Energy-Efficient Context-Aware Matching for Resource Allocation in Ultra-Dense Small Cells

IEEE Access
The journal for rapid open access publishing

Received July 31, 2015, accepted September 1, 2015, date of publication September 15, 2015, date of current version October 15, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2478863

Energy-Efficient Context-Aware Matching for Resource Allocation in Ultra-Dense Small Cells

ZHENYU ZHOU[1], (Member, IEEE), MIANXIONG DONG[2], (Member, IEEE), KAORU OTA[2], (Member, IEEE), AND ZHENG CHANG[3], (Member, IEEE)

[1]State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China
[2]Department of Information and Electric Engineering, Muroran Institute of Technology, Muroran 050-8585, Japan
[3]Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä FIN-40014, Finland

Corresponding author: M. Dong (mx.dong@ieee.org)

This work was supported in part by the Grants-in-Aid for Scientific Research through the Japan Society for the Promotion of Science (JSPS) under Grant 15K15976 and Grant 26730056, in part by the JSPS A3 Foresight Program, in part by the National Natural Science Foundation of China under Grant 61203100, in part by the Fundamental Research Funds for the Central Universities under Grant 13MS19, 14MS08, and 15MS04, in part by the Academy of Finland under Grant 284748 and Grant 288473, and in part by the China Electric Power Research Institute through the State Grid Corporation of China.

ABSTRACT With the explosive growth of mobile data traffic and rapidly rising energy price, how to implement caching at small cells in an energy-efficient way is still an open problem and requires further research efforts. In this paper, we study the energy-efficient context-aware resource allocation problem, which falls into the category of mixed integer nonlinear programming (MINLP) and is NP-hard. To provide a tractable solution, the MINLP problem is decoupled and reformulated as a one-to-one matching problem under two-sided preferences, which are modeled as the maximum energy efficiency that can be achieved under the expected matching. An iterative algorithm is developed to establish preference profiles by employing nonlinear fractional programming and Lagrange dual decomposition. Then, we propose an energy-efficient matching algorithm based on the Gale–Shapley algorithm, and provide the detailed discussion and analysis of stability, optimality, implementation issues, and algorithmic complexity. The proposed matching algorithm is also extended to scenarios with preference, indifference, and incomplete preference lists by introducing some tie-breaking and preference deletion rules. The simulation results demonstrate that the proposed algorithm achieves significant performance and satisfaction gains compared with the conventional algorithms.

INDEX TERMS Energy-efficient, context-aware, caching, ultra-dense, small cell.

I. INTRODUCTION

The explosive growth of mobile internet will lead to an avalanche of mobile data traffic, which is predicted to increase more than 1000 times over the next decade [1]. Along with the explosion of data traffic, it is also expected that almost 50 billion devices will be connected by 2020 [2]. Considering that air-interface spectrum efficiency (SE) is approaching its physical limit and new spectrum acquisition becomes more and more difficult, a further requirement of 1000-fold increase in capacity from the long term evolution (LTE) system is a very challenging task [3].

To tackle this challenge, deploying small cells (SCs) in an ultra-dense way to complement existing cellular infrastructures provides a promising approach to further increase area SE by shrinking cell size and *bringing contents closer to users* [4]. SCs represent a novel networking paradigm shift from conventional long-range, high-cost macro base stations (MBSs) to short-range, low-cost small cell base stations (SBSs), which can be ultra-densely deployed underlaying MBSs to enable localized communications and high-density spatial reuse of wireless resources [4]. Despite numerous benefits, the integration of ultra-dense SBSs with conventional MBSs poses new challenges in resource allocation design. High-speed backhaul links which connect SBSs with the core network are indispensable to guarantee harsh and stringent quality of service (QoS) requirements of numerous delay-sensitive applications. In a regime that the cell density is comparable to user density [4], [5], it will be too costly to deploy high-speed fiber backhaul for every SBS. Therefore, state-of-the-art SC architectures propose to handle highly predicable bulky traffic by implementing caching at the wireless edge. By such, the storage capacity is able to

replace the limited backhaul capacity, which is used only to refresh caches at a rate that is much slower than users' content request rates.

A few works have addressed resource allocation problems in caching based SCs. The idea of using caching to support mobility has been exploited in [6]–[8]. The main underlying theme behind this body of works is to cache contents at local access points to reduce the delay experienced by users moving from one cell to another. Another line of works aim at maximizing the *hit ratio* of content requests by proactive caching. Due to the limited storage capacity, it is critical to properly predict future requests and then decide which contents to cache based on *content popularity*. A common simplification for content popularity is to assume that the global content popularity follows the Zipf distribution [4], [9], [10]. More complicated models that employ user-specific content popularity and social information for future content request predictions were studied in [11]–[13]. Instead of focusing on individual client, collaborative frameworks for coordinating content retrieval in cooperative SCs and users were proposed in [14] and [15], respectively. In [4] and [6], the authors proposed coded content caching schemes to enable robust and fast content dissemination in large-scale dynamic networks at the cost of increased computation complexity. The joint optimization of content caching and pushing through broadcasting is shown to effectively alleviate cellular data bottleneck and resolve randomness of content request arrivals [10], [17]. In video streaming, while full caching would incur tremendous cost, partial caching that only stores part of the requested video stream based on viewing patterns provides a promising way for efficient content splitting [18], [19].

However, most of the previous studies mainly focus on SE optimization and ignore energy efficiency (EE) during the resource allocation process. It is difficult to directly extended these works to the domain of EE optimization because optimum EE and SE are not always achievable simultaneously and may sometimes even conflict with each other [20]. On the other hand, EE has already become a critical design factor in cellular networks due to rapidly rising energy price and environmental concerns [21]. The current mobile network operational expenditure (OPEX) for electricity globally is already more than 10 billion dollars, among which $60\% - 80\%$ of the energy is consumed by BSs [22]. The CO_2 emissions produced by cellular networks are as high as those from 8 million vehicles [21]. It is expected that the ultra-densely deployed SBSs and their corresponding network infrastructures will further incur significant increase in electrical energy consumption and CO_2 production. Furthermore, energy-efficient resource allocation is also important for UEs because UEs with limited battery capacity can quickly run out of battery if without careful energy optimization design. As a result, how to implement caching at the ultra-dense SCs in an energy-efficient way while satisfying various practical system constraints such as transmission power, backhaul capacity, storage capacity, QoS requirement, etc., is still an open problem and requires further research efforts.

In this paper, we propose an energy-efficient context-aware matching approach for resource allocation by exploiting properties of matching theory, nonlinear fractional programming, and Lagrange dual decomposition. Matching theory provides a low-complexity decentralized self-organizing solution to the *two-sided matching problem* in college admissions [23], marriage stability [24], labor markets [25], etc., and has been widely applied for solving resource allocation problems in cellular networks [6], [11], [26], cognitive radios [27], social networks [28], D2D communications [29], mobile energy-harvesting networks [30], etc.

The contributions of this paper are summarized in the following three main aspects:

- We consider the scenario that both SBSs and UEs seek to form proper SBS-UE partnerships in order to maximize EE under practical constraints of QoS, transmission power, backhaul capacity, and storage capacity. The energy-efficient context-aware resource allocation problem for SBSs is formulated as a *joint partner selection and power allocation problem*, which falls into the category of NP-hard mixed integer nonlinear programming (MINLP). To provide a tractable solution, we decouple the partner selection and power allocation subproblems by reformulating the MINLP problem as a *one-to-one matching* problem, which consists of two finite and disjoint sets (SBSs and UEs), and their preferences over each other.
- Preference profiles of SBSs and UEs are modeled based on maximum EE that can be achieved under the expected matching by taking into consideration dynamically varying channel states and aggregate interference levels. A SBS's preference is obtained by solving a distributed power allocation problem, which is nonconvex due to the objective function in fractional form. We transform the nonconvex problem into an equivalent convex one with an objective function in subtractive form and propose an iterative algorithm to solve it based on nonlinear fractional programming [31] and Lagrange dual decomposition [32].
- With the established preference profiles, we propose an matching algorithm based on the Gale-Shapley (GS) algorithm [23], and prove that the produced matching is not only stable for both SBSs and UEs, but also is weak Pareto optimal for SBSs. We also extend the proposed matching algorithm to the cases of *preference indifference* and *incomplete preference lists* by introducing some *tie-breaking* and *preference deletion* rules, and provide a detailed discussion and analysis for stability, optimality, implementation issues, and algorithmic complexity. Finally, the proposed algorithm is evaluated by simulations and compared with conventional algorithms from the perspective of EE performance and satisfaction gains.

The structure of this paper is organized as follows: Section II describes the system model and the problem formulation. Section III introduces the proposed energy-efficient context-aware matching algorithm and analyzes the stability, optimality, implementation issues and algorithmic complexity. Section IV presents simulation results and performance evaluations. Section V draws relevant conclusions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we firstly introduce the system model of caching-enabled ultra-dense small cells, and then presents the problem formulation.

A. SYSTEM MODEL

We consider the ultra-dense SC scenario that is composed of a conventional high-power MBS, K low-power SBSs such as picocells, microcells, or femtocells, and M UEs [33]. The density of SBSs is comparable to or even larger than UE density, i.e., $K > M$, which is shown in Fig. 1. Due to the ultra-dense deployment, we assume that each SBS can serve at most one UE. The MBS is used to provide downlink coverage of the overall network, which is generally equipped with advanced signal processing units and high-speed backhaul links. In contrast, SBSs with limited storage and backhaul capacities are deployed near to UEs to offload traffic loads from the MBS and deliver high QoS at low operation costs.

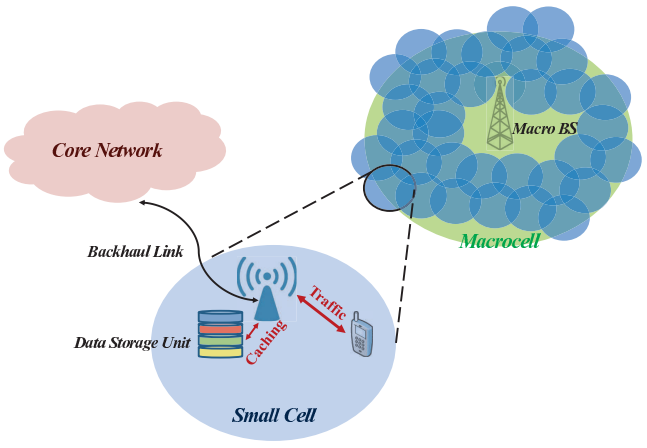


FIGURE 1. System model of ultra-dense small cells.

Throughout the paper, the words "file" and "content" are used interchangeably for the same meaning. In the network, the sets of SBSs and UEs are denoted as $\mathcal{S} = \{s_1, \cdots, s_k, \cdots, s_K\}$ and $\mathcal{U} = \{u_1, \cdots, u_m, \cdots, u_M\}$, respectively. Each SBS $s_k \in \mathcal{S}$ is equipped with a data storage of capacity D_k that contains a set of cached files $\mathcal{C}_k \subseteq \mathcal{C}$ from the total set of contents $\mathcal{C}$ in the system. For simplicity, all files are assumed to have the same size s. Each file is requested based on its popularity, which is assumed to follow a Zipf distribution [9], [10]. When s_k has available backhaul bandwidth, it can cache popular files via the backhaul until reaching the maximum storage capacity D_k. The macro-cell spectrum is divided into orthogonal frequency channels (e.g., an orthogonal resource block in LTE), and each SBS $s_k \in \mathcal{S}$ is allocated with a bandwidth w_k. To increase SE, the frequency spectrum is reused by SBSs after certain geographical distance with dynamic frequency allocation [34]. Both SBSs and UEs seek to form SBS-UE partnerships through proper partner selections in order to optimize EE while guaranteeing QoS requirements. The SBS and UE partner selection decisions are defined as follows.

Definition 1: The SBSs' partner selection matrix $\mathbf{X}$ is a $K \times M$ matrix with the (k, m)-th element $x_{k,m} \in \{0, 1\}$ indicating the SBS-UE partnership (s_k, u_m) for the SBS s_k, $\forall s_k \in \mathcal{S}$, $\forall u_m \in \mathcal{U}$. If $x_{k,m} = 1$, s_k is willing to form a partnership with u_m, and if $x_{k,m} = 0$, otherwise.

Definition 2: The UEs' partner selection matrix $\mathbf{Y}$ is a $M \times K$ matrix with the (m, k)-th element $y_{m,k} \in \{0, 1\}$ indicating the SBS-UE partnership (u_m, s_k) for the UE u_m, $\forall s_k \in \mathcal{S}$, $\forall u_m \in \mathcal{U}$. If $y_{m,k} = 1$, u_m is willing to form a partnership with s_k, and if $y_{m,k} = 0$, otherwise.

Remark 1: In general, the partner selection decisions of s_k and u_m may contradict with each other due to different preferences, i.e., $x_{k,m} \neq y_{m,k}$. A SBS-UE partnership (s_k, u_m) can be formed if and only if $x_{k,m} = y_{m,k} = 1$.

Over a time period of T, each UE $u_m \in \mathcal{U}$ requests N_m files $\mathcal{F}_m = \{f_1^m, \cdots, f_{N_m}^m\}$, $\mathcal{F}_m \subset \mathcal{C}$ from its serving SBS, e.g., s_k. If the requested files are available in the cache of s_k, i.e., $\mathcal{F}_m \cap \mathcal{C}_k$, then these files will be delivered directly from s_k to u_m without going through the core network. Defining t_c as the channel coherence time, there are a total of $\lfloor T/t_c \rfloor$ time instants, and the interval duration between any two consecutive time instants is t_c. At time instant t, the maximum achievable transmission rate between s_k and u_m is given by

$$r_{k,m}(t) = w_k \log_2 \left(1 + \frac{x_{k,m} y_{m,k} p_k(t) g_{k,m}(t)}{N_0 + \sum_{j \neq k, j \in \mathcal{S}_k} p_j(t) g_{j,m}(t)}\right), \quad (1)$$

where $g_{k,m}$ is the channel gain between s_k and u_m, p_k is the transmission power of s_k, and $\sum_{j \neq k, j \in \mathcal{S}_k} p_j(t) g_{j,m}(t)$ denotes the aggregate interference caused by the set of SBSs $\mathcal{S}_k$ that reuse the same channel. $g_{j,m}$ represents the interference channel gain between s_j and u_m. Fast fading due to multi-path propagation and slow fading due to shadowing and pathloss are considered in the channel model [35]. For example, $g_{k,m}$ is given by

$$g_{k,m} = \varpi \beta_{k,m} \zeta_{k,m} d_{k,m}^{-\alpha}, \quad (2)$$

where ϖ is the pathloss constant, $\beta_{k,m}$ is the fast-fading gain with exponential distribution, $\zeta_{k,m}$ is the slow-fading gain with log-normal distribution, α is the pathloss exponent, and $d_{k,m}$ is the transmission distance.

Denoting $\|\mathcal{F}_m \cap \mathcal{C}_k\|$ as the number of files that are available in s_k's cache, the estimated time duration to transmit

$\|\mathcal{F}_m \cap \mathcal{C}_k\|$ to u_m is given by

$$\tilde{T}_{k,m}^{cache} = \frac{\|\mathcal{F}_m \cap \mathcal{C}_k\| s}{\bar{r}_k(t)}, \tag{3}$$

where $\bar{r}_k(t)$ is the historical average transmission rate of s_k until time instant t. In case of $\tilde{T}_{k,m}^{cache} > T$, only a portion of cached data can be delivered to u_m within $[0, T]$. Therefore, the time duration during which cached data can be transmitted is defined as

$$T_{k,m}^{max} = \min\{\tilde{T}_{k,m}^{cache}, T\}. \tag{4}$$

If the requested files are not available in s_k's cache, i.e., $\mathcal{F}_m \backslash \{\mathcal{F}_m \cap \mathcal{C}_k\}$, s_k has to firstly retrieve these files from the core network through its backhaul link with capacity B_k, and then transmits them to u_m. The maximum transmission rate between s_k and u_m is given by

$$r_{k,m}^{max}(t) = \min\{r_{k,m}(t), x_{k,m} y_{m,k} B_k\}. \tag{5}$$

Note that, in case of limited backhaul capacity and user proximity, B_k is insufficient to keep up with $r_{k,m}(t)$, i.e., $B_k \ll r_{k,m}(t)$. As a result, u_m will experience a significantly increased delay, which is independent from the quality of wireless channels.

The total throughput (bits) between s_k and u_m during T is defined as

$$U_{k,m} = \left(\sum_{t=1}^{\left\lfloor T_{k,m}^{max}/t_c \right\rfloor} r_{k,m}(t) + \sum_{t=\left\lfloor T_{k,m}^{max}/t_c \right\rfloor}^{\lfloor T/t_c \rfloor} r_{k,m}^{max}(t) \right) t_c, \tag{6}$$

where $\lfloor x \rfloor$ denotes the largest integer not greater than x.

The total energy consumption (J) of s_k during T is given by

$$E_k = \sum_{t=1}^{\lfloor T/t_c \rfloor} \left(\frac{1}{\eta} p_k(t) + p_k^{cir} \right) t_c, \tag{7}$$

where η is the power amplifier (PA) efficiency, i.e., $0 < \eta < 1$. p_k^{cir} is the circuit power of s_k which represents the average energy consumption of device electronics such as mixers, filters, digital-to-analog/analog-to-digital converters, etc., and is assumed as a constant.

The *utility* EE (bits/J) is defined as the ratio of throughput to total energy consumption [36]. The EE of s_k is given by

$$\begin{aligned} U_k^{EE} &= \frac{\sum_{u_m \in \mathcal{U}} U_{k,m}}{E_k} \\ &= \frac{\sum_{t=1}^{\left\lfloor T_{k,m}^{max}/t_c \right\rfloor} r_{k,m}(t) + \sum_{t=\left\lfloor T_{k,m}^{max}/t_c \right\rfloor}^{\lfloor T/t_c \rfloor} r_{k,m}^{max}(t)}{\sum_{t=1}^{\lfloor T/t_c \rfloor} \frac{1}{\eta} p_k(t) + p_k^{cir}}. \end{aligned} \tag{8}$$

Regarding UE u_m, since the total downlink throughput between s_k and u_m during T is the same as $U_{k,m}$ defined in (6), we only need to consider the total energy consumption, which is calculated as

$$E_m = p_m^{cir} T, \tag{9}$$

where p_m^{cir} is the circuit power of u_m consumed for receiving data, and is assumed as a constant. The EE of u_m is given by

$$\begin{aligned} U_m^{EE} &= \frac{\sum_{s_k \in \mathcal{S}} U_{k,m}}{E_m} \\ &= \frac{\left(\sum_{t=1}^{\left\lfloor T_{k,m}^{max}/t_c \right\rfloor} r_{k,m}(t) + \sum_{t=\left\lfloor T_{k,m}^{max}/t_c \right\rfloor}^{\lfloor T/t_c \rfloor} r_{k,m}^{max}(t) \right) t_c}{p_m^{cir} T}. \end{aligned} \tag{10}$$

B. PROBLEM FORMULATION

In this subsection, firstly, we introduce the problem formulation for SBSs. During the resource allocation process, any SBS faces the following challenges:

- Which UE should be selected to form a SBS-UE partnership in order to maximize EE considering various practical constraints and factors such as local content availability, backhaul capacity, QoS requirement, channel state, and interference levels, etc?
- How much transmission power should be allocated for the expected SBS-UE partnership in order to maximize EE while satisfying QoS and transmission power constraints?
- Will the selected UE also be willing to form this partnership?
- Will the formed partnership be easily disrupted by other SBSs who also want to form partnerships with the selected UE?

Thus, the energy-efficient context-aware resource allocation problem for a SBS s_k consists of two subproblems, the first is denoted as the *power allocation subproblem*, and the second is denoted as the *partner selection subproblem*. The power allocation subproblem and the partner selection subproblem are solved in Subsection III-A, and III-B, respectively.

Denoting $\mathbf{x_k} = \{x_{k,1}, \cdots, x_{k,m}, \cdots, x_{k,M}\}$ and $\mathbf{p_k} = \{p_k(1), \cdots, p_k(t), \cdots, p_k(\lfloor T/t_c \rfloor)\}$ as s_k's partner selection and power allocation strategy sets, respectively. For any $s_k \in \mathcal{S}$, the corresponding EE optimization problem is formulated as

$$\begin{aligned} \max_{(\mathbf{x_k}, \mathbf{p_k})} \ & U_k^{EE}(\mathbf{x_k}, \mathbf{p_k}) \\ \text{s.t. } & C_1: 0 \le p_k(t) \le p_k^{max}, \quad t = \left[1, \left\lfloor T_{k,m}^{max}/t_c \right\rfloor\right] \\ & C_2: 0 \le p_k(t) \le p_k^B(t), \quad t = \left[\left\lfloor T_{k,m}^{max}/t_c \right\rfloor, \lfloor T/t_c \rfloor\right] \\ & C_3: U_{k,m}(x_{k,m}, \mathbf{p_k}) \ge x_{k,m} y_{m,k} U_{k,m}^{min}, \quad \forall u_m \in \mathcal{U}, \\ & C_4: x_{k,m} = \{0, 1\}, \quad \forall u_m \in \mathcal{U}, \\ & C_5: \sum_{u_m \in \mathcal{U}} x_{k,m} \le 1. \end{aligned} \tag{11}$$

C_1 and C_2 are the maximum transmission power constraints, i.e., the transmission power should be no greater than p_k^{max} when $t = \left[1, \left\lfloor T_{k,m}^{max}/t_c \right\rfloor\right]$, and no greater than $p_k^B(t)$ when $t = \left[\left\lfloor T_{k,m}^{max}/t_c \right\rfloor, \lfloor T/t_c \rfloor\right]$. $p_k^B(t)$ is the transmission power

that achieves exactly the backhaul capacity B_k and is calculated as

$$p_k^B(t) = \frac{\left(2^{\frac{B_k}{w_k}} - 1\right)\left(N_0 + \sum_{j\neq k, j\in\mathcal{S}_k} p_j(t) g_{j,m}(t)\right)}{g_{k,m}(t)}. \quad (12)$$

C_3 specifies the QoS requirement in terms of minimum throughput. C_4 and C_5 ensure that s_k serves at most one UE.

Secondly, we introduce the problem formulation for UEs. Similarly, any UE also faces the following challenges:

- Which SBS should be selected to form a UE-SBS partnership in order to maximize EE?
- Considering various practical constraints and factors, will the selected SBS also be willing to form this partnership?
- Will the formed partnership be easily disrupted by other UEs who also want to form partnerships with the selected SBS?

For each $u_m \in \mathcal{U}$, denoting $\mathbf{y_m} = \{y_{m,1}, \cdots, y_{m,k}, \cdots, y_{m,K}\}$ as u_m's partner selection strategy set, the corresponding EE optimization problem is formulated as

$$\begin{aligned} &\max_{\mathbf{y_m}} \ U_m^{EE}(\mathbf{y_m}) \\ &\text{s.t. } C_6: U_{k,m}(y_{m,k}) \geq x_{k,m} y_{m,k} U_{k,m}^{min}, \quad \forall s_k \in \mathcal{S}, \\ &\quad C_7: y_{m,k} = \{0, 1\}, \quad \forall s_k \in \mathcal{S}, \\ &\quad C_8: \sum_{s_k\in\mathcal{S}} y_{m,k} \leq 1. \end{aligned} \quad (13)$$

C_6 specifies the QoS constraint similarly as C_3. C_7 and C_8 ensure that u_m is served by at most one SBS.

Remark 2: Regarding the QoS constraints C_3 and C_6, from (6), we can see that the files $\mathcal{F}_m \cap \mathcal{C}_k$ are transmitted at data rate $r_{k,m}$ during the interval $[0, T_{k,m}^{max}]$, and the files $\mathcal{F}_m \backslash \{\mathcal{F}_m \cap \mathcal{C}_k\}$ are transmitted at data rate $r_{k,m}^{max}$ during the interval $[T_{k,m}^{max}, T]$. As a result, $\mathcal{F}_m \cap \mathcal{C}_k$ only depends on the instantaneous channel capacity, but $\mathcal{F}_m \backslash \{\mathcal{F}_m \cap \mathcal{C}_k\}$ are exposed to possible QoS degradation due to the limited backhaul capacity B_k.

Remark 3: From (11), it is noted that the partner selection subproblem and the power allocation subproblem are coupled with each other. The formulated problem falls into the category of MINLP, which is NP-hard and computationally intractable.

To solve (11), we decouple the partner selection and the power allocation subproblems by reformulating the MINLP problem as a *one-to-one matching problem*. The matching problem is denoted as the triple $(\mathcal{S}, \mathcal{U}, \mathcal{P})$, which consists of two finite and disjoint sets, i.e., $\mathcal{S}$, $\mathcal{U}$, and the set of their preferences $\mathcal{P}$. A *one-to-one matching* μ is defined as follows [24]:

Definition 3: In the matching problem $(\mathcal{S}, \mathcal{U}, \mathcal{P})$, a *matching* μ is a one-to-one correspondence from the set $\mathcal{S} \cup \mathcal{U}$ onto itself under preference $\mathcal{P}$ such that for each $s_k \in \mathcal{S}$ and $u_m \in \mathcal{U}$, $\mu(s_k) \in \mathcal{U} \cup \{s_k\}$ and $\mu(u_m) \in \mathcal{S} \cup \{u_m\}$. $\mu(s_k) = u_m$ if and only if $\mu(u_m) = s_k$.

If $\mu(u_m) = u_m$ or $\mu(s_k) = s_k$, u_m or s_k stays single. It means that either none SBS is able to satisfy the QoS requirement of u_k, or there is no UE to be served by s_k. In the former case, QoS requirement should be reduced to a lower level for s_k to be served by a SBS. While in the latter case, s_k can enter into sleeping mode to save energy. $\mu(s_k) = s_k$ and $\mu(s_k) \in \mathcal{U}$ cannot hold at the same time. The same property holds for $\mu(u_m)$. Either s_k or u_m designs a proposal to demonstrate its expected partner based on its preference. A total of K SBS (or M UEs) may jointly design K (or M) proposals, and the set of these proposals is defined as a matching. We assume that either s_k or u_m cares about whom it is matched with, but is not otherwise concerned with partners of other SBSs or UEs. In the following, we discuss how to establish the two-sided preferences $\mathcal{P}$ and how to produce an energy-efficient context-aware matching μ under $\mathcal{P}$.

III. THE ENERGY-EFFICIENT CONTEXT-AWARE MATCHING APPROACH

In this section, firstly, we show how to establish preference profiles by using an iterative algorithm. Then, the energy-efficient context-aware matching algorithm under the established two-sided preferences is introduced in detail. Finally, stability, optimality, implementation issues and algorithmic complexity are discussed and analyzed.

A. PREFERENCE PROFILE ESTABLISHMENT

Before solving the energy-efficient matching problem, the set of preference profiles $\mathcal{P}$ needs to be established. Taking the SBS s_k as an example, we model s_k's preference over u_m as the maximum EE that can be achieved when they are matched with each other, i.e., $\mu(s_k) = u_m$, $\mu(u_m) = s_k$, and $x_{k,m} = y_{m,k} = 1$. The maximum EE under this matching is obtained by solving the following power allocation problem:

$$\begin{aligned} &\max_{\mathbf{p_k}} \ U_k^{EE}(\mathbf{p_k})\Big|_{\mu(s_k)=u_m} \\ &\text{s.t. } C_1, C_2, C_3. \end{aligned} \quad (14)$$

To solve (14), we introduce a transformation to handle the nonconvex problem via nonlinear fractional programming [31], and propose an iterative algorithm to solve the transformed problem. The maximum EE under $\mu(s_k) = u_m$ is defined as

$$q_{k,m}^* = \max_{\mathbf{p_k}} U_k^{EE}(\mathbf{p_k})\Big|_{\mu(s_k)=u_m} = \frac{U_{k,m}(\mathbf{p_k^*})}{E_k(\mathbf{p_k^*})}, \quad (15)$$

where $\mathbf{p_k^*}$ is the optimum power allocation strategy set of s_k. The following theorem can be derived based on [31]:

Theorem 1: q_i^{d*} is achieved if and only if

$$\begin{aligned} &\max_{\mathbf{p_k}} \ U_{k,m}(\mathbf{p_k}) - q_{k,m}^* E_k(\mathbf{p_k}) \\ &\qquad = U_{k,m}(\mathbf{p_k^*}) - q_{k,m}^* E_k(\mathbf{p_k^*}) = 0. \end{aligned} \quad (16)$$

Theorem 1 shows that the transformed problem with an objective function in subtractive form is equivalent to the nonconvex problem with an objective function in fractional form.

Algorithm 1 Iterative Power Allocation Algorithm

Input: $\mathcal{F}_m$, $g_{k,m}$, $\sum_{j\in\mathcal{S}_k, j\neq k} g_{j,m}p_j$, $\mathcal{C}_k$, $\tilde{T}^{cache}_{k,m}$, p^{cir}_k, p^{max}_k, p^B_k, $U^{min}_{k,m}$.
Output: $q^*_{k,m}$.
Initialize: N_{max}, Δ, $q_{k,m}$
while $n < N_{max}$ **do**
  obtain $\hat{\mathbf{p}}_\mathbf{k}$ using (21) or (22)
  if $U_{k,m}(\hat{\mathbf{p}}_\mathbf{k}) - q_{k,m}E_k(\hat{\mathbf{p}}_\mathbf{k}) > \Delta$ **then**
    Update: $q_{k,m} = U_{k,m}(\hat{\mathbf{p}}_\mathbf{k})/E_k(\hat{\mathbf{p}}_\mathbf{k})$
  else
    $p^*_k = \hat{\mathbf{p}}_\mathbf{k}$, and $q^*_{k,m} = U_{k,m}(\mathbf{p}^*_\mathbf{k})/E_k(\mathbf{p}^*_\mathbf{k})$
  end if
  Update: $n = n + 1$
end while

Therefore, instead of solving (14), we can focus on the following equivalent problem:

$$\max_{\mathbf{p_k}} \; U_{k,m}(\mathbf{p_k}) - q^*_{k,m}E_k(\mathbf{p_k}) \quad \text{s.t. } C_1, C_2, C_3. \tag{17}$$

The new problem can be viewed as a weighted sum of $U_{k,m}$ and E_k, where the parameter $q^*_{k,m}$ acts as the *price* (negative weight) of the power consumption.

It can be easily seen that (17) is a convex optimization problem. However, the specific value of $q^*_{k,m}$ is still unknown. Thus, we propose an iterative algorithm based on Dinkelbach's method to find $q^*_{k,m}$. The iterative power allocation algorithm is summarized in Algorithm 1. The initial values of $q_{k,m}$ can be set as a small positive number, e.g., 10^{-4}. At each iteration, the following transformed problem is solved

$$\max_{\mathbf{p_k}} \; U_{k,m}(\mathbf{p_k}) - q_{k,m}E_k(\mathbf{p_k}) \quad \text{s.t. } C_1, C_2, C_3. \tag{18}$$

Karush-Kuhn-Tucker (KKT) conditions and Lagrange dual decomposition are used to solve the above problems. The Lagrangian associated with (18) is given by

$$\begin{aligned}\mathcal{L}^{EE}_k(\mathbf{p_k}, \Lambda_k, \Theta_k, \vartheta_k) &= U_{k,m}(\mathbf{p_k}) - q_{k,m}E_k(\mathbf{p_k}) - \sum_{t=1}^{\lfloor T^{max}_{k,m}/t_c \rfloor} \lambda_k(t)(p_k(t) - p^{max}_k) \\ &+ \vartheta_k\Big(U_{k,m}(\mathbf{p_k}) - U^{min}_{k,m}\Big) - \sum_{t=\lfloor T^{max}_{k,m}/t_c \rfloor}^{\lfloor T/t_c \rfloor} \theta_k(t)\big(p_k(t) - p^B_k(t)\big),\end{aligned} \tag{19}$$

where Λ_k and Θ_k are the Lagrange multiplier vectors associated with $C_1 \sim C_2$, respectively. ϑ_k is the Lagrange multiplier corresponding to C_3. The equivalent dual problem is decomposed as [32]

$$\min_{(\Lambda_k, \Theta_k, \vartheta_k \geq 0)} \; \max_{\mathbf{p_k}} \; \mathcal{L}^{EE}_k(\mathbf{p_k}, \Lambda_k, \Theta_k, \vartheta_k). \tag{20}$$

If $t = \left[1, \left\lfloor T^{max}_{k,m}/t_c \right\rfloor\right]$, the optimal value $\hat{p}_k(t)$ corresponding to $q_{k,m}(t)$ is given by

$$\hat{p}_k(t) = \left[\frac{\eta(1+\vartheta_k)w_k t_c \log_2 e}{q_{k,m}(t) + \eta\lambda_k(t)} - \frac{I_{k,m}(t) + N_0}{g_{k,m}(t)}\right]^+, \tag{21}$$

where $I_{k,m}(t) = \sum_{j\in\mathcal{S}_k, j\neq k} g_{j,m}(t)p_j(t)$, and $[x]^+ = \max\{0, x\}$. If $t = \left[\left\lfloor T^{max}_{k,m}/t_c \right\rfloor, \lfloor T/t_c \rfloor\right]$, $\hat{p}_k(t)$ is given by

$$\hat{p}_k(t) = \left[\frac{\eta(1+\vartheta_k)w_k t_c \log_2 e}{q_{k,m}(t) + \eta\theta_k(t)} - \frac{I_{k,m}(t) + N_0}{g_{k,m}(t)}\right]^+. \tag{22}$$

Equation (21) and (22) indicate a water-filling algorithm, where the water level is determined by the cost of satisfying the transmission power and QoS constraints, i.e., λ_k, θ_k, and ϑ_k, respectively, as well as the current cost of total power consumption given by $q_{k,m}$. Then, the Lagrange multipliers are updated by using the gradient method [37] as

$$\begin{aligned}\lambda_k(t, \tau+1) &= \left[\lambda_k(t,\tau) + \epsilon_{k,\lambda}(t,\tau)\left(\hat{p}_k(t,\tau) - p^{max}_k\right)\right]^+, \\ \theta_k(t, \tau+1) &= \left[\theta_k(t,\tau) + \epsilon_{k,\theta}(t,\tau)\left(\hat{p}_k(t,\tau) - p^B_k(t)\right)\right]^+, \\ \vartheta_k(\tau+1) &= \left[\vartheta_k(\tau) - \epsilon_{k,\vartheta}(\tau)\left(U_{k,m}(\tau) - U^{min}_{k,m}\right)\right]^+,\end{aligned} \tag{23}$$

where τ is the iteration index, and ϵ is the positive step size. We have adopted a constant step size to strike a balance between optimality and convergence speed.

Then, $q_{k,m}$ is updated for the next iteration as $q_{k,m} = U_{k,m}(\hat{\mathbf{p}}_\mathbf{k})/E_k(\hat{\mathbf{p}}_\mathbf{k})$. The iteration process will continue until either the stoping criteria Δ or the maximum iteration number N_{max} is reached. In the final iteration, we set $\mathbf{p}^*_\mathbf{k} = \hat{\mathbf{p}}_\mathbf{k}$, and calculate $q^*_{k,m}$ as (15).

Similar to the modeling of s_k's preference over u_m, the preference of u_m over s_k is also defined as the maximum EE that can be achieved under $\mu(u_m) = s_k$. Since u_m has no knowledge of the contents stored in the s_k's cache, and the denominator of (10) is a constant term over $[0, T]$, u_m's preference only depends on the total throughput that can be provided by s_k during $[0, T]$, i.e., $U_{k,m}$.

The preference profile establishment algorithm is summarized in Algorithm 2. Firstly, by using Algorithm 1, s_k is able to obtain the maximum EE that can be achieved for every possible matching with $u_m \in \mathcal{U}$. We introduce a preference relation $\succ$, which is a complete, reflexive, and transitive binary relation between any $s_k \in \mathcal{S}$ and $u_m \in \mathcal{U}$ [24]. We write $u_m \succ_{s_k} u_{m'}$ to mean s_k prefers u_m to $u_{m'}$, which is defined as

$$u_m \succ_{s_k} u_{m'} \Leftrightarrow q^*_{k,m} > q^*_{k,m'}. \tag{24}$$

Similarly, we write $s_k \succ_{u_m} s_j$ to mean u_m prefers s_k to s_j, which is defined as

$$s_k \succ_{u_m} s_j \Leftrightarrow U_{k,m}(\mathbf{p}^*_\mathbf{k}) > U_{j,m}(\mathbf{p}^*_\mathbf{j}). \tag{25}$$

Next, the preference profile of s_k is represented by an ordered list of preferences on the set $\mathcal{U}$, which is established by sorting every $u_m \in \mathcal{U}$ in descending order based on the

Algorithm 2 Preference Establishment Algorithm

Input: $\mathcal{S}, \mathcal{U}$.
Output: $\mathcal{P}$.
for $s_k \in \mathcal{S}$ **do**
  for $u_m \in \mathcal{U}$ **do**
    calculate mutual preferences of the SBS-UE pair (s_k, u_m) using Algorithm 1.
  end for
end for
for $s_k \in \mathcal{S}$ **do**
  establish $P(s_k)$ by sorting each $u_m \in \mathcal{U}$ in descending order based on $q^*_{k,m}$.
end for
for $u_m \in \mathcal{U}$ **do**
  establish $P(u_m)$ by sorting each $s_k \in \mathcal{S}$ in descending order based on $U_{k,m}$.
end for

Algorithm 3 Energy-Efficient Stable Matching Algorithm

Input: $\mathcal{S}, \mathcal{U}, \mathcal{P}$.
Output: an energy-efficient context-aware matching μ.
Initialize: $\mu = \phi$, $\Phi = \mathcal{S}$.
while $\Phi \neq \phi$ **do**
  for $s_k \in \Phi$ **do**
    s_k proposes to the most preferred UE among those who have not yet rejected it in $P(s_k)$.
  end for
  for $u_m \in \mathcal{U}$ **do**
    if u_m receives a proposal from s_k, and prefers s_k to its currently engaged s_j, i.e., $s_k \succ_{u_m} s_j$ **then**
      s_k is kept engaged, while s_j is rejected, i.e., $\mu(u_m) = s_k$;
      add s_j into Φ, and remove s_k from Φ.
    else
      s_j is continually kept engaged, while s_k is rejected, i.e., $\mu(u_m) = s_j$.
    end if
  end for
end while

criteria of maximum achievable EE $q^*_{k,m}$, e.g., $P(s_k) = \{\cdots, u_m, u_{m'} \cdots\}$. In a similar way, u_m's preference is represented by an ordered list of preferences on the set $\mathcal{S}$, e.g., $P_{u_m} = \{\cdots, s_k, s_j, \cdots\}$. The set of preference lists is denoted as $\mathcal{P} = \{P(s_1), \cdots, P(s_K), P(u_1), \cdots, P(u_M)\}$.

B. THE ENERGY-EFFICIENT CONTEXT-AWARE MATCHING ALGORITHM

After establishing preference profiles for each $s_k \in \mathcal{S}$ and $u_m \in \mathcal{U}$, a one-to-one matching between SBSs and UEs is produced based on the GS algorithm [23]. The proposed energy-efficient context-aware matching algorithm is summarized in Algorithm 3. To start, each SBS proposes to its most favorite UE, who is ranked as the first choice on its preference list. After receiving the proposal, each UE rejects the SBS if it already holds a better proposal. Any SBS who is not rejected at this point is kept "engaged". In the next step, any SBS who was rejected previously proposes to its next choice who is the most preferred UE among those who have not yet rejected it. If a SBS finds that it has been rejected by all of the UEs whom it has already proposed to, then it issues no further proposals, and enters into the sleeping mode for energy saving. Each UE receiving proposals rejects all but its most preferred among the group consisting of the new proposals together with any SBS it may have kept engaged from previous steps. The process continues until no SBS is rejected, when every SBS is either matched to some UE or has been rejected by every UE on its preference list. At the end of this process, we will have a stable matching between SBSs and UEs. The algorithm has a nature of *deferred acceptance* since UEs are able to keep the best available SBS at any step engaged without accepting it outright.

In the case of *preference indifference*, i.e., some SBS or UE is indifferent between two or more possible matching partners, Algorithm 3 can be extended to handle this problem by introducing some fixed *tie-breaking* rule. For example, at any step of the algorithm when s_k must indicate a choice between u_m and $u_{m'}$ whom are equally well liked, i.e., $q^*_{k,m} = q^*_{k,m'}$, the tie-breaking rule proceeds as if the preferences are according to the order of signal to noise plus interference ratio (SINR), or occupancy time of the backhaul link, i.e., $\min\{T - T^{max}_{k,m}, T - T^{max}_{j,m}\}$. Such a tie-breaking rule can be used to specify which UE a SBS will propose to when it is indifferent about its next proposal, or to specify which SBS a UE will keep engaged when it is indifferent between two or more SBSs.

In the case of *incomplete preference lists*, SBSs on one side of the matching market only have partial knowledge about UEs on the other side. Taking an example, in a large-scale network, it is sometimes infeasible for a SBS to obtain the complete knowledge of every UE due to scalability issues. If the knowledge of u_m is not available to s_k, then u_m will not appear on s_k's preference list $P(s_k)$. We assume that $P(s_k)$ and $P(u_m)$ are *consistent* for any $s_k \in \mathcal{S}$ and $u_m \in \mathcal{U}$, which represents that if deleting u_m from s_k's preference list $P(s_k)$ implies that s_k is also deleted from u_m's preference list $P(u_m)$ [25]. Algorithm 3 can be easily modified to handle the matching problem with incomplete preference list by introducing some *preference deletion* rule. For example, SBSs and UEs that cannot be involved in the matching process should be deleted from each other's preference list. If the SBS-UE pair (s_k, u_m) is deleted, it entails deleting s_k and u_m from $P(u_m)$ and $P(s_k)$, respectively. Then the algorithm can proceed as in the case of complete lists to produce an matching, which can be obtained in polynomial time.

C. DISCUSSIONS

In this subsection, the stability, optimality, implementation issues and algorithmic complexity of the proposed algorithm are discussed and analyzed in detail.

1) STABILITY

Before proving that a matching is stable, we firstly introduce the concepts of *blocking pair*. In an instance of the energy-efficient context-aware matching problem, we assume that there exists a SBS s_k and a UE u_m who are not matched to one another at μ, i.e., $\mu(s_k) \neq u_m$ and $\mu(u_m) \neq s_k$. If s_k and u_m prefer each other more than their current assignments at μ, i.e., $u_m \succ_{s_k} \mu(s_k)$ and $s_k \succ_{u_m} \mu(u_m)$, we say that (s_k, u_m) is the blocking pair who *blocks* the matching μ. Thus, μ is unstable since both s_k and u_m are eager to disrupt μ and to be matched with each other. The *stability* of a matching is defined as follows [24]:

Definition 4: A matching μ is stable if it is not blocked by any individual SBS-UE pair.

In order to show that the proposed energy-efficient matching is stable, we need to prove that any SBS-UE pair cannot improve its EE by disrupting the partner selection and power allocation decisions produced by μ.

Theorem 2: The matching μ produced by Algorithm 3 is stable.

Proof: The proof of ***Theorem 2*** is given in Appendix A. ■

2) OPTIMALITY

Regarding optimality, we derive the following theorems by exploiting properties of nonlinear fractional programming and matching theory.

Theorem 3: For every $s_k \in \mathcal{S}$, $q_{k,m}$ produced by Algorithm 1 in each iteration converges to the unique optimum EE $q^*_{k,m}$.

Proof: The proof of ***Theorem 3*** is given in Appendix B. ■

Theorem 4: The matching μ produced by Algorithm 3 is weak Pareto optimal to SBSs.

Proof: The proof of ***Theorem 4*** is given in Appendix C. ■

Taking s_k and u_m as an example, ***Theorem 2*** shows that if $\mu(u_m) = s_k$, then there is no such $s_j \in \mathcal{S}\backslash\{s_k\}$ that satisfies $s_j \succ_{u_m} \mu(u_m)$ and $u_m \succ_{s_j} \mu(s_j)$. ***Theorem 3*** ensures that the achieved EE performance under $\mu(u_m) = s_k$ is the optimum one under QoS and transmission power constraints. ***Theorem 4*** shows that there is no other matching, stable or not, that all UEs prefer to μ.

3) IMPLEMENTATION

Although the proposed algorithm is implemented in a distributed fashion, it can also be implemented in a centralized way by exploiting the powerful MBSs. In the centralized implementation, the MBS can serve as a matchmaker to perform a matching between SBSs and UEs under established preferences. Implementation procedures are explained below.

First of all, since the MBS does not know preference profiles of the two sides, it intends to build these profiles by asking each SBS and UE for necessary information, such as $\mathcal{F}_m$, $g_{k,m}$, $I_{k,m}$, $\mathcal{C}_k$, $\tilde{T}^{cache}_{k,m}$, p^{cir}_k, p^{max}_k, p^B_k, and $U^{min}_{k,m}$. After collecting enough information, the MBS establishes the preference profile $P(s_k)$ for each $s_k \in \mathcal{S}$ and $P(u_m)$ for each $u_m \in \mathcal{U}$ by using Algorithm 2. Finally, the MBS will employ Algorithm 3 to produce a stable matching by using the established preference profiles.

The centralized implementation is also suitable for future cloud-based architectures of cellular networks such as cloud radio access network (C-RAN) [38], [39]. In C-RAN, remote radio heads (RRHs) with edge storage capabilities are densely deployed and are managed by a centralized base baseband unit (BBU) pool to cooperatively support UEs. Information exchange and RRU-UE matching can be performed in the centralized BBU pool to provide flexible control through centralized network coordination, which further facilitates the implementation of the proposed algorithm.

4) COMPLEXITY

Algorithm 3 must eventually stop after a finite number of iterations because the number of SBSs is limited, and no SBS proposes more than once to any UE. For a SBS-UE pair such as (s_k, u_m), the complexity to establish preference profiles is mainly dominated by Algorithm 1. The algorithmic complexity of Algorithm 1 is in the order of $\mathcal{O}(L^{max}_{loop} L^{max}_{dual})$, where L^{max}_{loop} and L^{max}_{dual} are the maximum numbers of iterations required for reaching convergence and solving dual problems, respectively. In Algorithm 2, with K SBSs and M UEs, sorting algorithms that sort preferences in descending order have a known complexity of $\mathcal{O}\big(KM \log(KM)\big)$. The algorithmic complexity of Algorithm 3 is in polynomial time $\mathcal{O}(KM)$ [25].

IV. SIMULATION RESULTS

In this section, the proposed algorithm is compared with conventional context-unaware water-filling algorithms in which UE association is based on maximum SINR [35], [40]. The values of simulation parameters are based on [4], [6], [9], [11], and [14], and are summarized in Table 1. We consider a single macro cellular network with a cell radius of 500 m. The results are averaged over a total number

TABLE 1. Simulation parameters.

Parameter	Value
Cell radius	500 m
Pathloss exponent α	4
Pathloss constant ϖ	10^{-2}
Shadowing $\zeta_{k,m}$ (standard deviation of a log-normal distribution)	8 dB
Multi-path fading $\beta_{k,m}$ (the mean of an exponential distribution)	1
Maximum transmission power p^{max}_k	33 dBm
SBS circuit power p^{cir}_k	41.6 dBm
UE constant circuit power p^{cir}_m	20 dBm
Noise power N_0	-114 dBm
Number of SBSs K	$100 \sim 200$
Number of UEs M	$100 \sim 200$
PA efficiency η	35%
Bandwidth per SBS w_k	2 MHz
Total number of files $\| \mathcal{C} \|$	1.5×10^9
Number of files requested per UE $\| \mathcal{F}_m \|$	1.5×10^3
File size s	2 KB
Storage capacity per SBS D_k	$0 \sim 5$ TB
Backhaul capacity per SBS B_k	$0.5 \sim 10$ Mbps
Zipf parameter	0.4
QoS requirement U^{min}_k,	$0.5\times \| \mathcal{F}_m \|$

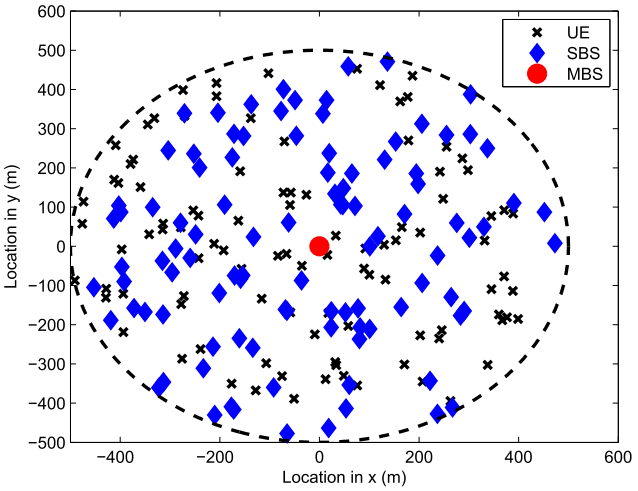


FIGURE 2. The locations of K SBSs and M UEs generated in one simulation ($K = M = 100$, the cell radius is 500 m).

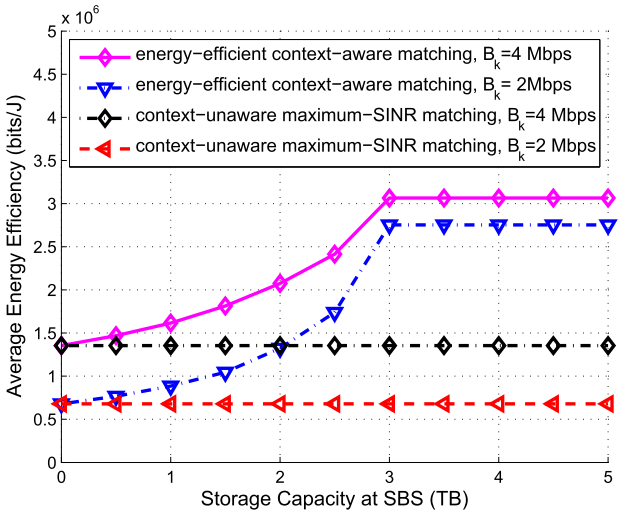


FIGURE 3. Average energy efficiency performance versus storage capacity per SBS ($K = M = 100$, $B_k = 2, 4$ Mbps $D_k = 0 \sim 5$ TB).

of 500 simulations. For each simulation, the locations of K SBSs and M UEs are generated randomly as shown in Fig. 2. The total bandwidth is 100 MHz and the assigned bandwidth per SBS is $w_k = 2$ MHz. Each UE $u_m \in \mathcal{U}$ requests $\|\mathcal{F}_m\| = 1.5 \times 10^3$ files, out of a set of $\|\mathcal{C}\| = 1.5 \times 10^9$ files. The files in $\mathcal{C}$ have the same size of $s = 2$ KB. Each SBS $s_k \in \mathcal{S}$ has a storage capacity chosen from an interval $D_k = [0, 5]$ TB, and a backhaul capacity chosen from an interval $B_k = [0.5, 10]$ Mbps. A training phase with a duration of 600 seconds is considered prior to performance evaluations, in which each s_k downloaded a set of popular files $\mathcal{C}_k$ through its backhaul link based on file popularity.

The proposed algorithm is verified from the aspects of EE performance and satisfaction. We evaluate satisfactions of SBSs based on cumulative distribution functions (CDFs) of their matched UEs. Taking s_k as an example, we define s_k's satisfaction threshold as $u_{m'}$, which is assumed to be ranked as the j-th choice on s_k's preference list $P(s_k)$. This threshold is the criteria used to evaluate s_k's satisfaction, that is, s_k is satisfied with the matching μ if it is matched to some UE that is preferred by s_k at least as well as $u_{m'}$, i.e., $\mu(s_k) \succeq_{s_k} u_{m'}$. Otherwise, s_k is unsatisfied with μ if it prefers $u_{m'}$ to its matched partner, i.e., $u_{m'} \succ_{s_k} \mu(s_k)$. The CDF of the satisfaction is defined as $\Pr\{\mu(s_k) \succeq_{s_k} u_{m'}\}$, which describes the probability that $\mu(s_k)$ will be found to have a higher ranking than the satisfaction threshold $u_{m'}$.

Fig. 3 shows the average EE performance as a function of the storage capacity D_k in a network with $K = M = 100$ SBSs and UEs, for different backhaul capacities $B_k = 2, 4$ Mbps. Simulation results show that the proposed algorithm is mostly beneficial during a storage-capacity limited regime (i.e., $D_k \leq 3$ TB). In this regime, the proposed approach yields a performance gain that increases exponentially with storage capacity since higher capacity allows SBSs to cache more popular files and increase the probability of having requested files cached closer to UEs. Note that, locally cached files can be transmitted at data rates much higher than the backhaul capacity, which is able to substantially reduce transmission duration and corresponding circuit energy consumption. Comparing to the context-unaware maximum-SINR matching algorithm, the proposed approach achieves maximum performance gains of 305% and 126% for $B_k = 2$ Mbps and $B_k = 4$ Mbps, respectively. The proposed algorithm dramatically improves EE performance by exploiting local content availability notably in backhaul-capacity limited scenarios. Finally, the gains achieved from caching saturate when storage capacity is already large enough to cache all of the files available in the network (i.e., $D_k > 3$ TB).

Fig. 4 shows the average EE performance as a function of the backhaul capacity B_k in a network with $K = M = 100$ SBSs and UEs, for different storage capacities $D_k = 1, 2, 3$ TB. Simulation results show that the maximum EE performance gains for $D_k = 1, 2$ and 3 TB are 133%, 370%, and 928%, respectively. The performance gains reach the

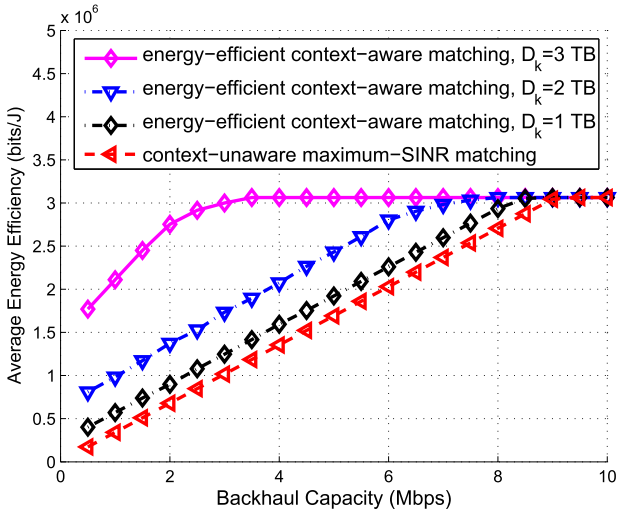


FIGURE 4. Average energy efficiency performance versus backhaul capacity per SBS ($K = M = 100$, $B_k = 0.5 \sim 10$ Mbps $D_k = 1, 2, 3$ TB).

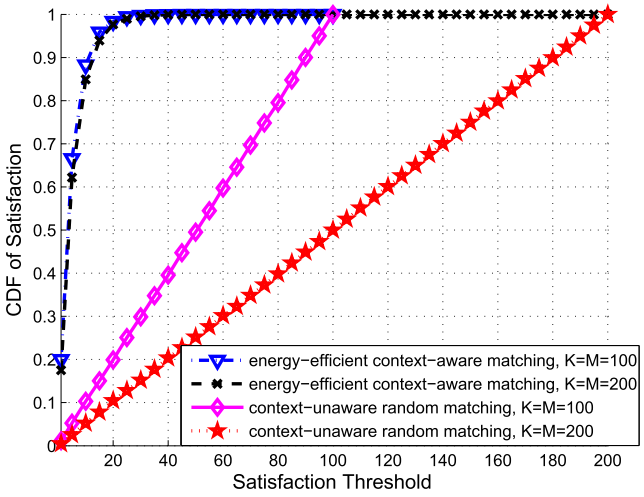


FIGURE 5. CDF of SBSs' satisfactions versus satisfaction threshold ($N = K = 100, 200$, $B_k = 2$ Mbps, $D_k = 2$ TB, 10^4 simulations).

maximum value when $B_k = 0.5$ Mbps, and decrease monotonically as the backhaul capacity increases. For example, the maximum performance gain with $D_k = 2$ TB is reduced from 370% to 26% when B_k is increased form 0.5 to 7 Mbps. When the backhaul capacity is large enough, the performance gains brought by caching reduce to zero because backhaul capacity no longer represents a bottleneck to data delivery.

Fig. 5 compares SBSs' satisfactions of the proposed approach with context-unaware random matching under $B_k = 2$ Mbps, $D_k = 2$ TB, $K = M = 100,\ 200$. We have not used maximum SINR based matching for comparison because it has not taken SBSs' preferences into consideration at all. In Fig. 5, the CDF is obtained based on a Monte-Carlo approach by repeating simulations for 10^4 times. When $K = M = 100$, simulation results show that 50.87% of SBSs are matched with their first three choices. In comparison, only 2.71% of SBSs are matched with their first three choices under the random matching. As cell and UE densities increase (i.e., K and M are increased from 100 to 200), the probability that SBSs are matched with their first three choices is still as high as 44.96% under the proposed matching, while under the random matching the probability is only 1.52%. The proposed approach achieves a satisfaction performance that is an order of magnitude higher than the conventional approach. Furthermore, the proposed approach is able to exploit the spatial diversity of ultra-dense deployment and enhance satisfaction performance in a wide range of satisfaction thresholds.

V. CONCLUSION

In this paper, we investigated the energy-efficient context-aware resource allocation problem in caching-enabled ultra-dense small cells. We formulated the joint partner selection and power allocation problem as a one-to-one matching problem, and took both SBSs' and UEs' preferences into consideration. We developed an iterative algorithm to establish preference profiles by employing nonlinear fractional programming and Lagrange dual decomposition. An energy-efficient context-aware stable matching algorithm was proposed based on the GS algorithm, and was extended into the cases of preference indifference and incomplete preference lists. Stability, optimality, implementation issues and algorithmic complexity were discussed and analyzed. Simulation results show that the proposed algorithm is able to overcome the backhaul-capacity limitations, and yield significant gains in terms of EE performance and SBS satisfaction with respect to conventional non-cache based algorithms. In future works, we will deal with heterogeneous data traffic, and evaluate the proposed matching algorithm with some other performance metrics such as file delay distributions, etc.

APPENDIX A
PROOF OF THEOREM 2

Proof: Suppose $u_m \succ_{s_k} \mu(s_k)$, then s_k must have proposed to u_m before proposing to $\mu(s_k)$ according to Algorithm 3. Since $\mu(s_k) \neq u_m$ when the algorithm stops, s_k must have been rejected by u_m in favor of some $\mu(u_m)$, i.e., $\mu(u_m) \succ_{u_m} s_k$. Thus, u_m and s_k do not block the matching μ. Since the matching μ is not blocked by any $s_k \in \mathcal{S}$ and any $u_m \in \mathcal{U}$, it is stable. ■

APPENDIX B
PROOF OF THEOREM 3

Proof: Firstly, we prove that $q^*_{k,m}$ produced by Algorithm 1 is unique. According to ***Theorem 1***, we have

$$\max_{\mathbf{p_k}} U_{k,m}(\mathbf{p_k}) - q^*_{k,m}E_k(\mathbf{p_k}) = U_{k,m}(\mathbf{p}^*_\mathbf{k}) - q^*_{k,m}E_k(\mathbf{p}^*_\mathbf{k}) = 0. \tag{26}$$

Define $F(q_{k,m}) = U_{k,m}(\mathbf{p}^*_\mathbf{k}) - q_{k,m}E_k(\mathbf{p}^*_\mathbf{k})$, $F(q_{k,m})$ is only an function of $q_{k,m}$ given $\mathbf{p}^*_\mathbf{k}$ is fixed. Thus, we have $\lim_{q_{k,m}\to-\infty} F(q_{k,m}) = +\infty$, and $\lim_{q_{k,m}\to+\infty} F(q_{k,m}) = -\infty$. Since $F(q_{k,m})$ is monotonically decreasing as $q_{k,m}$ increases and continuous for $q_{k,m}$, $F(q_{k,m}) = 0$ must have a unique solution $q^*_{k,m}$.

Secondly, we prove that $q_{k,m}$ produced by Algorithm 1 converges to $q^*_{k,m}$. In order to prove the convergence, we need to show that $q_{k,m}$ obtained by solving the equivalent problem (18) increases in each iteration of Algorithm 1.

We denote $\hat{\mathbf{p}}_\mathbf{k}(n)$ as the optimum power allocation in the n-th iteration, and $q_{k,m}(n)$ and $q_{k,m}(n+1)$ as the EE in the n-th iteration and $(n+1)$-th iteration, respectively. We assume that $q_{k,m}(n) \neq q^*_{k,m}$, and $q_{k,m}(n+1) \neq q^*_{k,m}$. Otherwise, the iterative algorithm terminates due to the stopping criteria. After $\hat{\mathbf{p}}_\mathbf{k}(n)$ is obtained, $q_{k,m}(n+1)$ is updated as $q_{k,m}(n+1) = U_{k,m}\big(\hat{\mathbf{p}}_\mathbf{k}(n)\big)/E_k\big(\hat{\mathbf{p}}_k(n)\big)$. The optimization problem (18) in the n-th iteration can be rewritten as

$$\begin{aligned}
&\max_{\mathbf{p_k}(n)} U_{k,m}\big(\mathbf{p_k}(n)\big) - q_{k,m}(n)E_k(\mathbf{p_k}(n)) \\
&= U_{k,m}\big(\hat{\mathbf{p}}_\mathbf{k}(n)\big) - q_{k,m}(n)E_k\big(\hat{\mathbf{p}}_\mathbf{k}(n)\big) \\
&= q_{k,m}(n+1)E_k\big(\hat{\mathbf{p}}_\mathbf{k}(n)\big) - q_{k,m}(n)E_k\big(\hat{\mathbf{p}}_\mathbf{k}(n)\big) \\
&= E_k\big(\hat{\mathbf{p}}_\mathbf{k}(n)\big)\big(q_{k,m}(n+1) - q_{k,m}(n)\big) \overset{(a)}{>} 0. \\
&\overset{(b)}{\Longrightarrow} q_{k,m}(n+1) > q_{k,m}(n)
\end{aligned} \tag{27}$$

Defining $\tilde{\mathbf{p}}_{\mathbf{k}}(n)$ such that $q_{k,m}(n) = U_{k,m}\big(\tilde{\mathbf{p}}_{\mathbf{k}}(n)\big)/E_k\big(\tilde{\mathbf{p}}_{\mathbf{k}}(n)\big)$, (a) can be proved based on ***Theorem 1*** as

$$\max_{\mathbf{p}_{\mathbf{k}}(n)} U_{k,m}\big(\mathbf{p}_{\mathbf{k}}(n)\big) - q_{k,m}(n)E_k(\mathbf{p}_{\mathbf{k}}(n)) \geq U_{k,m}\big(\tilde{\mathbf{p}}_{\mathbf{k}}(n)\big) - q_{k,m}(n)E_k\big(\tilde{\mathbf{p}}_{\mathbf{k}}(n)\big) = 0. \quad (28)$$

(b) is available since if $E_k\big(\hat{\mathbf{p}}_{\mathbf{k}}(n)\big)\big(q_{k,m}(n+1) - q_{k,m}(n)\big) > 0$ and $E_k\big(\hat{\mathbf{p}}_{\mathbf{k}}(n)\big) > 0$, we must have $q_{k,m}(n+1) - q_{k,m}(n) > 0$. Therefore, $q_{k,m}$ is increased in each iteration and will eventually approaches $q^*_{k,m}$ as long as the number of iterations is large enough. ■

APPENDIX C
PROOF OF THEOREM 4

Proof: $\forall s_k \in \mathcal{S}$, if there exists a matching μ' such that $\mu'(s_k) \succ_{s_k} \mu(s_k)$, then μ' must match every $s_k \in \mathcal{S}$ to some UE who had rejected it under μ. Hence, all of these UEs, $\mu'(\mathcal{S})$, must have rejected some SBS under μ. However, according to Algorithm 3, any UE who gets a proposal in the last step of the algorithm has not rejected any SBS. Otherwise, the rejected SBS needs at least one further step to be matched, which contradicts with the assumption of the last step. Thus, the above assumptions do not hold, and no such μ' exists. ■

REFERENCES

[1] A. Osseiran *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[2] O. Bello and S. Zeadally, "Intelligent device-to-device communication in the Internet of Things," *IEEE Syst. J.*, to be published.

[3] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5G network capacity: Key elements and technologies," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 71–76, Mar. 2014.

[4] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[5] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[6] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *Proc. WiOpt*, Hammamet, Tunisia, May 2014, pp. 37–42.

[7] G. Xylomenos, X. Vasilakos, C. Tsilopoulos, V. A. Siris, and G. C. Polyzos, "Caching and mobility support in a publish-subscribe Internet architecture," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 52–58, Jul. 2012.

[8] T. Wang, L. Song, and Z. Han, "Dynamic femtocaching for mobile users," in *Proc. IEEE WCNC*, New Orleans, LA, USA, Mar. 2015, pp. 861–865.

[9] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, New York, NY, USA, Mar. 1999, pp. 126–134.

[10] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "GreenDelivery: Proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[11] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Match to cache: Joint user association and backhaul allocation in cache-aware small cell networks," in *Proc. IEEE ICC*, London, U.K., Jun. 2015, pp. 3082–3087.

[12] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[13] Z. Su and Q. Xu, "Content distribution over content centric mobile social networks in 5G," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 66–72, Jun. 2015.

[14] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "In-network caching and content placement in cooperative small cell networks," in *Proc. 5GU*, Levi, Finland, Nov. 2014, pp. 128–133.

[15] S. Nikolaou, R. V. Renesse, and N. Schiper, "Proactive cache placement on cooperative client caches for online social networks and applications," *IEEE Trans. Parallel Distrib. Syst.*, to be published.

[16] M. Gerami, M. Xiao, and M. Skoglund, "Partial repair for wireless caching networks with broadcast channels," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 145–148, Apr. 2015.

[17] K. Wang, Z. Chen, and H. Liu, "Push-based wireless converged networks for massive multimedia content delivery," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2894–2905, May 2014.

[18] G. Gao, W. Zhang, Y. Wen, Z. Wang, and W. Zhu, "Cost-efficient video transcoding in media cloud by leveraging user viewing pattern," *IEEE Trans. Multimedia*, to be published.

[19] Y. Bao, X. Wang, S. Zhou, and Z. Niu, "An energy-efficient client pre-caching scheme with wireless multicast for video-on-demand services," in *Proc. IEEE APCC*, Jeju Island, Korea, Oct. 2012, pp. 566–571.

[20] Y. Chen *et al.*, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.

[21] S. Bu, F. R. Yu, Y. Cai, and X. P. Liu, "When the smart grid meets energy-efficient communications: Green wireless cellular networks powered by the smart grid," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 3014–3024, Aug. 2012.

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

[23] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, no. 1, pp. 9–15, Jan. 1962.

[24] A. E. Roth and M. A. O. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[25] G. O'Malley, "Algorithmic aspects of stable matching problems," Ph.D. dissertation, Dept. Inf. Math. Sci., Univ. Glasgow, Glasgow, Scotland, 2007.

[26] A. M. El-Hajj, Z. Dawy, and W. Saad, "A stable matching game for joint uplink/downlink resource allocation in OFDMA wireless networks," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 5354–5359.

[27] X. Feng *et al.*, "Cooperative spectrum sharing in cognitive radio networks: A distributed matching approach," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2651–2664, Aug. 2014.

[28] Y. Gu, Y. Zhang, M. Pan, and Z. Han, "Student admission matching based content-cache allocation," in *Proc. IEEE WCNC*, New Orleans, LA, USA, Mar. 2015, pp. 2179–2184.

[29] Y. Gu, Y. Zhang, M. Pan, and Z. Han, "Cheating in matching of device to device pairs in cellular networks," in *Proc. IEEE GLOBECOM*, Austin, TX, USA, Dec. 2014, pp. 4910–4915.

[30] D. Niyato, P. Wang, H. Tan, W. Saad, and D. I. Kim, "Cooperation in delay-tolerant networks with wireless energy transfer: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 64, no. 8, pp. 3740–3754, Aug. 2015.

[31] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[33] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.

[34] S. H. Ali and V. C. M. Leung, "Dynamic frequency allocation in fractional frequency reused OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4286–4295, Aug. 2009.

[35] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

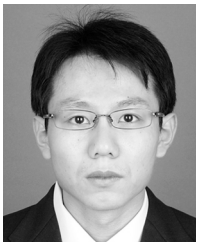
[36] H. Kwon and T. G. Birdsall, "Channel capacity in bits per joule," *IEEE J. Ocean. Eng.*, vol. OE-11, no. 1, pp. 97–99, Jan. 1986.

[37] K. T. K. Cheung, S. Yang, and L. Hanzo, "Achieving maximum energy-efficiency in multi-relay OFDMA cellular networks: A fractional programming approach," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 2746–2757, Jul. 2013.

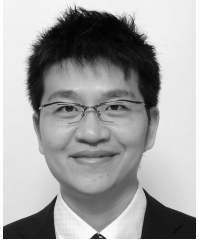
[38] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.

[39] J. Liu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "CONCERT: A cloud-based architecture for next-generation cellular systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 14–22, Dec. 2014.

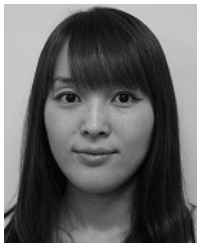
[40] G. Miao, N. Himayat, and G. Y. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Trans. Commun.*, vol. 58, no. 2, pp. 545–554, Feb. 2010.



ZHENYU ZHOU (S'06–M'11) received the M.E. and Ph.D. degrees from Waseda University, Tokyo, Japan, in 2008 and 2011, respectively. From 2012 to 2013, he was the Chief Researcher with the Department of Technology, KDDI, Tokyo. Since 2013, he has been an Associate Professor with the School of Electrical and Electronic Engineering, North China Electric Power University, China. He has been a Visiting Scholar with the Tsinghua–Hitachi Joint Laboratory on Environment Harmonious ICT, University of Tsinghua, Beijing, since 2014. His research interests include green communications and smart grid. He is currently a member of IEICE and CSEE. He received the Young Researcher Encouragement Award from the IEEE Vehicular Technology Society in 2009. He served as the Workshop Co-Chair of the IEEE ISADS 2015, the Session Chair of the IEEE Globecom 2014, and a TPC Member of the IEEE Globecom 2015, the ACM Mobimedia 2015, and the IEEE Africon 2015.



MIANXIONG DONG received the B.S., M.S., and Ph.D. degrees in computer science and engineering from The University of Aizu, Japan. He was a Researcher with the National Institute of Information and Communications Technology, Japan. He was a Japan Society for the Promotion of Sciences (JSPS) Research Fellow with the School of Computer Science and Engineering, The University of Aizu, and a Visiting Scholar with the BBCR Group, University of Waterloo, Canada, supported by the Excellent Young Researcher Overseas Visit Program from 2010 to 2011. He is currently an Assistant Professor with the Department of Information and Electronic Engineering, Muroran Institute of Technology, Japan. His research interests include wireless networks, big data, and cloud computing. He was selected as a Foreigner Research Fellow (a total of three recipients all over Japan) by the NEC C&C Foundation in 2011. He was the Best Paper Award Winner of the IEEE HPCC 2008, the IEEE ICESS 2008, ICA3PP 2014, and GPC 2015. He is an Associate Editor of the IEEE Access and *Cyber-Physical Systems* (Taylor & Francis), a Leading Guest Editor of *Peer-to-Peer Networking and Applications* (Springer), and a Guest Editor of *IEICE Transactions on Information and Systems*, *Mobile Information Systems*, and the *International Journal of Distributed Sensor Networks*. He is currently a Research Scientist with the A3 Foresight Program (2011–2016) funded by JSPS, the NSFC of China, and the NRF of Korea.



KAORU OTA received the M.S. degree in computer science from Oklahoma State University, USA, in 2008, and the Ph.D. degree in computer science and engineering from The University of Aizu, Japan, in 2012. From 2010 to 2011, she was a Visiting Scholar with the BBCR Group, University of Waterloo, Canada. She was also a Japan Society of the Promotion of Science (JSPS) Research Fellow with the Kato–Nishiyama Laboratory, Graduate School of Information Sciences, Tohoku University, Japan, from 2012 to 2013. She has been with the JSPS A3 Foresight Program as one of primary researchers since 2011, which is supported by the Japanese, Chinese, and Korean Government. She is currently an Assistant Professor with the Department of Information and Electronic Engineering, Muroran Institute of Technology, Japan. Her research interests include wireless sensor networks, vehicular ad hoc networks, and ubiquitous computing. She was a Guest Editor of the IEEE Wireless Communications and *IEICE Transactions on Information*, and serves as an Editor of *Peer-to-Peer Networking and Applications* (Springer), *Ad Hoc & Sensor Wireless Networks*, and the *International Journal of Embedded Systems* (Inderscience).

ZHENG CHANG received the B.Eng. degree from Jilin University, Changchun, China, in 2007, the M.Sc. (Tech.) degree from Aalto University, Espoo, Finland, in 2009, and the Ph.D. degree from the University of Jyväskylä, Jyväskylä, Finland, in 2013. In 2013, he was a Visiting Researcher with Tsinghua University, China. Since 2008, he has held various research positions with the Helsinki University of Technology, the University of Jyväskylä, and Magister Solutions Ltd., Finland. He is currently with the University of Jyväskylä. His research interests include signal processing, radio resource allocation, cross-layer optimizations for wireless networks, and green communications. He has been awarded by the Ulla Tuominen Foundation, the Nokia Foundation, and the Riitta and Jorma J. Takanen Foundation for his research work.