



A self-organizing base station sleeping and user association strategy for dense cellular networks

Bart Post¹ · Sem Borst¹ · Hans van den Berg²

© The Author(s) 2020

Abstract

Due to the rising concerns of energy consumption in wireless networks, base station (BS) sleeping strategies were introduced to save energy in low traffic scenarios. In this paper we analyse a weighted trade-off between energy consumption and user-perceived performance in dense cellular networks. We present an optimization problem representing this trade-off and derive properties of its optimal solutions. Using these properties we design a self-organizing strategy that dynamically (online) makes load-aware user association and BS operation decisions. Our strategy is self-organizing in the sense that it does not need any information or optimization beforehand, it simply relies on real-time load measurements at the BSs and user-reported SINR values. We furthermore present extensive simulation results, demonstrating the effectiveness of our self-organizing strategy and the impact of increased energy consumption on the user-perceived performance.

Keywords BS sleeping · Dense cellular networks · Load balancing · Self-organizing

1 Introduction

Wireless cellular networks have experienced immense growth in traffic loads over the last years as a consequence of the rapid proliferation of smartphones, tablets, and their bandwidth-hungry applications. A key option to further increase wireless network capacity is to deploy dense cellular networks (DCNs) since they allow for higher spectral reuse and efficiency (shorter communication range, and thus lower path loss).

The denser concentration of base stations (BSs) raises new and challenging issues compared with the traditional

macro cellular networks (MCNs), especially with regard to cell planning and traffic engineering [5]. Physical constraints will typically make it even harder to arrange BSs in an ideal hexagonal pattern, which causes the coverage areas to significantly overlap, and the natural cell regions to be irregularly shaped. As a result, the nominal traffic loads will tend to exhibit not only more spatial variation but also stronger temporal fluctuations. This variability in traffic could potentially result in severe load imbalances and performance degradation under existing BS sleeping strategies and traditional user association schemes.

BS sleeping strategies were introduced as a result of the rising concerns of energy consumption of wireless networks, both in terms of environmental impact and economic cost. In MCNs, BSs are responsible for about 60–80% of the total energy consumption [28], where a single BS may consume up to 90% of its peak energy consumption in the absence of any traffic due to cooling and pilot signalling [24]. In terms of economic costs, Nokia corporation recently estimated [19] that the global energy bill of radio access networks is currently over 72 billion Euros. These costs and the environmental impact caused by the massive energy consumption of cellular networks drives the need to improve their energy efficiency. A

✉ Bart Post
bart.post@cqm.nl

Sem Borst
s.c.borst@tue.nl

Hans van den Berg
j.l.vandenberg@tno.nl

¹ Eindhoven University of Technology, Eindhoven, The Netherlands

² Centrum Wiskunde and Informatica (Amsterdam), TNO (Dutch Scientific Research Institute, The Hague), University of Twente in Enschede, Enschede, The Netherlands

common approach to save energy is to switch BSs into low-power operational modes in the absence of traffic, e.g. sleep modes.

Although DCNs are expected to experience more variability in traffic, the high density of BSs also offers more flexibility than traditional MCNs to deal with this increased variability. With the overlapping cell areas in DCNs, switching BSs into sleep mode does not directly lead to coverage holes, as is often the case in MCNs, forcing the latter to be more conservative in switching off BSs. That means DCNs can potentially react to traffic dynamics on a much smaller timescale than MCNs. For example, DCNs may be able to react to locally appearing (and disappearing) hotspots of traffic demands on a timescale of (several) minutes, while MCNs can typically only react to day/night traffic patterns appearing on a timescale of hours to days due to the severe coverage degradation when switching off a macro BS. To fully harvest the potential energy savings and capacity gains in dense cell deployments, more refined load-aware BS sleeping strategies must be developed [2, 22].

An important issue is that reducing energy consumption by switching BSs into sleep mode basically reduces the system capacity. This is at odds with the primary goal of dense cellular networks: increasing network capacity. The latter is most important for optimizing user-perceived performances, which is typically done by applying load balancing schemes. From an energy perspective we wish to have a minimum number of active BSs, while optimizing the user-perceived performance ideally activates as many BSs as possible. Hence—as has also been mentioned by Zhou et al. [35]—we have two opposite objectives and a trade-off has to be made.

To complicate matters further, in DCNs, traffic conditions are typically strongly varying over time and hard to estimate, making manual traffic engineering highly impractical. This motivates the need for self-organizing strategies: measurement-based algorithms that realize excellent performance without requiring explicit prior knowledge of system parameters like traffic conditions. In this paper we present a self-organizing, load-aware strategy that makes a trade-off between energy consumption and user-perceived performance for DCNs, using a pre-specified trade-off parameter.

1.1 Discussion and related work

In the past few years “green cellular networks” has been an active research field providing many approaches to reduce energy consumption. In this section we give an overview of different approaches proposed in literature and point out how our approach is different from existing solutions.

In principle energy can be saved in two ways: reducing the downlink transmission power (e.g. [3, 33]) yet keeping the BSs active, and/or switching BSs into sleep mode in situations of low load conditions. We focus on the second option, i.e. switching BSs into sleep mode.

There are roughly three network modelling perspectives in the existing literature on BS sleeping strategies: models that focus on a single BS (e.g. [12, 13, 32]), models that focus on a single HetNet cell with a macro BS and several pico BSs (e.g. [10]), and models that consider the entire network (e.g. [1, 3, 9, 11, 14–17, 23, 25, 27, 29, 31, 34]). We will briefly discuss the first two approaches and then focus on network wide models since our approach belongs to the latter category.

First, when considering a single BS, energy can be saved by switching the BS into sleep mode when no more users are in service. Several papers exploit an M/G/1 queueing model to derive asymptotically (locally) optimal activation strategies [12, 13, 32] such as activation after a pre-optimized sleep period or when the number of users awaiting the activation hits a certain threshold. These queueing based methods are easy to implement but require a priori information to operate optimally and also place users in a queue when the BS is not active, potentially leading to unnecessary delays.

Secondly, in the case of a single HetNet cell the macro BS is typically always on [10, 21]. In such a scenario the authors optimize the operational modes of the pico BSs in the macro cell. Their approaches have the advantage that users are always directly placed in service, but limit the potential for saving energy as they do not consider the traffic conditions in neighbouring (HetNet) cells, nor allow for the macro BS to be turned off.

For the remainder of this section we focus on the cellular network at system level. In this setting, user assignments can specifically focus on active BSs, which eliminates delays experienced by users assigned to a sleeping BS. Moreover, BS sleeping strategies with system-wide awareness may better recognize when to (de-)activate a BS as they can account for traffic offloading to neighbouring active BSs, potentially leading to increased energy efficiency compared to local strategies. There are many results on models that consider the entire network and we further discuss them from three different perspectives: the considered network topology (regular or arbitrary), the proposed decision type of the algorithms (randomized or not randomized), and the considered user population (fixed or dynamic). We briefly consider each perspective, using them to position our work in relation to existing literature.

In terms of the network topology, several papers focus on the traditional macro cellular hexagonal BS positioning, presenting specific case studies [2], dealing with both

transmission power and operation modes [3], or using detailed user position information [31]. These results rely on regular BS positioning, which is no longer a fair assumption in DCNs. The works outlined below (including our work) do apply to arbitrary network topologies.

Several papers exploit stochastic geometry to find an asymptotically optimal number of active BSs, or probabilities that BSs are active [9, 23, 25, 29]. These all result in randomized strategies that can make a different decision when presented with the same load conditions. Other approaches [1, 11, 14–17, 27, 34] and—as we will see later—our approach consistently make the same decision under the same conditions, making them more reliable in the sense that the algorithms do not suffer from “unlucky tosses”.

Looking at the user population dynamics, we find many approaches that optimize BS operation modes for a specific (fixed) user population [1, 11, 14, 27, 34]. Jahid et al. [15, 16] on the other hand focus on user association specifically and on the use of on-site renewable energy sources (e.g. solar panels), and simply switch any BS into sleep mode during off-peak hours. We particularly mention the work of Zheng et al. [34], which applies game theory to include the effect that switching a BS into sleep mode leads to load increases at other BSs. All of these approaches are optimized for a static user population, potentially requiring a new optimization every time the user population changes. Considering the fast flow-level dynamics of DCNs, strategies optimized for static user populations may need to change operation modes faster than the optimal operation modes can be determined.

Although the literature overview given above is by no means exhaustive, it does paint a broader picture: there are little to no known strategies in green cellular networks that cover arbitrary network topologies with dynamic user populations and that make consistent decisions. One of the few exceptions is the work of Klessig et al. [17], which proposes a BS activation strategy inspired by the activation of cytotoxic killer cells in the immune system of mammals. However, even though their method is self-organizing in the sense that it does not need manual intervention during operation, it does require detailed information on the BS coverage areas and an a priori chosen BS hierarchy.

1.2 Main contributions

In this paper we analyse a weighted trade-off between energy consumption through switching BSs on and off and user-perceived performance in dense cellular networks. Optimizing the user performance is realized by applying load balancing user association schemes, and we furthermore introduce a trade-off parameter. To the best of our knowledge we are the first to analyse such a tunable trade-

off for dense cellular wireless networks: typically the user-perceived performance is taken as a hard constraint. We present an optimization problem representing the above-mentioned trade-off and derive properties of its optimal solutions. Using these properties we design a self-organizing strategy that dynamically (online) makes load-aware user association and BS operation decisions. Our strategy is self-organizing in the sense that it does not need any information or optimization beforehand, it simply relies on real-time load measurements at the BSs and user-reported signal-to-interference-plus-noise ratio (SINR) values. We furthermore present extensive simulation results, demonstrating the effectiveness of our self-organizing strategy and the role of the trade-off parameter.

1.3 Organization of this paper

The remainder of this paper is organized as follows. In Sect. 2 we give the model description and introduce some useful notations. In Sect. 3 we present our optimization problem and derive conditions and properties of optimal strategies. Then, in Sect. 4 we propose our dynamic approach which is based on the properties derived in Sect. 3. In Sect. 5 we show results of simulations we performed to gain insight in the performance characteristics of our proposed approach. Finally, in Sect. 6 we make some concluding remarks and mention directions for future research.

2 Problem statement and model description

The challenge is to dynamically adapt the set of active BSs and the user association to changing traffic demands such that a specific desired trade-off is realized between energy consumption and user-perceived performance. In our proposed approach, we first study this trade-off in a stationary regime: a period in time where the traffic demands (i.e. file transfer initiation rate, mean download file size) are constant. For such a stationary regime we formulate an optimization problem that represents the desired trade-off using a trade-off parameter. We then analyse the structure of the optimization problem to obtain properties for optimal user association and optimal BS (de-)activation. These properties are then used to design an optimal *dynamic* user association algorithm and sufficient conditions for dynamic BS (de-)activation. The user association algorithm and BS (de)activation rules will no longer depend on the stationary scenario and can react to changing traffic demands.

We will now proceed as follows. In the remainder of this section, we introduce the system model for a stationary scenario and we specify the power consumption model. In Sect. 3 we formulate the optimization problem for the

stationary scenario and derive properties of optimal user association and BS (de-)activation. In Sect. 4 we will describe how these properties are turned into a dynamic algorithm, the Green Shadow Price Assignment (GSPA) algorithm.

2.1 System model

We consider a system with L BSs, and we focus on downlink communication only. Within the considered area, BSs provide service to a time-varying set of users. For convenience, we assume that there is a discrete set of N user locations, which may be interpreted as a suitable discretization of the overall coverage area. A location represents a class of users that all have (approximately) equal physical transmission rate characteristics with respect to the BSs. At location n , users initiate file transfers (downloads) at a rate v_n . The sizes of the file transfers initiated by users in location n are independent and have mean β_n Mbit. As soon as a user has downloaded its file it leaves the system. We do not explicitly consider user mobility.

Remark 2.1 Together, the rates v_n and the mean file sizes β_n determine the traffic demands of the system and they may change over time. However, in the stationary regime that we currently consider they are constant (by definition).

The bit rate at which users are served depends on their experienced SINR values. We do not account for fast fading and consider only average SINR values. A BS l transmits at a fixed power in the time slots it is serving users. A user i does not receive the full transmission power of the BS, but a reduced portion due to path loss factors. For convenience we do not consider frequency-selective fading, which means that the path loss is independent of the frequency and only depends on the distance between the location and the BS. The Shannon formula implies that user i at location n can receive a maximum communication rate $R_{n,l}$ (in bits per second) from BS l given by

$$R_{n,l} = w \log(1 + \text{SINR}(n, l)), \quad (1)$$

where w is the fixed bandwidth available to the BS in Hz, and where $\text{SINR}(n, l)$ is the signal-to-interference-plus-noise ratio that user i experiences at location n when served by BS l . We assume that there is an interference-free spectrum allocation to the BSs such that each BS has a fixed bandwidth w . BSs apply a proportional fair scheduling policy as is also common in 4G LTE networks, such that the service rate experienced by users depends on the number of users in service at the BS as well as the SINR of that user.

Let $x_{n,l}$ be the fraction of the users that initiate a file transfer at location n that is allocated to BS l , and let $\mathbf{x}$ be a

vector representing all the individual values of $x_{n,l}$. Then the long term load, or resource utilization, of BS l can be expressed by (see e.g. [26])

$$\rho_l(\mathbf{x}) = \sum_{n=1}^N \frac{v_n \beta_n}{R_{n,l}} x_{n,l}. \quad (2)$$

We assume that the network has a centralized control unit, which may be realized by using Radio-over-Fiber (RoF) technology [18]. That means that at the BS site there is only a simple remote radio head, and all BS intelligence is located at a centralized entity. This has the advantage that important state information—specifically load estimates and user-reported SINR values—is known for the entire system, and can be used in the dynamic operation of the network.

Remark 2.2 At this point we wish to note that, although in our model and analysis we use a discrete set of locations, we aim to design decision rules that do not rely on these locations since the underlying discretization is most probably not available in practice. For the same reason, we also aim to avoid the specific use of v_n and β_n in the final decision rules. Instead we use estimates of the resource utilization at BSs to obtain load proxies and use these in our decisions. As a result the decision rules become self-organizing in the sense that they dynamically react to changing load conditions at the BSs. In other words, when the system changes from one stationary scenario to another, the decision rules automatically adjust to the new situation.

2.2 BS modes and power consumption model

A BS can be in three operational modes: active, sleep and start-up mode. A BS can directly go from active to sleep mode, but when a BS is switched from sleep to active mode, there is a start-up delay of T_{On} seconds before the BS is operational and can start serving users. The start-up delay is typically small (e.g. 1 second [12]) compared with the time-scale at which BSs are activated or de-activated (minutes). When a BS is switched to sleep mode, the users in service at that BS will be handed over to other BSs. Similarly, an activated BS may take over users of other BSs.

We adopt the widely-used (e.g. [10, 11, 25, 30, 32, 34]) load-dependent power consumption model of Auer et al. [7], where the power consumption P_l in Watt of a BS l is given by

$$P_l = P_0 + P \cdot \rho_l(\mathbf{x}). \quad (3)$$

Here, P_0 is the constant power consumption of an operational BS with no traffic, and $P \cdot \rho_l(\mathbf{x})$ is the load-dependent power consumption term for given load $\rho_l(\mathbf{x})$. When a BS is in sleep mode, we assume its power consumption

P_{Off} satisfies $0 \leq P_{\text{Off}} < P_0$. Moreover, in start-up mode, a BS has a power consumption of P_{ST} (e.g. $P_{\text{ST}} = 2P_0$ [12]).

3 Optimization problem and analysis

In this section we first formulate an optimization problem for a stationary regime of the system, where the stationary regime represents a period in which the rates v_n and mean file sizes β_n are constant. Thereafter we break down the optimization problem into two separate parts: user association and BS (de-)activation. In the first part we will show how to realize an optimal (in the sense of the trade-off objective) user association for a given set of active BSs. In the second part we derive sufficient conditions for activating a BS or putting a BS into sleep mode. The results presented in this section will serve as input for Sect. 4, where we will use the optimal user association and the sufficient conditions for (de-)activating BSs to design a dynamic control algorithm.

Let $s_l = 1$ if BS l is in active mode, and $s_l = 0$ if it is in sleep mode. The setup mode is not considered as it is a very short temporary mode preceding activation, having little effect on the overall power consumption. The objective is a trade-off between minimizing the total power consumption and optimizing the user perceived-performance. For the latter we specifically choose load balancing as is common in cellular networks [4, 6, 8]. Load balancing is realized by minimizing the highest BS load, and hence we can formulate the following minimization problem, where α is the desired trade-off factor between power minimization and load balancing, and U represents the maximum BS load (further explanation of the constraints is given below).

$$\min_{\mathbf{x}, \mathbf{s}, U} h(\mathbf{x}, \mathbf{s}, U) = \sum_{l=1}^L P_l(\mathbf{x}, \mathbf{s}) + \alpha \cdot U \quad (4a)$$

$$\text{sub: } P_l(\mathbf{x}, \mathbf{s}) = s_l(P_0 + P \cdot \rho_l(\mathbf{x})) + (1 - s_l)P_{\text{Off}}, \quad \forall l, \quad (4b)$$

$$\rho_l(\mathbf{x}) = \sum_{n=1}^N \frac{v_n \beta_n}{R_{n,l}} x_{n,l} \leq U, \quad \forall l, \quad (4c)$$

$$\sum_{l=1}^L x_{n,l} \cdot s_l = 1, \quad \forall n, \quad (4d)$$

$$s_l \in \{0, 1\}, \quad \forall l, \quad (4e)$$

$$x_{n,l} \geq 0, \quad \forall n, \forall l. \quad (4f)$$

The objective (4a) is minimized over the operation modes of the BSs, over the user association $\mathbf{x}$, and over the maximum BS load U . The user association $\mathbf{x}$ is included since users can only be assigned to active BSs. The user

association in combination with constraint (4c) and the minimization of U give rise to a load balancing problem, which is then weighted by a factor α with the power consumption. Constraint (4d) makes sure that (exactly all) traffic is only assigned to active BSs (all locations have coverage).

The problem (4a)–(4f) is a non-convex [due to constraints (4d)], mixed-integer, quadratically constrained [also due to constraints (4d)] quadratic program (QCQP—the objective contains quadratic terms), and in particular the non-convexity makes it hard to find (provably) globally optimal solutions. However, we do not wish to find optimal solutions for this formulation directly since we aim for decision rules that can be applied dynamically and in particular without knowledge of the values of v_n and β_n . We will now consider the optimization problem (4a)–(4f) in two separate parts: user association and the operation modes of the BSs.

Remark 3.1 The optimization problem (4a)–(4f) applies to a single-tier network. Multiple tiers may be incorporated by splitting the sum in the objective function over different sets of BSs, where each set of BSs represents a network tier with possibly different values for P_0 , P , and P_{Off} .

3.1 Optimal user association

To gain insight in the optimal user association, let us fix the operation modes of the BSs and consider the sub-problem of load balancing for the active BSs $\hat{\mathcal{L}}(\mathbf{s}) = \{\hat{l} \in \mathcal{L} \mid s_{\hat{l}} = 1\}$ for any given operational mode $\mathbf{s}$:

$$\min_{\mathbf{x}, U} \hat{L}P_0 + \sum_{\hat{l}=1}^{\hat{L}} P \cdot \rho_{\hat{l}}(\mathbf{x}) + \alpha \cdot U \quad (5a)$$

$$\text{sub: } \rho_{\hat{l}}(\mathbf{x}) = \sum_{n=1}^N \frac{v_n \beta_n}{R_{n,\hat{l}}} x_{n,\hat{l}} \leq U, \quad \forall \hat{l}, \quad (5b)$$

$$\sum_{\hat{l}=1}^{\hat{L}} x_{n,\hat{l}} = 1, \quad \forall n, \quad (5c)$$

$$x_{n,\hat{l}} \geq 0, \quad \forall n, \forall \hat{l}. \quad (5d)$$

where $|\mathcal{L}(\mathbf{s})| = \hat{L}$. The problem (5a)–(5d) is a linear programming problem (LP) with continuous decision variables $\mathbf{x}$ and U . This problem is a variation on the LP for the user association problem presented by Post and Borst [26], where the only difference is in the objective function, which is originally $\min_{\mathbf{x}, U} U$. Post and Borst [26] presented a dynamic user association algorithm—the shadow price assignment (SPA) algorithm—that realizes an optimal user association for the original objective function. We can

modify the SPA-algorithm to also find optimal user assignment fractions $\mathbf{x}^*$ for (5a)–(5d).

The idea of the original SPA-algorithm was to assign users to BSs using shadow prices y_i^1 for the BSs. The shadow prices are adapted over time depending on load proxies observed at the BSs, eventually leading to optimal user associations. We briefly explain the two most important differences with the original SPA-algorithm if we wish to apply it to find optimal solutions for the LP (5a)–(5d).

Let y_i be the shadow prices of BS $\hat{l}$, and let $\mathbf{y}$ denote the vector containing all shadow prices (including the BSs in sleep mode). The first, and perhaps most important difference appears in the optimality condition of user assignments. Using the Karush-Kuhn-Tucker optimality conditions, we can derive that an optimal user assignment $\mathbf{x}^*$ satisfies

$$x_{n,l}^*(\mathbf{y}) > 0 \quad \Rightarrow \quad \hat{l} \in \arg \min_{\{l': s_{l'}=1\}} \left\{ \frac{y_{l'} + P}{R_{n,l'}} \right\}, \quad (6)$$

breaking ties at random when the minimizer is not unique. Condition (6) may be interpreted as follows: given (optimal) shadow prices $\mathbf{y}$, the optimal user assignment allocates users to BSs with either a low shadow price or a high service rate. If all shadow prices are equal, users will simply be assigned to the BS l that provides them with the highest rate $R_{n,l}$. The only difference compared with the original SPA-algorithm is the added power consumption term P . Notice that condition (6) may be implemented without relying on locations by using user-dependent service rates $R_{i,l}$ for user i . This can be thought of as if each user has its own unique location. In practice, the service rates $R_{i,l}$ can be obtained by using user-reported SINR values.

The second most important difference is that the optimal shadow prices $\mathbf{y}$ satisfy $\sum_{i=1}^{\hat{L}} y_i^* = \alpha$ (instead of summing to one as in the original SPA-algorithm). More details about the modification of the original SPA-algorithm, and specifically the modified update step for the shadow prices, can be found in the “Appendix”. The practical implementation is summarized in Sect. 4.3.

3.2 Sufficient conditions for changing operation modes

We will now study the operational modes of BSs. For notational convenience we take $P_{\text{Off}} = 0$. The analysis for $P_{\text{Off}} > 0$ only leads to one added term P_{Off} in the conditions that we derive, and P_{Off} is often assumed negligible compared to P_0 .

¹ We continue to use identifiers $\hat{l}$ for BSs to stress that we are only dealing with active BSs in this subsection.

Considering that we are dealing with a non-convex, mixed-integer QCQP problem, we are not planning to find the optimal set of BSs for a given situation, but rather take a different approach. We consider the current BS loads and test if we can (expect to) improve the objective (4a) by activating BSs or by switching BSs into sleep mode. Then, after we have made a change we measure the new realized BS loads, at which point we start to repeat the process. Although we do not expect to always find optimal operation modes, the advantage of this approach is that it can dynamically react to changing load conditions and that the choices are clearly motivated and consistent: given the same situation, this approach will take the same decision.

Recall that we wish to avoid relying on the discretization into locations. In Sect. 3.1 we have seen a dynamic algorithm that realizes an optimal user association for a given set of active BSs without relying on the discretization. Let $\rho_l(\mathbf{s})$ denote the resulting load of BS l under the optimal user association realized by the modified SPA-algorithm for the set of active BSs represented by $\mathbf{s}$. This allows us to only focus on the operation modes $\mathbf{s}$ of BSs and the corresponding optimal BS loads $\rho_l(\mathbf{s})$, without having to worry about the user assignments too much.

Switching to sleep mode Suppose the system is currently using the operation modes $\mathbf{s}$. Then the objective value (4a) can be written as $h(\rho(\mathbf{s}))$, with

$$h(\rho(\mathbf{s})) = \sum_{l=1}^L s_l (P_0 + P \cdot \rho_l(\mathbf{s})) + \alpha \max_l \{\rho_l(\mathbf{s})\}. \quad (7)$$

Let us focus on any BS $\tilde{l}$ with operation mode $s_{\tilde{l}}$. Changing the operation mode of BS $\tilde{l}$ leads to the new mode $\tilde{s}_{\tilde{l}} = 1 - s_{\tilde{l}}$ (when we disregard the setup mode). This gives a new vector of operation modes $\tilde{\mathbf{s}}$ which is only different from $\mathbf{s}$ in the $\tilde{l}$ th coordinate. Then, it pays off to change the operation mode of BS $\tilde{l}$ if and only if

$$h(\rho(\mathbf{s})) > h(\rho(\tilde{\mathbf{s}})). \quad (8)$$

Although we may obtain the values $\rho(\mathbf{s})$ by using load measurements at the active BSs, it is unclear what the new loads $\rho(\tilde{\mathbf{s}})$ will become. However, if we can find estimates $\hat{\rho}_l(\tilde{\mathbf{s}})$ for the load values $\rho_l(\tilde{\mathbf{s}})$ such that $h(\rho(\mathbf{s})) > h(\hat{\rho}(\tilde{\mathbf{s}}))$, then for sure we have $h(\rho(\mathbf{s})) > h(\rho(\tilde{\mathbf{s}}))$ by optimality of $\rho_l(\tilde{\mathbf{s}})$ for $\tilde{\mathbf{s}}$. Hence, if we can obtain reliable estimates $\hat{\rho}_l(\tilde{\mathbf{s}})$ we can deduce a sufficient condition for changing the operation mode of BS $\tilde{l}$: BS $\tilde{l}$ is a candidate for changing its operation mode if

$$h(\rho(\mathbf{s})) > h(\hat{\rho}(\tilde{\mathbf{s}})). \quad (9)$$

We will now explain how to use the shadow prices of the modified SPA-algorithm to obtain estimates $\hat{\rho}(\tilde{\mathbf{s}})$ for the new load values $\rho(\tilde{\mathbf{s}})$. Considering that we are applying the

modified SPA-algorithm as described in Sect. 3.1, the load proxies used by this algorithm can represent the values $\rho_l(\mathbf{s})$. Then, to obtain values for $\hat{\rho}_l(\tilde{\mathbf{s}})$ we will use the instantaneous user population at the BS $\tilde{l}$. Let us focus on the active BS $\tilde{l}$ which serves the set of users $\mathcal{I}_{\tilde{l}}$ and which experiences a load of $\rho_{\tilde{l}}(\mathbf{s})$. Then we can associate this load proportionally to the users in service at BS $\tilde{l}$ as follows: user $i \in \mathcal{I}_{\tilde{l}}$ is responsible for a fraction $q_{i,\tilde{l}}$ of the load, where the fraction $q_{i,\tilde{l}}$ is given by

$$q_{i,\tilde{l}} = \frac{R_{i,\tilde{l}}^{-1}}{\sum_{j \in \mathcal{I}_{\tilde{l}}} R_{j,\tilde{l}}^{-1}}. \quad (10)$$

Suppose now that after BS $\tilde{l}$ has been switched to sleep mode, user i is handed over from BS $\tilde{l}$ to its new serving BS $l \in \arg \min_{l' \neq \tilde{l}} \{(\tilde{y}_{l'} + P)/R_{i,l'}\}$, where $\tilde{y}_{l'}$ are the shadow prices directly after BS $\tilde{l}$ has been switched into sleep mode (or directly after activation, see Sect. 4.2). Then the load that user i adds to BS l is the load it induced to BS $\tilde{l}$ multiplied by a factor $R_{i,\tilde{l}}/R_{i,l}$. Hence, if we handover user i from BS $\tilde{l}$ to BS l , the new load $\tilde{\rho}_l(\mathbf{s})$ of BS l can be estimated by

$$\tilde{\rho}_l = \rho_l(\mathbf{s}) + q_{i,\tilde{l}} \cdot \rho_{\tilde{l}}(\mathbf{s}) \frac{R_{i,\tilde{l}}}{R_{i,l}}. \quad (11)$$

Following these lines of reasoning, we can obtain values for $\hat{\rho}_l(\tilde{\mathbf{s}})$ (where $l \neq \tilde{l}$ and $s_l = 1$) by taking

$$\hat{\rho}_l(\tilde{\mathbf{s}}) = \rho_l(\mathbf{s}) + \sum_{i \in \mathcal{I}_{\tilde{l}}} \mathbb{1} \left[l = \arg \min_{l': \tilde{s}_{l'}=1} \left\{ \frac{\tilde{y}_{l'} + P}{R_{i,l'}} \right\} \right] q_{i,\tilde{l}} \cdot \rho_{\tilde{l}}(\mathbf{s}) \frac{R_{i,\tilde{l}}}{R_{i,l}}. \quad (12)$$

Activation Now suppose BS $\tilde{l}$ is in sleep mode, and we consider activating it. Then in a similar way as described above, the load of candidate BS $\tilde{l}$ after activation can be approximated by

$$\hat{\rho}_{\tilde{l}}(\tilde{\mathbf{s}}) = \sum_{l: s_l=1} \sum_{i \in \mathcal{I}_{\tilde{l}}} \mathbb{1} \left[\tilde{l} = \arg \min_{l': \tilde{s}_{l'}=1} \left\{ \frac{\tilde{y}_{l'} + P}{R_{i,l'}} \right\} \right] q_{i,l} \cdot \rho_l(\mathbf{s}) \frac{R_{i,l}}{R_{i,\tilde{l}}}. \quad (13)$$

Furthermore, the loads $\rho_l(\tilde{\mathbf{x}})$ of other (active) BSs after activation of BS l can be approximated by

$$\hat{\rho}_l(\tilde{\mathbf{s}}) = \rho_l(\mathbf{s}) \left(1 - \sum_{i \in \mathcal{I}_{\tilde{l}}} \mathbb{1} \left[\tilde{l} = \arg \min_{l': \tilde{s}_{l'}=1} \left\{ \frac{\tilde{y}_{l'} + P}{R_{i,l'}} \right\} \right] q_{i,l} \right). \quad (14)$$

The new load estimates given by (12) and (14) can be used to obtain sufficient conditions for switching to sleep mode and activation respectively. However, the sufficient condition for activation can be further improved. Since the

load estimates are based on instantaneous user populations, the following situation is very likely. By using the modified SPA-algorithm there are multiple BSs that maximize the load, i.e. $|\arg \max_l \{\rho_l(\mathbf{x}(\mathbf{s}))\}| > 1$. When we activate a currently sleeping BS $\tilde{l}$ it may attract load from some of the BSs in $\arg \max_l \{\rho_l(\mathbf{x}(\mathbf{s}))\}$, but very likely not all. This means that the maximum load among the BSs is not decreased according to the load estimates given by (14). Even though we indeed do not expect the newly activated BS $\tilde{l}$ to be able to alleviate all maximum loaded BSs, we can expect a cascading effect: BS $\tilde{l}$ takes over some load from BS $l \in \arg \max_l \{\rho_l(\mathbf{x}(\mathbf{s}))\}$, which in turn allows the BS l to take over some load from another BS $l' \in \arg \max_l \{\rho_l(\mathbf{x}(\mathbf{s}))\}$. This effect is eventually realized by the modified SPA-algorithm, but it is not captured by the load estimates given by Eq. (14).

To account for the cascading effect described above, we propose an extra step in determining load estimates for BS activation. First we determine the load estimates according to Eq. (14). If $\hat{\rho}_{\tilde{l}}(\tilde{\mathbf{s}})$ is higher than the old maximum load, then we do not expect to gain in the objective, and hence we can assume that $\hat{\rho}_{\tilde{l}}(\tilde{\mathbf{s}}) < \max_l \{\rho_l(\mathbf{s})\}$. Next we consider the set of BSs $\mathcal{L}^\circ$ that had a load equal to the maximum load and that did not offload any users to the newly activated BS $\tilde{l}$. We will then pretend that BS $\tilde{l}$ will take some load by averaging the loads of the BSs in $\mathcal{L}^\circ$ with BS $\tilde{l}$ in a weighted manner. First, the weights for BSs $l' \in \mathcal{L}^\circ$ are given by $w_{l'} = \min_{i \in \mathcal{I}_{\tilde{l}}} \{R_{i,l'}/R_{i,\tilde{l}}\}$, and $w_{\tilde{l}} = 1$, such that the weights represent the best possible ratio in which load from BS l' can be offloaded to BS $\tilde{l}$. Let $W = w_{\tilde{l}} + \sum_{l'} w_{l'}$ be the total weight, then the improved load estimates are given by

$$\tilde{\rho}_l(\tilde{\mathbf{s}}) = \begin{cases} \sum_{l' \in \mathcal{L}^\circ \cup \{\tilde{l}\}} \frac{w_{l'}}{W} \hat{\rho}_{l'}(\tilde{\mathbf{s}}) & \text{if } l \in \mathcal{L}^\circ \cup \{\tilde{l}\}, \\ \hat{\rho}_l(\tilde{\mathbf{s}}) & \text{otherwise.} \end{cases} \quad (15)$$

In short, we use load estimates (12) to obtain a sufficient condition for switching a BS into sleep mode. We use the load estimates (15) to obtain a sufficient condition for activating a BS, resulting in a set of candidate BSs for which a change in operation mode improves the objective.

The condition (9) considers changing the operation mode of a single BS at the time. Theoretically we can simply consider a set of BSs rather than a single BS, but when we change the operation modes of more than one BS it immediately becomes unclear how the new loads will behave. Therefore we have chosen to focus on one BS at a time. Nevertheless, condition (9) may still present several candidates of BSs for which a change in operation mode realizes a better objective. In the next section we will discuss which BSs are selected.

4 Dynamic control

In this section we propose a self-organizing strategy which dynamically makes load-aware user association and BS operation decisions. These decisions are based on the optimality conditions for the optimization problem (4a)–(4f) derived in the previous section and use load proxies observed at the active BSs.

For the user assignments we apply a modified version of the SPA-algorithm as described in Sect. 3.1, which requires to frequently update the shadow prices associated with the BSs. For the operation modes we use periodic decision moments in which we will change at most two BSs operation modes at each decision moment: at most one activation and at most one into sleep mode. The proposed strategy will therefore perform two types of updates, each on a different timescale:

1. Updates of the shadow prices, where at shadow price update moment $t_y^{(j)}$ we determine the j th iterate of the shadow prices denoted by $\mathbf{y}^{(j)}$.
2. Changes in the operation modes of the BSs, where at operation update moment $t_s^{(k)}$ we determine the k th iterate of the operation modes denoted by $\mathbf{s}^{(k)}$.

The shadow price updates are fully determined by the modified SPA-algorithm as described in Sect. 3.1, but we still need to specify how we update the operation modes. In addition, we will also specify what happens to the shadow prices $\mathbf{y}^{(j)}$ when the set of active BSs changes. These two issues will be covered in the next two subsections. In Sect. 4.3 we will give a precise description of our proposed strategy.

4.1 Operation mode updates

At operation update moments we treat the set of active and sleeping BSs separately. For the set of active BSs we check which ones are candidates for de-activation (switching into sleep mode). To do this we use the load estimates (12) and compute $\Delta_{\tilde{l}}(\mathbf{s}^{(k)}) = h(\rho(\mathbf{s}^{(k)})) - h(\hat{\rho}(\mathbf{s}^{(k)}))$ for each active BS $\tilde{l}$. From all active candidate BSs with $\Delta_{\tilde{l}}(\mathbf{s}^{(k)}) > 0$ we choose the BS $l^* = \max_{\tilde{l}: s_{\tilde{l}}^{(k)}=1} \{\Delta_{\tilde{l}}(\mathbf{s}^{(k)})\}$ and switch it into sleep mode.

Similarly, for all sleeping BSs we also compute $\Delta_{\tilde{l}}(\mathbf{s}^{(k)})$, but now using the load estimates (15). Then, from all sleeping candidate BSs with $\Delta_{\tilde{l}}(\mathbf{s}^{(k)}) > 0$ we activate BS $l^* = \max_{\tilde{l}: s_{\tilde{l}}^{(k)}=0} \{\Delta_{\tilde{l}}(\mathbf{s}^{(k)})\}$.

Remark 4.1 We allow for a BS activation and another BS to switch into sleep mode simultaneously. In a small

network, we do not expect this to happen, however in a large network the two BSs may be separated by enough distance that they locally do not influence each other. Moreover, in large networks the local traffic demands may vary a lot, where one area experiences a high load, whereas other areas are better off reducing their number of active BSs.

Remark 4.2 The activation and sleep rules presented above do not take into account any activation cost or deactivation costs. These costs may be implemented by setting different thresholds for $\Delta_{\tilde{l}}(\mathbf{s}^{(k)})$ that allow activation or deactivation.

4.2 Adjusting shadow prices after operation mode changes

In Sect. 3.1 we showed that for an optimal user assignment, the shadow price iterates $y_l^{(j)}$ sum up to α . If we activate a BS or put a BS into sleep mode, the number of active BSs changes and we either gain or lose a shadow price respectively. Consequently we have to adjust the shadow prices such that the sum over the shadow prices of active BSs equals α . The easiest way to do this is to reset all shadow prices to $\alpha/|\mathcal{L}(\mathbf{s}^{(k)})|$, where $\mathcal{L}(\mathbf{s}^{(k)})$ is the set of active BSs under operation mode vector $\mathbf{s}^{(k)}$. However, this method loses the information that the modified SPA-algorithm has already learned on the shadow prices, and therefore we will introduce updates for the shadow prices that maintain their mutual ratios.

First we will describe how the shadow prices $y_l^{(j)}$ are changed when we activate a BS. Suppose we have shadow prices $\mathbf{y}^{(j)}(\mathbf{s}^{(k)})$ for a given operation mode vector $\mathbf{s}^{(k)}$ where BS $\tilde{l}$ is in sleep mode (i.e. $s_{\tilde{l}}^{(k)} = 0$), and at decision time t_{k+1} we activate BS $\tilde{l}$ so that we get the new operation modes $\mathbf{s}^{(k+1)} = \mathbf{s}^{(k)} + \mathbf{e}_{\tilde{l}}$, where $\mathbf{e}_{\tilde{l}}$ is the $\tilde{l}$ th unit vector. Then the new shadow prices $\mathbf{y}^{(j)}(\mathbf{s}^{(k)} + \mathbf{e}_{\tilde{l}})$ are given by

$$y_l^{(j)}(\mathbf{s}^{(k)} + \mathbf{e}_{\tilde{l}}) = \begin{cases} y_l^{(j)}(\mathbf{s}^{(k)}) \frac{|\mathcal{L}(\mathbf{s}^{(k)})|}{|\mathcal{L}(\mathbf{s}^{(k)} + \mathbf{e}_{\tilde{l}})|} = y_l^{(j)}(\mathbf{s}^{(k)}) \frac{|\mathcal{L}(\mathbf{s}^{(k)})|}{|\mathcal{L}(\mathbf{s}^{(k)})| + 1} & \text{if } l \neq \tilde{l}, \\ \frac{\alpha}{|\mathcal{L}(\mathbf{s}^{(k)} + \mathbf{e}_{\tilde{l}})|} & \text{if } l = \tilde{l}, \end{cases} \quad (16)$$

which indeed sum up to α given that the original shadow prices sum up to α . Moreover, from (16) we can see that for any pair $\{l, l'\}$ of active BSs under $\mathbf{s}^{(k)}$ the ratios of the shadow prices are preserved.

Secondly we consider the situation where we put the BS $\tilde{l}$ into sleep mode. Hence we now assume that under $\mathbf{s}^{(k)}$ the BS $\tilde{l}$ is active, and switching it into sleep mode results in

operation mode vector $\mathbf{s}^{(k)} - \mathbf{e}_{\tilde{l}}$. Then the new shadow prices are given by

$$y_l^{(j)}(\mathbf{s}^{(k)} - \mathbf{e}_{\tilde{l}}) = \begin{cases} y_l^{(j)}(\mathbf{s}^{(k)}) \left(1 + \frac{y_{\tilde{l}}^{(j)}(\mathbf{s}^{(k)})}{\alpha - y_{\tilde{l}}^{(j)}(\mathbf{s}^{(k)})} \right) & \text{if } l \neq \tilde{l}, \\ 0 & \text{if } l = \tilde{l}, \end{cases} \quad (17)$$

where again the new shadow prices sum up to α and their mutual ratios are maintained.

We can now fully specify how the shadow prices are adapted. In case of assignment update moments, the shadow prices are updated according to the SPA-algorithm, as described in Sect. 3.1. In the case of operation update moments, the shadow prices $\mathbf{y}^{(j)}(\mathbf{s}^{(k)})$ are updated according to update step (16) in case of a BS activation or

The modified SPA-algorithm uses load proxies $\sigma_l^{(j)}$ to update the shadow prices. The formal definition of the proxies $\sigma_l^{(j)}$ as necessary to obtain theoretical optimality results is given in Sect. 1. In practice they can be defined as the fractional resource utilization of BS l between time $t_y^{(j-1)}$ and time $t_y^{(j)}$. The same kind of proxies are used to obtain load estimates $\rho_l^{(k)}$ for the operation mode updates, where the loads are estimated as

$$\rho_l^{(k+1)} = (1 - \varepsilon_s) \rho_l^{(k)} + \varepsilon_s \sigma_l^{(k)}. \quad (18)$$

The loads are hence estimated by a moving average principle, where ε_s determines the size of the updates, and thus how sensitive the load estimates are to the realized load proxies. We now have introduced all ingredients for our self-organizing scheme: the GSPA-algorithm, which is summarized in Algorithm 1.

GSPA-algorithm

Initialization:	Set $y_l^{(0)} = \alpha/L$, $\rho_l^{(0)} = 0$, $\sigma_l^{(0)} = 0$, $\forall l = 1, \dots, L$.
Shadow price updates	At decision times $t_y^{(j+1)}$, calculate $\mathbf{y}^{(j+1)}$ by applying update (20).
Operation mode updates	At decision times $t_s^{(k+1)}$, determine new load estimates (18), change operation modes of BS as in Section 4.1 by using the load estimates. Adjust shadow prices as in Section 4.2.
Assignment step:	Initiating user $i + 1$ is assigned (uniformly at random) to a BS $l \in \arg \min_{l'} \{(y_{l'}^{(i+1)} + P)/R_{i,l'}\}$.

according to update step (17) when a BS has been switched to sleep mode. In the situation where the update moment of the shadow prices and the update moment of the operation modes coincide, we first apply the shadow price updates given in Sect. 3.1 and then (16) or (17).

4.3 Algorithm specification

We will now give a formal algorithm description for the GSPA-algorithm: the Green Shadow Price user Association algorithm. There are two types of decision epochs: $t_y^{(j)}$ for the shadow price updates, and $t_s^{(k)}$ for the operation mode updates. The time between two same-type decision epochs is deterministic, i.e. we always have the same number of decision epochs per time unit (second). However, as the SPA-algorithm needs time to find new user associations, the rate of shadow price updates is higher than the rate of operation mode updates. Also, operation modes should not be updated too often as from an operational point of view it is undesirable to have a large number of BSs switching on and off on a fast time scale. However, operation modes should be updated often enough to follow statistical changes in the traffic demands.

5 Numerical results

In this section we present various results of numerical experiments we conducted to gain insight in the performance of the GSPA-algorithm. We consider an area of $1000 \text{ m} \times 500 \text{ m}$ with 10 BSs and used three different traffic scenarios:

1. *Uniform* The times between two users initiating a file transfer is independent and exponentially distributed with mean $1/5$ s. The positions of the users are independent, uniformly at random in the considered area.
2. *Moving Hotspot* Similar to the Uniform scenario, except that there are additional users initiating file transfers in the form of a non-stationary hotspot. The hotspot is a $200 \text{ m} \times 100 \text{ m}$ area, and moves to a new position after every 1000 s. It starts with its south-west corner at $(200, 100)$, then it moves to $(400, 100)$, $(600, 100)$, back to $(400, 100)$ and finally returns to $(200, 100)$, after which this pattern repeats. The hotspot has a relative file transfer initiation rate of 10 times the normal rate. This scenario is designed to test

if our algorithm can cope with (rather extreme) spatial and temporal variation in the traffic demands.

3. *Rush hour* This scenario represents rush hours, and basically switches between two Uniform scenarios of different file transfer initiation rates. During the first two hours the initiation rate is of a high intensity $\nu = 50$ file transfers per second, and during the next four hours it is of low intensity with $\nu = 5$ file transfer initiations per second. This pattern repeats over time.

The file sizes of users are independent and exponentially distributed with a mean of $\beta = 5$ Mbit. These chosen initiation and files size distributions are not essential for the GSPA-algorithm to operate, but are primarily used for convenience in the simulations. The load proxies $\sigma_l^{(j)}$ and $\rho_l^{(k+1)}$ are obtained by measuring the fractional resource utilization of BSs, as we suggested for practical systems in Sect. 4.3. The shadow prices update moments occur every second, and the operation mode update moments occur every ten seconds.

All simulations are based on 500,000 user arrivals. The number of users that can be in service at a BS simultaneously is limited by 100 users. If there are 100 users in service at a BS l and a new user initiates a connection and is assigned to BS l , then that user will be denied service and leaves the system directly without receiving service.

The values for P_0 , P are derived from the work of Auer et al. [7], leading to $P_0 = 13.6$ W and $P = 1$ W for Pico BSs. Furthermore, each BS transmits with equal power of 24 dBm over the spectrum it has available. The signal propagation and path loss follows the 3GPP urban micro model defined in 3GPP 36.814 v9.0.0, where the path loss (in dB) from BS l to user i is given by $PL(i, l) = 140.7 + 36.7 \log_{10}(d(i, l)/1000)$, and $d(i, l)$ is the distance in meters between user i and BS l . Furthermore, each BS has available spectrum of bandwidth 5 MHz, and we assume a thermal noise of -174 dBm/Hz.

The BS positions are generated uniformly at random, and for the Rush Hour scenario they are shown in Fig. 1.

5.1 Benchmarks

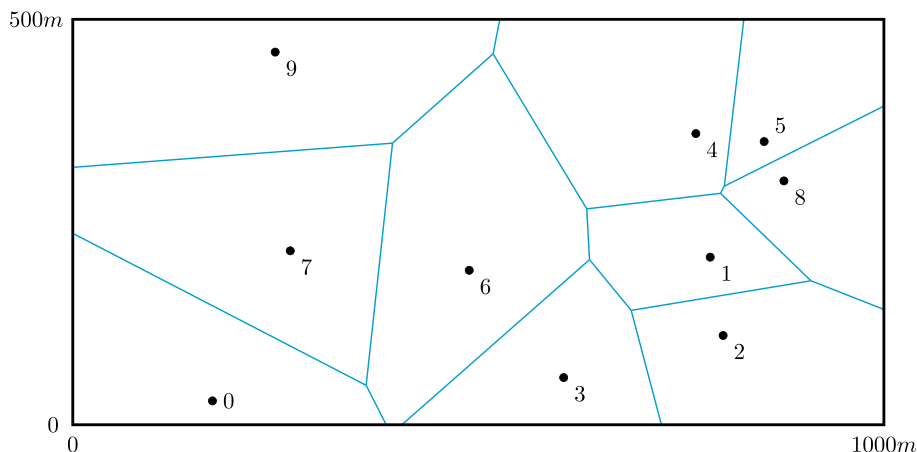
We compare our GSPA-algorithm to three benchmarks. Two benchmarks are based on queues with vacation times and were proposed by Guo et al. [12]. In both these benchmarks, a BS directly goes into sleep mode when it has no more users to serve. The activation policies are different:

- *SISL* Single Sleep. The BS is activated after a deterministic time since it was switched into sleep mode, regardless of if there are users to serve.
- *NLIM* N-limited. The BS is activated when there are N users in the queue at the BS.

In both SISL and NLIM benchmark systems, users are always assigned to the BS that provides the strongest received reference signal, even if the BS is in sleep mode. The SISL and NLIM systems treat BSs on an individual basis, and do not take into account that other BSs can take over the users of a BS that was switched off. The advantage of the SISL and NLIM systems is that they are easy to operate as they have very simple and intuitive activation and de-activation policies. However, the optimal number of users in the queue before a BS is activated in the NLIM system depends on the arrival rate that a BS is experiencing [12]. In practice this arrival rate may be unknown and time-varying. For the purpose of the simulations we have averaged the arrival rates at BSs over time (considering the Hotspot and Rush Hour scenarios) to determine the optimal number of users waiting in the queue before a BS is activated. The arrival rates per BS are obtained by considering the cell sizes of BSs as shown in e.g. Fig. 1.

The third benchmark that we consider will be referred to as the OPT system. The OPT system assigns users to BSs according to the same rule as the GSPA-algorithm, but it

Fig. 1 The BS positions with their natural cell areas in the Rush Hour scenario



uses predetermined optimal operation modes and shadow prices. These optimal values are obtained by discretizing the $500m \times 1000m$ area into $5m \times 5m$ squares, where each square represents a location. Then we use Cplex to find optimal solutions to (5a)–(5d) for each state that the user arrival process can be in (Uniform has only one state, Hotspot has 3 states, Rush Hour has 2 states). When the user arrival process changes state, the OPT system applies the corresponding optimal operation modes and shadow prices for that new state.

Remark 5.1 For each scenario we generated a sequence of users (file sizes, locations, and times between file transfer initiations), and each system was presented with the same sequence of users to obtain fair comparisons.

Remark 5.2 We chose the benchmarks SISL and NLIM because of their consistency: the evaluation of the systems is completely determined by the sequence of user as described in Remark 5.1 and does not include any probabilistic mechanism for BS activation or user association. It would be interesting to compare the GSPA algorithm against the policy proposed by Zhen et al. [34] or Klessig et al. [17], as these approach also takes network-wide effects into account. However, it is difficult to make a fair comparison since Zhen et al. do not provide a mechanism for practically obtaining an interaction graph, and Klessig et al. do not provide a mechanism for choosing a good BS hierarchy. These details are crucial for the respective approaches.

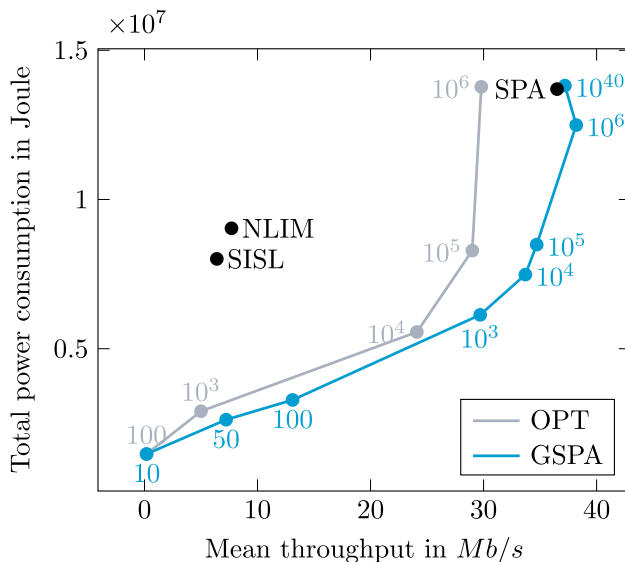


Fig. 2 Uniform scenario, mean perceived throughput versus total power consumption

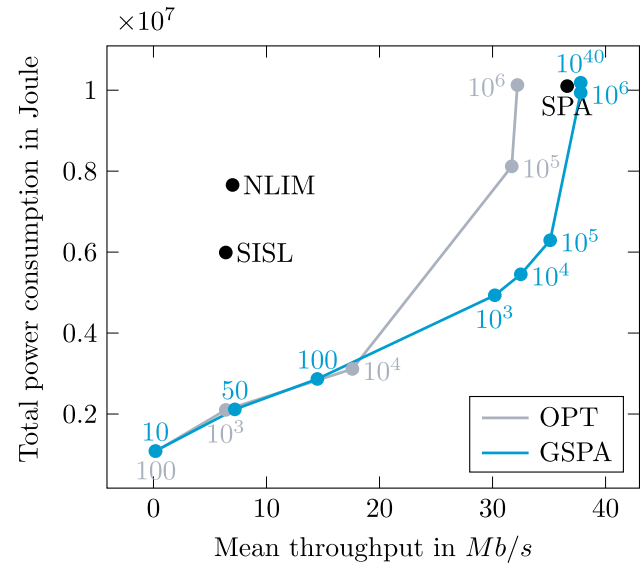


Fig. 3 Moving Hotspot scenario, mean perceived throughput versus total power consumption

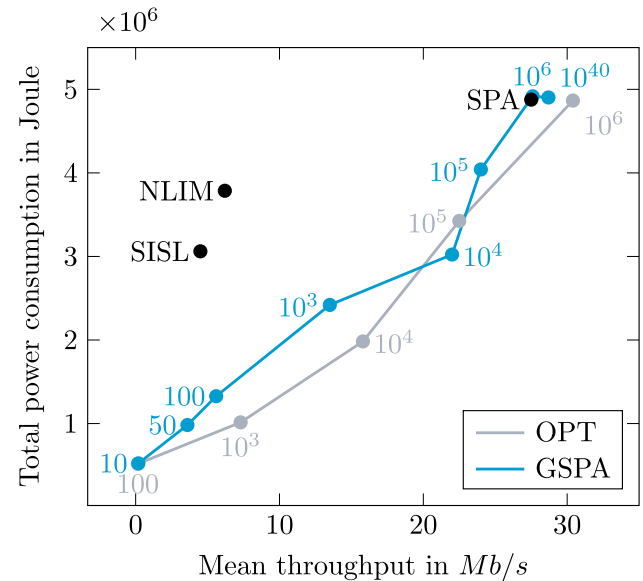


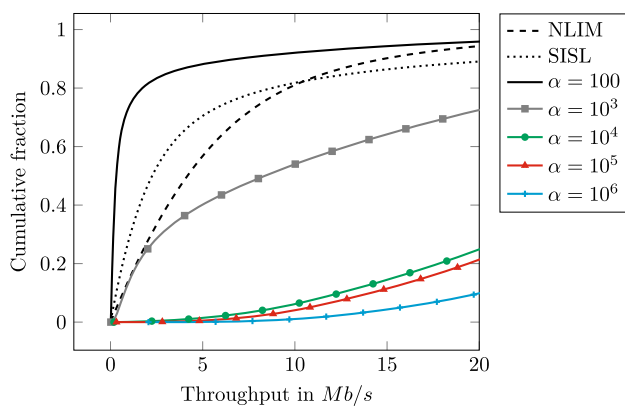
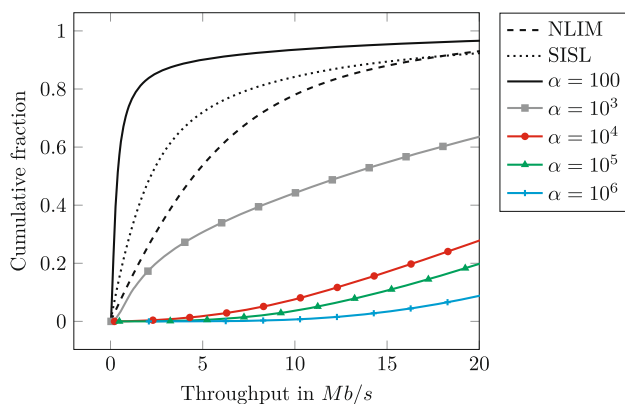
Fig. 4 Rush Hour scenario, mean perceived throughput versus total power consumption

5.2 Performance

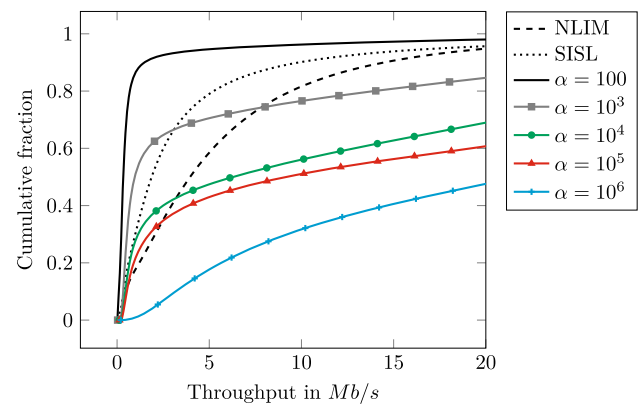
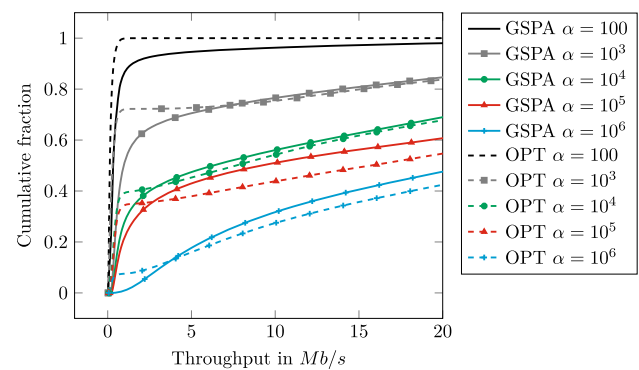
As described in Sect. 1 we are primarily interested in the power consumption and user-perceived performance. For the latter we consider two performance metrics: the number—or fraction—of service denials and the user-perceived throughput, where we define user-perceived throughput as the file size of a user divided by its total time spent in the system (hence it includes the time a user may be waiting in the queue of the SISL or NLIM systems for its BS to activate). Under equal power consumption, a lower fraction

Table 1 Percentage of service denials per system, with various values for α

System	Uniform	M Hotspot	Rush H
SISL	0	0	1
NLIM	0	0	1
GSPA 100	0	0	28
OPT 100	52	53	86
GSPA 10^3	0	0	17
OPT 10^3	6	29	69
GSPA 10^4	0	0	1
OPT 10^4	0	0	43
GSPA 10^5	0	0	1
OPT 10^5	0	0	29
GSPA 10^6	0	0	0
OPT 10^6	0	0	29

**Fig. 5** Uniform scenario, GSPA versus NLIM and SISL**Fig. 6** Moving Hotspot scenario, GSPA versus NLIM and SISL

of service denials and/or a higher user-perceived throughput implies a more efficient user association. In Figs. 2, 3 and 4 we plot the realized total power consumption (in Joule) versus the realized mean user-perceived throughput

**Fig. 7** Rush Hour scenario, GSPA versus NLIM and SISL**Fig. 8** Rush Hour scenario, GSPA versus OPT

(in Mbit) for the GSPA and OPT systems, where the plot marks are labelled with the respective values for α . The SISL and NLIM benchmarks are also included in these plots as single nodes. The SPA-nodes represent a system that applies the original SPA-algorithm and always has all BSs active, and can be thought of as the GSPA-system with $\alpha \rightarrow \infty$. Furthermore, Table 1 shows the realized service denials and we present plots of the user-perceived throughput in Figs. 5, 6, 7 and 8.

We can clearly see that as α increases, the power consumption of the GSPA and OPT systems is increasing, and simultaneously the user-perceived performance is improving: the percentage of service denials is decreasing and the user-perceived throughput is increasing. For $\alpha = 100$, the realized power consumption of the GSPA and OPT systems seem extremely favourable, but they have to be weighed against the high number of service denials.

In the Rush Hour scenario with $\alpha = 10^4$, the GSPA-algorithm has comparable service denials as the NLIM and SISL systems, against a slightly improved power consumption and a significantly higher mean user-perceived throughput. In this case, the user-perceived throughput is worse in the low-throughput region as users in the GSPA system may be offloaded to BSs that provide them with a

weaker signal, with the benefit of avoiding the power consumption of an extra BS.

The GSPA-algorithm outperforms the SISL and NLIM systems in the Uniform and Moving Hotspot scenarios when $\alpha = 10^3$ or $\alpha = 10^4$: it has a lower power consumption, less service denials and higher user-perceived throughputs. This suggests that optimal trade-offs may be expected for some α in between 10^3 and 10^4 , although Quality-of-Service constraints may require higher values for α . Also, it shows that we can improve both user-perceived performance and power consumption by considering the system as a whole and accounting for traffic offloading, instead of looking at each BS individually.

Curiously, the OPT system is not outperforming the GSPA system on all levels: for the Uniform and Moving Hotspot scenario with $\alpha = 10^5$ and $\alpha = 10^6$ the GSPA-algorithm realizes a lower power consumption. Moreover, in Fig. 8 we see that the GSPA system has significantly less users with very low throughputs (≤ 1 Mbit/s) for all investigated values of α . This can be explained by the dynamic behaviour of the GSPA-algorithm. Although the OPT system applies optimal shadow prices and operation modes for each specific (statistically different) state of the user file transfer initiation process, it does not respond to inherent variations in of this stochastic process. The GSPA-algorithm on the other hand may not directly have the optimal shadow prices, nor have an optimal set of active BSs, but it does respond to variations in the user arrival process and clearly that comes with some gains.

The plots in Figs. 2, 3 and 4 can act as a guide for operators to choose the best value for the trade-off parameter α . For both the Uniform and Moving Hotspot scenarios, a trade-off value $\alpha = 10^4$ appears a very good choice: the increase in power consumption compared to $\alpha = 10^3$ also comes at a significant improvement in mean throughput, but for $\alpha > 10^4$ the increase in power consumption only comes at a marginal improvement in the mean user-perceived throughput. We have also considered plots where the (arithmetic) mean user-perceived throughput is replaced by the geometric mean to put more weight on users with low experienced throughputs, and the conclusions drawn in this paper also apply to the geometric mean user-perceived throughputs.

Finally, observe that the GSPA-algorithm operates without any a priori information, in contrast to the OPT, NLIM and SISL systems, and solely bases its decisions on load proxies determined at the BSs and SINR values reported by the users. In light of this property, the performance of the GSPA-algorithm is remarkably favourable compared to the considered benchmarks.

6 Conclusion

In this paper we presented a self-organizing green load balancing algorithm, the GSPA-algorithm, specifically designed to deal with the many overlapping cells and the cell load fluctuations appearing in dense cellular networks. We formulated an optimization problem for a trade-off between power consumption and user-perceived performance and derived sufficient conditions for activating BSs and switching BSs into sleep mode. Furthermore, we constructed a user assignment strategy that realizes an optimal user assignment in terms of the trade-off for a given set of active BSs. These results were then used to design the GSPA-algorithm. The GSPA-algorithm relies on load measurements at BSs and SINR measurements reported by users, to make a tunable trade-off between power consumption and user-experienced performance by activating BSs or putting BSs into sleep mode and also by adapting the user assignment.

Extensive simulations demonstrated the effectiveness of the GSPA-algorithm to dynamically react to changing load conditions without other information than load proxies at the BSs and SINR measurements from users. The GSPA-algorithm realized both a lower power consumption and better user-perceived performances (fewer service denials, higher perceived throughput) than two considered benchmarks. Moreover, by tuning the trade-off, the simulations clearly show a change from minimizing power consumption towards optimizing user-perceived performance.

To the best of our knowledge, this is the first self-organizing BS sleeping strategy designed for dense cellular networks. We wish to stress the fact that the GSPA-algorithm realizes good performance without the need of prior optimization. An interesting direction for future research is to improve the performance of the GSPA-algorithm for large systems by locally (geographically) clustering the BSs in smaller sub-systems and hence increasing the rate of local self-organization.

Acknowledgements This research is partly funded by NWO Gravitation project Networks, Grant Number 024.002.003.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix: Modified SPA-algorithm

In this section we provide further details about the modification of the SPA-algorithm [26] so that it can be applied to the optimization problem (5a)–(5d). In the setting of the original SPA-algorithm, all loads were equal in the optimal solution. In the current setting however, this complete load balancing property is lost due to the “price” of extra power consumption when offloading users to a less favourable BS (with lower user-experienced SINR values). That means that in the optimal solution $(\mathbf{x}^*(\mathbf{s}), U^*(\mathbf{s}))$ to (5a)–(5d) some BSs $\hat{l}$ have an optimal load $\rho_{\hat{l}}(\mathbf{x}^*(\mathbf{s}))$ strictly lower than the optimal maximum load $U^*(\mathbf{s})$. This influences the way in which we have to update the shadow prices. Further analysis of the Lagrangian dual problem to (5a)–(5d) gives us the complementary slackness condition: $y_i^* > 0$ implies $\rho_{\hat{l}}(\mathbf{x}^*) = U^*$. In other words, in the optimal solution, BSs with a positive shadow price have an optimal load equal to the maximum load. BSs with an optimal load lower than the maximum load have their shadow prices equal to 0. The update step for the modified SPA-algorithm has to take into account that shadow prices may become zero, and furthermore reflect that for all BSs $\hat{l}$ with optimal shadow price $y_i^* > 0$, the loads should be equal to the maximum load, and hence all loads of BSs with positive optimal shadow price have equal loads. Hence, rather than looking at the system wide average load, we will use the system wide average load conditioned on the shadow price being positive:

$$\rho_+(\mathbf{x}, \mathbf{y}) = \frac{1}{|\mathcal{L}_+(\mathbf{y})|} \sum_{\hat{l} \in \mathcal{L}_+(\mathbf{y})} \rho_{\hat{l}}(\mathbf{x}), \quad (19)$$

where $\mathcal{L}_+(\mathbf{y})$ is the set of BSs with positive shadow price y_i : $\mathcal{L}_+(\mathbf{y}) = \{\hat{l} \in \mathcal{L} : y_i > 0\}$.

We will now present the modified update step. The update step for the shadow prices will only have to balance the loads of BSs with strictly positive shadow prices. Let $B^{(i)}$ be the file size of user i , let the instantaneous load $\sigma_0^{(i)}$ that user i brings to BS $\hat{l}_i$ (and hence into the system) be defined by $\sigma_0^{(i)} = \frac{B^{(i)}}{R_{n,\hat{l}_i}}$, and furthermore $\sigma_i^{(i)} = \sigma_0^{(i)} \mathbb{1}[\hat{l} = \hat{l}_i]$ is the load brought to BS $\hat{l}$. The modified mean load proxy $\sigma_+^{(i)}(\mathbf{y})$ is then given by $\sigma_+^{(i)}(\mathbf{y}) = \sigma_0^{(i)} \mathbb{1}[y_i > 0]$. Let $\mathcal{Y}_\alpha = \{\mathbf{y} \mid \sum_{i=1}^{\hat{L}} y_i = \alpha, y_i \geq 0\}$ be the set of feasible shadow prices. Then the update step of the modified SPA-algorithm for the shadow price iterates $\mathbf{y}^{(i)}$ is given by

$$\mathbf{y}^{(i+1)} = \Pi_{\mathcal{Y}_\alpha} \left[\mathbf{y}^{(i)} + \varepsilon^{(i)} \left(\boldsymbol{\sigma}^{(i)} - \frac{\sigma_+^{(i)}(\mathbf{y}^{(i)})}{|\mathcal{L}_+(\mathbf{y}^{(i)})|} \mathbf{1}_L \right) \right], \quad (20)$$

where $\Pi_{\mathcal{Y}_\alpha}[\mathbf{y}]$ is the projection of $\mathbf{y}$ to the closest vector in $\mathcal{Y}_\alpha$ in the Euclidean sense. The projection is needed because we can not guarantee by the updates themselves that shadow prices remain non-negative. We even suspect that when α is small compared to P many optimal shadow prices will be zero, reflecting that only a few BSs will have their loads equal to the maximum load. If we then do not actively correct the shadow prices, they turn negative when a BS has a load lower than the average load.

With the above-described modifications to the original SPA-algorithm, we have obtained the modified SPA-algorithm for the updates of the shadow prices. The framework of Kushner and Yin [20, Thm. 8.2.5] can be used to conclude that the modified SPA-algorithm realizes optimal user assignments in the long run (a formal proof of this statements is not the main focus of this paper and is omitted due to space restrictions).

References

1. Aissaoui Ferhi, L., Sethom, K., Choubani, F., & Bouallegue, R. (2018). Multiobjective self-optimization of the cellular architecture for green 5G networks. *Transactions on Emerging Telecommunications Technologies*, 29(10), e3478.
2. Ajmone Marsan, M., Chiaraviglio, L., Ciullo, D., & Meo, M. (2013). On the effectiveness of single and multiple base station sleep modes in cellular networks. *Computer Networks*, 57(17), 3276–3290.
3. Alam, A. S., Dooley, L. S., & Poulton, A. S. (2013). Traffic-and-interference aware base station switching for green cellular networks. In *Proceedings IEEE international workshop on computer aided modeling and design of communication links and networks* (pp. 63–67). IEEE.
4. Ali, M. S., Coucheney, P., & Coupechoux, M. (2016). Load balancing in heterogeneous networks based on distributed learning in near-potential games. *IEEE Transactions on Wireless Communications*, 15(7), 5046–5059.
5. Andrews, J. G., Claussen, H., Dohler, M., Rangan, S., & Reed, M. C. (2012). Femtocells: Past, present, and future. *IEEE Journal on Selected Areas in Communications*, 30(3), 497–508.
6. Andrews, J. G., Singh, S., Ye, Q., Lin, X., & Dhillon, H. S. (2014). An overview of load balancing in HetNets: Old myths and open problems. *IEEE Wireless Communications Magazine*, 21(2), 18–25.
7. Auer, G., Giannini, V., Desset, C., Godor, I., Skillermark, P., Olsson, M., et al. (2011). How much energy is needed to run a wireless network? *IEEE Wireless Communications Magazine*, 18(5), 40–49.
8. Borst, S. C., Hampel, G., Saniee, I., & Whiting, P. (2005). Load balancing in cellular wireless networks. In M. G. C. Resende & P. M. Pardalos (Eds.), *Handbook of optimization in telecommunications* (pp. 941–978). Boston: Springer.
9. Di Renzo, M., Zappone, A., Lam, T. T., & Debbah, M. (2018). System-level modeling and optimization of the energy efficiency in cellular networks—A stochastic geometry framework. *IEEE Transactions on Wireless Communications*, 17(4), 2539–2556.
10. Dini, P., Miozzo, M., Bui, N., & Baldo, N. (2013). A model to analyze the energy savings of base station sleep mode in LTE

- HetNets. In *IEEE international conference on green computing and communications and IEEE internet of things and IEEE cyber, physical and social computing* (pp. 1375–1380). IEEE.
11. Gong, J., Thompson, J. S., Zhou, S., & Niu, Z. (2014). Base station sleeping and resource allocation in renewable energy powered cellular networks. *IEEE Transactions on Communications*, 62(11), 3801–3813.
 12. Guo, X., Niu, Z., Zhou, S., & Kumar, P. (2016). Delay-constrained energy-optimal base station sleeping control. *IEEE Journal on Selected Areas in Communications*, 34(5), 1073–1085.
 13. Guo, X., Zhou, S., Niu, Z., & Kumar, P. (2013). Optimal wake-up mechanism for single base station with sleep mode. In *Proceedings 25th International Teletraffic Congress (ITC)* (pp. 1–8). IEEE.
 14. Hossain, F., Munasinghe, K. S., & Jamalipour, A. (2018). Multi-operator cooperation for green cellular networks with spatially separated base stations under dynamic user associations. *IEEE Transactions on Green Communications and Networking*, 3(1), 93–107.
 15. Jahid, A., Shams, A. B., & Hossain, M. F. (2018). Dynamic point selection CoMP enabled hybrid powered green cellular networks. *Computers & Electrical Engineering*, 72, 1006–1020.
 16. Jahid, A., Shams, A. B., & Hossain, M. F. (2018). Green energy driven cellular networks with JT CoMP technique. *Physical Communication*, 28, 58–68.
 17. Klessig, H., Ohmann, D., Reppas, A. I., Hatzikirou, H., Abedi, M., Simsek, M., et al. (2016). From immune cells to self-organizing ultra-dense small cell networks. *IEEE Journal on Selected Areas in Communications*, 34(4), 800–811.
 18. Koonen, A. M. J., Larrode, M. G., Ng'Oma, A., Wang, K., Yang, H., Zheng, Y., et al. (2008). Perspectives of radio-over-fiber technologies. In *Conference on optical fiber communication/national fiber optic engineers conference*. IEEE.
 19. Kuosa, H. (2019). *Nokia has ambitious plans to reduce network power consumption*. Retrieved April 23, 2019 from <https://www.nokia.com/blog/nokia-has-ambitious-plans-reduce-network-power-consumption>.
 20. Kushner, H., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications* (Vol. 35). Berlin: Springer.
 21. Lassila, P., Gebrehiwot, M. E., & Aalto, S. (2019). Optimal energy-aware load balancing and base station switch-off control in 5G HetNets. *Computer Networks*, 159, 10–22.
 22. Li, Y., Zhang, Y., Luo, K., Jiang, T., Li, Z., & Peng, W. (2018). Ultra-dense HetNets meet big data: Green frameworks, techniques, and approaches. *IEEE Communications Magazine*, 56(6), 56–63.
 23. Liu, C., Natarajan, B., & Xia, H. (2016). Small cell base station sleep strategies for energy efficiency. *IEEE Transactions on Vehicular Technology*, 65(3), 1652–1661.
 24. Oh, E., Krishnamachari, B., Liu, X., & Niu, Z. (2011). Toward dynamic energy-efficient operation of cellular network infrastructure. *IEEE Communications Magazine*, 49(6), 56–61.
 25. Peng, J., Hong, P., & Xue, K. (2014). Stochastic analysis of optimal base station energy saving in cellular networks with sleep mode. *IEEE Communications Letters*, 18(4), 612–615.
 26. Post, B., & Borst, S. C. (2017). Load-driven cell assignment algorithms for dense pico-cell networks. In *Proceedings 29th International Teletraffic Congress (ITC)* (Vol. 1, pp. 37–45). IEEE.
 27. Renga, D., Hassan, H. A. H., Meo, M., & Nuaymi, L. (2018). Energy management and base station on/off switching in green mobile networks for offering ancillary services. *IEEE Transactions on Green Communications and Networking*, 2(3), 868–880.
 28. Rengarajan, B., Rizzo, G., & Ajmone Marsan, M. (2011). Bounds on QoS-constrained energy savings in cellular access networks with sleep mode. In *Proceedings 23rd International Teletraffic Congress (ITC)* (Vol. 1, pp. 47–54). IEEE.
 29. Soh, Y. S., Quek, T. Q. S., & Kountouris, M. (2013). Dynamic sleep mode strategies in energy efficient cellular networks. In *Proceedings IEEE international conference on communications* (pp. 3131–3136). IEEE.
 30. Tabassum, H., Siddique, U., Hossain, E., & Hossain, M. J. (2014). Downlink performance of cellular systems with base station sleeping, user association, and scheduling. *IEEE Transactions on Wireless Communications*, 13(10), 5752–5767.
 31. Vereecken, W., Deruyck, M., Colle, D., Joseph, W., Pickavet, M., Martens, L., et al. (2012). Evaluation of the potential for energy saving in macrocell and femtocell networks using a heuristic introducing sleep modes in base stations. *EURASIP Journal on Wireless Communications and Networking*, 2012(1), 170.
 32. Wu, J., Zhou, S., & Niu, Z. (2013). Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks. *IEEE Transactions on Wireless Communications*, 12(8), 4196–4209.
 33. Zhai, X., Guan, X., Zhu, C., Shu, L., & Yuan, J. (2018). Optimization algorithms for multiaccess green communications in internet of things. *IEEE Internet of Things Journal*, 5(3), 1739–1748.
 34. Zheng, J., Cai, Y., Chen, X., Li, R., & Zhang, H. (2015). Optimal base station sleeping in green cellular networks: A distributed cooperative framework based on game theory. *IEEE Transactions on Wireless Communications*, 14(8), 4391–4406.
 35. Zhou, S., Gong, J., Yang, Z., Niu, Z., & Yang, P. (2009). Green mobile access network with dynamic base station energy saving. *ACM MobiCom*, 9, 10–12.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Bart Post studied mathematics at Utrecht University and obtained his Bachelor degree in 2012. He then pursued a Master degree in applied mathematics at Eindhoven University of Technology (TU/e), specializing in Discrete Mathematics and Applications. In 2014 he graduated with a thesis on Verifying and optimising disjoint paths in ISP networks, written in cooperation with TNO. In November 2014 Bart started a Ph.D. project at TU/e of which the results are

presented in his dissertation entitled “Load-driven self-organization of Radio-over-Fibre enabled dense cellular networks”. The Ph.D. project was a cooperation between the Department of Mathematics and Computer Science and the Department of Electrical Engineering, and was part of the Dutch research consortium NETWORKS. As of January 2020, Bart is employed at CQM B.V. in Eindhoven, The Netherlands.



Sem Borst has been a Full Professor in Stochastic Operations Research in the Department of Mathematics and Computer Science at Eindhoven University of Technology (TU/e) since 1998. He also had a (part-time) position in the Mathematics of Systems Research Department at Bell Laboratories in Murray Hill, USA, from 1995 to 2019, and was a Senior Researcher at the Center for Mathematics and Computer Science (CWI) in Amsterdam from 1998 to 2006.

His main research interests are in the area of performance evaluation and resource allocation algorithms for large-scale stochastic networks, in particular computer-communication systems. Sem was (co-)recipient of the best-paper awards at SIGMETRICS/Performance 1992 and IEEE Infocom 2003, the 2001 Yosef Levy Prize, the 2005 Van Dantzig Prize, and the 2017 ACM SIGMETRICS Achievement Award. He serves as co-editor-in-chief of the ACM Transactions on Modeling and Performance Evaluation of Computing Systems (ToMPECS) and as associate editor of Queueing Systems and Stochastic Models.



Prof. Dr. Hans van den Berg (M.Sc. and Ph.D. degree in Mathematics from Utrecht University, The Netherlands, in 1986 and 1990, respectively) has been working for more than 25 years in the field of planning, performance optimization and reliability of mobile communication networks, currently focusing on (Beyond) 5G. He has been active in many national and European research projects and programs. He was co-founder and vice-chair of COST

Services' (COST ACROSS, 2014–2018). Hans van den Berg has co-organized and -chaired many international conferences and workshops, and he has been active as TPC member of numerous international conferences. He has published more than 175 refereed papers in international journals and conference proceedings, and is co-editor of four books. In 2017 he received the ITC Arne Jensen lifetime award for his outstanding contributions in the field. Hans van den Berg started his professional career in 1990 at the research division of KPN, the Dutch incumbent telco. Since 2003 he is with TNO and holds a part-time position as full professor within the research group Design and Analysis of Communication Systems at the University of Twente.

Action IC1304 'Autonomous Control for a Reliable Internet of