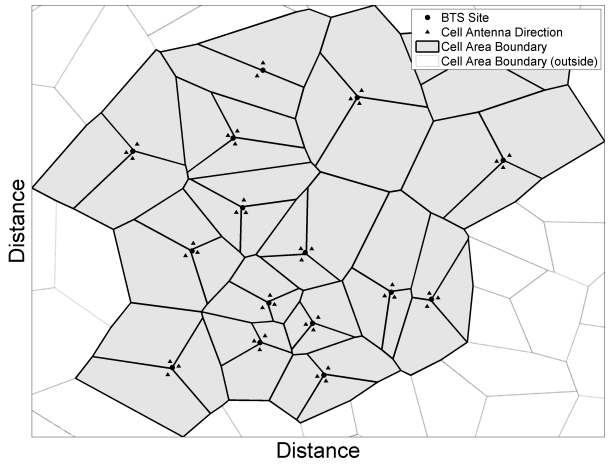

Fig. 1. Visualization of the network topology used in the experiments.

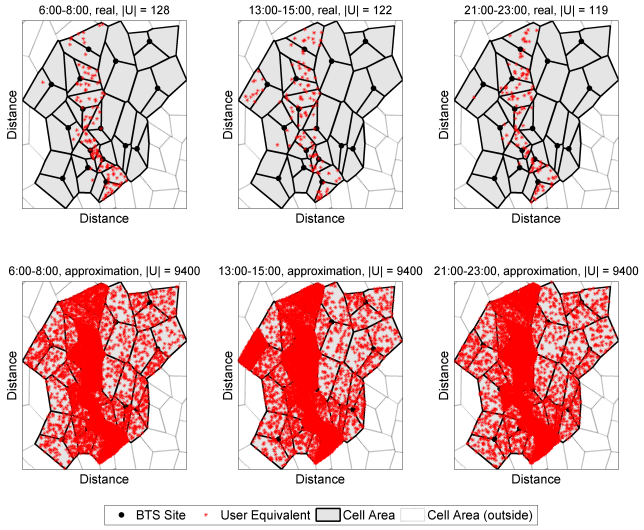

Fig. 2. Visualization of user equivalent sets used in the experiments.

invariant during the experiment as we decided to deal with the stationary case first.

To approximate the real distribution of users in the network, we base the distribution of user equivalents to cells on traces obtained by active tracking of selected users' cell associations, using the platform from [21]. The platform allows to periodically poll and store cell association of a set of users in a real-time manner and without user cooperation. The problem with the traces is that they can never contain data about all users subscribed to the network. We address this by using the absolute amount of users in the cells as an approximation of spatial multinomial probability distribution from which we draw a sample of desired size. Inside each cell area, the user equivalents are placed randomly. Statistics of used user equivalent sets are summarized in Table I, some of the sets are visualized in Figure 2.

C. User-Cell Affinity Graph

We model the affinity between user equivalents and cells by means of bipartite graphs. Let $G(U, C)$ be the bipartite,

TABLE I
USER EQUIVALENT SETS STATISTICS, FOR DIFFERENT TIMES AND SIZES. THE NAMING CONVENTION IS $U^{\text{TYPE}}_{\text{TIME}}(\text{SIZE})$.

Time	Real	Approximation of real distribution	
6:00–8:00	$U^R_6(128)$	$U^A_6(470)$	$U^A_6(9400)$
13:00–15:00	$U^R_{13}(122)$	$U^A_{13}(470)$	$U^A_{13}(9400)$
21:00–23:00	$U^R_{21}(119)$	$U^A_{21}(470)$	$U^A_{21}(9400)$

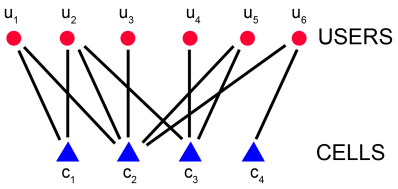

Fig. 3. Example of an affinity graph. User equivalents are in the upper part of the graph, cells in the bottom part. Edges connect user equivalents and cells that see each other. Physical location of vertices is ignored for simplicity.

undirected graph with vertex set $V = \{C \cup U\}$ and such that there is an edge between $u \in U$ and $c \in C$ if and only if user equivalents u sees the signal from cell c (i.e. $l_u \in o_c$).

We call this type of graph the *affinity graph* because it describes affinity (and thus possible associations) between user equivalents and cells. Figure 3 gives an example of small size, Figure 4 visualizes the affinity graph for one of the user equivalent sets used in experiments.

The same type of graph can be used to model the association of user equivalent to one of the cells it sees. We call *association graph* a subgraph of the affinity graph such that each vertex $u \in U$ has degree exactly one. This corresponds to the constraint that one user equivalent is always associated to exactly one cell.

D. Traffic

As discussed in the introduction, we deal here with data traffic only. To obtain a realistic data traffic for each of the user

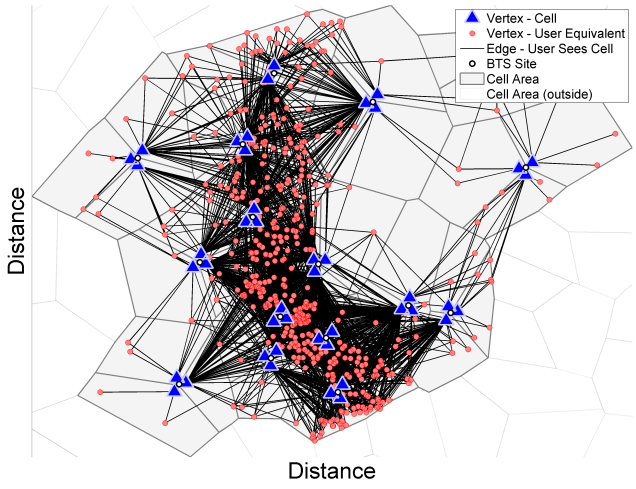

Fig. 4. Visualization of affinity graph $G(U^A_6(470), C)$. The graph shows all possible association of 470 user equivalents to 47 cells, given the assumption that all cell radii are 8 km.

equivalents we use the SURGE tool [17] capable of generating representative Web workloads. We leave for further studies a more complex traffic generation taking into account both data and voice traffic, and the fact that data traffic generated by mobile phone users differs from that generated by computer users (and SURGE), e.g. by the mean size of downloaded files.

Generated user requests are fed to an uplink queue with service rate $R_{UP} = 2$ Mbps, experience the emulated network delay, and requested data are delivered to user equivalents through downlink queue with service rate $R_{DOWN} = 14.4$ Mbps (service rates based on the HSDPA technology). All user equivalents in one cell share a single uplink and downlink processor sharing queue. Network delay is modeled as a fixed one of 100 ms, and we measure the overall delay between the uplink and downlink processes, which is the delay users perceive. With these values, the load offered by a single user equivalent is order of 1% of the uplink. The values we use are based on today's technology, and make a consistent set. However, our method is independent of these values; in the future we expect that uplink and downlink bit rates will increase dramatically, as will probably the traffic per user.

Not all user equivalents present in the network are going to be active, reflecting the fraction of mobile operator customers who do not produce any traffic. In our model, we introduce the *network activity level* A_N to denote the fraction of the remaining *potentially active user equivalents*.

To account for the fact that the traffic load varies heavily during the day, we then introduce the *Time-of-Day (ToD) activity level*, $A_{ToD} = P(\text{potentially active user is active within the time interval ToD})$. A_{ToD} thus changes over the day to reflect daily network traffic patterns. Such daily activity levels can be derived from aggregate network measurements (see Table II).

Next, we introduce parameter K dependent on the traffic activity levels, which denotes the maximum number of user equivalents allowed to associate to a single cell. To derive the value of K, we establish a threshold K_{active}, the maximum number of simultaneously *active* user equivalents that can be associated to a single cell. K_{active} is limited by the cell traffic capacity, dependent on network technology used. Then, K must satisfy, for a particular Time-of-Day interval, the equation $K \cdot A_{ToD} \cdot A_N = K_{\text{active}}$, to respect the cell traffic capacity limit.

E. Energy Profiles

By the term *energy profile* we consider the dependency of total power P (in watts) consumed by switched on BTS, on its current traffic load A (in erlangs),

$$P = f(A). \tag{1}$$

A switched off BTS is considered to consume 0 watts.

One might argue that it is more correct to consider the consumed energy (in watthours) instead of power, but since we are concerned with optimum saving strategy, this will not influence our findings.

Let us note that the term *total power* includes not only the radiated power, but also the power consumed for the proper operation of all other devices and equipment usually stored at the BTS site, e. g. power supply, signal processing, line transmitting, remote monitoring and cooling equipment.

For our research we use two energy profiles, constant and linear. *Constant profile* assumes that the power consumption of BTS is virtually independent on its traffic load, i. e.

$$P(A) = P_{0c} = const. \tag{2}$$

Based on the typical parameters for the existing three-sectors BTS sites, we adopt the value $P_{0c} = 800$ W. The constant model may apply well to the existing technology, with transmitting equipment using class A power amplifiers.

Linear profile assumes that the consumed power increases linearly with the traffic load,

$$P(A) = aA + P_0. \tag{3}$$

From the measurement results reported by Corliano and Hufschmid in [12] we adopt the following parameter values: $a = 3.5$ W/E, $P_0 = 750$ W. We expect the linear model to be suitable for the future technologies, with dominant data traffic and exploiting digital power amplifiers. As for the varying distance of the mobile unit from the BTS, in this work we choose not to consider it as influencing the power consumption, based on the low traffic influence on the total BTS power budget in general.

IV. Optimization of User Associations

We define the problem of *optimal association of user equivalents to cells* as follows. Let $G_A(U, C)$ be the affinity graph as defined above. Let K be the maximum number of user equivalents allowed to associate to a single cell. The value of K depends on the technology and on the traffic generated by users (see Section III-D).

We seek to find a subgraph X of $G_A(U, C)$ which satisfies the following constraints:

- $\forall u \in U : \text{degree}(u) = 1$
- $\forall c \in C : \text{degree}(c) \leqslant K$

The first item expresses that X is a valid association graph (one user is associated to exactly one cell), while the second item is a constraint that we introduce.

Among those graphs, we seek to find a solution graph X that maximizes the number of empty cells that can be switched off, i.e. we want to maximize $f(X)$ with

$$f(X) = |\{c \text{ cell of } X | \text{degree}(c) = 0\}|. \tag{4}$$

A. Formulation as Integer Linear Program

We solve this problem by transforming it into an instance of *binary integer programming*. The main difficulty here is with the objective function, as the above problem formulation leads to maximization of the incommodious function $f(X)$.

To be able to use a more convenient objective function, we first turn the original graph $G_A(U, C)$ into a weighted graph $G'(U, C)$ by assigning weight 1 to all edges in $G_A(U, C)$. Second, we add a "special user equivalent" u^s, that is connected to each cell $c \in C$ with an edge of weight K (see Figure 5).

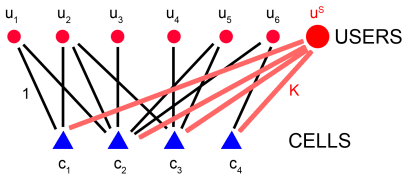

Fig. 5. Example of an affinity graph extended with the "special user equivalent" u^s.

Our problem now becomes to find a subgraph X of $G'(U,V)$ which satisfies the constraints

(a) $\forall u \in U : \text{degree}(u) = 1$
This remains unchanged, with the exception that the user equivalent u^s is allowed to associate to any number of cells at the same time.

(b) $\forall c \in C : \sum_e \text{weight}(e) \leqslant K$
This means that on any cell can be associated either the user equivalent u^s, or up to K regular user equivalents.

and which maximizes

$$g(X) = \sum_e \text{weight}(e). \tag{5}$$

We claim that this is equivalent to our original problem. Indeed, for any subgraph X of $G'(U,C)$:

$$g(X) = \sum_{e=(u,c),\, u\in U} \text{weight}(e) + \sum_{e=(u^s,c)} \text{weight}(e) = \\ = \sum_{e=(u,c),\, u\in U} 1 + \sum_{e=(u^s,c)} K. \tag{6}$$

Should X satisfy the condition (a), then $\sum_{e=(u,c),\, u\in U} 1 = |U|$ and thus the problem boils down to maximization of $\sum_{e=(u^s,c)} K$ which corresponds to the number of cells associated to the user equivalent u^s. Obviously, the cells associated to the user equivalent u^s in graph $G'(U,C)$ are exactly the empty cells in the graph $G_A(U,C)$.

Finally, the formulation of binary integer programming problem is as follows. Let E be the set of edges of the modified affinity graph $G'(U,C)$. Let $x = [x_e] \in \{0,1\}$ be a yet unknown vector of length $|E|$, with $x_e = 1$ if and only if the edge e is present in the solution graph X. Let $w = [w_e] \in \{1,K\}$ be the vector of length $|E|$ with $w_e = \text{weight}(e)$.

The problem can be defined as:

$$\max w^T \cdot x, \text{ constraint to } \begin{cases} A \cdot x \leqslant b \\ A' \cdot x = b' \end{cases},$$

where A [resp. A'] is a $|C| \times |E|$ matrix [resp. $|U| \times |E|$] with

$$A_{c,e} = \begin{cases} w_e, & \text{if } e = (u,c) \text{ for some } u \in U \\ 0, & \text{otherwise} \end{cases},$$

$$A'_{u,e} = \begin{cases} 1, & \text{if } e = (u,c) \text{ for some } c \in C \\ 0, & \text{otherwise} \end{cases},$$

$$b = [K,\ldots,K]^T,\ b' = [1,\ldots,1]^T.$$

The constraint $A \cdot x \leqslant b$ enforces condition (b), while $A' \cdot x = b'$ enforces condition (a).

Binary integer programming is NP-hard (its decision version was one of Karp's 21 NP-complete problems [22]). For medium size user sets we use *lpsolve* [23], a state of the art C-software for solving integer linear programming problems. For problems of large size (several thousands of users) the computing time of lpsolve is prohibitive. This is why we propose a heuristic, the greedy algorithm, as explained next.

B. The Greedy Algorithm

Let $a : U \to C$ be a function that for each user equivalent $u \in U$ returns the cell where the user equivalent is associated at the moment. Let $S : C \to P(U)$ be a function that for each cell $c \in C$ returns the set of user equivalents associated to the cell c. The algorithm can be described as follows:

1) Start with the original association graph.
2) While the value of objective function (as defined above) improves, repeat:
   a) Randomly choose a user $u_m \in U$. The probability of choosing each user u is inversely proportional to the total number of user equivalents, that are in the same cell as the user u. Thus
   $$P(u) = \frac{1}{|S(a(u))|}. \tag{7}$$
   b) Change the association of the user u_m to some other cell c that will have at least the same or higher number of users associated after the change. This condition ensures that the objective function value over time is a non-decreasing function. We choose the new association of the user randomly with uniform distribution from the set
   $$I = \{c | E(u_m, c) = 1 \wedge |S(c)| \geqslant S(a(u_m)) - 1\}. \tag{8}$$
3) Output the final graph and the value of objective function.

The complexity of the greedy algorithm is polynomial in $|U|$ and $|C|$, times the complexity of functions a and S.

V. Experimental Results

A. Simulation Setup

In this section we present results of the user allocation optimization using the greedy algorithm, and provide discussion of the impact of the optimization on energy consumption and network performance. The cell radii used are an estimate of real cell sizes, appropriate for the technology of the network whose topology we work with — GSM 900 MHz[2]. For other technologies, e.g. UMTS, cell sizes differ.

We show results for three user equivalent sets (U_6^A, U_{13}^A, U_{21}^A), each set contains 9400 user equivalents and represents different time of the day (morning 6:00–8:00, afternoon 13:00–15:00, evening 21:00–23:00) with different traffic intensity. For each of the user equivalent sets, we generate traffic using the procedure described in section III-D, using activity levels for the appropriate time of day.

[2] For GSM, maximum cell radius is 35 km. Typical cell radius can be obtained e.g. using Okumura-Hata path loss model, with typical parameters being BTS radiated power $\in \{10, 20, 40, 80\}$ watts, received signal level threshold -100 dBm, antenna height 20 m, antenna gain 10 dB for BTS, resp. 3 dB for mobile unit, suburban environment.

TABLE II
TIME-OF DAY ACTIVITY LEVELS, VODAFONE UK.

Time	A_{ToD} (%)
6:00–8:00	25
13:00–15:00	60
21:00–23:00	100

In the absence of reliable operator data, we have experimented in our simulations with *network activity level* A_N set to the values ranging from 50% across 25% and 10% to 2%. For the results presented in this section, $A_N = 0.50$ is considered, to represent the least favorable alternative, i.e. the largest fraction of users being active, allowing for the lowest amount of re-allocations.

The ToD activity levels (A_{ToD}) used in the simulations are derived from measurements of 3G smartphone usage performed recently in London [24], see Table II, and they reflect the different relative levels of smartphone traffic as measured in the network during selected time intervals (morning 6:00–8:00, afternoon 13:00–15:00, evening 21:00–23:00).

We alternate K_{active}, the maximum number of simultaneously *active* user equivalents that can be associated to a single cell, among values typical for GPRS technology i.e. $K_{\text{active}} \in \{60, 120, 180, 240\}$, with 240 being the default value.

Before generating the traces using the traffic generator, we first decide about each user equivalent if it is active or not, and then generate traces just for the active ones.

To compare the optimization results we use following measures: number of empty cells, power consumption (in watts), delay (in milliseconds), throughput (in bits per second) and utilization (in percents).

B. Traffic Generator

In this subsection, we give the details of the traffic generator used in our simulation. We generate the user traffic using the SURGE tool where each user is modeled as an ON-OFF source. ON time is defined as the time interval from when a user initiates a request for a Web object to when she finishes to receive this object from a Web server. This is then followed by OFF time. A Web object consists of N files, where N follows Pareto distribution with $\kappa = 2$ and $\alpha = 1.245$. The file size S follows a mixture of lognormal distribution with $\mu = 7.63$ (log of bytes) and $\sigma = 1.001$ and Pareto distribution with $\kappa = 10000$ (bytes) and $\alpha = 1.2$.

We wish to compute the offered load generated by single user if she were alone in a cell. This is a figure of merit to calibrate the optimization parameter K. We calculate the utilization ratio[3] when there is only a single user in the cell.

For simplicity, we assume N and S are mutually independent, and there is only one user in the cell. Then, the average duration of ON time, i.e., $E[ON]$ can be calculated as follows:

$$E[ON] = \frac{E[N]B}{C_{UL}} + D + \frac{E[N]E[S]}{C_{DL}}$$

[3]The utilization ratio is given by the ratio of the offered load to link capacity. Offered load is defined as the number of bytes per time unit generated by one user on the uplink (downlink, respectively), if the user were alone in the cell.

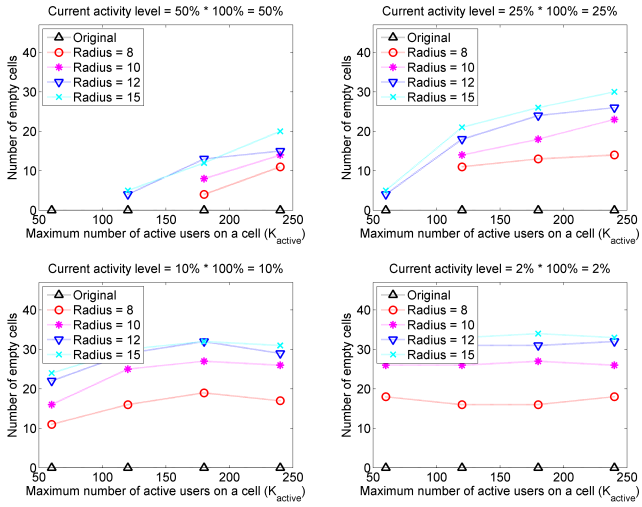

Fig. 7. Optimization results obtained using greedy algorithm approximation. User equivalent set U_{21}^A with 9400 users, Time-of-Day activity level $A_{ToD} = 100\%$, different network activity levels. On each graph, the x–axis is the maximum number of active users allowed to associate to a single cell, the y–axis is the number of empty cells out of 47 total number of cells.

where B is request packet size in bytes, D is fixed network delay in seconds, and C_{UP} and C_{DL} uplink/downlink capacity respectively in bytes per seconds. As expected, these are nothing but uplink queuing delay, network delay, and downlink queuing delay. The OFF intervals follow Pareto distribution with $\kappa = 1$ (seconds) and $\alpha = 1.4$. Then, the offered load is easily calculated by $Load_{UL} = \frac{E[N]B}{E[ON]+E[OFF]}$ and $Load_{DL} = \frac{E[N]E[S]}{E[ON]+E[OFF]}$ in the uplink and downlink respectively. Finally, we can get the utilization ratio of the uplink (U_{UL}) and downlink (U_{DL}) as follows:

$$U_{UL} = \frac{Load_{UL}}{C_{UL}}, \quad U_{DL} = \frac{Load_{DL}}{C_{DL}}.$$

In our simulation setting, $U_{UL} = 0.0054$ and $U_{DL} = 0.0092$. But note that the SURGE model is a closed loop model, i.e. when a user undergoes a delay, her offered load decreases.

C. Optimized User Allocation

Figure 6 visualizes on a graph the change in user allocation before and after the optimization. Figure 7 summarizes the influence of the parameter values on the results. The results indicate that the method is able to find an association graph $G_O(U, C)$ with significantly higher number of empty cells over the original association graph $G_A(U, C)$. Better results can be achieved for higher cell radii R and higher maximum number of *active* users allowed to associate to a single cell K_{active}. This is an expected outcome, as growing both parameters gives the algorithm more freedom to choose final association of the users. From now on, we use the parameter combination cell radius $R = 10$ km, maximum number of active users allowed to associate to a single cell $K_{\text{active}} = 240$ and network activity level $A_N = 50\%$.

D. Energy Savings

We apply the energy profiles described in subsection III-E to evaluate the energy savings achievable by user allocation opti-

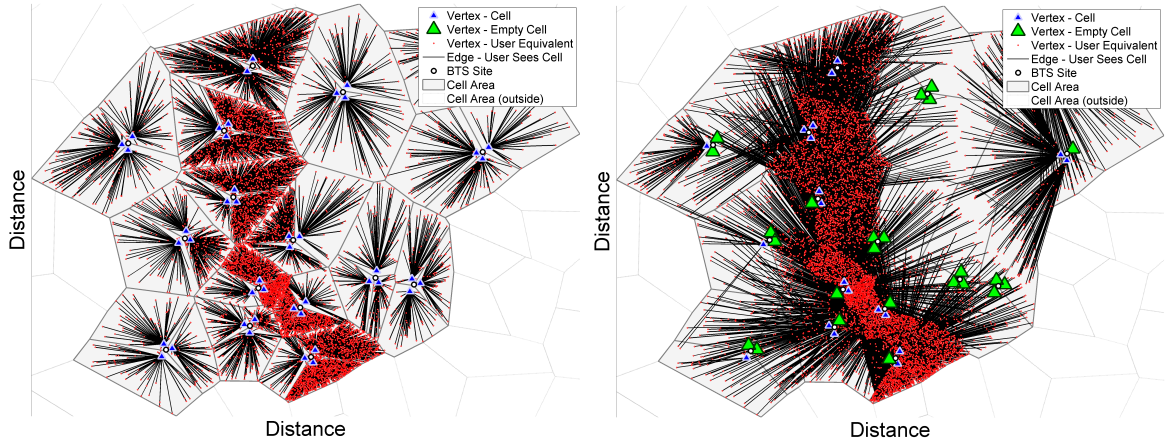

Fig. 6. Visualization of optimization results on the user equivalent set $U_6^R(9400)$. Cell radius $R = 10$ km, maximum number of active users allowed to associate to a single cell $K_{\text{active}} = 240$. As can be seen, the number of empty cells rises from 0 to 24, as a consequence of the fact that some of the formerly under-utilized cells started to serve user equivalents from neighbouring cells. Left: Original user-cell association. Right: Optimized user-cell association.

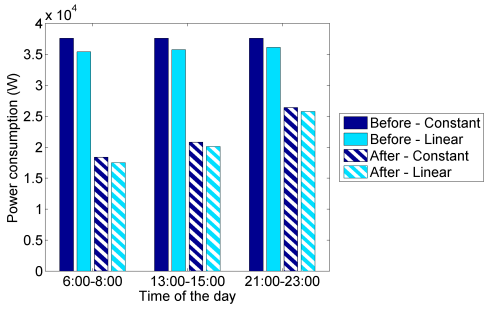

Fig. 8. Energy savings summary — the impact of user allocation optimization and cell switch-off on power consumption. On the x–axis is the time of day, for each time there are results for constant and linear energy profile, both for the situations before and after the user allocation optimization. On the y–axis is the total network power consumption in watts.

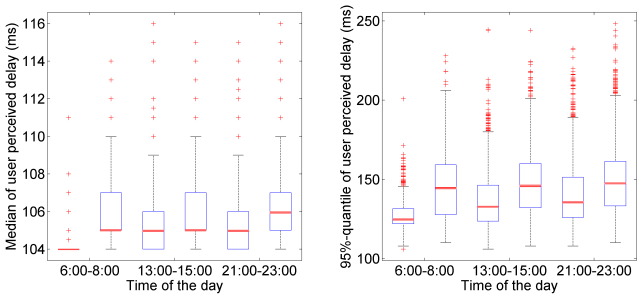

Fig. 9. User perceived delay — the impact of user allocation optimization. On the x–axis is the time of the day, for each time there are two boxplots, the one on the left for the situation before user allocation optimization and the one on the right for the situation after. On the y–axis is the delay in milliseconds. On a boxplot, the central thick line is the median, the box edges are the 25th and 75th percentiles, the whiskers mark data points not considered outliers, outliers plotted as dots. Left: Median delay of all requests of a user equivalent. Right: 95% quantile delay of all requests of a user equivalent.

mization and cell switch-off. Figure 8 summarizes the network power consumption both before and after the optimization. In all cases, BTS sectors corresponding to empty cells are considered switched off. As can be seen, for both constant and linear energy profile the potential energy savings are up to 50% for low traffic conditions (6:00–8:00), up to 40% for moderate traffic conditions (13:00–15:00) and up to 25% for peak traffic conditions (21:00–23:00). Savings are caused by the increased number of empty cells after the optimization.

E. Network Performance

Obviously, when some of the cells are switched off, the mean number of user equivalents per cell increases which can negatively impact the quality of service. We assess this impact by measuring the delay and throughput perceived by the user equivalents before and after the user allocation optimization. The Figure 9 shows the measured delay, both for median and 95% quantile case. The graph shows that the difference in median delay is insignificant for all cases, after the optimization the results are about 1 millisecond worse. The 95% quantile of delay differs more, but still in an acceptable extent of roughly 20 milliseconds. In Figure 10 we show the dependence of the number of active users in a cell on the delay users perceive, which justifies the usage of the number of users in a cell as an optimization constraint.

Figure 11 shows the measured throughput, both for median and 5% quantile case. The graph shows that the difference in median throughput before and after user allocation optimization is negligible. For the 5% quantile of throughput we see a slight degradation in order of hundreds of bits per second.

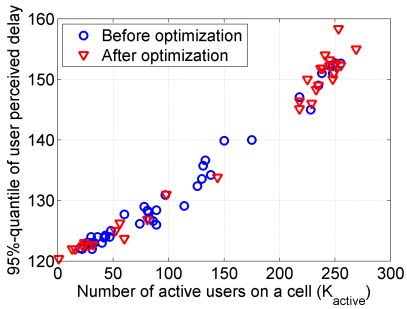

Fig. 10. User perceived delay (95% quantile) and its correlation with the number of active users in a cell.

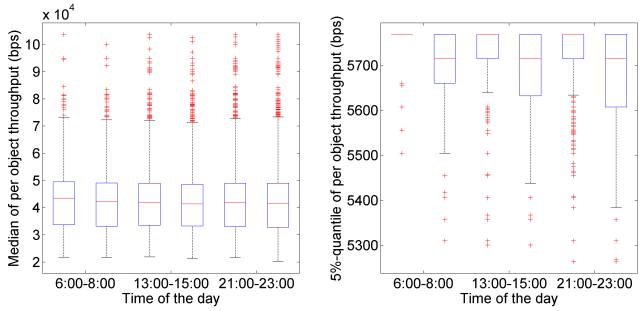

Fig. 11. User perceived per object throughput — the impact of user allocation optimization. On the x–axis is the time of the day, for each time there are two boxplots, the one on the left for the situation before user allocation optimization and the one on the right for the situation after user allocation optimization. On the y–axis is the throughput in bits per second. Left: Median throughput of all requests of a user equivalent. Right: 5% quantile throughput of all requests of a user equivalent.

TABLE III
IMPACT OF USER ALLOCATION OPTIMIZATION ON FAIRNESS

Jain's index (%)		6:00–8:00		13:00–15:00		21:00–23:00	
		Before	After	Before	After	Before	After
Delay	median	99.99	99.96	99.97	99.97	99.97	99.96
	95% q.	99.31	97.59	98.11	97.88	97.98	97.82
Throughput	median	92.29	92.39	92.59	92.62	92.20	92.28
	5% q.	100.00	99.97	99.98	99.97	99.98	99.97

Finally, Table III summarizes the impact of user allocation optimization on fairness (measured using the Jain's fairness indexes [25]) of delay and throughput distribution among all user equivalents. It can be seen that the user association optimization influence on the fairness is negligible.

VI. CONCLUSION

We have presented an optimization method, which can be used to save energy in cellular networks. In this work we have taken the network-centric view, aiming especially to reduce the energy consumption of the operator.

We focused on data traffic, as we expect it to become the lion's share in the future. We gave a concrete algorithm to perform user to cell association. In conjunction with prior findings on the predictability of user location, it shows the feasibility of our method, and also shows that the savings potential are far from negligible, even at the busy hour. Furthermore, it shows that the elasticity of data traffic is such that the impact on quality of service of packing more users into cells is very small over a large range of users per cell; this could be interpreted as a further incentive to use our optimization method. The exact impact of the cell re-association on the user power budget remains to be evaluated.

Future research should focus on transforming these theoretical findings into practical ones, by utilizing immediate user-location and mobility patterns and developing real-time distributed algorithms for solving the optimization problem.

REFERENCES

[1] BusinessGreen Blog. http://blog.businessgreen.com.

[2] Global Action Plan. http://www.globalactionplan.org.uk.

[3] G. Fettweis and E. Zimmermann. ICT Energy consumption – Trends and challenges. In *The 11th International Symposium on Wireless Personal Multimedia Communications (WPMC 2008)*, 2008.

[4] Eurostat web page. http://epp.eurostat.ec.europa.eu.

[5] Kateřina Dufková, Jean-Yves Le Boudec, Lukáš Kencl, and Milan Bjelica. Predicting user-cell association in cellular networks from tracked data. In *MELT '09: Mobile entity localization and tracking in GPS-less environments*, pages 19–33, Berlin, Germany, 2009. Springer.

[6] G. Anastasi, M. Conti, E. Gregori, and A. Passarella. A performance study of power-saving polices for Wi-Fi hotspots. *Computer Networks*, 45(3):295 – 318, 2004.

[7] G. Anastasi, M. Conti, and A. Passarella. Power management in mobile and pervasive computing systems. In A. Boukerche, editor, *Algorithms for Wireless and Mobile Networks*, chapter 24, pages 535–576. CRC-Hall Publisher, 2005.

[8] Eugene Shih, Paramvir Bahl, and Michael J. Sinclair. Wake on wireless: An event driven energy saving strategy for battery operated devices. In *MobiCom '02: Proceedings of the 8th annual international conference on Mobile computing and networking*, pages 160–171, New York, NY, USA, 2002. ACM.

[9] G. Anastasi, M. Conti, E. Gregori, and A. Passarella. Performance comparison of power saving strategies for mobile web access. *Performance Evaluation*, 53:273–294, 2003.

[10] G.P. Perrucci, F.H.P Fitzek, G. Sasso, and M. Katz. Energy saving strategies for mobile devices using wake-up signals. In *4th International Mobile Multimedia Communications Conference (MobiMedia 2008)*, Oulu, Finland, July 2008. ICTS/ACM.

[11] J.C.C. Restrepo, C.G. Gruber, and C.M. Machuca. Energy profile aware routing. In *IEEE International Conference on Communications Workshops (ICC Workshops 2009)*, pages 1–5, June 2009.

[12] A. Corliano and M. Hufschmid. Energieverbrauch der mobilen Kommunikation — Schlussbericht. Technical Report 280030, Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation, Bundesamt für Energie, Bern, Switzerland, 2008. In German.

[13] J.T. Louhi. Energy efficiency of modern cellular base stations. In *29th International Telecommunications Energy Conference INTELEC 2007*, pages 475–476, 30 2007-Oct. 4 2007.

[14] M.A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo. Optimal energy savings in cellular access networks. In *IEEE International Conference on Communications Workshops (ICC Workshops 2009)*, pages 1–5, June 2009.

[15] Luca Chiaraviglio, Marco Mellia, and Fabio Neri. Energy-aware networks: Reducing power consumption by switching off network elements. In *FEDERICA-Phosphorus tutorial and workshop (TNC2008)*, Brugges, BE, 2008.

[16] Luca Chiaraviglio, Marco Mellia, and Fabio Neri. Energy-aware UMTS core network design. In *The 11th International Symposium on Wireless Personal Multimedia Communications*, Lapland, Finland, 2008.

[17] Paul Barford and Mark Crovella. Generating representative web workloads for network and server performance evaluation. In *SIGMETRICS '98/PERFORMANCE '98: Proceedings of the 1998 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 151–160, New York, NY, USA, 1998. ACM.

[18] A. J. Fehske, F. Richter, and G. Fettweis. Energy efficiency improvements through micro sites in cellular mobile radio networks. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'09)*, 2009.

[19] Franz Aurenhammer. Voronoi diagrams — a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3):345–405, 1991.

[20] Michal Ficek and Lukáš Kencl. Improving roamer retention by exposing weak locations in GSM networks. In *CoNEXT 2009: Proceedings of the 5th ACM International Conference on emerging Networking EXperiments and Technologies*, New York, NY, USA, Dec 2009. ACM.

[21] Kateřina Dufková, Michal Ficek, Lukáš Kencl, Jakub Novák, Jan Kouba, Ivan Gregor, and Jiří Danihelka. Active GSM cell-id tracking: "Where did you disappear?". In *MELT 2008: Proceedings of the first ACM international workshop on mobile entity localization and tracking in GPS-less environments*, pages 7–12, New York, NY, USA, 2008. ACM.

[22] Richard M. Karp. Reducibility among combinatorial problems. *R. E. Miller and J. W. Thatcher (editors), Complexity of Computer Computations, New York: Plenum*, pages 85–103, 1972.

[23] LPSOLVE project. http://lpsolve.sourceforge.net/.

[24] Vodafone UK London Smartphone Traffic Measurements 2009. Private communication.

[25] Rajendra K. Jain, Dah-Ming W. Chiu, and William R. Hawe. A quantitative measure of fairness and discrimination for resource allo-

cation in shared computer systems. Technical report, Digital Equipment Corporation, September 1984.