Belief Propagation for Energy Efficiency Maximization in Wireless Heterogeneous Networks

Sang Hyun Lee, *Member, IEEE*, Mintae Kim, Hunmin Shin, and Inkyu Lee, *Fellow, IEEE*

***Abstract*— In this article, we study an energy efficient management of two-tier heterogeneous cellular networks (HetNets) which consist of one macro base station (BS) and multiple micro base stations. This article presents a distributed user association algorithm that maximizes the network-wide energy efficiency (EE) in HetNets. A subset of BSs that support only a small number of users can be turned off to save the energy consumption. By turning off BSs in the HetNet and offloading serving users to adjacent active BSs, the network-wide energy consumption is minimized, while the sum throughput is maximized. To solve the problem efficiently, we introduce a new approach based on a message-passing framework and derive a distributed load balancing algorithm. The proposed method provides a very efficient solution with reduced computational complexity compared to existing schemes. Simulation results verify that the proposed algorithm outperforms other conventional load balancing strategies.**

***Index Terms*— Energy-efficiency maximization, belief propagation, wireless heterogeneous networks.**

I. Introduction

OVER the past decades, demands for a variety of services in wireless mobile communication systems have dramatically increased [1]. Meanwhile, this has also led to growing energy consumption in communication infrastructures, which raises a serious concern from both operators and environments' perspectives. The 5th generation (5G) wireless system is predicted to incur 150-170% increases in total network energy consumption by 2026, with the largest increases in cellular networking operations [2]. Therefore, the network operators are interested in technologies that can enhance energy efficiency (EE) to reduce energy costs of the wireless network [3]. Recent research has been focused on the optimization of the EE towards environmental and economic sustainability. In response to this, there have been extensive efforts to design the communication networks from the EE's point of view [4]–[8]. The EE is defined as the ratio of the achievable throughput to the energy consumption. In order to enhance the EE, the network throughput is maximized, while the energy consumption is minimized.

In the mean time, a heterogeneous network (HetNet) has been regarded as an enabling technology for enhancing the EE [9]–[11]. A typical HetNet consists of a macro cell and a number of small cells. A multiuser multiple input multiple output (MU-MIMO) HetNet has been examined [12] where a joint linear precoder design maximizes the EE. The authors in [13] have provided simple models for energy consumption with different BS types and derived characteristics for micro BSs. In [14]–[22], several EE maximization strategies have been presented by turning off a subset of unused BSs to reduce the energy consumption. Similar optimization based techniques handling the BS switching also include cell zooming [23], [24], self-organizing networks [25], and energy harvesting [26], [27].

Furthermore, such switching-off policies have been addressed in various aspects of applications. In [28], cell zooming and sleeping mechanisms have been analyzed for small cells using stochastic geometry. Also, a node-degree centrality oriented graph-theory based approach has been developed for power savings in HetNets [29]. In addition, financial models among multiple operators have been considered in multi-objective optimization [30] and game theory [31].

Some other studies have taken into account link qualities between BSs and users for energy efficient BS-user association [15], [32]. The aforementioned conventional works have mostly focused on maximizing the sum of individual EEs. It may be often important to directly consider a global metric of the network-wide EE, which is defined as the total sum rate divided by the total energy consumption from the whole network point of view [7]. However, the network-wide EE has a quite complicated structure which requires a special mathematical consideration of its optimization for energy efficient network management.

To handle this, we aim to develop a distributed load balancing algorithm that maximizes the network-wide EE of the HetNet using the message-passing framework based on belief propagation [33], [34]. A distributed algorithm has the following advantages: First, it allows to solve a computationally challenging problem by decomposing it to obtain local solutions and by combining them to yield the global solution. Second, it readily adapts to a local topology in the network by updating the corresponding local configuration.

A simple way for reducing the total energy consumption is to turn off some BSs. However, it may lead to a throughput loss, and thus it is critical to identify a right set of turned

Manuscript received April 1, 2020; revised July 23, 2020; accepted September 5, 2020. Date of publication September 17, 2020; date of current version January 8, 2021. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) under Grant 2017R1A2B3012316, in part by the Information Technology Research Center (ITRC) supervised by the Institute of Information Communications Technology Planning and Evaluation (IITP) under Program IITP-2020-0-01749, and in part by the Korea University Grant. The associate editor coordinating the review of this article and approving it for publication was J. Yang. *(Corresponding author: Inkyu Lee.)*

Sang Hyun Lee, Mintae Kim, and Inkyu Lee are with the School of Electrical Engineering, Korea University, Seoul 02841, South Korea (e-mail: sanghyunlee@korea.ac.kr; wkd2749@korea.ac.kr; inkyu@korea.ac.kr).

Hunmin Shin is with Samsung Electronics, Suwon 16677, South Korea (e-mail: semo0617@gmail.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2020.3023079

off BSs for maximizing the network-wise EE. The proposed algorithm appropriately configures the on-off states of the BSs with the joint consideration of the user association that achieves the best network sum throughput. To this end, a belief propagation algorithm conducts a clustering task to obtain the user association in a distributed way. Individual BSs exchange the information about whether it should be in an operating mode via a real number quantity called a message. Based on the messages, the total number of active BSs is identified to optimize the EE performance.

Since the network-wide energy consumption can be estimated with this information, the optimization of the EE performance suffices to ensure the maximum network throughput. Thus, each BS associates users to maximize the total throughput of its serving users. All decisions are made by BSs and users via message exchanges in an autonomous way. By virtue of the message passing operation, the HetNet load balancing can be optimized with low complexity and, more importantly, in a distributed manner. As compared to the previous work [22], which has also applied belief propagation to determine the switch-off state of BSs with the objective of minimizing solely the total energy consumption, the proposed approach enables to jointly compromise nontrivial trade-off relationship between the energy consumption and the sum throughput to maximize the network-wide EE in heterogeneous networks.

Major contributions of this article are summarized as follows:

- An optimization is formulated to maximize the network-wide EE and readily poses a mixed-integer nonlinear formulation. This article develops a graphical model that enables to tackle the total EE in an autonomous way so that each BS determines its operating state.
- In two-tier HetNets, a distributed algorithm is derived to associate users with operating BSs with the objective of obtaining the largest total EE via a belief-propagation based message passing framework. The beauty of the algorithm lies in the fact that the computational loads are distributed over all network nodes. This means that the combinatorial optimization with the EE maximization can be broken down into multiple subproblems without compromising the optimality of a solution. The proposed algorithm exhibits over a 26% performance improvement with respect to existing techniques.

The remaining of the paper is presented as follows: the system model is illustrated in Section II. In Section III, we formulate the problem and propose the distributed message passing algorithm for maximizing the EE. The complexity and the convergence of the algorithm are analyzed in Section IV. Simulation results are presented in Section V to evaluate the proposed algorithm. Finally, Section VI concludes the paper.

II. System Model

Consider a two-tier downlink HetNet consisting of a single macrocell overlaid with several dense microcells as shown in Fig. 1. The sets of BSs and users are denoted by $\mathcal{I}$ and $\mathcal{A}$, respectively. There are one macro BS, $B-1$ micro BSs and

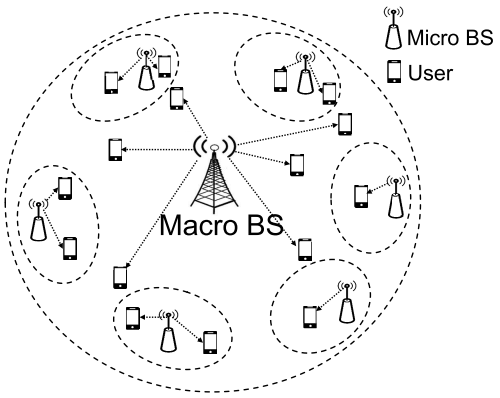


Fig. 1. System model.

N users in the HetNet, i.e., $B = |\mathcal{I}|$ and $N = |\mathcal{A}|$. The macro BS, which is assumed to have larger transmit power and greater user capacity of M_1, is designated as BS 1. Another BS denoted as BS i ($i = 2, \ldots, B$) operates at an individual microcell and has a limited capacity supporting up to M_i users with smaller transmit power. Furthermore, it is assumed that each BS is connected with adjacent BSs via a gateway, which is realized by reliable wired links such as X2-interfaces in 3GPP-LTE networks [35].

Although the BS network is constructed in an one-branch tree for simplicity, the shape of the network can be arbitrary as long as the network forms a tree shape. Users are uniformly distributed around the macro BS which is located at the center of the cell. Each BS is equipped with a single antenna and equal transmit power is allocated to support its users. Each user can be associated with only a single BS.

For concreteness of the formulation, a binary variable x_{ia} is introduced to represent the association status between user a ($a = 1, \ldots, N$) and BS i, i.e., $x_{ia} = 1$ indicates that user a is associated with BS i. BS i transmits the power P_i. All macrocells and microcells are assumed to operate in an open access mode, i.e., no priority for the BS association. User a receives a message signal from BS i given by

$$y_{ia} = \sqrt{P_i} d_{ia}^{-\lambda} h_{ia} s_{ia} + \textstyle\sum_{j \in \mathcal{I} \setminus i} \sqrt{P_j} d_{ja}^{-\lambda} h_{ja} s_{ja} + n_{ia}, \quad (1)$$

where d_{ia}, h_{ia}, and s_{ia} are the distance between BS i and user a, the small scale fading coefficient, and the corresponding transmitted signal, respectively, and n_{ia} denotes the additive Gaussian noise with zero mean and variance σ^2. Given a path loss exponent λ, the corresponding path loss is proportional to $d_{ia}^{-\lambda}$. The first and second term in (1) correspond to a message and interference, respectively. Thus, the throughput dominantly depends on the interference. Since the throughput and the energy consumption are jointly considered for the EE, the reduction of the interference and the power dissipation is essential for the network management.

To do so, some BSs in the network can be turned off. If a BS comes into a turn-off state, some of the hardware components in the BS are either completely switched off or operated in low-power modes so that it disables the pilot transmission and the associated radio processing for the users within its coverage [14], which implies that s_{ia} becomes zero for BS i in the idle state. To characterize the on-off state of the i-th

BS, another binary variable denoted by u_i represents the ON state of the BS. If a single user in $\mathcal{A}$ accessing to BS i results in a nonzero u_i, i.e., $u_i = 1$, it holds that $u_i = \max_{a\in\mathcal{A}} x_{ia}$. If at least one user is served by BS i, it needs to be turned on, i.e., $u_i = 1$. Note also that u_1 corresponds to the on-off state of the macro BS and can be thought of as a global parameter. To consider a group of active BSs, the number of ON-state BSs in the network, denoted by v, is estimated. This is a global parameter that has the relationship with local parameters $\{u_i\}$ as $v = \sum_{i=1}^{B} u_i$.

In consequence, the throughput of the link between BS i and user a is obtained by

$$R_{ia} = \log\left(1 + \frac{P_i d_{ia}^{-2\lambda}|h_{ia}|^2}{\sigma^2 + \sum_{j\in\mathcal{I}\setminus i} P_j d_{ja}^{-2\lambda}|h_{ja}|^2 u_j}\right). \quad (2)$$

For user association of a BS, limited resource sharing is employed. All users supported by the same BS share the system bandwidth of the BS uniformly. Thus, the total sum throughput is expressed as a function of $\{x_{ia}\}$ as

$$\mathcal{R}(\{x_{ia}\}) = \sum_{i\in\mathcal{I}} \left(\frac{\sum_{a\in\mathcal{A}} R_{ia}x_{ia}}{\sum_{b\in\mathcal{A}} x_{ib}}\right). \quad (3)$$

Furthermore, BS i serves up to M_i users at the same time, and this constraint can be described by $\sum_{a\in\mathcal{A}} x_{ia} \le M_i$.

When computing the network-wide EE, the total power of the network can be separated into two portions as the transmit power P_T and the operating power P_O of a BS. Thus, the transmission of message information requires P_T, while keeping the BS in operating mode consumes P_O. If the number of active BSs and the on-off state of the macro BS are known, the total energy consumption of the network is given by $P_{all}(v, u_1) = (v-u_1)(P_T^m + P_O^m) + u_1(P_T^M + P_O^M)$, where the superscripts m and M of P_T and P_O account for a micro BS and a macro BS, respectively. Since $v \ge u_1$, we set $P_{all}(0, u_1) = P_{all}(0, 0) = 0$.

Then, the EE maximization is formulated as

$$\max_{x_{ia}\in\{0,1\}^{B\times N}} \quad \frac{\sum_{i\in\mathcal{I}} \frac{\sum_{a\in\mathcal{A}} R_{ia}x_{ia}}{\sum_{b\in\mathcal{A}} x_{ib}}}{P_{all}(v, u_1)} \quad (4a)$$

$$\text{subject to} \quad \sum_{i\in\mathcal{I}} x_{ia} = 1 \ \forall a, \quad (4b)$$

$$\sum_{a\in\mathcal{A}} x_{ia} \le M_i \ \forall i, \quad (4c)$$

$$u_i = \max_{a\in\mathcal{A}} x_{ia} \ \forall i, \quad (4d)$$

$$v = \sum_{i=1}^{B} u_i, \quad (4e)$$

where the first constraint indicates that each user is associated with only one BS, the second constraint corresponds to the limit of the number of users that a BS can serve simultaneously, and the third and fourth constraints specify the on-off state of each BS and the total number of active BSs, respectively. Thus, the resulting space for feasible configurations is readily seen to be highly nonlinear and combinatorial. The denominator of the objective requires the estimation of a global parameter without centralized management for a distributed control. Furthermore, the numerator has a sum-of-ratio form which is well known that its optimization is NP-complete [36]. Therefore, the overall optimization falls into a class of nonlinear integer programming, which becomes intractable as the network dimension scales. To handle this challenging task, a state-of-the-art optimization technique based on a belief propagation strategy is employed to obtain a low complexity distributed solution for associations among users and BSs.

III. Distributed Algorithm

A. Belief Propagation

In this section, a distributed user association algorithm is derived based on belief propagation, also known as a message-passing algorithm. These two terms are called interchangeably in the sequel, in that message passing is a popular way of realizing a distributed solution based on this framework. To find an efficient solution, we provide a graphical modeling approach such as a factor graph, which facilitates to address a constrained optimization with belief propagation.

According to [33], [34], the graphical model can be constructed to characterize a local structure of the multivariate function $\omega(\mathbf{x})$, which is factorized into a product of several factor functions $\nu_f(\cdot)$ defined for some subset of variables $\mathbf{x}_f$ as $\omega(\mathbf{x}) = \prod_f \nu_f(\mathbf{x}_f)$. Furthermore, this technique allows to find the best assignment of this function based on maximum-a-posteriori (MAP) principle, i.e., $\mathbf{x}^* = \arg\max_{\mathbf{x}} \omega(\mathbf{x})$. The belief propagation techniques exploit the factorized structure to realize it in a way of exchanging messages among nodes in the constructed graph.

A max-product based message-passing algorithm can solve this MAP optimization problem for a factor graph consisting of variable and factor nodes [33], [34]. It can be thought of as an deterministic counterpart of a sum-product-based message-passing algorithm, which calculates marginal probabilities in a probabilistic model. Since the probability weighs the objective function, the assignment of variables with the highest corresponds to the solution of the optimization. Each node transfers a message to each of its neighbors connected by an edge. A message transferred from origin κ to destination τ about variable χ is denoted by $\mu_{\kappa\to\tau}(\chi = \chi_0)$. This message can be thought of as origin κ tells destination τ how much it desires variable χ to become χ_0 in the optimal assignment. Basically, two different messages are transferred in two opposite directions along an edge, since there are two nodes at both ends of the edge. All outgoing messages are updated using all incoming messages received from neighboring nodes and sent backward to them. Thus, the message from variable x_i to factor function ν_f and from factor function ν_f to variable x_i at the t-th iteration can be expressed, respectively, as $\mu_{x_i\to\nu_f}^{(t)}(x_i)$ and $\mu_{\nu_f\to x_i}^{(t)}(x_i)$. One cycle of forward and backward transfers of all messages form a single iteration of the algorithm. These message transfers continue until all messages converge to certain fixed values. The corresponding solution is identified with those values.

The max-product based message computation rules are given by

$$\mu^{(t)}_{\nu_f \to x_i}(x_i) = \max_{\mathbf{x}_f \backslash x_i} \left[\nu_f(x_i, \mathbf{x}_f \backslash x_i) \prod_{j \neq i} \mu^{(t)}_{x_j \to \nu_f}(x_j) \right], \quad (5a)$$

$$\mu^{(t+1)}_{x_i \to \nu_f}(x_i) = \prod_{b \neq f} \mu^{(t)}_{\nu_f \to x_i}(x_i), \quad (5b)$$

where $\nu_f(x_i, \mathbf{x}_f \backslash x_i)$ denotes the function with only x_i fixed. Messages are normally expressed in a log-domain representation, which helps avoiding various numerical issues and allowing to address a constrained formation, by denoting the logarithm of messages in (5) by new messages along with $\bar{\nu}_f(\mathbf{x}_f) = \log \nu_f(\mathbf{x}_f)$. The resulting max-sum based computation rules are written as

$$\mu^{(t)}_{\bar{\nu}_f \to x_i}(x_i) = \max_{\mathbf{x}_f \backslash x_i} \left[\bar{\nu}_f(x_i, \mathbf{x}_f \backslash x_i) + \sum_{j \neq i} \mu^{(t)}_{x_j \to \bar{\nu}_f}(x_j) \right], \quad (6a)$$

$$\mu^{(t+1)}_{x_i \to \bar{\nu}_f}(x_i) = \sum_{b \neq f} \mu^{(t)}_{\bar{\nu}_f \to x_i}(x_i). \quad (6b)$$

Upon the convergence of messages in (6), the final estimate to the MAP inference problem is found by computing the belief $\bar{\omega}(x_i)$ given by $\bar{\omega}(x_i) = \sum_f \mu^{(t)}_{\bar{\nu}_b \to x_i}(x_i)$. If each x_i is assigned with the value that maximizes its belief $\bar{\omega}(x_i)$ for a set of possible values for x_i, the best estimate is obtained using $x_i^* = \arg\max_{x_i} \omega(x_i)$. Then, the solution $\mathbf{x}^*$ is constructed by collecting all x_i^* for the MAP optimization.

B. Graphical Representation

To handle a constrained optimization, a careful design of graphical representation is essential. A factor graph proves viable in various network optimizations [22], [37]. It is a bipartite graph with two classes of variable and factor nodes, which are associated with optimization variables and their constraints in the optimization problem, respectively. Factor functions are introduced to define individual component terms appropriately in objective and constraint functions.

For the constrained formulation in (4), five factor functions are defined to address the objective (4a) and the constraints in (4b)-(4e). Each function is defined either to evaluate the objective value for the maximization or to penalize the violation of individual constraints. First, $F_a(\{x_{ia}\})$ is defined to enforce the constraint in (4b) so that each user is driven to choose a single BS as

$$F_a(\{x_{ia}\}) = \begin{cases} -\infty & \text{if } \sum_{i \in \mathcal{I}} x_{ia} \neq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

In addition, $G_i(\{x_{ia}\}, u_i)$ is responsible for determining an individual additive term of the objective function along with constraints in (4c) and (4d) together, since these two functions have the same input variables, which are associated with BS i. Thus, $G_i(\{x_{ia}\}, u_i)$ permits the output to satisfy user capacity constraints along with the values of the objective function as

$$G_i(\{x_{ia}\}, u_i) = \begin{cases} -\infty & \text{if } \sum_{a \in \mathcal{A}} x_{ia} > M_i \\ & \text{or } u_i \neq \max_{a \in \mathcal{A}} x_{ia}, \\ 0 & \text{if } \sum_{a \in \mathcal{A}} x_{ia} = 0, \\ \frac{\sum_{a \in \mathcal{A}} R_{ia} x_{ia}}{P_{all}(v, u_1) \sum_{b \in \mathcal{A}} x_{ib}} & \text{otherwise.} \end{cases} \quad (8)$$

On the other hand, the constraint in (4e) requires multiple factor functions for its distributed processing. To be precise, v in (4e) is a global parameter that corresponds to the number of active BSs in the network. For a fully distributed operation, individual BSs are aware of this value and estimates the value denoted by $\hat{v}$ via the cooperation among BSs. To facilitate it, u_i is split into two parts as $z_{i1} = \sum_{k=1}^{i} u_k$ and $z_{i2} = \sum_{k=i+1}^{B} u_k$. Since $v = z_{i1} + z_{i2} = \sum_{k=1}^{B} u_k$ for all i with $z_{0,1} = 0$ and $z_{B,2} = u_B$, $H_i(z_{i1}, z_{i2}, u_i)$ is defined as

$$H_i(z_{i1}, z_{i2}, \hat{v}) = \begin{cases} -\infty & \text{if } v_i \neq z_{i1} + z_{i2}, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Furthermore, two factors $H_{i1}(z_{i1}, z_{i-1,1}, u_i)$ and $H_{i2}(z_{i2}, z_{i-1,2}, u_i)$ enforce relationships of $z_{i1} = z_{i-1,1} + u_i$ and $z_{i-1,2} = z_{i2} + u_i$ and are defined, respectively, by

$$H_{i1}(z_{i1}, z_{i-1,1}, u_i) = \begin{cases} -\infty & \text{if } z_{i1} \neq z_{i-1,1} + u_i, \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

$$H_{i2}(z_{i2}, z_{i-1,2}, u_i) = \begin{cases} -\infty & \text{if } z_{i-1,2} \neq z_{i2} + u_i, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

The constrained problem in (4) can now be reformulated into an unconstrained formulation for handling only with factor functions (7)-(11) as

$$\max_{\{x_{ia}\}, \{z_{i1}, z_{i2}\}, \{u_i\}} \sum_{i \in \mathcal{I}} \Big(G_i(\{x_{ia}\}, u_i) + H_{i1}(z_{i1}, z_{i-1,1}, u_i) + H_{i2}(z_{i2}, z_{i-1,2}, u_i) + H_i(z_{i1}, z_{i2}, u_i) \Big) + \sum_{a \in \mathcal{A}} F_a(\{x_{ia}\}). \quad (12)$$

Since the unconstrained formulation in (12) has purely an additive objective function of several factor functions associated with only a subset of variables, it is favorable to proceed with a distributed solving strategy by finding maximizers of those factor functions independently and matching them to form a consistent global solution over the graph. Note that the output of any factor function is not finite unless the associated constraint is satisfied, i.e., the input assignment incurring the infinite output value is useless and automatically excluded from the candidate for the solution. Likewise, any assignment of variables resulting in a finite objective value can be a feasible solution, and the largest finite objective value of (12) consequently becomes the optimal solution. This sheds a light on a decentralized strategy for the optimal solution. The maximizers of individual factor functions are obtained separately, and the final solution is identified by choosing a consistent set of those solutions. Since the solutions of (4)

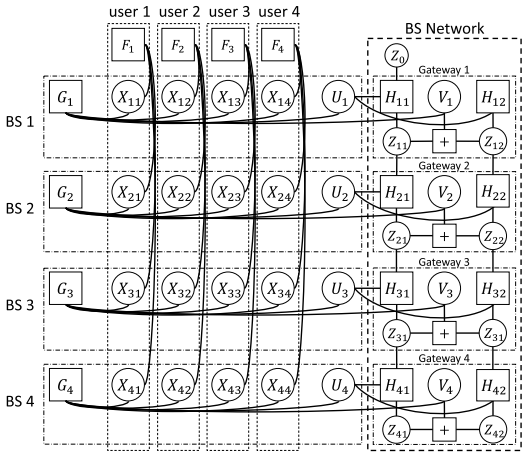


Fig. 2. An example of the factor graph associated with HetNets of 4 BSs and 4 users.

and (12) are identical, solving (12) is equivalent to find the original solution.

The visualization of the factor graph is also important in that it is involved with a practical deployment of the distributed algorithm. For visual representation of the factor graph, a circle normally denotes a variable, while a square stands for a factor function. An edge is connected between a circle and a square to indicate the membership of the associated variable with the associated factor function. Fig. 2 illustrates an example of the factor graph for the case of four BSs and four users using the above factor functions. The factor graph consists of two subgraphs, each representing a user-BS network and a BS network. Factor functions $G_i(\cdot)$ and $F_a(\cdot)$ of the left-hand side are connected to variables in the i-th row and the a-th column indicating user a and BS i, respectively. In contrast, the right-hand side subgraph corresponds to the BS network interconnected with neighboring BSs, where individual BSs use their internal states and external information transferred from neighboring BSs to obtain their own estimates on the number of active BSs in the overall network. It is noted that the factor graph in Fig. 2 is, in fact, not a physically deployed structure. To describe a physical deployment of the network, boxes represented in dashed line are introduced in the factor graph, i.e., a network node associated with each box shares states and messages related to all graph nodes that it contains. Thus, variables x_{ia} are shared by two nodes, while all other variables such as u_i and v_i are independently managed.

C. Message Definition

To derive a distributed algorithm using a belief propagation framework, we first define messages. Since x_{ia} is binary and associated with two factor functions $F_a(\cdot)$ and $G_i(\cdot)$, four different types of messages $\mu_{x_{ia}\to F_a}(x_{ia}=x)$, $\mu_{F_a\to x_{ia}}(x_{ia}=x)$, $\mu_{x_{ia}\to G_i}(x_{ia}=x)$, and $\mu_{G_i\to x_{ia}}(x_{ia}=x)$ are necessary for two different values of $x\in\{0,1\}$. Furthermore, additional six messages $\mu_{u_i\to G_i}(u_i=u)$, $\mu_{G_i\to u_i}(u_i=u)$, $\mu_{u_i\to H_{i1}}(u_i=u)$, $\mu_{H_{i1}\to u_i}(u_i=u)$, $\mu_{u_i\to H_{i2}}(u_i=u)$, and $\mu_{H_{i2}\to u_i}(u_i=u)$ are denoted for another binary variable u_i with $u\in\{0,1\}$. On the other hand, integer variables z_{i1} and z_{i2} are associated with six messages denoted by $\mu_{z_{i1}\to H_{i1}}(z_{i1}=z)$, $\mu_{H_{i1}\to z_{i1}}(z_{i1}=z)$, $\mu_{z_{i2}\to H_{i2}}(z_{i2}=z)$,

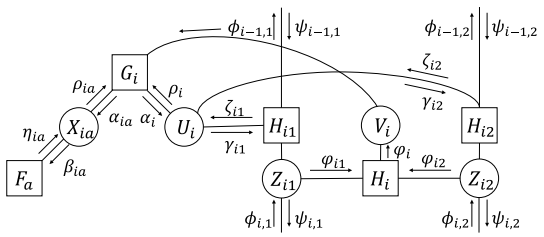


Fig. 3. Message definitions in the factor graph.

$\mu_{H_{i2}\to z_{i2}}(z_{i2}=z)$, $\mu_{z_{i1}\to H_i}(z_{i1}=z)$, and $\mu_{z_{i2}\to H_i}(z_{i2}=y)$ for $z_{i1},z_{i2}\in\{1,\ldots,B\}$. Note that the physical deployment leads to a reduction of some messages for unnecessary internal message exchange. In addition, since variable $\hat{v}$ is used as a parameter, no message is associated with it.

For compact representations, we redefine the messages. Fig. 3 depicts new definition of messages in the factor graph. For a binary variable x_{ia}, the corresponding messages are expressed as the differences between two messages associated with the cases whether $x_{ia}=1$ or not. Accordingly, the redefined messages are given by

$$\begin{aligned}\beta_{ia}&=\mu_{x_{ia}\to F_a}(x_{ia}=1)-\mu_{x_{ia}\to F_a}(x_{ia}=0),\\ \eta_{ia}&=\mu_{F_a\to x_{ia}}(x_{ia}=1)-\mu_{F_a\to x_{ia}}(x_{ia}=0),\\ \rho_{ia}&=\mu_{x_{ia}\to G_i}(x_{ia}=1)-\mu_{x_{ia}\to G_i}(x_{ia}=0),\\ \alpha_{ia}&=\mu_{G_i\to x_{ia}}(x_{ia}=1)-\mu_{G_i\to x_{ia}}(x_{ia}=0).\end{aligned}\tag{13}$$

These messages are used for identifying user association. In contrast, the following messages for variable u_i represent the switch-off state given by

$$\begin{aligned}\rho_i&=\mu_{u_i\to G_i}(u_i=1)-\mu_{u_i\to G_i}(u_i=0),\\ \alpha_i&=\mu_{G_i\to u_i}(u_i=1)-\mu_{G_i\to u_i}(u_i=0).\end{aligned}\tag{14}$$

Also, the following messages about the switch-off state for variable u_i are exchanged with neighboring BSs in two ways for $k\in\{1,2\}$ as

$$\begin{aligned}\gamma_{ik}&=\mu_{u_i\to H_{ik}}(u_i=1)-\mu_{u_i\to H_{ik}}(u_i=0),\\ \zeta_{ik}&=\mu_{H_{ik}\to u_i}(u_i=1)-\mu_{H_{ik}\to u_i}(u_i=0).\end{aligned}\tag{15}$$

According to the underlying idea of belief propagation, the above messages encode the information about whether BS i is well-suited to serve users in the optimal configuration or not. Since there are only two states for variables, it suffices to know the difference between messages associated with two opposite states for the decision on those variables. The compact representations of the remaining messages are denoted for $k\in\{1,2\}$ by

$$\begin{aligned}\psi_{ik}(z)&=\mu_{z_{ik}\to H_{i+1,k}}(z_{ik}=z),\\ \phi_{ik}(z)&=\mu_{H_{i+1,k}\to z_{ik}}(z_{ik}=z),\\ \varphi_{ik}(z)&=\quad\mu_{z_{ik}\to H_i}(z_{ik}=z).\end{aligned}\tag{16}$$

In contrast to the messages associated with binary variables, variables associated with (16) can take on nonbinary values, and the differences are not used for simplification. The physical meaning of messages in (16) accounts for the contribution of a subset of BSs to the objective value. In particular, messages $\psi_{i1}(z)$, $\phi_{i1}(z)$, and $\varphi_{i1}(z)$ can be combined to

estimate the objective value expected when z BSs out of the BS group from BS 1 to BS i are in the operating state, while messages $\psi_{i2}(z)$, $\phi_{i2}(z)$, and $\varphi_{i2}(z)$ correspond to the contributions to the objective value by z active BSs from BS i to BS B.

D. Derivation of Algorithms

We develop the message update rules based on belief propagation. The user association algorithm is derived first. To this end, we introduce the update rules for the messages sent from a variable to a factor function first since their calculation is mostly simpler than their counterparts. At a variable node, an outgoing message is updated as the sum of all incoming messages according to the max-sum based message-passing algorithm in (6b). According to the factor graph in Fig. 2, there are only two edges connected to variable x_{ia}, and the outgoing messages are simply duplicates of the corresponding incoming messages given by

$$\beta_{ia} = \alpha_{ia}, \qquad \rho_{ia} = \eta_{ia}. \tag{17}$$

Next, we derive the message update rules for messages transferred from a factor function to a variable. These message update rules involve the maximization of sums of the factor function and the incoming sum message according to the max-sum based message-update rule in (6a). The outgoing message indicates the difference between the maximum objective value for the case where the associated constraint is satisfied with variable x_{ia} set to 1 and for the same case but x_{ia} set to 0, i.e., its sign can determine the preference of the associated variable x_{ia}. For simple representation, two variable sets $\mathcal{X}_a = \{x_{ia}\colon i \in \mathcal{I}\}$ and $\mathcal{X}_i = \{x_{ia}\colon a \in \mathcal{A}\}$ are defined, and the message update rule for message η_{ia} is derived as

$$\begin{aligned}
\eta_{ia} &= \mu_{F_a \to x_{ia}}(x_{ia}=1) - \mu_{F_a \to x_{ia}}(x_{ia}=0) \\
&= \max_{\mathcal{X}_a \backslash x_{ia}} \Big(F_a(x_{ia}=1, \mathcal{X}_a \backslash x_{ia}) + \sum_{j \in \mathcal{I} \backslash i} \mu_{x_{ja} \to F_a}(x_{ja}) \Big) \\
&\quad - \max_{\mathcal{X}_a \backslash x_{ia}} \Big(F_a(x_{ia}=0, \mathcal{X}_a \backslash x_{ia}) + \sum_{j \in \mathcal{I} \backslash i} \mu_{x_{ja} \to F_a}(x_{ja}) \Big) \\
&= \sum_{j \in \mathcal{I} \backslash i} \mu_{x_{ja} \to F_a}(0) \\
&\quad - \max_{j \in \mathcal{I} \backslash i} \Big(\mu_{x_{ja} \to F_a}(1) + \sum_{k \in \mathcal{I} \backslash \{i,j\}} \mu_{x_{ka} \to F_a}(0) \Big) \\
&= - \max_{j \in \mathcal{I} \backslash i} \Big(\mu_{x_{ja} \to F_a}(1) - \mu_{x_{ja} \to F_a}(0) \Big) \\
&= - \max_{j \in \mathcal{I} \backslash i} \beta_{ja}.
\end{aligned} \tag{18}$$

The output of $F_a(x_{ia}, \mathcal{X}_a \backslash x_{ia})$ is evaluated by fixing x_{ia} to either 1 or 0 first and choosing other variables appropriately. Its value is valid only when it takes a finite value. Since $x_{ia} = 1$ indicates that user a is associated with BS i, no other BS can be chosen, i.e., the corresponding incoming message sum is obviously $\sum_{j \in \mathcal{I} \backslash i} \mu_{x_{ja} \to F_a}(0)$. On the other hand, in case of $x_{ia} = 0$, user a chooses another BS instead of BS i. The corresponding valid incoming message sum is different from $\sum_{j \in \mathcal{I} \backslash i} \mu_{x_{ja} \to F_a}(0)$ by a single message $\mu_{x_{ka} \to F_a}(1)$ for some k. Accordingly, its subtraction from the remaining part simplifies the overall computation for η_{ia}. The message update rule in (18) provides a rule that user a chooses its serving BS. If all incoming messages are negative, i.e., $\beta_{ia} < 0$, the corresponding value of η_{ia} becomes positive, implying that no other BS can serve user a and the user accesses to BS i.

We next develop the message update rule for α_{ia} that addresses the factor function $G_i(\cdot)$. This message evaluates an expected advantage in the objective function value when user a is connected to BS i over the case when it is not connected. Thus, the message value closely depends on the energy consumption of the network, and a special consideration about the total energy consumption can help the improvement of the performance. To this end, we distinguish two cases whether the macro BS is turned on or not, since the operating energy consumption of the macro BS is much larger than micro BSs. Let $P(v, u_1)$ denote the estimate on the total energy consumption expressed in terms of the total number of active BSs v and the status of the macro BS u_1. If the macro BS is turned on, i.e., $u_1 = 1$, we have $P(v,1) = (v-1)(P_O^m + P_T^m) + P_O^M + P_T^M$. Otherwise, for $u_1 = 0$, the energy consumption is $P(v,0) = v(P_O^m + P_T^m)$.

Defining a new operator $\overline{\max}_{s \in \mathcal{S}}(l, \theta_s)$ as the l-th largest value among message input θ_s indexed with $s \in \mathcal{S}$ for node set $\mathcal{S}$, the message update rule for α_{ia} is presented in (19), shown at

$$\begin{aligned}
\alpha_{ia} &= \mu_{G_i \to x_{ia}}(x_{ia}=1) - \mu_{G_i \to x_{ia}}(x_{ia}=0) \\
&= \max_{u_1} \max_{\mathcal{X}_i \backslash x_{ia}} \Big(G_i(x_{ia}=1, \mathcal{X}_i \backslash x_{ia}, u_1) + \sum_{b \in \mathcal{A} \backslash a} \mu_{x_{ib} \to G_i}(x_{ib}) + \mu_{u_1 \to G_i}(u_1) \Big) \\
&\quad - \max_{u_1} \max_{\mathcal{X}_i \backslash x_{ia}} \Big(G_i(x_{ia}=0, \mathcal{X}_i \backslash x_{ia}, u_1) + \sum_{b \in \mathcal{A} \backslash a} \mu_{x_{ib} \to G_i}(x_{ib}) + \mu_{u_1 \to G_i}(u_1) \Big) \qquad (19a) \\
&= \max_{u_1} \max \Big(\frac{R_{ia}}{P(v,u_1)}, \frac{R_{ia}}{2P(v,u_1)} + \overline{\max}_{b \in \mathcal{A} \backslash a} \Big(1, \frac{R_{ib}}{2P(v,u_1)} + \rho_{ib}\Big), \ldots, \frac{R_{ia}}{M_i P(v,u_1)} + \sum_{l=1}^{M_i - 1} \overline{\max}_{b \in \mathcal{A} \backslash a} \Big(l, \frac{R_{ib}}{M_i P(v,u_1)} + \rho_{ib}\Big) \Big) \\
&\quad - \max_{u_1} \max \Big(-\rho_i, \overline{\max}_{b \in \mathcal{A} \backslash a} \Big(1, \frac{R_{ib}}{P(v,u_1)} + \rho_{ib}\Big), \sum_{l=1}^{2} \overline{\max}_{b \in \mathcal{A} \backslash a} \Big(l, \frac{R_{ib}}{2P(v,u_1)} + \rho_{ib}\Big), \ldots, \sum_{l=1}^{M_i} \overline{\max}_{b \in \mathcal{A} \backslash a} \Big(l, \frac{R_{ib}}{M_i P(v,u_1)} + \rho_{ib}\Big) \Big) \qquad (19b) \\
&= \max(\mathcal{A}_{ia}(1, P(v,1)), \mathcal{A}_{ia}(1, P(v,0))) - \max(\mathcal{A}_{ia}(0, P(v,0)), \mathcal{A}_{ia}(0, P(v,0))), \qquad (19c)
\end{aligned}$$

the bottom of the previous page, and $\mathcal{A}_{ia}(x_{ia}, p)$ is a function of incoming messages with two input parameters of the binary state x_{ia} and the energy consumption p, introduced for simple representation, given by (20), shown at the bottom of the page. Note that, in (19a), the definition of message α_{ia} is given according to (13) as the difference between the contributions of the objective value whether the association between BS i and user a is made or not. The objective value for an individual case is calculated using the max-sum based message-update rule in (6a). Although the total energy consumption $P(v, u_1)$ is a parameter rather than a message, its value may change abruptly depending on whether the macro BS is turned on or not. For proper message update operations, these two cases are considered separately in the message calculation. Thus, two distinct message values are evaluated according to the state of the macro BS and compared to find the most likely message value in (19b).

Note that the simplified form of message α_{ia} in (19c) is evaluated by the difference between preferences that BS i serves user a or not, regardless of the state of the macro BS. However, the state of the macro BS proves crucial for the optimal decision since the total energy consumption highly depends on the status of the macro BS. To manifest this, message α_{ia} is modified to distinguish the cases whether the macro BS is turned on or not separately and compare them for a new message.

Two differences associated with the states of the macro BS are computed in terms of

$$\begin{aligned}\Delta_{ia}(1) &= \mathcal{A}_{ia}(1, P(1)) - \mathcal{A}_{ia}(0, P(1)),\\ \Delta_{ia}(0) &= \mathcal{A}_{ia}(1, P(0)) - \mathcal{A}_{ia}(0, P(0)).\end{aligned} \tag{21}$$

These quantities can be interpreted as the preferences of the link between BS i and user a over the unconnected link when the macro BS is turned on or off, respectively. Thus, it is proper to choose a new form of message α_{ia} as the preference of the largest absolute value, and the resulting message is obtained as

$$\alpha_{ia} = \mathrm{sgn}(\Delta_{ia}(1) + \Delta_{ia}(0)) \max(|\Delta_{ia}(1)|, |\Delta_{ia}(0)|). \tag{22}$$

This simplification of message α_{ia} leads to the comparison between two out of four possible cases and diminishes the range of the message value, i.e., the decreased variation in α_{ia}. However, it helps obtaining a final solution with better reliability, since it provides more conservative evidences to make a decision on the state of x_{ia}.

The messages transferred over the interface between user association and the BS network are derived in the following. As with the previous case, the outgoing message from variable u_i is the sum of all incoming messages simply given as

$$\rho_i = \zeta_{i1} + \zeta_{i2}, \qquad \gamma_{ik} = \alpha_i, \quad k \in \{1, 2\}. \tag{23}$$

Note that γ_{ik} should have been defined as $\gamma_{ik} = \alpha_i + \zeta_{i,3-k}$ since u_i has three neighboring nodes, all having associated messages of α_i, $\zeta_{i,1}$, and $\zeta_{i,2}$. Since γ_{ik} is interpreted as an estimate on the state of BS i, it strongly depends on α_i rather than $\zeta_{i,3-k}$. Thus, adding $\zeta_{i,3-k}$ to α_i contributes little to estimate γ_{ik}, and identifying u_i benefits from independent estimates of z_{i1} and z_{i2}, which remain uncorrelated with each other. Therefore, $\zeta_{i,3-k}$ can be ignored in computing γ_{ik}.

The outgoing message α_i emanating from the factor function $G_i(\cdot)$ is evaluated using

$$\begin{aligned}\alpha_i &= \mu_{G_i \to u_i}(u_i = 1) - \mu_{G_i \to u_i}(u_i = 0)\\ &= \max_{\mathcal{X}_i}\Big(G_i(\mathcal{X}_i, u_i = 1) + \sum_{a \in \mathcal{A}} \mu_{x_{ia} \to G_i}(x_{ia})\Big)\\ &\quad - \max_{\mathcal{X}_i}\Big(G_i(\mathcal{X}_i, u_i = 0) + \sum_{a \in \mathcal{A}} \mu_{x_{ia} \to G_i}(x_{ia})\Big)\\ &= \max_{u_1} \max\Bigg(\overline{\max_{a \in \mathcal{A}}}\Big(1, \frac{R_{ia}}{P(v, u_1)} + \rho_{ia}\Big),\\ &\qquad \sum_{l=1}^{2} \overline{\max_{a \in \mathcal{A}}}\Big(l, \frac{R_{ia}}{2P(v, u_1)} + \rho_{ia}\Big), \ldots,\\ &\qquad \sum_{l=1}^{M_i} \overline{\max_{a \in \mathcal{A}}}\Big(l, \frac{R_{ia}}{M_i P(v, u_1)} + \rho_{ia}\Big)\Bigg)\\ &= \max(\mathcal{A}_i(P(v, 1)), \mathcal{A}_i(P(v, 0))),\end{aligned} \tag{24}$$

where $\mathcal{A}_i(p)$ is a new function defined as

$$\begin{aligned}\mathcal{A}_i(p) = \max\Bigg(&\overline{\max_{a \in \mathcal{A}}}\Big(1, \frac{R_{ia}}{p} + \rho_{ia}\Big), \sum_{l=1}^{2} \overline{\max_{a \in \mathcal{A}}}\Big(l, \frac{R_{ia}}{2p} + \rho_{ia}\Big),\\ &\ldots, \sum_{l=1}^{M_i} \overline{\max_{a \in \mathcal{A}}}\Big(l, \frac{R_{ia}}{M_i p} + \rho_{ia}\Big)\Bigg),\end{aligned} \tag{25}$$

and $\mu_{G_i \to u_i}(u_i = k)$ is also calculated for $k = 0, 1$ using the max-sum based message update rule in (6a). Note that message $\mu_{G_i \to u_i}(u_i = 0)$ simply becomes zero since BS i is turned off and no user is allowed to associate with BS i. Similarly with α_{ia}, α_i addresses two different cases based on the state of the macro BS which may make the total energy consumption change abruptly. Message α_i conveys the preference between the operating and sleeping modes of BS i. Such information is simply duplicated into γ_{i1} and γ_{i2} and transferred to the BS network for the identification of the operating BS population.

Finally, we examine the message-update rules for the census of active BSs in the BS network. The BS network includes two trellises that expand in the opposite directions with the two-fold objective of estimating the total number of active BSs in the entire network and forcing each BS to switch on and off for achieving the best energy efficient network configuration. To this purpose, each BS exchanges two types

$$\begin{aligned}\mathcal{A}_{ia}(x_{ia}, p) = \max\Bigg(&\Big(\frac{R_{ia}}{p} + \rho_i\Big)x_{ia} - \rho_i, \frac{R_{ia}}{(x_{ia}+1)p} + \overline{\max_{b \in \mathcal{A} \setminus a}}\Big(1, \frac{R_{ib}}{(x_{ia}+1)p} + \rho_{ib}\Big),\\ &\frac{R_{ia}}{(x_{ia}+2)p} + \sum_{l=1}^{2} \overline{\max_{b \in \mathcal{A} \setminus a}}\Big(l, \frac{R_{ib}}{(x_{ia}+2)p} + \rho_{ib}\Big), \ldots, \frac{R_{ia}}{M_i p} + \sum_{l=1}^{M_i - x_{ia}} \overline{\max_{b \in \mathcal{A} \setminus a}}\Big(l, \frac{R_{ib}}{M_i p} + \rho_{ib}\Big)\Bigg).\end{aligned} \tag{20a}$$

of messages corresponding to two opposite directions with neighboring BSs. Since this challenge falls into an optimization task, we resort to the max-sum based message-passing algorithm in (6). Therefore, the message calculation for the BS network boils down to a forward-backward algorithm such as BCJR algorithm [33]. Once the input messages obtained by user association are applied from individual BSs, each BS calculates new messages, such as $\psi_{i1}(\cdot)$, $\psi_{i2}(\cdot)$, $\phi_{i1}(\cdot)$, and $\phi_{i2}(\cdot)$, and passes them to neighboring BSs in two directions of interconnecting edges. Two messages $\psi_{i1}(\cdot)$ and $\psi_{i2}(\cdot)$ are referred to as the forward messages since those are passed to the adjacent BS with the index of the increasing order, while ϕ_{i1} and ϕ_{i2} are called the backward messages for their transfer direction corresponding to the index of the decreasing order. Since the relationship between adjacent BSs are $z_{i1} = z_{i-1,1} + u_i$ and $z_{i-1,2} = z_{i2} + u_i$, two trellises are simply constructed to represent two different accumulation operations for the number of active BSs. The resulting outputs of this forward-backward algorithm are the estimates on z_{i1} and z_{i2} along with the preference for the value of each u_i fed back to the user association network. Also, the networks that estimate z_{i1} and z_{i2} are referred to as the forward and backward network, respectively, according to the directions of the state transition.

The corresponding forward and backward messages are expressed by

$$\begin{aligned}
\psi_{i1}(z) &= \max(\psi_{i-1,1}(z), \psi_{i-1,1}(z-1) + \gamma_{i1}),\\
\phi_{i1}(z) &= \max(\phi_{i+1,1}(z), \phi_{i+1,1}(z+1) + \gamma_{i+1,1}),\\
\psi_{i2}(z) &= \max(\psi_{i-1,2}(z), \psi_{i-1,2}(z+1) + \gamma_{i-1,2}),\\
\phi_{i2}(z) &= \max(\phi_{i+1,2}(z), \phi_{i+1,2}(z-1) + \gamma_{i2}). \qquad (26)
\end{aligned}$$

Thus, $\psi_{i1}(z)$ and $\phi_{i2}(z)$ indicate the state transition caused by the increment of the number of active BSs, while $\phi_{i1}(z)$ and $\psi_{i2}(z)$ account for the decrement of the number of active BSs according to the state of the ith BS.

To estimate the total number of active BSs using the information in (26), BS i computes the preferences for z_{i1} and z_{i2}, representing the number of active BSs in the index ranges of $\{1, \ldots, i\}$ and $\{i+1, \ldots, B\}$, respectively, using

$$\begin{aligned}
\varphi_{i1}(z) &= \max(\phi_{i+1,1}(z), \phi_{i+1,1}(z+1) + \gamma_{i+1,1}),\\
\varphi_{i2}(z) &= \max(\phi_{i+1,2}(z), \phi_{i+1,2}(z-1) + \gamma_{i2}). \qquad (27)
\end{aligned}$$

Note that $\varphi_{i1}(z)$ and $\varphi_{i2}(z)$ are the objective values expected from the information transferred from the forward and backward networks. Based on these messages in (27), the total number of active BSs is predicted at BS i as

$$\hat{v} = \arg\max_{1 \le v \le B} \max_{1 \le z \le v} (\varphi_{i1}(z) + \varphi_{i2}(v - z)), \qquad (28)$$

and becomes identical among all BSs. Furthermore, messages ζ_{i1} and ζ_{i2}, which inform BS i whether it should be turned on or off in the optimal solution, are calculated at BS i as

$$\begin{aligned}
\zeta_{i1} &= \max_z(\phi_{i1}(z+1) + \psi_{i-1,1}(z))\\
&\quad - \max_z(\phi_{i1}(z) + \psi_{i-1,1}(z)),\\
\zeta_{i2} &= \max_z(\phi_{i2}(z) + \psi_{i-1,2}(z+1))\\
&\quad - \max_z(\phi_{i2}(z) + \psi_{i-1,2}(z)). \qquad (29)
\end{aligned}$$

To determine the switch-off state of BS i, message ρ_i is simply evaluated as their sum given by $\rho_i = \zeta_{i1} + \zeta_{i2}$. Thus, its positive value indicates that it is better to turn on BS i to assure that all network users can be served properly.

Although the above technique provides an efficient way of identifying the total number of operating BSs v and the state of BS i, it does not necessarily optimize the objective function directly. To maximize the network-wide EE, the algorithm minimizes the value of the denominator, while maximizing the numerator. Although the value of v is kept as small as possible, the total throughput is not likely to have a monotonic change with respect to v, and the network-wide EE becomes sensitive to the value of v. To this end, trellises, which prove versatile in finding an efficient discrete value for distributed optimization problems [22], [34], are employed for identifying the optimal value of v in the BS network. The trellis operations are configured for each $v \in \{1, \ldots, B\}$ to force each BS to switch on and off to satisfy the total number of active BSs. Meanwhile, the best value of $\hat{v}$ is searched simultaneously among the corresponding results. To enforce such an optimization feature, the initialization of (26) at both ends of the trellises is applied as

$$\begin{aligned}
\psi_{B1}(z), \phi_{B1}(z) &= \begin{cases} -\infty & \text{if } z \neq v \\ 0 & \text{otherwise,} \end{cases}\\
\psi_{12}(z), \phi_{12}(z) &= \begin{cases} -\infty & \text{if } z \neq B \\ 0 & \text{otherwise.} \end{cases} \qquad (30)
\end{aligned}$$

Although message updates and transfers are designed to conduct normally one by one at each single iteration, we can construct a protocol that schedules the overall algorithm as the BCJR algorithm proceeds: The forward and backward trellis searches run first in a row from the macro BS (BS 1) to BS B and vice versa. Subsequently, each BS collects the corresponding messages to estimate the total number of active BSs and its optimal state u_i for the use of the optimal user association without the aid of any centralized management. A major advantage of this approach is that the global optimality of the solution is guaranteed with the current input from BSs since the BS network is tree-like. According to the factor graph structure in Fig. 2, variable nodes associated with v_i and ρ_i separate the BS network from the user association graph. Once $\hat{v}$ and ρ_i are reliably determined, the conditional independence [34] holds for the results between BS network and user association. This is likely to make outgoing messages from BS robust and the convergence behavior of the belief propagation for user association is improved.

The overall messages are updated in each single iteration at the network using

$$\begin{aligned}
\alpha_{ia}^{(t)} &= \operatorname{sgn}\Big(\mathcal{A}_{ia}^{(t)}(1, P(1)) - \mathcal{A}_{ia}^{(t)}(0, P(1))\\
&\quad + \mathcal{A}_{ia}^{(t)}(1, P(0)) - \mathcal{A}_{ia}^{(t)}(0, P(0))\Big) \times\\
&\quad \max\ \Big(|\mathcal{A}_{ia}^{(t)}(1, P(1)) - \mathcal{A}_{ia}^{(t)}(0, P(1))|,\\
&\quad |\mathcal{A}_{ia}^{(t)}(1, P(0)) - \mathcal{A}_{ia}^{(t)}(0, P(0))|\Big), \qquad (31)
\end{aligned}$$

$$\rho_{ia}^{(t+1)} = -\max_{j\in\mathcal{I}\setminus i} \alpha_{ja}^{(t)}. \quad (32)$$

Also, the BS network updates the remaining messages as

$$\alpha_i^{(t)} = \max(\mathcal{A}_i^{(t)}(P(1)), \mathcal{A}_i^{(t)}(P(0))), \quad (33)$$

$$\zeta_{ik}^{(t)} = \max_{z_{ik}}(\phi_{ik}^{(t)}(z_{ik}+1) + \psi_{i-1,k}^{(t)}(z_{ik})) - \max_{z_{ik}}(\phi_{ik}^{(t)}(z_{ik}) + \psi_{i-1,k}^{(t)}(z_{ik})), \quad (34)$$

$$\psi_{ik}^{(t)}(z_{ik}) = \max_{z_{ik}}(\psi_{i-1,k}^{(t)}(z_{ik}), \psi_{i-1,k}^{(t)}(z_{ik}+(-1)^k) + \gamma_{i+1-k,k}^{(t)}), \quad (35)$$

$$\phi_{ik}^{(t)}(z_{ik}) = \max_{z_{ik}}(\phi_{i+1,k}^{(t)}(z_{ik}), \phi_{i+1,k}^{(t)}(z_{ik}-(-1)^k) + \gamma_{i+2-k,k}^{(t)}), \quad (36)$$

$$\rho_i^{(t)} = \zeta_{i1}^{(t)} + \zeta_{i2}^{(t)}, \quad (37)$$

where $\mathcal{A}_{ia}^{(t)}(\cdot)$ and $\mathcal{A}_i^{(t)}(\cdot)$ denote the value obtained using the set of input message $\rho_{ia}^{(t)}$ updated at the t-th iteration. Note that all message values are evaluated with simple combinations of arithmetic operations only.

E. Implementation Issues

This subsection briefly discusses the complexity and convergence behaviors of the developed algorithm. A glance at the computational rules presented in (31)-(37) reveals that the evaluation of $\alpha_{ia}^{(t)}$ in (31) has dominant contributions to the overall computational efforts, although the overall computational rules are divided into three parts: user message updates, BS message updates, and BS trellises. To see this more carefully, at users' side, each user utilizes at most B different input messages from neighboring BSs to compute $\rho_{ia}^{(t+1)}$, and the complexity becomes $O(B)$ which is strictly less than $O(N)$. At BSs' side, each BS calculates $\alpha_{ia}^{(t)}$ to determine the user association. This operation involves sorting at most M_i values obtained by adding the corresponding different messages and takes $O(M_i \log M_i)$ operations. Since there are M_i different values to be compared in the calculation, the overall message computation requires $O(M_i^2 \log M_i)$ at each BS. When predicting the number of active BSs, each BS is responsible for the calculation of messages $\psi_{ik}^{(t)}(z_{ik})$ and $\phi_{ik}^{(t)}(z_{ik})$ using the incoming messages from neighboring BSs. Accordingly, the complexity of estimating the number of active BSs becomes $O(M)$ for each BS. Thus, the overall complexity mainly depends on the calculation of $\alpha_{ia}^{(t)}$ and amounts to $O(M_i^2 \log M_i)$ for a single node in a single iteration. Considering the distributed nature of the algorithm, this amount of computational efforts is readily seen affordable.

In the proposed algorithm, to help improving the convergence property and numerical stability, a damping technique called the tree-reweighted message-passing algorithm [37] is employed. In this technique, a new version of messages $\tilde{\alpha}_{ia}^{(t)}$ and $\tilde{\rho}_{ia}^{(t)}$ are obtained by evaluating the linear combinations of the result of the message updated rule and the previous input messages with coefficient ε $(0 \le \varepsilon \le 1)$ as

$$\tilde{\alpha}_{ia}^{(t)} = \varepsilon\alpha_{ia}^{(t)} - (1-\varepsilon)\tilde{\rho}_{ia}^{(t)} \quad (38)$$

$$\tilde{\rho}_{ia}^{(t+1)} = \varepsilon\rho_{ia}^{(t+1)} - (1-\varepsilon)\tilde{\alpha}_{ia}^{(t)}. \quad (39)$$

Algorithm 1 Distributed Energy Efficiency Management Algorithm

Set $t \leftarrow 0$ and $\tilde{\rho}_{ia}^{(t)} = 0$ for all (i, a). All BSs are turned on.
Repeat
BS i updates $\tilde{\alpha}_{ia}^{(t)}$ using (38) for each $a \in \mathcal{N}(i)$, and sends it back to user a.
User a updates $\tilde{\rho}_{ia}^{(t)}$ using (39) for each $i \in \mathcal{N}(a)$, and sends it back to BS i.
All BSs run forward-backward updates using (33)-(36).
Individual BSs calculate the sum of (34) to obtain $\rho_i^{(t)}$ and use (35) to obtain $\hat{v}$.
Individual BSs turn off and release users if the sign of $\rho_i^{(t)}$ is negative.
Set $t \leftarrow t+1$.
Until convergence or t reaches t_{iter}.
User a uses (40) to determine its serving BS $\tilde{C}_a^{(t)}$.

This approach enhances the convergence behavior by accepting some information about the previous values. At each iteration, user a makes a tentative choice for BS as

$$\tilde{C}_a^{(t)} = \text{argmax}_{i\in\mathcal{I}}(\tilde{\alpha}_{ia}^{(t)} + \tilde{\rho}_{ia}^{(t)}). \quad (40)$$

Thus, upon convergence of the sum of messages, user a is associated with BS $\tilde{C}_a^{(t)}$. Algorithm 1 overviews the overall distributed energy efficiency management algorithm.

The optimality of the solution with the converged messages is closely related to the structure of the factor graph. In fact, by definition, the message updates rules of the developed algorithm ensure a necessary condition that any optimal solution of the problem in (4) satisfies. This indicates that, once each message takes a fixed value, i.e., converges to a limit, the equations representing the message updates rules hold and the solution for the messages becomes the optimum. Therefore, it is important to make the algorithm converge within a finite number of iterations. The optimality established in this step is valid only when the problem in (4) has a unique solution, since multiple solutions of the problem may make the messages oscillate. Since largely scattered throughput values of R_{ia} may lead to distinct objective values for different feasible assignments of variables, the corresponding solution is highly likely to be unique. In addition, the trellis-based optimization solution for BS network provides the best solution for a given configuration. By virtue of the factor graph structure and the quality of those solutions, the convergence property of user association can be further enhanced. A detailed description on the convergence behavior will be discussed with simulation results in the following section.

We consider a practical deployment protocol for the distributed energy efficiency management. Initially, all BSs broadcast their specific reference signals at their periodical cell selection process [38]. A BS in an idle state awakes to do this step. Users receiving multiple reference signals measure the signal quality to report the received signal received power (RSRP) [38] from their neighboring BSs. Since individual BSs and users can calculate system parameters based on the RSRP, Algorithm 1 can be carried out in a distributed

TABLE I
SIMULATION SETUP

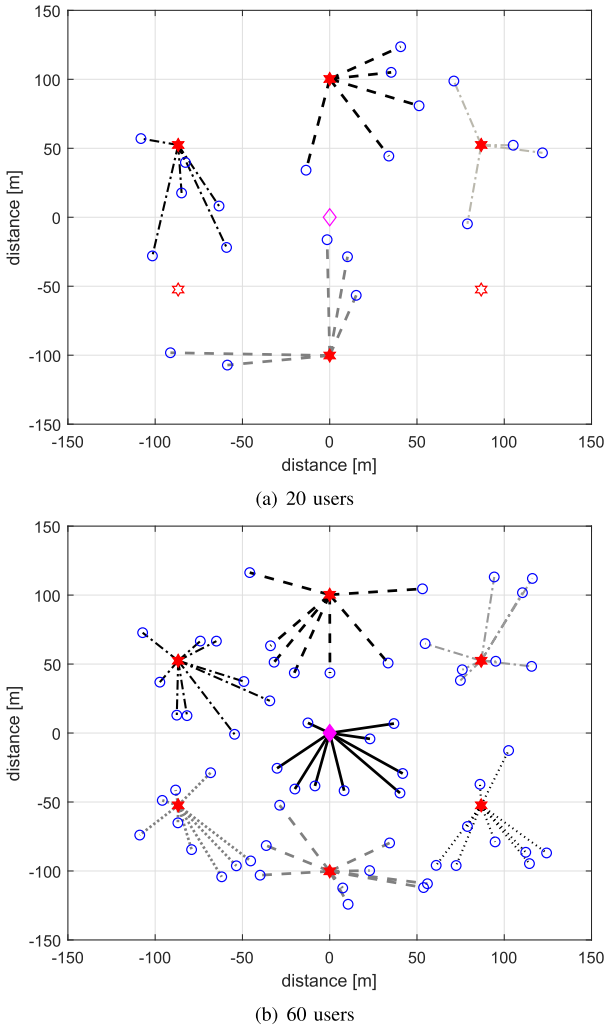
Description	Value
Path loss exponent λ	4.0
Macro BS transmit power P_T^M	37.99 dBm
Micro BS transmit power P_T^m	21.14 dBm
Macro BS operating power P_O^M	47.53 dBm
Micro BS operating power P_O^m	38.33 dBm
Thermal noise power σ^2	-174 dBm
Maximally number of users at micro BS M_i	9
Damping parameter ε	0.4
Maximal number of iterations t_{iter}	30
Number of BS M	7

manner over the network. Thus, message-passing operations for user association commence at individual BSs and users using (38) and (39), respectively. In addition, each BS runs the BS trellis operation using (33)-(36) over the BS network. For this trellis operation, messages are exchanged through a dedicated backhaul network [35]. Once the algorithm reaches the convergence, individual BSs determine whether to remain active or switched off. A BS which decides to switch off releases its users and deactivates some functions for energy-saving. This procedure repeats periodically.

IV. NUMERICAL RESULTS

In this section, we evaluate the developed algorithm in a two-tier HetNet consisting of one macrocell and multiple microcells in a rectangular region. The macro BS is located at the center of the region. The microcells are distributed at random so that mean coordinates of individual microcells are equally separated and apart from a given distance from the macro BS. Their actual positions are determined by Gaussian distributions with 5 m deviation. To reflect a relatively dynamic user distribution, users are initially distributed uniformly and are in a Brownian motion with the average displacement of 10 m for consecutive simulation instances. User coordinates are wrapped around so that the population within the simulation domain is preserved. We set M_1 to be twice greater than M_i, $(i = 2, \ldots, B)$.

The proposed message passing (MP) algorithm is compared with other existing algorithms which determine operating/sleeping BSs and user association. The conventional algorithms are briefly described here. A naive approach to find the user association relies on the use of channel state information itself by utilizing the fact that the channel gain is somewhat proportional to the throughput and inversely proportional to the energy consumption. Thus, the sum of channel states algorithm denoted by SC [21] lets each BS select users so that the sum of its supporting users' channel states is maximized, i.e., a BS updates the sum of downlink channel states by incrementing the number of supported users until reaching the limit. Then, a subset of BSs with the largest values is chosen to associate their nearby users. Also, the user channel selection algorithm denoted by UC [14] lets each user decide the association based on channel state information. If the selected BS is fully occupied by other users, the user chooses the next best BS. In addition, the individual energy efficiency algorithm denoted by IE evaluates the sum of individual BS's EEs to identify the user association in the algorithm [6], [15], [16]. Finally, the



(a) 20 users

(b) 60 users

Fig. 4. Snapshots of user association for different numbers of users.

primal-dual decomposition algorithm denoted by PD is based on a distributed technique of primal-dual decomposition [6], [39], [40]. The simulation is carried out with 1000 independent instances of user location configurations, and the results are averaged out. Table I summarizes the detailed description of the simulation setup.

Fig. 4 illustrates two instances of user association results of the proposed algorithm for 20 and 60 user populations. The macro BS, micro BSs and users are represented by a diamond, stars and circles, respectively. When the number of users is 20, the proposed MP algorithm decides to turn off the macro BS and two micro BSs, since nearby micro BSs can support all users. In such a case, active micro BSs accept relatively a large number of users. The resulting EE amounts to 6 bits/Hz/J. In contrast, for the case of 60 users, the MP algorithm achieves the network-wide EE of 3.6 bits/Hz/J. All BSs including the macro BS are turned on to provide connections to nearby users almost up to their user limits to avoid the inefficiency in energy consumption. As a result, the EE decreases when all BSs are turned on, since the operating energy of the macro BS is considerable. This indicates that the management of the number of active BSs is crucial to the EE maximization.

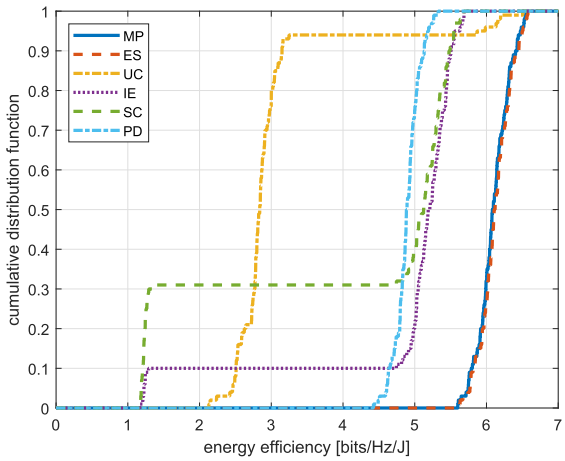


Fig. 5. Cumulative distribution functions of EE ($N = 20$).

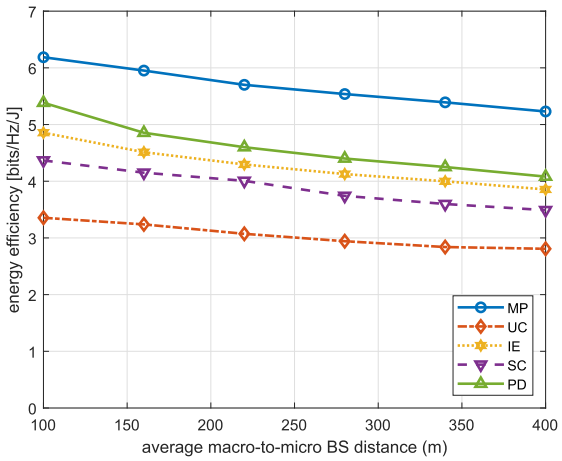


Fig. 7. Average EE with respect to the cell coverage.

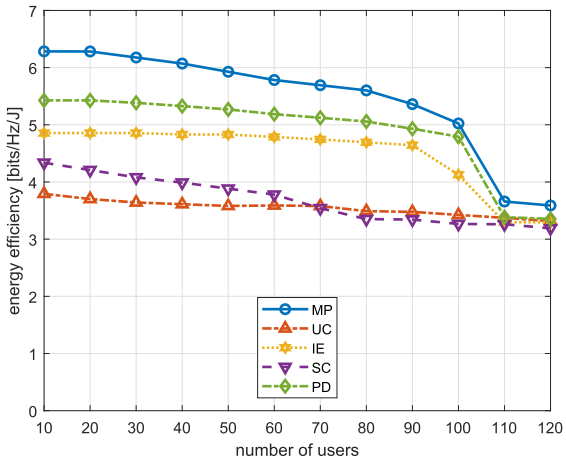


Fig. 6. Average EE with respect to the number of users in the network (micro BSs supporting total of 100 users).

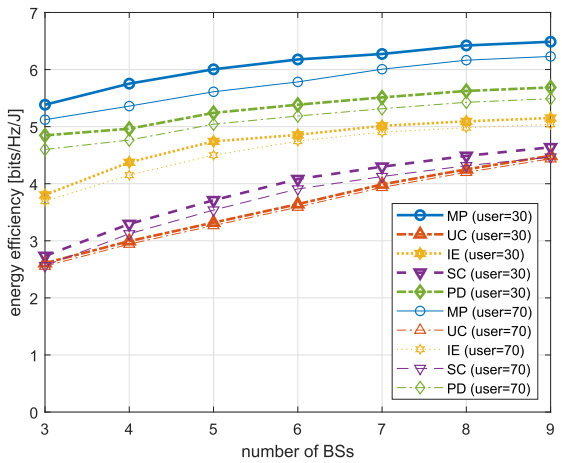


Fig. 8. Average EE with respect to the BS density.

Fig. 5 compares the cumulative distribution function (CDF) of the EE obtained by various algorithms in the case of $N = 20$. Since the rightmost curve obviously indicates the best performance, it is clear that the proposed algorithm outperforms other schemes consistently. The UC algorithm produces a moderate worst-case EE of 2.2 bits/Hz/J, since the macro BS is mostly turned on and most users try to access to the macro BS, thereby resulting in a large population group with relatively uniform EE. On the other hand, the SC and IE algorithms exhibit the worst-case EE of 1.2 bits/Hz/J. In contrast, the proposed MP algorithm shows the minimum EE value of 5.6 bits/Hz/J which is the highest among all schemes. To show how good the MP algorithm performs, the global optimal solution evaluated using exhaustive search (ES) is presented together. The result reveals that the MP has identical performance with the global optimum.

Fig. 6 presents the average EE performance in terms of the number of users with maximum users, and the number of users that the micro BS can support is set to 100. The EE decreases as the number of user connections increases. The proposed algorithm exhibits more than 17% performance gain over the PD algorithm which shows the best performance among existing algorithms. The gap narrows gradually with the increased users since a larger user population naturally reduces the EE value for most schemes. If the number of users goes beyond the micro BS's supporting capacity, the proposed algorithm turns on the macro BS and the energy consumption increases abruptly, thereby diminishing the resulting EE. Nevertheless, the MP algorithm still demonstrates a performance gain over other methods since a very efficient user association result is identified.

Fig. 7 compares the relationship between the EE and the cell coverage characterized by the average distance between the macro BS and micro BSs. The results show consistent trends of a monotonic EE decrease for the cell radius growth with distinct gaps between the proposed algorithm and other schemes. As the cell coverage increases, the propagation distance between a user and a BS becomes large. Thus, the sum throughput decreases while the energy consumption increases slightly faster, thereby reducing the total EE with a slow decay. This, in fact, indicates that sophisticated algorithms including the MP and PD algorithm can achieve better performance as the cell coverage becomes smaller, since there are greater chances for optimization with the average user-BS distance.

Fig. 8 depicts the average EE performance in terms of the number of BSs. The average distance between the macro BS

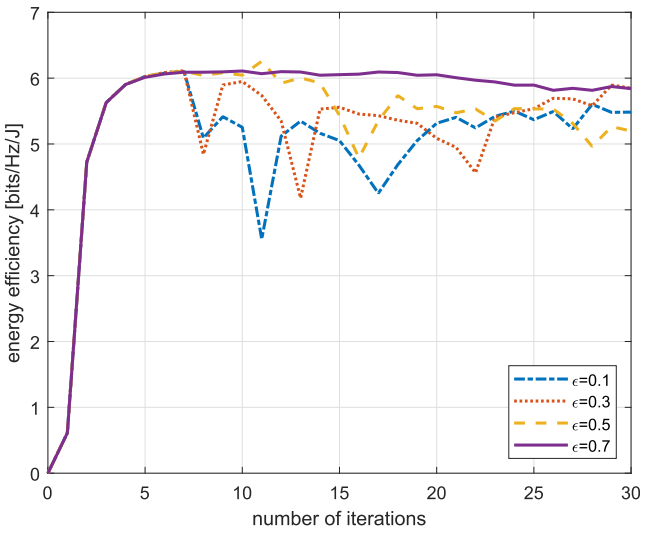


Fig. 9. Convergence behavior of the average objective function.

and micro BS is set to 100 m. For the fixed number of users, as the number of BSs increases, the network EE increases monotonically since interference is managed efficiently and the average distance of a user-BS pair is shortened if additional BSs are available. If the user population grows in the network, the interference among links becomes large and additional BSs including the macro BS begin to operate. Thus, the sum throughput decreases and the energy consumption increases, thereby dropping the network EE. The proposed MP algorithm shows performance gains over other techniques for both user population cases.

Fig. 9 exhibits the convergence property of the MP algorithm by tracking the change of the average objective value with respect to the number of iterations. The difference calculated at consecutive iterations is often considered as an importance measure in the analytical guarantee for convergent algorithms [41]. Instead of a mathematical proof, the average objective value is examined to assure the existence of a fixed point in the limit. The objective functions are evaluated with various values of the damping factor at each iteration and averaged over 1000 random instances. Initially, the objective values for all cases reach within the first 5 iterations up to the overestimated objective point and then show different decreasing dynamics with damping parameters. As it becomes small, the objective value has frequent changes. This indicates that the choice of a small value provides an improved quality of a solution, while a large value leads to accelerated convergence. We have obtained satisfactory performance by setting $\epsilon = 0.7$ in convergence and performance.

V. Conclusion

This article investigates a two-tier HetNet system to maximize the EE. A distributed algorithm that handles the user association by jointly optimizing the sum throughput and power saving is developed based on belief propagation. The developed algorithm appropriately configures the on-off states of the BSs with the joint consideration of the user association that achieves the best network sum throughput. To do so, the belief propagation algorithm conducts a clustering task to obtain the user association. Based on the association results, the BSs compete for their on-off states via trellis-based optimization over the BS network. We demonstrate that the proposed approach for the HetNet is practical in terms of implementation issues, the convergence and the distributed operation. Simulation results verify that the proposed algorithm improves the average EE performance by more than 25% over other conventional strategies.

References

[1] C. Han *et al.*, "Green radio: Radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, Jun. 2011.

[2] "Bringing 5G to power," Ericsson, Stockholm, Sweden, White Paper, Mar. 2020. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/industrylab/reports/bringing-5g-to-power

[3] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

[4] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, "Energy efficiency in the future Internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 2, pp. 223–244, 2nd Quart., 2011.

[5] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 30–38, Oct. 2011.

[6] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan, "Distributed energy efficient fair user association in massive MIMO enabled HetNets," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1770–1773, Oct. 2015.

[7] P. He, S. Zhang, L. Zhao, and X. Shen, "Multichannel power allocation for maximizing energy efficiency in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5895–5908, Jul. 2018.

[8] C. Li, J. Zhang, and K. B. Letaief, "Energy efficiency analysis of small cell networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 4404–4408.

[9] A. Damnjanovic *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–11, Jun. 2011.

[10] T. Nakamura *et al.*, "Trends in small cell enhancements in LTE advanced," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 99–107, Feb. 2013.

[11] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 823–831, Apr. 2016.

[12] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939–3952, Aug. 2014.

[13] A. J. Fehske, F. Richter, and G. P. Fettweis, "Energy efficiency improvements through micro sites in cellular mobile radio networks," in *Proc. IEEE Globecom Workshops*, Honolulu, HI, USA, Nov. 2009, pp. 1–5.

[14] I. Ashraf, F. Boccardi, and L. Ho, "SLEEP mode techniques for small cell deployments," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 72–79, Aug. 2011.

[15] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in Millimeter-Wave-Based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, Sep. 2017.

[16] F. Han, Z. Safar, W. S. Lin, Y. Chen, and K. J. R. Liu, "Energy-efficient cellular network operation via base station cooperation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Ottawa, ON, Canada, Jun. 2012, pp. 4374–4378.

[17] H. Ghazzai, M. J. Farooq, A. Alsharoa, E. Yaacoub, A. Kadri, and M.-S. Alouini, "Green networking in cellular HetNets: A unified radio resource management framework with base station ON/OFF switching," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 5879–5893, Jul. 2017.

[18] P.-H. Huang, S.-S. Sun, and W. Liao, "GreenCoMP: Energy-aware cooperation for green cellular networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 1, pp. 143–157, Jan. 2017.

[19] M. Baianifar, S. M. Razavizadeh, H. Akhlaghpasand, and I. Lee, "Energy efficiency maximization in mmWave wireless networks with 3D beamforming," *J. Commun. Netw.*, vol. 21, no. 2, pp. 125–135, Apr. 2019.

[20] J. Jung, S.-R. Lee, and I. Lee, "Saturation power-based simple energy efficiency maximization schemes for MISO broadcast channel systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6022–6031, Sep. 2017.
[21] G. Lee and H. Kim, "Green small cell operation of ultra-dense networks using device assistance," *Energies*, vol. 9, no. 12, p. 1065, Dec. 2016.
[22] S. H. Lee and I. Sohn, "Affinity propagation for energy-efficient BS operations in green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4534–4545, Aug. 2015.
[23] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, Nov. 2010.
[24] X. Weng, D. Cao, and Z. Niu, "Energy-efficient cellular network planning under insufficient cell zooming," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, Yokohama, Japan, May 2011, pp. 1–5.
[25] Z. Niu, S. Zhou, Y. Hua, Q. Zhang, and D. Cao, "Energy-aware network planning for wireless cellular system with inter-cell cooperation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1412–1423, Apr. 2012.
[26] S. Jang, H. Lee, S. Kang, T. Oh, and I. Lee, "Energy efficient SWIPT systems in multi-cell MISO networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8180–8194, Dec. 2018.
[27] M. Ashraf, J. Jung, H. M. Shin, and I. Lee, "Energy efficient online power allocation for two users with energy harvesting," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 24–28, Jan. 2019.
[28] R. Tao, W. Liu, X. Chu, and J. Zhang, "An energy saving small cell sleeping mechanism with cell range expansion in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2451–2463, May 2019.
[29] M. J. Daas, M. Jubran, and M. Hussein, "Energy management framework for 5G ultra-dense networks using graph theory," *IEEE Access*, vol. 7, pp. 175313–175323, 2019.
[30] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Multiobjective auction-based switching-off scheme in heterogeneous networks: To bid or not to bid?" *IEEE Trans. Veh. Tech.*, vol. 65, no. 11, pp. 9168–9180, Nov. 2016.
[31] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Game-theoretic infrastructure sharing in multioperator cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3326–3341, May 2016.
[32] F. Meshkati, H. V. Poor, and S. C. Schwartz, "Energy-efficient resource allocation in wireless networks," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 58–68, May 2007.
[33] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[34] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[35] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2*, document TS36.330, 3rd Generation Partnership Project, Sophia Antipolis Cedex, France, 2015.
[36] S. Schaible and J. Shi, "Fractional programming: The sum-of-ratios case," *Optim. Methods Softw.*, vol. 18, no. 2, pp. 219–229, Apr. 2003.
[37] M. Moretti, A. Abrardo, and M. Belleschi, "On the convergence and optimality of reweighted message passing for channel assignment problems," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1428–1432, Nov. 2014.
[38] *Evolved Universal Terrestrial Radio Access Physical Layer Measurements*, document TS36.214, 3rd Generation Partnership Project, Sophia Antipolis Cedex, France, 2015.
[39] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[40] Y. Chen, J. Li, Z. Lin, G. Mao, and B. Vucetic, "User association with unequal user priorities in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7374–7388, Sep. 2016.
[41] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

Sang Hyun Lee (Member, IEEE) received the B.S. and M.S. degrees from the Korea Advanced Institute of Science and Technology in 1999 and 2001, respectively, and the Ph.D. degree from The University of Texas at Austin in 2011. Since 2017, he has been with the School of Electrical Engineering, Korea University, Seoul, South Korea. His research interests include coding, optimization, inference, learning and their applications to wireless communication systems, and cross-disciplinary areas in natural and social sciences.

Mintae Kim received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the School of Electrical Engineering. His research interests include information theory, optimization techniques, and machine learning for the next-generation wireless communications.

Hunmin Shin received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2012, 2014, and 2018, respectively. Since 2018, he has been with Samsung Electronics as a Staff Engineer. In 2014, he visited the Imperial College London, London, U.K., to conduct collaborative research. His research interests include signal processing and optimization for wireless communication systems, such as interference network and green communications.

Inkyu Lee (Fellow, IEEE) received the B.S. (Hons.) degree in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1990, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1992 and 1995, respectively. From 1995 to 2001, he was a Member of the Technical Staff with Bell Laboratories, Lucent Technologies, where he studied high-speed wireless system designs. From 2001 to 2002, he was a Distinguished Member of the Technical Staff with Agere Systems (formerly the Microelectronics Group, Lucent Technologies), Murray Hill, NJ, USA. Since 2002, he has been with Korea University, Seoul, South Korea, where he is currently the Department Head of the School of Electrical Engineering. In 2009, he was a Visiting Professor with the University of Southern California, Los Angeles, CA, USA. He has authored or coauthored more than 180 journal articles in IEEE publications and has 30 U.S. patents granted or pending. His research interests include digital communications, signal processing, and coding techniques applied for next-generation wireless systems. He has served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 2001 to 2011 and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2007 to 2011. In addition, he was the Chief Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (Special Issue on 4G Wireless Systems) in 2006. He currently serves as the Co-Editor-in-Chief for the *Journal of Communications and Networks*. He was a recipient of the IT Young Engineer Award at the IEEE/IEEK Joint Award in 2006 and of the Best Paper Award at the Asia-Pacific Conference on Communications in 2006, the IEEE Vehicular Technology Conference in 2009, and the IEEE International Symposium on Intelligent Signal Processing and Communication Systems in 2013. He was a recipient of the Best Research Award from the Korean Institute of Communications and Information Sciences (KICS) in 2011, the Best Young Engineer Award from the National Academy of Engineering, South Korea, in 2013, and the Korea Engineering Award from the National Research Foundation of Korea in 2017. He has been elected as a member of National Academy of Engineering in Korea in 2015. He is an IEEE Distinguished Lecturer.