

QoS-Aware Greening of Interference-Limited Cellular Networks

Balaji Rengarajan*, Gianluca Rizzo*, Marco Ajmone Marsan*[†], Barbara Furletti[‡]
{balaji.rengarajan, gianluca.rizzo, marco.ajmone}@imdea.org, barbara.furletti@isti.cnr.it

*Institute IMDEA Networks, Madrid, Spain – [†]Politecnico di Torino, Italy – [‡]ISTI-CNR, Pisa, Italy

Abstract—We consider the problem of minimizing the energy consumed in a cellular access network, under loads that slowly vary over space and time, while guaranteeing quality of service (QoS). In particular, we formalize the problem of jointly optimizing the base stations (BS) power levels and the association of users to BSs, while guaranteeing a minimum throughput to each user, and a target value of blocking probability. We propose abstractions that enable tracking of long-term spatial load distributions, and a practical algorithm for energy efficient user association and base station power allocation. Our algorithm is applicable to arbitrary (planar) BS layouts, to settings with interference, to different BS energy models, and to arbitrary user distributions over the service area. Through extensive simulations using measured data, and realistic BS deployments, we show that our algorithm leads to substantial energy savings both with traditional BS designs and with energy-proportional equipment, and we demonstrate the potential of BS sleep modes to achieve network-level energy proportionality.

I. INTRODUCTION

Some portions of the cellular networks of today have a complex structure, which overlays two or three generations of cells operating with different technologies in different frequency bands, designed to provide users with the throughput they need during peak traffic periods. However, periods of peak traffic normally account for a small fraction of the day, and much lower capacities are necessary during long periods of time. This fact has been accounted for by operators in the definition of their tariffs, but it has now come into the spotlight because the excess capacity implies a very significant cost in terms of consumed (actually, wasted) energy. As a result, the problem of energy-efficient cellular networking has come to the attention of the international research community.

Research efforts in this field can be broadly grouped into two main approaches. The first consists in the design of energy parsimonious, “energy proportional” BS equipment, whose energy consumption is, in some way, proportional to the amount of traffic served. Indeed, the energy consumption of present day BSs can be described by a function that rises very steeply when a BS transitions from the *off* state to the *on* state at very low loads, and then grows linearly to the maximum as load increases [1]. Studies show that typically 60% to 80% of the maximum power is accounted for by the steep increase at low load [2]. Research efforts here aim at reducing the step-like increase of the energy curve around zero load mostly through a modular BS architecture, which allows some hardware components to be switched to low

power state in periods of low traffic [1], [3]. Exploiting such techniques for energy proportionality, several strategies have been proposed to adapt the operating point (and the associated power consumption) of a BS to its traffic load, independently for each BS [3]–[5].

Indeed, if perfect energy-proportionality were achieved at the device level, the adaptation of the energy consumption of the whole network to variable loads would be automatic. However, from these research efforts it emerges clearly that, at least in the near future, there will always be a significant component of consumed power which is independent of load. For this reason, another direction of research acts at the system (network) level, trying to adapt the energy consumption of the whole access network to traffic variations. This is achieved by reducing the number of active BSs through *sleep modes*, decreasing in this way the available capacity as well as the energy consumed by the network. This approach has been shown to be very effective, and leads to substantial energy savings [6]. Sleep modes usually involve rearranging the user-cell association so as to allow shutting down BSs with low traffic loads [7], [8], and enlarging the area served by the remaining BSs, possibly by means of BS cooperation techniques and antenna tilt tuning [8].

One of the main problems associated with these strategies, and in general with all energy efficiency techniques, is their impact on the quality of the service delivered to the users. A requisite of such techniques that would greatly simplify their adoption is transparency to the users, who should ideally not experience any degradation in QoS. A strong limitation of all the strategies mentioned so far is that they consider their effect on QoS only a posteriori, and they leave open the problem of determining a system configuration which allows to achieve a given QoS target. Given the difficulty in explicitly taking QoS into account, several proposed algorithms take QoS into account only implicitly [7], [9], [10]. In [11], greedy heuristics are provided in order to choose the BSs to turn off, and the user association problem is solved independently as a load balancing problem. The authors show that, higher energy savings can be achieved through increasing penalties on average delay performance. However, no mechanism to achieve a given performance criterion is provided. Thus, the potential exists for such solutions to choose an operating point that results in the network sacrificing its basic functionality in order to save energy. Further, none of the approaches

above take into account the impact of inter-cell interference on BS sleep strategies. While careful network planning ensures acceptable performance when all BSs are turned on, changing the transmit power of the BSs and setting some to sleep could have severe consequences on user performance.

The work reported in this paper assumes that a network of BSs is deployed, and given a QoS requirement in terms of minimum individual user throughput and blocking probability, jointly optimizes the power emitted by each BS (which influences cell size and capacity), and the policy for associating users to cells in order to minimize network power consumption. As we explain in the sequel, the optimization of the user association policy is tightly coupled with the optimization of the BS transmit power, and our iterative approach explores the joint space. This approach is not limited to current BSs, but is designed keeping in mind future devices expected to exhibit more graceful scaling of power consumption with load.

We consider adapting to spatial traffic loads that vary slowly (on the time scale of hours, as seen in real traces) over time. We propose *user classes* as an abstraction to enable this changing load to be tracked in a scalable manner that is transparent to users. This is inspired by what is normally done in the network planning phase, when coverage is computed with a granularity corresponding to fixed size squares with sides of tens to hundreds meters. We then formulate the problem of jointly optimizing BS transmit powers and user association to minimize overall power consumption while ensuring that a given QoS (blocking probability) target is met, and propose a tractable iterative algorithm for the problem solution. A key feature of the proposed approach is that the impact of inter-cell interference on performance is taken into account, which is a crucial consideration for today's cellular networks and likely to remain one in the future.

Clearly, as some BSs are put to sleep, the distance between users and their serving BS increases, leading to increased power consumption at the mobile devices. While it is BSs rather than mobile devices that dominate in the overall power consumption, as part of the performance evaluations, we also examine the impact of our proposed approach on mobile users. Note that sleep modes would only be used in regions where the density of installed BSs is high (such as urban centers), thus reducing the impact on users. Further, considering that cellular transmissions only form a small part of overall device power consumption [12] and that the majority of the traffic is on the downlink rather than on the uplink, we can conclude that sleep modes induce only a limited increase in mobile terminal power consumption.

We rigorously evaluate the proposed approach under realistic propagation conditions, using data from a real-world BS deployment in the city of Pisa in Italy. Our extensive numerical results indicate that a careful optimization of the BS powers, and user association through the proposed algorithm can lead to very substantial energy savings in both the case of traditional BS designs, as well as futuristic energy-proportional BSs. Further, we compare the results obtained using the proposed algorithm with an approach based on implicitly

controlling blocking, and show that our algorithm significantly outperforms such an approach.

The rest of the paper is organized as follows. In Sections II, III, and IV we present our model for the system, and user association, and we state the main assumptions underlying our approach. In Sec. V, we present the formulation of the optimization problem and present a computationally tractable iterative solution approach. In Sec. VI, we present and discuss numerical and simulation results, and we finally conclude the paper in Sec. VII.

II. SYSTEM MODEL

We assume that large wireless networks can be split into a number of independent groups of BSs, with each group tasked with serving the traffic in an allotted area. For example, the base stations under a single base station controller could be a natural grouping. We focus here on one such group of BSs and the associated service area. Denote by $\mathcal{N} = \{1, \dots, N\}$, the set of N BSs in a group. We denote by $\mathbf{P} = (P_n | n = 1, \dots, N)$, the vector of transmit powers of the N BSs, where P_n is the transmit power of BS n . We denote by P^M , the maximum permitted transmit power. We propose an algorithm that adapts to long-term (on the time scale of hours) spatial traffic loads, and adjusts BS transmit powers (including putting some BSs into sleep modes) and the user association policy to minimize power consumption without degrading performance.

A. The performance metric

We consider the case of QoS sensitive traffic such as voice or video on the downlink of a cellular access network. User requests are assumed to arrive randomly in space with independent locations across requests, and as a Poisson process (in time) with rate λ . User requests remain in the system for a holding time that is assumed to be exponentially distributed with mean μ^{-1} . We assume that users require a minimum throughput of R_0 bits/sec. While we make this choice for ease of exposition, note that our formulation can be easily modified to include the case of heterogeneity in holding times and minimum throughput requirements. A user request that cannot be served at the required rate is blocked, and the system performance metric that we use is the blocking probability. Note that the blocking probability is also a good indicator of performance even if the system does not have explicit admission control. It is an indicator of the tail probabilities on how poor the user experienced performance could be, though in this case the performance of the entire system would be degraded rather than few users being blocked. Our choice of performance metric is also aligned with traditional approaches to dimensioning the overall system.

The service discipline at the BS ensures that all users receive identical throughput, R_0 . The utilization of BS n at time t is denoted by $\rho_n(t)$, and is the fraction of time that the BS is required to transmit in order to meet the above objective. A new user request arriving at time t can be associated to BS n only if the ratio of R_0 to the BS's capacity to the new user is

less than $1 - \rho_n(t)$, i.e., if its throughput requirements can be met.

B. Propagation and capacity model

We only consider large-scale propagation effects in our optimization framework, as we model the long-term data rates received by users, and shorter time-scale effects such as small-scale fading are averaged out. As we will see in the sequel, the proposed measurement-based algorithm is not dependent on a particular path loss model. We denote by $g^n(u)$ the path gain between user u and BS n . We assume that each base station is assigned a frequency band, and two base stations using the same frequency band interfere with each other. For each base station n , the set of co-channel, interfering base stations is denoted $\mathcal{I}_n \subseteq \mathcal{N} \setminus n$. We assume additive white Gaussian noise with power spectral density N_0 and model through our propagation model, both the received power and interference power perceived by a user.

The maximum rate at which a base station can transmit to a user is modeled as a function of the received signal to interference plus noise ratio (SINR). We assume a Shannon's capacity-like formula to map received SINR to data rate. While the rates predicted by Shannon's formula are not presently achievable, currently available modulation and coding schemes do result in a similar log-like mapping between SINR and data rate, that can be modeled well using Shannon's formula applied to a scaled SINR value.

C. Power consumption model

In currently deployed BSs, there is a big jump in power consumption between a BS in sleep mode that consumes low power, and one that is transmitting. Measurements reported in [13] for macro and micro base stations demonstrate this fact, and also clearly show the concave increasing relationship between base station transmit power and power consumption. We propose the following model that is inspired by these results above as well as prior models such as [14], where BS power consumption varies with the transmit power, P_n , as:

$$\text{Power consumption} = \theta_n^0 + \theta_n^1 P_n + \theta_n^2 \log(d_n P_n + c_n) \quad (1)$$

Here, θ_n , c_n , and d_n parametrize the model. The parameters are base-station specific, allowing us to model deployments with heterogeneous base stations such as mixed deployments of macro, and micro cells. The above model can be used to model power consumption characteristics ranging from that of current BSs to more futuristic energy-proportional designs.

Fig. 1 depicts the measured power consumption of a micro base station as reported in [13] as well as the duly parametrized model. We set $\theta^1 = 0$ to mimic the on-off behavior with a sharp step at 0. However, note that this model differs from the measurements in that it is continuous even at zero. This is to facilitate the formulation of the optimization problem in the sequel. However, in order to model current BSs, we assume that any base station not in sleep mode has to transmit at a minimum power P^m (here, at 20% of the maximum), which is not unrealistic. The model is then parametrized such

that the power consumption at this point corresponds to the actual power consumed, and accounts for the jump in power consumption from 0 to 60-80% of the maximum at near-zero transmit power levels. As we explain in Sec. V-A, the algorithm ensures that no BS operates at power between 0 and P^m . Thus, we are able to account for the discontinuity in power consumption at zero. At the other extreme, in order to model futuristic energy-proportional base stations, we set $\theta^2 = 0$

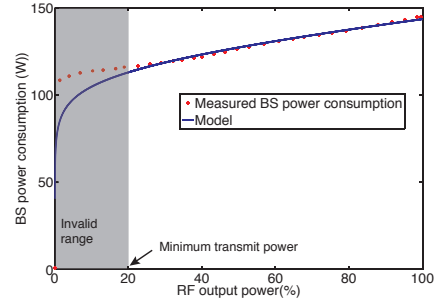


Fig. 1. Fit of the power consumption model to measured data for a micro base station from [13].

III. SPATIO-TEMPORAL LOADS AND USER CLASSES

The performance of a wireless system depends strongly on the spatial distribution of load which in turn determines the capacities that users perceive to the BSs. Since accurately tracking the exact distribution of user requests is intractable, we propose the abstraction of user classes similar to [15] in order to characterize the relevant properties of the traffic. The user classes form the backbone of the proposed user association policy, and also allow the estimation of blocking probability as a function of the BS transmit powers and the user association policy. Note that the classes are structures maintained by the base stations, and completely transparent to users. They are built by base stations from the measurements reported by users during regular operation, and thus capture spatio-temporal load and capacity variations.

A. Defining user classes

User classes aggregate users that perceive similar propagation characteristics to the various BSs in the system. The intuition behind user classes is to group users that can be served at similar rates by BSs, and can thus form a quantized building block using which a scalable optimization methodology can be built. The BSs monitor the user population in the network over a training period, which can be a sliding window into the recent past, or based on historical data. Each user i measures the channel gain (averaging out fast fading) to each BS in the system, and reports the vector of gains, $\mathbf{g}^n(u) = (g^n(u)|n = 1, \dots, N)$ to their serving base station. The path gain vectors are then aggregated into K classes, such that member of a class share similar path gains and capacities to the different base stations. We denote by $\mathcal{K} = \{1, \dots, K\}$, the set of user classes. For each class k , we define a representative gain vector $\mathbf{g}_k = \{g_k^n|n = 1, \dots, N\}$ where g_k^n is the estimate of the path gain between a user

in class k and BS n . Future users can be classified by the BS based on the path gain vector of the users relative to the representative path gains of the classes.

Classes can be constructed using various criteria, and the proposed optimization framework is not dependent on a particular choice. The technique we use is the k -means clustering algorithm [16] on the user path gain vectors. The algorithm produces as output, K class centroids that correspond to the representative path gains of the classes. Future users can be classified based on the distance of their gain vector to the centroids. As K increases, both complexity and the accuracy of the estimates increase, and its choice depends on the tradeoff that is acceptable.

B. Estimating spatial load and capacity profiles

Once the classes are defined, they form the basis for tracking the evolving spatial loads relative to the propagation environment. Note that the proposed methodology does not use geographical information, and propagation effects such as shadowing would be reflected in the structure of the user classes. We denote by λ_k , the estimate of arrival rate of user requests into class k ; with the sum over all classes, i.e., the overall arrival rate equal to λ . The mean number of user requests in the system belonging to class k (class load) is then denoted by $\nu_k = \lambda_k \mu^{-1}$. The capacity from BS n to a user in class k is denoted by $C_k^n(\mathbf{P})$. Since we will operate in the regime where SINR $\gg 1$, guaranteed by a suitable coverage constraint (see Sec. V), we use Shannon's formula approximated for the high SINR regime as

$$C_k^n(\mathbf{P}) = B \log_2 \left(\eta \frac{P_n g_k^n}{N_0 + \sum_{m \in \mathcal{I}_n} P_m g_k^m} \right).$$

Here, $\eta \leq 1$ is a scaling factor to model the gap between current modulation and coding schemes and the Shannon upper bound. The estimate for the fraction of time that BS n would have to devote to a user from class k is given by $R_0/C_k^n(\mathbf{P})$. Note that we could easily include class-dependent holding times and throughput requirements. The combination of the class capacity estimates, and the user arrival rate and service requirements per class allow us to track temporal changes, and are key to the optimization framework described in the sequel.

IV. A CLASS-BASED USER ASSOCIATION POLICY

We use a class-based user association policy which is characterized by the association vector $\mathbf{f} = (f_k^n | n = 1, \dots, N; k = 1, \dots, K)$, where f_k^n denotes the fraction of user requests belonging to class k that are associated to and served by BS n . A valid association vector must satisfy the constraint $\sum_1^N f_k^n = 1, \forall k = 1, \dots, K$. The user association policy can be implemented by further dividing each user class into subclasses corresponding to each BS with the appropriate sizes based on a performance metric. However, when sufficiently many classes are used, a simple randomized user association policy that associates a user of class k to a random BS according to the distribution induced by the association fractions

$f_k^n, n = 1, \dots, N$ suffices, and is used in the simulations described in this paper. The above user association policy could be augmented with a complementary dynamic policy that allows a user who is initially blocked to attempt to join one of the other BSs. The proposed methodology searches over the space of valid user association vectors to determine the optimized operating point as explained in the following section.

V. THE OPTIMIZATION FRAMEWORK

Our objective is to minimize the amount of energy consumed by the BSs that cover a given service area, while at the same time optimizing the associations of users with BSs in such a way that a target blocking probability is achieved. The blocking probability in such a system is difficult to characterize in closed form as users at different locations perceive different blocking probabilities. For example, in the case of a single BS, it is easy to see that users with low capacities to the BS require a larger amount of BS resources and are more likely to be blocked. Thus, directly optimizing the blocking probability is a difficult task. In this paper, we propose indirectly controlling blocking probability by allowing only those operating points that ensure that the average utilization, $\bar{\rho}_n$ of each BS is less than a threshold $1 - \epsilon$. Below, we describe how the BS transmit powers and the user association vectors are optimized, and also how ϵ is chosen in order to achieve the objective.

The formulation of the joint optimization problem subject to the BS average utilization constraints is shown below.

Problem 5.1:

$$\min_{\mathbf{P}, \mathbf{f}} \sum_{n=1}^N \theta_n^0 + \theta_n^1 P_n + \theta_n^2 \log(P_n + c_n)$$

$$\text{Subject to: } \sum_{k=1}^K \frac{R_0}{C_k^n(\mathbf{P})} \nu_k f_k^n \leq 1 - \epsilon, n \in \mathcal{N} \quad (2)$$

$$(C_k^n(\mathbf{P}) - R_0) f_k^n \geq 0, n \in \mathcal{N}; k \in \mathcal{K} \quad (3)$$

$$(P_n - P^m) f_k^n \geq 0, n \in \mathcal{N}; k \in \mathcal{K} \quad (4)$$

$$\sum_{n=1}^N f_k^n = 1, k \in \mathcal{K} \quad (5)$$

$$f_k^n \geq 0, n \in \mathcal{N}; k \in \mathcal{K} \quad (6)$$

$$0 \leq P_n \leq P^M, n \in \mathcal{N} \quad (7)$$

Our objective function is the total power consumption across all BSs and is derived from (1). The average base station utilizations, given by

$$\bar{\rho}_n(\mathbf{P}, \mathbf{f}) = \sum_{k=1}^K \frac{R_0}{C_k^n(\mathbf{P})} \nu_k f_k^n$$

are constrained in (2). Here, $\nu_k f_k^n$ is the average number of users from class k associated to BS n under the association policy \mathbf{f} , and $\frac{R_0}{C_k^n(\mathbf{P})}$ is the estimated fraction of time required for BS n to serve a user of class k . Constraint (3) is a coverage constraint that ensures a BS can indeed serve at least

a single user from any class assigned to it, while constraint (5) ensures that all class loads are completely assigned among base stations. Constraints (6) and (7) define the valid range of values for the association vector and transmit powers respectively. Constraint (4) enforces the condition that a base station that is non-empty uses at least the minimum power level P_m . Thus, any base station with transmit power less than P_m can enter sleep mode.

Problem 5.1 has non-convex, non-separable constraints, and to the best of our knowledge, such a problem cannot be solved efficiently. In the following subsection, we describe our approximate iterative approach to solve the above problem.

A. A Practical Algorithm

Our approach relies on breaking problem 5.1 into two sub-problems that can be efficiently solved, and using an iterative approach where in each iteration, the two sub-problems are solved in tandem. First, we focus on the two sub-problems that i) determine optimal BS transmit powers given a user association policy and ii) find a complementary user association policy that enables the reduction of overall power consumption. Then, we formally specify our iterative algorithm to find a joint solution for a specific choice of ϵ . Finally, we address how ϵ should be chosen in order to achieve a target blocking probability. Let $\mathbf{P}(i)$ and $\mathbf{f}(i)$ represent the optimized values of the transmit power vector and the association vector respectively after the i^{th} iteration. We focus in the sequel on the optimization carried out in the $(i+1)^{\text{th}}$ iteration.

1) *Optimizing BS powers:* The optimized total power consumption at iteration $i+1$, is given by the value of the objective function after solving the sub-problem formulated below, and is denoted by $\Phi(i+1)$.

Problem 5.2 (Sub-problem 1):

$$\Phi(i+1) = \min_{\mathbf{P}(i+1)} \sum_{n=1}^N \theta_n^0 + \theta_n^1 P_n(i+1) + \theta_n^2 \log(P_n(i+1) + c_n)$$

Subject to:

$$\bar{\rho}_n(\mathbf{P}(i+1), \mathbf{f}(i)) \leq 1 - \epsilon, \quad n \in \mathcal{N} \quad (8)$$

$$R_0 + R_\Delta - C_k^n(\mathbf{P}(i+1)) \leq \frac{R_\Delta f_\Delta}{f_k^n(i)}, \quad n \in \mathcal{N}; k \in \mathcal{K} \quad (9)$$

$$P^m + P_\Delta - P_n(i+1) \leq \frac{P_\Delta f_\Delta}{f_k^n(i)}, \quad n \in \mathcal{N}; k \in \mathcal{K} \quad (10)$$

$$0 \leq P_n(i+1) \leq P^M, \quad n \in \mathcal{N} \quad (11)$$

Note that constraints (3) and (4) in problem 5.1 are either active if the corresponding fraction is non-zero or inactive if it is zero. In order to obtain estimates on the sensitivity of the power consumption to the user association policy (which we use in sub-problem 2), we modify these constraints as shown in constraints (9) and (10), where $R_\Delta, f_\Delta, P_\Delta$ are small constants. The iterative algorithm in Sec. V-A3 ensures that none of the fractions $f_k^n(i)$ are less than f_Δ at the input to the above sub-problem. The modified constraints are equivalent to the original ones when $f_k^n(i) = f_\Delta$, and get more

stringent as $f_k^n(i)$ grows. When $f_k^n(i) = 1$, the requirements on the class capacity and power exceed the original ones by $R_\Delta(1 - f_\Delta)$ and $P_\Delta(1 - f_\Delta)$ respectively. While the above problem is a non-convex optimization in $\mathbf{P}(i+1)$, we use a transformation of variables similar to [17], and use the logarithmically transformed power vector $\hat{\mathbf{P}} = \log(\mathbf{P}(i+1))$ as the decision variables. Under this transformation, it is shown in [17] that $C_k^n(\mathbf{P}(i+1))$ is a concave function of $\hat{\mathbf{P}}$. Since $C_k^n(\mathbf{P}(i+1)) > 0$ whenever $f_k^n(i) > 0$ as constrained by (9), i.e., within the feasible region, and the transform is monotonic, it is easy to see that the feasible region is convex. It can also be easily verified that the objective function is a convex function of $\hat{\mathbf{P}}$. Thus, the above problem is convex in $\hat{\mathbf{P}}$, and can be solved optimally. We denote the Lagrange multipliers (assumed to all be positive) at the optimum associated with constraints (8) as l_ρ^n , with the coverage constraints (9) as l_{Cov}^{nk} and with the minimum power constraints (10) as l_{Pow}^{nk} .

2) *Optimizing user association policy:* The purpose of the following sub-problem is to optimize the user association policy in a way that complements sub-problem 1, i.e., to enable base station power consumption to be further reduced in the following iteration. We scale up the BS transmit powers by a factor $1+\delta$, where $\delta \geq 0$ is a small constant, in order to enlarge the feasibility region of the optimization problem defined below. This small perturbation allows the association vector to be nudged in a direction that enables BS power consumption to be further reduced. Letting $\mathbf{Q}(i+1) = (1+\delta)\mathbf{P}(i+1)$, the objective function to be minimized is chosen as:

$$L(\mathbf{f}(i+1)) = \sum_{n=1}^N \sum_{k=1}^K f_k^n(i+1) \left(\frac{-l_\rho^n R_0 \nu_k}{C_k^n(\mathbf{P}(i+1))} + \frac{l_{\text{Cov}}^{nk} R_\Delta f_\Delta}{(f_k^n(i))^2} + \frac{l_{\text{Pow}}^{nk} P_\Delta f_\Delta}{(f_k^n(i))^2} \right) \quad (12)$$

Here, we are interpreting the Lagrange multipliers as the shadow prices, i.e., the sensitivity of the minimal power consumption to relaxing the corresponding constraint. The objective function above is a weighted sum of the fractions, where the weights correspond to the sensitivity of the total power consumption to the respective fraction. Hence, we are optimizing the user association policy in a way that allows the optimization over BS powers in the following iteration to make the fastest progress.

Problem 5.3 (Sub-problem 2):

$$\min_{\mathbf{f}(i+1)} L(\mathbf{f}(i+1))$$

Subject to:

$$\bar{\rho}_n(\mathbf{P}(i+1), \mathbf{f}(i+1)) \leq 1 - \epsilon, \quad n = 1, \dots, N$$

$$(C_k^n(\mathbf{P}(i+1)) - R_0) f_k^n(i+1) \geq 0, \quad n \in \mathcal{N}; k \in \mathcal{K}$$

$$(P_n(i+1) - P^m) f_k^n(i+1) \geq 0, \quad n \in \mathcal{N}; k \in \mathcal{K}$$

$$f_k^n(i+1) \geq 0, \dots, N; k \in \mathcal{K}$$

$$\sum_{n=1}^N f_k^n(i+1) = 1, \quad k \in \mathcal{K}$$

Note that the above is a linear problem in the fractions, and can be solved efficiently. Sub-problem 2 is guaranteed to have a feasible solution when parametrized with the optimal BS transmit powers obtained from sub-problem 1, since the user association vector used to define sub-problem 1 remains a feasible solution.

3) *The Iterative algorithm:* As the initial starting point of the algorithm, we set $\Phi(0) = \inf$ and choose a random (feasible) user association vector, $\mathbf{f}(0)$ that satisfies the condition that no non-zero fractions, f_k^n are less than f_Δ . For example, the user association policy that maps each class to the base station with the maximum path gain is a potential choice. If no feasible starting point is found, we conclude that no satisfactory policy exists for the given value of ϵ . In each

Algorithm 1 BS powers and user association vector given ϵ

```

1:  $i = 1$ ;
2: repeat
3:   Solve sub-problem 1 for  $\mathbf{P}(i)$  and  $\Phi(i)$  given  $\mathbf{f}(i-1)$ ;
4:   if  $\Phi(i) > \Phi(i-1)$  or step 3 is infeasible then
5:      $i = i - 1$ ;  $\delta = \delta/2$ ;
6:   end if
7:    $\mathbf{Q}(i) = (1 + \delta)\mathbf{P}(i)$ ;
8:   Solve sub-problem 2 to find  $\mathbf{f}(i)$  given  $\mathbf{Q}(i)$ ;
9:   Set  $f_k^n(i) = 0$ , if  $f_k^n(i) < f_\Delta$ ;
10:   $\forall k$ , scale fractions, such that  $\sum_{n=1}^N f_k^n(i+1) = 1$ ;
11:   $i = i + 1$ ;
12: until  $\frac{\Phi(i-1) - \Phi(i)}{\Phi(i-1)}$  and  $\delta$  are each above their thresholds
13: return  $\mathbf{P}(i)$ ,  $\mathbf{f}(i-1)$ ;
```

iteration, if the overall power consumption does not improve after solving sub-problem 1, or if sub-problem 1 is infeasible, the current solution is rolled back to the previous state. This occurs as we use a scaled-up value of base station powers to solve sub-problem 2. Thus, sometimes, the perturbation is too large and induces too much error in the sensitivity estimates that form the weights in the objective function of sub-problem 2. In this case, we reduce the constant δ that is the scaling factor and proceed with the iterations. The iterations stop when the improvement in power consumption between iterations falls below a preset threshold, or if the value of δ grows too small. Note that after iteration, either the value of overall power consumption decreases or the value of δ decreases. Since both of the above are bounded quantities, the algorithm is guaranteed to converge.

Since the overall problem is not a convex problem, the algorithm above does not necessarily converge to a global optimum. However, considering that the algorithm we propose is designed to be used to optimize the network over long durations, the algorithm can be tried with few random starting points and the best solution can be chosen to obtain a better approximation of the global optimum. In Section VI we present results on the computational time requirements of the proposed approach.

4) *Calibrating ϵ :* Denote by \mathbf{P}_ϵ , the base station powers and by \mathbf{f}_ϵ , the user association vector obtained at the output of the heuristic run with utilization threshold $1-\epsilon$. We estimate the system blocking probability using a multidimensional

version of the Erlang B formula [18] leveraging the estimated class loads and gain vectors. We consider each base station in turn, and discretize the time requirements of each class as well as the total capacity of the base station. We denote the total number of (discrete) channels available to BS n by T_n , which is chosen to be a suitably large constant, and the number of channels required for BS n to serve class k by $\tau_k^n = \lceil R_0/C_k^n(\mathbf{P}_\epsilon) \rceil$. The offered traffic from class k to BS n is denoted $A_k^n = f_k^n \nu_k$. The probability that j of the base station channels are occupied is proportional to $q(j)$, which is defined recursively as $q(j) = (\sum_{i=1}^K A_i \tau_i q(j - \tau_i)) / j$, with $q(0) = 1$. The probability that j channels are occupied can be computed efficiently in a recursive manner, allowing us to estimate the blocking probability of each class at the BS. Using our estimates of the fraction of a BS's traffic that originates from each class, and the fraction of the total traffic served by each BS, we can estimate the overall blocking probability.

In a practical setting, valid values for ϵ will lie between 0 and 1. Since the blocking probability is a decreasing function of the utilization gap ϵ , we can efficiently find the value of ϵ that induces the target blocking probability. We perform a binary search over the valid range of ϵ in order to determine the final operating point of the system. As we show in the following section, our approach allows us to accurately estimate the resultant blocking probabilities, and the proposed algorithm is able to find power-conserving solutions while achieving the target blocking probability.

VI. SIMULATION RESULTS

We consider two scenarios in order to evaluate the proposed algorithm. We first consider a group of five base stations, four of which are at the vertices of a square of side 1 km, with the fifth one at the center. We then apply it to a realistic setting using details of the base station deployment in Pisa along with measured traffic loads. The maximum BS transmit power is assumed to be 10 W. The carrier frequency and bandwidth of each basestation are assumed to be 1 GHz and 5 MHz respectively, both values being representative of cellular access networks of today. The noise power spectral density is chosen such that the received SNR equals 10 dB at a distance of 1 km from a BS transmitting at maximum power. We assume that the capacity perceived by users is given by Shannon's formula, with a 3 dB SINR backoff. We use event-driven simulations of the system with the powers and user association policy that result from the optimization to measure blocking probability. All blocking probability values reported here were estimated with uncertainty of 5%, and with confidence level 95%

A. A five BS Network

In this section, we consider the scenario where user requests arrive according to a uniform distribution in space, and the propagation model includes only path loss. We use a log distance path loss model, with path loss at a reference distance of 1 m calculated using Friis equation, with a path loss exponent $\alpha = 3.5$. The mean holding time, μ^{-1} , of a call/streaming session is assumed to be 300 s, with throughput requirements

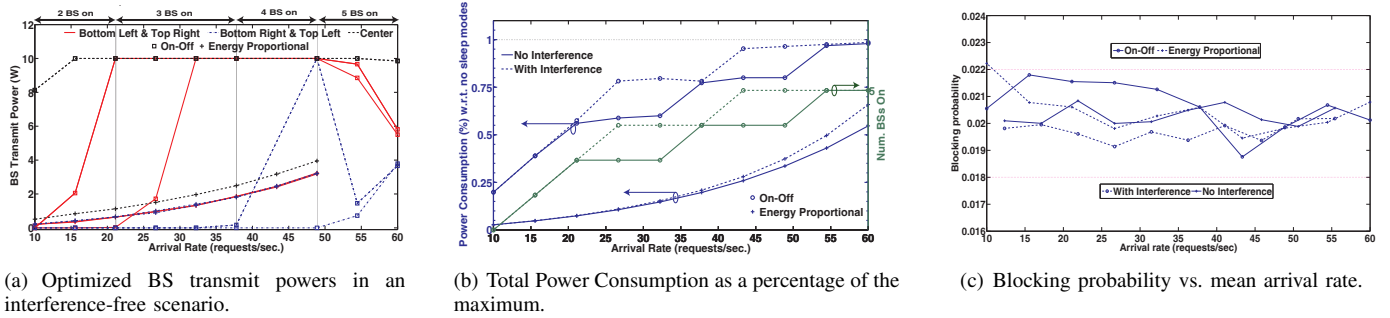


Fig. 2. Algorithm performance in the five BS Network, target blocking probability = 2%

of 10 Kbps (to model voice traffic). We consider two different choices for the parameters of the BS power consumption model in (1), while always keeping the total power consumed by a BS when transmitting at maximum power at 500 W. In the *on-off* setting, the base station consumes 75% of its maximum power consumption when it is out of sleep mode and transmitting at 0.1 W, which is also the value chosen for the minimum power P^m . This choice is to approximately model BSs that are currently deployed. We also consider a futuristic *energy proportional* (EP) setting where power consumed is proportional to transmit power. Additionally, we examine the impact of interference on power consumption. Fig. 2(a) exhibits the optimized BS transmit powers versus the mean arrival rate of user requests for the on-off and energy proportional models with a target blocking probability of 2%. In the on-off case, the dominance of the fixed cost results in choosing an operating point that is highly asymmetric, with the minimum required BSs operating at high transmit power levels and the others in sleep mode. As the load on the system increases, an additional BS is ‘turned on’ when it is needed to achieve the desired performance, and this additional BS quickly increases to maximum power in an almost step-like fashion. When the last base station turns on, we see that after the high turn-on cost has been paid, there is a re-balancing of transmit powers. This is due to the interaction between the power consumption curve and concave capacity function which results in diminishing returns as transmit power is increased. Conversely, in the energy proportional setting all BSs transmit at near-equal power levels, irrespective of the load. In this case, the center base station that is more advantageously placed to serve users always uses a slightly higher transmit power than the ones at the corners. The overall convex shape is again due to the nature of the capacity function. This behavior demonstrates the suitability of sleep modes for networks where the BS power consumption model follows an on-off pattern. On the other hand, the results demonstrate that sleep modes become irrelevant, as expected, when the BS power consumption model is perfectly proportional to load. Fig. 2(b) shows the power consumed as a percentage with respect to the case in which all BSs are permanently on and transmitting at the maximum power, with a target blocking probability of 2%. Here, in addition to the two models for power consumption, we examine the impact of interference on the system. In the

scenario with interference, the BSs on diagonals transmit on the same frequencies, i.e., the base station on the bottom left vertex, and top right vertex interfere with each other, as do the one on the bottom right and top left of the square. The step-like result under the on-off model vs. the smooth increase in power consumption under energy proportional base stations are expected consequences of the power consumption characteristics. The overall power consumption in the energy-proportional case is significantly lower than the on-off case. This is because base stations are more efficient when they serve users with high path gains at moderate power, while the on-off power consumption forces the system to use few base stations at high power inefficiently serving users with low path gains. In general, interference leads to higher power consumption due to the reduction in network capacity. This is visible, both under the energy proportional model as well as in the on-off model where base stations are forced to switch on earlier in order to satisfy the QoS requirements. Clearly energy-proportional hardware can play a crucial role in greening cellular access networks. However, in their absence, sleep modes along with intelligent management do introduce a certain level of network-level energy proportionality as demonstrated by the results in Fig. 2(b) which show a roughly linear trend between load and overall power consumption. At low loads, significant power savings are possible in both cases. Fig. 2(c) exhibits the estimated blocking probability versus the mean arrival rate for all the cases presented earlier. The results demonstrate that the proposed algorithm is able to achieve blocking performance that is very close to the desired target. In nearly all the cases, and under all loads, the simulated blocking probability is within 10% of our target. This demonstrates that the abstractions of user classes allow us to characterize the spatial load, and the optimization framework is effective at controlling the blocking probability through the utilization gap ϵ . Thus, the power savings observed previously do not come at the expense of a degradation in the quality of service. In the sequel, we further evaluate the performance of the algorithm under realistic BS positions, propagation conditions, and measured, non-homogeneous spatial loads.

B. Evaluation on a realistic scenario

To evaluate the performance of the proposed algorithm in a more realistic setting, we apply it to a geographical area corresponding to the center of the city of Pisa in Italy, shown

in Fig. 3. The area chosen contains 16 BSs (actually, BS sectors, but we will refer to them just as BS), distributed over 6 locations. BS locations, antenna orientations, and angular beamwidths, are derived from the data of a large Italian operator. Further, we use traffic information in terms of the number of active calls over time handled by each base station. As seen in the figure, the traffic load shows marked variations both in intensity and in the spatial distribution over time.

We use the COST 231-Hata model which is widely used to model urban environments along with lognormal shadowing with a standard deviation of 4 dB to model the propagation conditions. In order to have a conservative estimation of the effect of interference on sleep modes performance, all BSs are assumed to transmit in the same frequency band, thus causing inter-cell interference. The power consumption model is assumed to be the on-off model of Sec.II, with maximum power consumed equal to 500 W, as before.

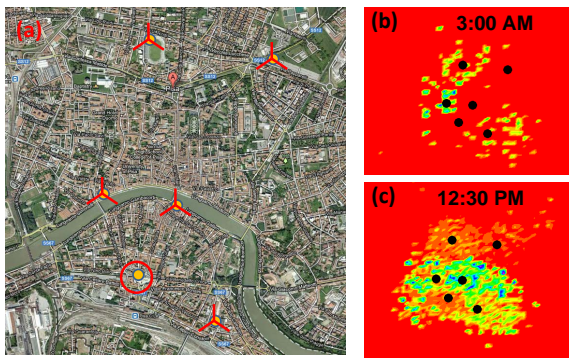


Fig. 3. Locations of the base stations in Pisa (a), and heatmaps of voice traffic at 3 : 00 (b) and 12 : 30 (c).

We use a spatially inhomogeneous Poisson process to model the spatial load at any given time. Note that the distribution of the average intensity of arrivals over space also varies over time, see Fig. 3. We overlay a fine grid on the geographical area and map each square of the grid to an intensity of arrivals at different points in time. We used the call detail records of each BS during a whole day (Wednesday) to build a picture of the spatio-temporal load. We first averaged the number of active calls for each BS over intervals of 30 min to obtain the average load of the base station in each 30 min, interval. In order to translate this measure to a spatial measure of load, we use in conjunction our propagation model and the chosen BSs, and their neighbors, and determine the area served by each BS. In order to do this, we assume that all BSs transmit at the same power and users join the base station with the strongest signal. We further assume that a call handled by the base station is uniformly distributed within its cell. This is a reasonable assumption given that the area served by the base stations is a fairly homogeneous urban centre. This allows us to derive the spatial intensity of the inhomogeneous Poisson process in every 30 minute interval. We assume that the throughput requirement of each user is $120kb/s$ (corresponding to high quality video calls), and that calls are exponentially distributed with average duration of 5 minutes.

We chose seven points of time in the 24 hour period, and

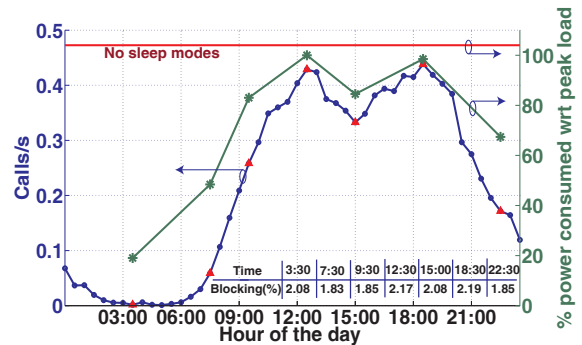


Fig. 4. One day traffic profile(left y axis), and consumed power (right yaxis).

applied our algorithm at each point to compute base station powers (including sleep modes) and the user association policy using a target blocking probability of 2%. Fig. 4 depicts the percentage power consumption with respect to the power consumption at peak load (always under the proposed algorithm, and at 2% blocking probability), along with the overall call arrival rate to the area under consideration. Note that, this call arrival pattern does not present the complete picture since the spatial distribution of load varies over time, and not just the mean as shown in Fig. 3. At peak load, the total power consumed under our algorithm is close to the scenario with all base stations transmitting at maximum power that consumes about 5% more. Using sleep modes, the power savings at non-peak hours can be very large, exceeding 90% in the early morning hours. This is due to the ability of the proposed algorithm to turn off a significant number of BSs at low loads, and concentrate the load in few active BSs. The resulting power consumption closely tracks the varying overall system load, demonstrating that a measure of network-level energy proportionality can be achieved through sleep modes with current hardware. Further, our simulation results (see inset in Fig. 4) confirm that the achieved blocking probabilities are always within 10% of the target, demonstrating that the power savings do not come at the expense of service degradation.

C. Comparison to prior art

We compare the performance of the proposed algorithm with the cell zooming approach proposed in [5], one of the most widely cited papers exploring base station sleep modes. The objective in [5] is similar to ours, i.e., to control blocking probability while minimizing power consumption. The set of base stations that are sleeping is adapted to the load through regularly sampling the system, and using a sample along with a safety margin called the reservation parameter in order to determine the set of active BSs. We simulate the blocking probability and power consumption resulting from the cell zooming approach in the scenario of Sec. VI-B, and compare the results to that of our algorithm. Using the cell zooming, the blocking probability achieved can vary with different samples even with the identical reservation parameter. Hence, we use 50 different samples for each case, and choose the minimum reservation parameter such that the maximum blocking probability over these samples does not exceed the target blocking probability of 2%. This ensures that the probability of violating

TABLE I
PERFORMANCE COMPARISON

Time	Reservation parameter	Blocking (std. dev./mean)	Power Con.) (% increase)
07:30	0.7	0.923	26.34%
15:00	0.5	2.373	14.55%

the blocking probability constraint is acceptable. The results are encapsulated in table I.

The results clearly demonstrate that the cell zooming approach has to accept a much higher level of power consumption (up to 26%) compared to the algorithm presented in this paper. A larger number of active base stations is required to overcome the high variability in the resultant blocking probability. Note further that the value of the reservation parameter varies quite widely with the load, and the choosing this crucial parameter automatically in systems serving diverse loads is a challenging task.

D. Computation Times

All the simulations reported were executed under Matlab on a laptop with a 2.80 GHz. Intel core 2 duo processor and 4 GB of RAM. The heuristic is computationally efficient, requiring on average 6 iterations to converge for a given value of ϵ . In the case of the five base station scenario reported in Sec. VI-A, each joint optimization required, on average, 135 seconds of CPU time. While In the case of realistic, 16 BS scenario reported in Sec. VI-B, each joint optimization required, on average, 600 seconds. of CPU time. Since the network configuration is expected to be adapted to track loads varying on the timescale of hours, the proposed algorithm is indeed practical.

E. Impact on Users

We use the increase in the distance between the users and the serving base station as an indicator of the energy burden imposed on users due to the use of BS sleep modes. The maximum increase is in the case of BSs that are not energy proportional, and in scenario with low traffic loads when the highest number of base stations are put to sleep. In the case of the five BS scenario of Sec. VI-A, both the average as well as the maximum distance between users and the serving base station increases by 30-35% in the scenarios with low arrival rates as compared to the peak-load scenario where all BSs are on. In the case of the realistic scenario of Sec. VI-B, we observe that the average distance from users to the serving BS does not change much even at low loads, when many BSs are in sleep mode. This is due to the spatial traffic pattern that shifts over time, with users geographically clustered even at low-load periods. Thus, we can reasonably expect that the energy cost imposed on users due to using power-saving approaches such ours would indeed be acceptable.

VII. CONCLUSIONS

In this paper, we presented a novel approach to minimize the power consumption of a cellular access network by adapting the BS transmit powers and the user association policy to

the long-term spatial distribution of traffic, while maintaining QoS at the desired level. Thus, the proposed approach enables power savings, while being transparent to the users being served. The proposed algorithm accounts for the impact of interference, a critical concern in cellular networks, and works both with the on-off power consumption characteristic of currently deployed BSs as well as the more energy proportional BSs of the future. We demonstrated through extensive simulations that the proposed method works well under both homogeneous spatial loads, and realistic spatial loads, under realistic propagation conditions (in particular, interference, which is a critical aspect of today's access networks). The simulation results clearly show that BS sleep modes are the correct approach to obtain significant energy savings (over 80%) with current BS designs, exhibiting limited proportionality of energy consumption to load.

REFERENCES

- [1] J. T. Louhi, "Energy efficiency of modern cellular base stations," in *INTELEC '07*, (Rome, Italy), pp. 475–476, october 2007.
- [2] J. Lorincz, T. Garma, and G. Petrovic, "Measurements and modelling of base station power consumption under real traffic loads," *Sensors*, vol. 12, no. 4, pp. 4281–4310, 2012.
- [3] I. Humar, J. Zhang, Z. Wu, and L. Xiang, "Energy savings modeling and performance analysis in multi-power-state base station systems," in *GreenCom-CPSCoM, 2010*, pp. 474–478, dec. 2010.
- [4] L. Saker, S.-E. Elayoubi, and T. Chahed, "Minimizing energy consumption via sleep mode in green base station," in *WCNC'10*, pp. 1–6, april 2010.
- [5] J. Wu, Z. Yang, S. Zhou, and Z. Niu, "A traffic-aware dynamic energy-saving scheme for cellular networks with heterogeneous traffic," in *ICCT'11*, pp. 357–361, sept. 2011.
- [6] B. Rengarajan, G. Rizzo, and M. Marsan, "Bounds on QoS-constrained energy savings in cellular access networks with sleep modes," in *ITC*, pp. 47–54, sept. 2011.
- [7] K. Dufkova, M. Bjelica, B. Moon, L. Kencl, and J.-Y. Le Boudec, "Energy Savings for Cellular Network with Evaluation of Impact on Data Traffic Performance," in *European Wireless*, 2010.
- [8] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *Communications Magazine, IEEE*, vol. 48, pp. 74–79, november 2010.
- [9] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *GreenComm'09*, June 2009.
- [10] L. Saker, S. Elayoubi, and H. Scheck, "System selection and sleep mode for energy saving in cooperative 2g/3g networks," in *VTC 2009-Fall*, pp. 1–5, sept. 2009.
- [11] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE JSAC*, vol. 29, pp. 1525–1536, Sep. 2011.
- [12] N. Sklavos and K. Toulou, "A system-level analysis of power consumption and optimizations in 3g mobile devices," in *1st International Conference on New Technologies, Mobility and Security*, 2007.
- [13] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?," *Wireless Communications, IEEE*, vol. 18, pp. 40–49, Oct. 2011.
- [14] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Future Network and Mobile Summit, 2010*, june 2010.
- [15] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic-aware system-level coordination of wireless networks," *Networking, IEEE/ACM Transactions on*, vol. 19, pp. 721–734, june 2011.
- [16] M. Anderberg, *Cluster Analysis for Applications*. Academic Press, 1973.
- [17] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE JSAC*, vol. 23, pp. 104–116, Jan. 2005.
- [18] Kleinrock, *Queueing Systems, Vo. I: Theory*. John Wiley and Sons, 1975.