Energy-Efficient Radio Resource Management

Zhiqiang Wei, Yuanxin Cai, Derrick Wing Kwan Ng, and Jinhong Yuan

Abstract

Energy-efficient resource management is a promising solution to enable environment friendly and cost-effective wireless communication networks. This chapter presents the basic principle of the energy-efficient resource allocation design in wireless networks. Two types of energy efficiency (EE) definitions, system-centric EE and user-centric EE, are introduced, discussed, and analyzed. We reveal that when the circuit power consumption is not negligible, there is a non-trivial trade-off between the EE and spectral efficiency which should be taken into account for the optimal resource allocation algorithm design. In general, the system-centric EE maximization and the user-centric EE maximization can be classified as a single-ratio problem and a sum-of-ratios problem, respectively, and they can be solved via various iterative parametric algorithms. As an illustrative example, we have also presented the energy-efficient resource allocation design in an orthogonal frequency division multiple access (OFDMA) system. The design is formulated as a non-convex optimization problem. By exploiting the factional programming and dual decomposition, the globally optimal solution is obtained. Furthermore, our simulation results demonstrate the fast convergence and the superior EE achieved by our proposed resource allocation design. In addition, some future research extensions for realizing energy-efficient wireless communication networks are identified and discussed.

Index Terms

Energy efficiency, resource allocation, non-convex optimization, fractional programming

Zhiqiang Wei, Yuanxin Cai, Derrick Wing Kwan Ng, and Jinhong Yuan are with the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia (email: zhiqiang.wei@unsw.edu.au; yuanxin.cai@unsw.edu.au; w.k.ng@unsw.edu.au; j.yuan@unsw.edu.au).

I. Introduction

A. Background

Wireless communications have become one of the disruptive technologies in modern societies and offered one of the best business opportunities across industrial, public, and government sectors [1], [2]. A widely held view is that the fifth-generation (5G) is not just an evolutionary version of the current fourth-generation (4G) communication systems [1], due to not only the exponentially increasing demand of data traffic but also the energy-hungry emerging services and functionalities. To be more specific, three main envisioned usage scenarios have been proposed for 5G which is expected to revolutionize our future daily life:

- Enhanced Mobile Broadband (eMBB) [3]: high-resolution video streaming, virtual reality (VR), augmented reality (AR), etc.
- Massive Machine Type Communications (mMTC) [3]–[6]: Internet-of-Things (IoT) services, metering, monitoring, and measuring, smart agriculture, smart city, smart port, etc.
- Ultra-Reliable and Low Latency Communications (URLLC) [3]: vehicle-to-vehicle (V2V) & vehicle-to-everything (V2X) communications, autonomous driving, remote control surgery, etc.

These new services impose unprecedentedly challenges for the development of 5G wireless communication systems, such as the requirement of ultra-high data rates ($100 \sim 1000\times$ of current 4G technology), lower latency (1 ms for a roundtrip latency), massive connectivity (10^6 $\mathrm{devices/km^2}$), and the support of diverse quality of service (QoS) [1]. To fulfill the challenging requirements and to cope with the tremendous demand for wireless communications, particular for eMBB usage scenario, three fundamental paradigms have been proposed to boost the capacity for wireless networks:

- Spectral efficiency improvement: Enhance the system spectral efficiency via using the recently proposed advanced transceiver techniques, such as massive multiple-input multiple-output (MIMO) [7], [8] and non-orthogonal multiple access (NOMA) [9]–[15], etc.
- Spectrum extension: Extend the spectrum usage to a higher and wider unlicensed frequency band, such as millimeter wave (mmWave) bands [16]–[20] from 30 GHz to 300 GHz.
- Traffic offloading: Improve the area spectral efficiency via network densification, such

as small-cells [21], device-to-device (D2D) [22]–[24] communications, and unmanned aerial vehicle (UAV) communications [25]–[28], etc.

However, the advantages of these technologies for enhancing the system data rate do not come for free. They raise significant concerns of financial implications to the service providers due to the huge power consumption in wireless communication networks. For instance, in mmWave communication systems, the tremendous energy consumption associated with radio frequency (RF) chains, including analog-digital converters/digital-analog converters (ADC/DAC), power amplifiers, mixers, and local oscillators, etc. [29], constitutes a large part of the total system energy consumption. It is predicted that billions of information and communication technology (ICT) devices could create up to 3.5% of global emissions by 2020 and up to 14% by 2040 [30]. In 2025, it is expected that the communications industry will be responsible for 20% of all the worlds electricity [30]. The escalating energy costs and the associated global carbon dioxide (CO_2) emission of information and communication technology (ICT) devices have stimulated the interest of researchers in an emerging area of energy-efficient radio management. In particular, the Global e-Sustainability Initiative (GeSI) shows that the ICT sector's emissions "footprint" is expected to decrease to 1.97% of global emission by 2030, compared to 2.3% in 2020 [31]. Also, the next generation mobile networks (NGMN) alliance also declared energy saving as a top priority. To this end, studying energy-efficient radio management via exploiting limited system resources is critical to strike a balance between system energy consumption and throughput [32], which is also the main focus of this article.

Organization. The rest of the chapter is organized as follows. In the remaining part of Section I, we discuss some basic concepts in energy-efficient resource management, including the EE performance metrics and the inherent trade-off between the EE and spectral efficiency. Also, two iterative parametric algorithms to maximize the system-centric EE and to maximize the user-centric EE are presented in Section I-D, respectively. In Section II, we formulate the energy-efficient resource allocation design in an orthogonal division multiple access system as a non-convex optimization problem and solve the problem via the fractional programming and the dual decomposition methods. Section III draws a conclusion of this chapter and discusses some promising research extensions for energy-efficient resource management in future wireless communication networks.

Notation. Notations used in this chapter are as follows. Boldface capital and lower case letters are reserved for matrices and vectors, respectively. $\mathbb{C}^{M\times N}$ denotes the set of all $M \times N$

matrices with complex entries; $\mathbb{R}^{M\times N}$ denotes the set of all $M\times N$ matrices with real entries; $\mathbb{B}^{M\times N}$ denotes the set of all $M\times N$ matrices with binary entries. $(\cdot)^{\mathrm{T}}$ denotes the transpose of a vector or a matrix and $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose of a vector or a matrix; $|\cdot|$ denotes the absolute value of a complex scalar; $\|\cdot\|$ denotes the Euclidean vector norm; $[\cdot]^{-1}$ denotes the inverse of a matrix; and $[\cdot]^{+}=\max(\cdot,0)$. The circularly symmetric complex Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathcal{CN}(\mu,\sigma^2)$.

B. Energy Efficiency Metrics

The system energy efficiency (EE) has emerged as a new prominent and fundamental figure of merit for resource management in current system designs when energy consumptions and related problems become a major issue. In this article, two representative definitions of system EE, i.e., system-centric EE and user-centric EE, are given to measure the trade-off between energy consumption and throughput from different perspectives. In general, they are both essentially in the form of a benefit-cost ratio to evaluate the amount of data delivered by utilizing the limited system energy resource (bits/Joule). In particular, the system-centric EE is motivated by the desire of reducing the total operating cost, i.e., the energy consumed for each bit of information delivered, of wireless communication systems. In contrast, the user-centric EE is defined to balance the energy efficiencies among users and is beneficial to prolong the lifetime of user equipment. Before embarking the EE definitions in this chapter, we need to note that there are alternative types of EE definitions such as from facility level, equipment level, and network level, respectively [33], depending on the design of particular systems. Interest readers are referred to Table I in [33] for more details. However, in this article, we adopt the two most representative and commonly used EE definitions in the literature for illustrating the resource management for achieving energy-efficient communication systems.

Definition 1: **System-centric EE** [8], [34], [35]

The system-centric EE (bit/Joule) is given as:

$$\mathrm{EE}_{\mathrm{Sys}}=\frac{\text{System throughput}}{\text{Total system power consumption}}=\frac{\sum\limits_{k=1}^{K}W_k\log_2\left(1+\gamma_k\right)}{\sum\limits_{k=1}^{K}\left(\delta p_k+P_{\mathrm{C},k}\right)}, \tag{1}$$

where K denotes the total number of users and k is the user index. The occupied bandwidth of user k is W_k and the transmit power for user k is p_k. The received signal-to-noise ratio (SNR) or signal-to-interference-plus-noise (SINR) at user k is denoted as γ_k, depending on whether there exists co-channel interference to user k. In addition, $P_{\mathrm{C},k}$ denotes the static

circuit power consumption associated with user k and $\delta > 1$ captures the inefficiency of the transmit power amplifier[1].

It is worth to note that the running index of user k in (1) can be interpreted as different physical entities depending on the types of the considered communication systems. For example, the indices can be treated as the subcarrier indices in a multi-carrier communication system. It can be observed that the $\mathrm{EE}_{\mathrm{Sys}}$ is the ratio between the total amount of data rate (bit/s) produced by per Watt of consumed power[2]. In addition, maximizing the system-centric EE can guarantee the most efficient utilization of system energy resources for communication. However, it does not take into account the diverse features or EE requirements of different user equipment. In particular, it may result in imbalance resource allocation since good users may dominate the system performance. Therefore, the user-centric EE should be introduced to balance the EE among users.

Definition 2: **User-centric EE** [36]

The user-centric EE (bit/Joule) is defined as:

$$\mathrm{EE}_{\mathrm{User}} = \sum_{k=1}^{K} \omega_k \frac{W_k \log_2\left(1+\gamma_k\right)}{\delta p_k + P_{\mathrm{C},k}}, \tag{2}$$

where $\omega_k > 0$ denotes the weight for EE of user k with $\sum_{k=1}^{K} \omega_k = 1$. Compared to the system-centric EE in (1), the user-centric EE metric can assign different priorities to balance the EE of different users through the selection of weights[3]. Also, the user-centric EE is in favour of heterogeneous features and specifications on EE of diverse user terminals. Comparing to (1), we can observe that the system-centric EE is a single-ratio function, while the user-centric EE metric has a sum-ratio form. These major differences yield different solution structures which will be detailed in Section I-D.

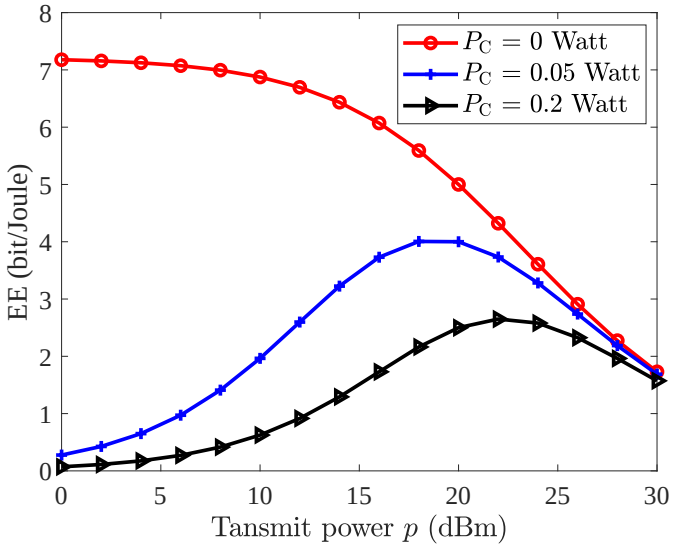
C. The Trade-off Between Energy Efficiency and Spectral Efficiency

In this section, we provide some discussions on the fundamental trade-off between EE and spectral efficiency (SE). It can be observed that maximizing the system-centric EE in

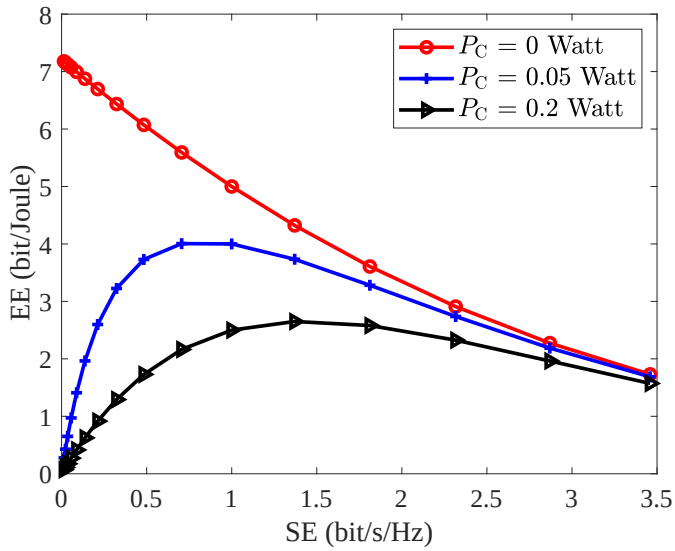
[1]Here, we assume that the power amplifiers equipped at the BS operate in its linear region and the hardware power consumption associated with user k, $P_{\mathrm{C},k}$, is a constant.

[2]Since we consider the system performance in a unit of time duration, the time duration factors in both denominator and numerator of (1) have canceled each other.

[3]Note that we can have a weighted system throughput to balance the spectral efficiencies of different users rather than their energy efficiencies.



(a) EE (bit/Joule) versus transmit power p (dBm).



(b) EE (bit/Joule) versus SE (bit/s/Hz).

Fig. 1. An illustration of the trade-off between EE, transmit power, and SE. The simulation setups are $W = 1$ Hz, $\frac{|h|^2}{N_0} = 10dB$, $P_{\mathrm{C}} = [0, 0.05, 0.2]$ Watt, and $\delta = 2$.

(1) is equivalent to maximizing the user-centric EE in (2) when $K = 1$. In the following, for notation simplicity, we will drop the subscript k and focus on the EE with $K = 1$ when co-channel interference does not exist, i.e.,

$$\mathrm{EE} = \frac{W \cdot \mathrm{SE}}{\delta p + P_{\mathrm{C}}}, \tag{3}$$

where $\mathrm{SE} = \log_2\left(1+\frac{p|h|^2}{WN_0}\right)$ denotes the spectral efficiency (bit/s/Hz), $|h|^2$ denotes the channel gain and N_0 is the noise power spectral density. It is clear that the SE monotonically increases with an increased transmit power p, but with a diminishing return due to the logarithmic nature of the achievable rate function. On the other hand, the denominator of (3) is a linear function of p. As a result, there is a non-trivial trade-off between the EE and SE which should be taken into account for resource allocation algorithm design.

Fig. 1 illustrates the trade-off between EE, transmit power and SE. When the total circuit power consumption is negligibly small, i.e., $P_{\mathrm{C}} = 0$ Watt, the EE is a monotonically decreasing function of both the transmit power and the SE, as illustrated in Fig. 1(a) and Fig. 1(b), respectively. In other words, transmission with an arbitrarily low power, i.e., $p \to 0$, is the optimal operation point for maximizing the system EE and the resulting system EE is $\lim\limits_{p\to 0, P_{\mathrm{C}}=0} = \frac{|h|^2}{\delta N_0}$. In addition, we can observe that when $P_{\mathrm{C}} > 0$, the system EE first increases with increasing the transmit power and then decreases with increasing the transmit power. In fact, in the low SNR regime, the EE is mainly limited by the fixed circuit power consumption P_{C} and SE scales almost linearly with respect to (w.r.t.) the transmit power p. Hence, increasing the transmit power can effectively increase both the SE and the EE. On the other hand, in the high SNR regime, the transmit power, p, dominates the total power consumption and there is an only marginal gain in SE when increasing the transmit power. As a result, after reaching the maximum system EE, as shown in Fig. 1(a), further increasing the transmit power decreases the system EE. Furthermore, in Fig. 1(a), we can observe that with increasing the circuit power consumption, the optimal operation point is pushed towards the high SNR regime. It is due to the fact that the larger the circuit power consumption, the higher transmit power is needed to outweigh the impact of the circuit power consumption on the EE. As a result, for a practical case of $P_{\mathrm{C}} > 0$, there is always a non-trivial trade-off between EE and SE. Hence, finding the optimal operation point to maximize the system EE has attracted significant attention in the literature in the past few years [8], [34], [35].

D. Energy-efficient Resource Allocation

Energy-efficient resource allocation is the concept of making the best use of limited communication resources based on the information available at the resource allocator to improve the system performance. In general, the system resources are the transmit power, the available bandwidth and time, as well as the available space if multiple antennas are employed at transmitters. The available information at the resource allocator usually includes

channel state information (CSI) and QoS requirements requested by the users. In particular, the CSI can be obtained from user feedback in frequency division duplex (FDD) systems or from uplink channel estimation in time division duplex (TDD) systems. Besides, the QoS requirements, such as the minimum data rate requirement and outage probability requirement, act as constraints in resource allocation optimization framework. To elaborate a bit further, energy-efficient resource allocation design relies on the application of the optimization theory to maximize the EE taking into account certain QoS constraints.

System-centric energy-efficient resource allocation design

The system-centric energy-efficient resource allocation design can be formulated as the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{x}}{\text{maximize}} \;\; \mathrm{EE}_{\mathrm{Sys}}\left(\mathbf{x}\right) = \frac{f\left(\mathbf{x}\right)}{g\left(\mathbf{x}\right)} \\ &\text{s.t.} \;\; \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{4}$$

where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{C}^n$ is the optimization variables and $\mathcal{X} \subseteq \mathbb{C}^n$ is the feasible solution set which is usually spanned by the system resource limitations and the QoS constraints. The numerator $f\left(\mathbf{x}\right) : \mathcal{X} \subseteq \mathbb{C}^n \rightarrow \mathbb{R}$ denotes the system data-rate produced by the resource allocation strategy $\mathbf{x}$. The denominator $g\left(\mathbf{x}\right) : \mathcal{X} \subseteq \mathbb{C}^n \rightarrow \mathbb{R}$ denotes the total system power consumption.

The formulated problems in (4) can be classified as factional programming [37]. Without loss of generality, we define the maximum EE of the problem in (4) as follows:

$$\mathrm{EE}^{*}_{\mathrm{Sys}} = \underset{\mathbf{x}}{\text{maximize}} \;\; \frac{f\left(\mathbf{x}\right)}{g\left(\mathbf{x}\right)}, \;\text{s.t.} \;\; \mathbf{x} \in \mathcal{X}. \tag{5}$$

Now, the following theorem can transform the fractional objective function in (4) to an equivalently subtractive form.

Theorem 1: The maximum EE $\mathrm{EE}^{*}_{\mathrm{Sys}}$ is achieved if and only if

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{maximize}} \;\; f\left(\mathbf{x}\right) - \mathrm{EE}^{*}_{\mathrm{Sys}} g\left(\mathbf{x}\right) = f\left(\mathbf{x}^{*}\right) - \mathrm{EE}^{*}_{\mathrm{Sys}} g\left(\mathbf{x}^{*}\right) = 0, \tag{6}$$

for $f\left(\mathbf{x}\right) \geq 0$ and $g\left(\mathbf{x}\right) > 0$.

Proof: The proof can be found in Appendix A. ■

In the literature, the Dinkelbach method has been proposed to find $\mathrm{EE}^{\mathrm{iter}}_{\mathrm{Sys}}$ iteratively [37]. In particular, in each iteration of the main loop, one needs to solve (4) for a given temporary

Algorithm 1 Dinkelbach's Algorithm

Initialization

Initialize the convergence tolerance $\epsilon \to 0$, the maximum number of iterations $\mathrm{iter}_{\max}$, the iteration index $\mathrm{iter} = 1$ and the initial system EE $\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}} = 0$

repeat {**Main loop**}

  Solve (6) for the given temporary $\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}$ to obtain the resource allocation strategy $\mathbf{x}^{\mathrm{iter}}$

  if $f\left(\mathbf{x}^{\mathrm{iter}}\right) - \mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}} g\left(\mathbf{x}^{\mathrm{iter}}\right) < \epsilon$ **then**

    Convergence = **true**

    return $\mathbf{x}^* = \mathbf{x}^{\mathrm{iter}}$ and $\mathrm{EE}_{\mathrm{Sys}}^* = \frac{f\left(\mathbf{x}^{\mathrm{iter}}\right)}{g\left(\mathbf{x}^{\mathrm{iter}}\right)}$

  else

    Set $\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}+1} = \frac{f\left(\mathbf{x}^{\mathrm{iter}}\right)}{g\left(\mathbf{x}^{\mathrm{iter}}\right)}$ and $\mathrm{iter} = \mathrm{iter} + 1$

    Convergence = **false**

  end if

until Convergence = **true** or $\mathrm{iter} = \mathrm{iter}_{\max}$

$\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}$, as shown in **Algorithm** 1. The convergence of Dinkelbach's algorithm is stated in the following theorem.

Theorem 2: The Dinkelbach's algorithm convergences to the globally optimal solution of the problem in (4) if the problem in (6) can be solved optimally for a given $\mathrm{EE}_{\mathrm{Sys}}$.

Proof: The proof can be found in Appendix B. ■

We note that Theorem 2 only requires that the problem in (6) can be solved but it does not impose any assumptions on the convexity or concavity of the function $f(\mathbf{x})$, $g(\mathbf{x})$ and the feasible solution set $\mathcal{X}$. However, it is clear that the Dinkelbach's algorithm can be implemented with a lower computational complexity when $f(\mathbf{x})$ is a concave function, $g(\mathbf{x})$ is a convex function and $\mathcal{X}$ is a compact convex set. The solution for the user-centric energy-efficient resource allocation design is presented in the following.

User-centric energy-efficient resource allocation

The user-centric energy-efficient resource allocation design can be formulated as the following optimization problem:

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{k=1}^{K} \omega_k \frac{f_k(\mathbf{x})}{g_k(\mathbf{x})} \tag{7}$$
$$\text{s.t.} \quad \mathbf{x} \in \mathcal{X},$$

where $f_k : \mathcal{X} \subseteq \overbrace{\mathbb{C}^n \rightarrow R}$ denotes the data-rate of user k produced by the resource allocation strategy $\mathbf{x}$ and $g_k : \mathcal{X} \subseteq \mathbb{C}^n \rightarrow R$ denotes the power consumption associated with user k.

The objective function in (7) is a sum-of-ratios function which cannot be solved by the Dinkelbach's algorithm. Until very recently, the author in [38] proposed a parametric solution, such that the globally optimal solution can be successfully found through an iterative algorithm. The key idea is to transform the original sum-of-ratio function into a parametric subtractive form, which is stated in the following theorem.

Theorem 3: If $\{\mathbf{x}^*\}$ is the solution of the problem in (7), there exist two parameter vectors α_k^* and β_k^*, $k \in \{1, 2, \ldots, K\}$, such that $\mathbf{x}^*$ is a solution of the following convex optimization problem for given α_k^* and β_k^*:

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{maximize}} \ \sum_{k=1}^{K} \alpha_k^* \left(\omega_k f_k\left(\mathbf{x}\right) - \beta_k^* g_k\left(\mathbf{x}\right)\right), \tag{8}$$

for $f_k\left(\mathbf{x}\right) \geq 0$ and $g_k\left(\mathbf{x}\right) > 0$ and $\{\mathbf{x}^*, \alpha_k^*, \beta_k^*\}$ satisfy the following equations:

$$\omega_k f_k\left(\mathbf{x}\right) - \beta_k^* g_k\left(\mathbf{x}\right) = 0, \text{ and} \tag{9}$$

$$\alpha_k^* g_k\left(\mathbf{x}\right) - 1 = 0, \forall k. \tag{10}$$

Proof: The proof can be found in Appendix C. ■

Compared to Theorem 2, we can observe that to solve the sum-of-ratios problem in (7), more restrictive constraints are imposed on functions $f_k\left(\mathbf{x}\right)$, $g_k\left(\mathbf{x}\right)$ and the feasible solution set $\mathcal{X}$. In particular, we need to have (i) $f_k\left(\mathbf{x}\right)$ is a concave function w.r.t. $\mathbf{x}$; (ii) $g_k\left(\mathbf{x}\right)$ is a convex function w.r.t. $\mathbf{x}$; and (iii) $\mathcal{X}$ is a convex and compact set.

Now, similar to the Dinkelbach's algorithm, an iterative resource allocation algorithm can be employed for solving the problem in (8). The key challenge of solving the problem in (7) is to obtain an update for α_k^* and β_k^* in the problem in (8). For notational simplicity, we introduce the parameter $\boldsymbol{\rho} = [\rho_1, \rho_2, \ldots, \rho_{2K}] = [\alpha_1, \ldots \alpha_K, \beta_1, \ldots, \beta_K] = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and functions

$$\varphi_i\left(\rho_i\right) = \rho_i g_i\left(\mathbf{x}\right) - 1 \text{ and} \tag{11}$$

$$\varphi_{K+i}\left(\rho_{K+i}\right) = \rho_{K+i} g_{K+i}\left(\mathbf{x}\right) - \omega_{K+i} f_{K+i}\left(\mathbf{x}\right), \tag{12}$$

where $i = \{1, \ldots, K\}$. According to Theorem 3, the optimal solution $\boldsymbol{\rho} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ is achieved if and only if

$$\boldsymbol{\varphi}\left(\boldsymbol{\rho}\right) = [\varphi_1, \ldots, \varphi_{2K}] = \mathbf{0} \tag{13}$$

Algorithm 2 Iterative Resource Allocation Algorithm for Sum-of-ratios Problems

Initialization

Initialize the maximum convergence tolerance $\epsilon \to 0$, $(\xi, \tau) \in [0, 1]$, the maximum number of iterations $\mathrm{iter}_{\max}$, the iteration index $\mathrm{iter} = 1$ and the initial parameter $\boldsymbol{\rho}^{\mathrm{iter}} = (\boldsymbol{\alpha}^{\mathrm{iter}}, \boldsymbol{\beta}^{\mathrm{iter}})$.

repeat {**Main loop**}

  Solve (8) for the given $\boldsymbol{\rho}^{\mathrm{iter}}$ to obtain the resource allocation strategy $\mathbf{x}^{\mathrm{iter}}$

  if $\|\boldsymbol{\varphi}(\boldsymbol{\rho})\| < \epsilon$ **then**

    Convergence = **true**

    return $\mathbf{x}^* = \mathbf{x}^{\mathrm{iter}}$ and $\mathrm{EE}^*_{\mathrm{User}} = \sum_{k=1}^{K} \omega_k {\beta_k}^{\mathrm{iter}}$

  else

    Update $\boldsymbol{\rho}^{\mathrm{iter}}$ via (14) and set $\mathrm{iter} = \mathrm{iter} + 1$

    Convergence = **false**

  end if

until Convergence = **true** or $\mathrm{iter} = \mathrm{iter}_{\max}$

is satisfied. Following the method in [38], in the iter-th iteration, we update $\boldsymbol{\rho} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ by

$$\boldsymbol{\rho}^{\mathrm{iter}+1} = \boldsymbol{\rho}^{\mathrm{iter}} + \lambda_{\mathrm{iter}} \mathbf{q}^{\mathrm{iter}}, \tag{14}$$

where $\mathbf{q}^{\mathrm{iter}} = \left[\boldsymbol{\varphi}'(\boldsymbol{\rho})\right]^{-1} \boldsymbol{\varphi}(\boldsymbol{\rho})$ denotes the update direction. Here, $\boldsymbol{\varphi}'(\boldsymbol{\rho})$ is the Jacobian matrix of $\boldsymbol{\varphi}(\boldsymbol{\rho})$ w.r.t. $\boldsymbol{\rho}$. The updating step λ_{iter} is obtained by the largest ξ^l that satisfies:

$$\left\|\boldsymbol{\varphi}\left(\boldsymbol{\rho}^{\mathrm{iter}} + \xi^l \mathbf{q}^{\mathrm{iter}}\right)\right\| \leq \left(1 - \tau\xi^l\right) \left\|\boldsymbol{\varphi}\left(\boldsymbol{\rho}^{\mathrm{iter}}\right)\right\|, \tag{15}$$

where $l \in \{1, 2, \ldots\}$, $0 < \xi < 1$, $0 < \tau < 1$. The iterative resource allocation algorithm is presented in Algorithm 2. The convergence of Algorithm 2 actually follows the convergence analysis of the Newton method, which can be found in [38].

II. Illustrative Example: Energy-efficient Resource Allocation in an OFDMA System

In this section, we use an illustrative example to demonstrate the energy-efficient radio management to improve the system EE.

A. System Model

We consider a single-cell downlink orthogonal frequency division multiple access (OFDMA) network with a cell radius of D meters, as shown in Fig. 2. The BS is located at the

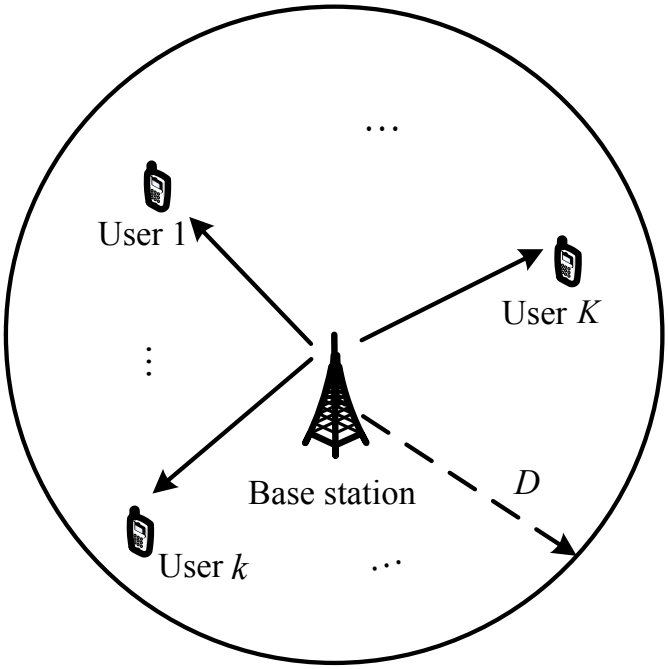


Fig. 2. A single-cell downlink OFDMA network with a base station (BS) and K users.

center of the cell serving K mobile users. The system bandwidth is divided equally into N_{F} subcarriers and $N_{\mathrm{F}} \geq K$. All transceivers are single-antenna devices. We assume that the channel state information (CSI) is perfectly known at the BS and the channel impulse response is time-invariant within each frame. Note that the energy consumption incurred by estimating CSI or CSI feedback is not included here since it is relatively insignificant compared to the energy consumed for payload transmission.

Let $p_{k,i}$ be the power allocation for user k on subcarrier $i \in \{1, \ldots, N_{\mathrm{F}}\}$. Binary variable $u_{k,i}$ denotes the user scheduling variable, which is $u_{k,i} = 1$ if subcarrier i is assigned to user k. Otherwise, $u_{k,i} = 0$. The transmitted signal from the BS to all the users on subcarrier i is given by

$$x_i = \sum_{k=1}^{K} u_{k,i} \sqrt{p_{k,i}} s_{k,i}, \tag{16}$$

where $s_{k,i} \in \mathbb{C}$ is the transmitted information symbol for user k on subcarrier i. The received signal from the BS to at user k on subcarrier i is given by

$$y_{k,i} = h_{k,i} \sum_{k=1}^{K} u_{k,i} \sqrt{p_{k,i}} s_{k,i} + z_{k,i}, \tag{17}$$

where $z_{k,i} \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) in subcarrier i at user k. The variable $h_{k,i} \in \mathbb{C}$ represents the channel coefficient between the BS and user k on

subcarrier i.

To avoid the inter-user interference, at most one user can be allocated on each subcarrier, i.e.,

$$\sum_{k=1}^{K} u_{k,i} \leq 1, \forall i. \tag{18}$$

Then, the achievable rate of user k on subcarrier i is obtained by

$$R_{k,i} = u_{k,i} \log_2 \left(1 + \frac{p_{k,i} \left|h_{k,i}\right|^2}{\sigma^2}\right). \tag{19}$$

For simplicity, in (19), we consider the subcarrier bandwidth is normalized, i.e., $W_i = 1$ Hz, $\forall i$. The individual data-rate of user k is given by

$$R_k = \sum_{i=1}^{N_{\mathrm{F}}} u_{k,i} \log_2 \left(1 + \frac{p_{k,i} \left|h_{k,i}\right|^2}{\sigma^2}\right) \tag{20}$$

and the system sum-rate is given by

$$R_{\mathrm{Sys}}\left(\mathbf{p}, \mathbf{u}\right) = \sum_{k=1}^{K} \sum_{i=1}^{N_{\mathrm{F}}} u_{k,i} \log_2 \left(1 + \frac{p_{k,i} \left|h_{k,i}\right|^2}{\sigma^2}\right), \tag{21}$$

where $\mathbf{p} \in \mathbb{R}^{KN_{\mathrm{F}} \times 1}$ and $\mathbf{u} \in \mathbb{B}^{KN_{\mathrm{F}} \times 1}$ collect all the power allocation variables $p_{k,i}$ and user scheduling variables $u_{k,i}$, respectively.

The total system power consumption mainly consists of the transmit power and circuit power. Therefore, we model the system power dissipation as follows:

$$U_{\mathrm{p}}\left(\mathbf{p}, \mathbf{u}\right) = \delta \sum_{k=1}^{K} \sum_{i=1}^{N_{\mathrm{F}}} u_{k,i} p_{k,i} + P_{\mathrm{C}}, \tag{22}$$

where P_{C} is the circuit power consumption of the BS. In addition, the total transmit power should be constrained by the limited dynamic range of the power amplifiers equipped at the BS, i.e.,

$$\sum_{k=1}^{K} \sum_{i=1}^{N_{\mathrm{F}}} u_{k,i} p_{k,i} \leq p_{\max}, \tag{23}$$

where $p_{\max}$ denotes the maximum input power of the power amplifier at the BS such that it operates in its linear operation region. Now, the system EE can be defined as follows:

$$\mathrm{EE}_{\mathrm{Sys}}\left(\mathbf{p}, \mathbf{u}\right) = \frac{R_{\mathrm{Sys}}\left(\mathbf{p}, \mathbf{u}\right)}{U_{\mathrm{p}}\left(\mathbf{p}, \mathbf{u}\right)}. \tag{24}$$

To maximize the system EE of the considered OFDMA system, we formulate the resource allocation design as an optimization problem and employ the fractional programming and the dual decomposition methods to solve the problem.

B. Energy-efficient Resource Allocation Design

The energy-efficient resource allocation of the considered OFDMA system can be formulated as the following optimization problem:

$$\begin{aligned} \underset{\mathbf{p},\mathbf{u}}{\text{maximize}} \;\; & \mathrm{EE}_{\mathrm{Sys}}\left(\mathbf{p},\mathbf{u}\right)=\frac{R_{\mathrm{Sys}}\left(\mathbf{p},\mathbf{u}\right)}{U_{\mathrm{p}}\left(\mathbf{p},\mathbf{u}\right)} \\ \text{s.t.} \;\; & \text{C1: } u_{k,i}\in\{0,1\},\forall k,i, \\ & \text{C2: } p_{k,i}\geq 0,\forall k,i, \\ & \text{C3: } \sum_{k=1}^{K}u_{k,i}\leq 1,\forall i, \\ & \text{C4: } \sum_{k=1}^{K}\sum_{i=1}^{N_{\mathrm{F}}}u_{k,i}p_{k,i}\leq p_{\max}, \\ & \text{C5: } R_k\geq R_{\min},\forall k, \end{aligned} \tag{25}$$

where C1 and C2 in (25) are the definitions of $u_{k,i}$ and $p_{k,i}$, respectively. Constraints C3 and C4 in (25) follow (18) and (23), respectively. Constraint C5 in (25) imposes a minimum data rate requirement for each user.

According to Theorem 1, the problem can be equivalently transformed into the following parametric optimization problem:

$$\begin{aligned} \underset{\mathbf{p},\mathbf{u}}{\text{maximize}} \;\; & R_{\mathrm{Sys}}\left(\mathbf{p},\mathbf{u}\right)-\mathrm{EE}_{\mathrm{Sys}}U_{\mathrm{p}}\left(\mathbf{p},\mathbf{u}\right) \\ \text{s.t.} \;\; & \text{C1-C5}, \end{aligned} \tag{26}$$

where the intermediate parameter $\mathrm{EE}_{\mathrm{Sys}}$ can be updated iteratively via the Dinkelbach's algorithm in **Algorithm** 1. However, the problem in (26) is a mixed combinatorial non-convex optimization problem. The combinatorial constraint C1 on the user scheduling variables creates a disjoint feasible solution set which is a hurdle for solving the problem via computationally efficient tools from convex optimization. Besides, the coupling between binary variables $u_{k,i}$ and continuous variables $p_{k,i}$ in the objective function and constraint in C5 yields a generally intractable problem.

C. Dual Decomposition Method for Solving the Main Loop Problem

In this section, we employ the dual decomposition method [8], [34], [35] to solve the main loop problem in (26) for a given $\mathrm{EE}_{\mathrm{Sys}}$. Firstly, to handle the binary constraint C1, we

relax the binary user scheduling variable $u_{k,i}$ to be a real between zero and one instead of a Boolbean, i.e.,

$$0 \leq u_{k,i} \leq 1, \forall k, i. \tag{27}$$

In fact, $u_{k,i}$ can be interpreted as a time-sharing factor in allocating subcarrier i to all the K users for information delivery. In the following, we prove that the optimal user scheduling variable is located on the boundary, i.e., $u_{k,i}^* = 0$ or $u_{k,i}^* = 1$. In other words, the time-sharing relaxation in (27) does not lose the optimality. In addition, we introduce the auxiliary time-shared power allocation variables $\widetilde{p}_{k,i} = u_{k,i} p_{k,i}$. Subsequently, we can rewrite the problem in (26) for a given parameter $\mathrm{EE}_{\mathrm{Sys}}$ as

$$\begin{aligned}
\underset{\mathbf{p},\mathbf{u}}{\text{maximize}} \;\; & \widetilde{R}_{\mathrm{Sys}}\left(\mathbf{p},\mathbf{u}\right) - \mathrm{EE}_{\mathrm{Sys}} \widetilde{U}_{\mathrm{p}}\left(\mathbf{p},\mathbf{u}\right) \\
\text{s.t.} \;\; & \text{C1: } 0 \leq u_{k,i} \leq 1, \forall k, i, \\
& \text{C2: } \widetilde{p}_{k,i} \geq 0, \forall k, i, \\
& \text{C3: } \sum_{k=1}^{K} u_{k,i} \leq 1, \forall i, \\
& \text{C4: } \sum_{k=1}^{K} \sum_{i=1}^{N_{\mathrm{F}}} \widetilde{p}_{k,i} \leq p_{\max}, \\
& \text{C5: } \widetilde{R}_k \geq R_{\min}, \forall k,
\end{aligned} \tag{28}$$

where $\widetilde{R}_{\mathrm{Sys}}\left(\mathbf{p},\mathbf{u}\right) = R_{\mathrm{Sys}}\left(\mathbf{p},\mathbf{u}\right)\Big|_{p_{k,i}=\frac{\widetilde{p}_{k,i}}{u_{k,i}}}$, $\widetilde{U}_{\mathrm{p}}\left(\mathbf{p},\mathbf{u}\right) = U_{\mathrm{p}}\left(\mathbf{p},\mathbf{u}\right)\Big|_{p_{k,i}=\frac{\widetilde{p}_{k,i}}{u_{k,i}}}$ and $\widetilde{R}_k = R_k\Big|_{p_{k,i}=\frac{\widetilde{p}_{k,i}}{u_{k,i}}}$.

Now, the transformed problem in (28) is convex w.r.t. to $\mathbf{p}$ and $\mathbf{u}$ while satisfying the Slaters constraint qualification condition [39]. Therefore, the duality gap is zero. As a result, to solve the primal problem in (28), we focus on solving its dual problem. To this end, we first define the Lagrangian function of the primal problem in (28) which can be written as

$$\begin{aligned}
\mathcal{L}\left(\mathbf{p},\mathbf{u},\boldsymbol{\zeta},\boldsymbol{\nu},\lambda\right) = & \sum_{k=1}^{K}\sum_{i=1}^{N_{\mathrm{F}}} \left(\nu_k + 1\right) u_{k,i} \log_2\left(1 + \frac{\widetilde{p}_{k,i}|h_{k,i}|^2}{u_{k,i}\sigma^2}\right) \\
& - \left(\delta \mathrm{EE}_{\mathrm{Sys}} + \lambda\right) \sum_{k=1}^{K}\sum_{i=1}^{N_{\mathrm{F}}} \widetilde{p}_{k,i} - \mathrm{EE}_{\mathrm{Sys}} P_{\mathrm{C}} \\
& - \sum_{i=1}^{N_{\mathrm{F}}}\sum_{k=1}^{K} \zeta_i u_{k,i} + \sum_{k=1}^{K} \zeta_i + \lambda p_{\max} - \sum_{k=1}^{K} \nu_k R_{\min},
\end{aligned} \tag{29}$$

where $\zeta_i \geq 0$, $\lambda \geq 0$ and $\nu_k \geq 0$ are the Lagrange multipliers corresponding to constraints C3, C4 and C5, respectively. Constraints C1 and C2 will be captured in the Karush-Kuhn-Tucker

(KKT) conditions when deriving the optimal resource allocation policy of the problem (28) in the following. Therefore, the dual problem for the primal problem in (28) is given by

$$\underset{\boldsymbol{\zeta},\boldsymbol{\nu},\lambda}{\text{minimize}} \ \ \underset{\mathbf{p},\mathbf{u}}{\text{maximize}} \, \mathcal{L}\left(\mathbf{p},\mathbf{u},\boldsymbol{\zeta},\boldsymbol{\nu},\lambda\right). \tag{30}$$

Since the dual problem is convex, the Lagrange dual decomposition can be employed to solve the dual problem in (30) iteratively. In particular, the dual problem in (30) is decomposed into two nested layers: Layer 1, maximizing the Lagrangian $\mathcal{L}\left(\mathbf{p},\mathbf{u},\boldsymbol{\zeta},\boldsymbol{\nu},\lambda\right)$ in (30) over the power allocation $\mathbf{p}$ and the user scheduling $\mathbf{u}$ for given Lagrangian multipliers $\boldsymbol{\zeta}$, λ and $\boldsymbol{\nu}$; Layer 2, minimizing $\mathcal{L}\left(\mathbf{p},\mathbf{u},\boldsymbol{\zeta},\boldsymbol{\nu},\lambda\right)$ over $\boldsymbol{\zeta}$, λ and $\boldsymbol{\nu}$ for given $\mathbf{p}$ and $\mathbf{u}$.

For a fixed set of Lagrange multipliers $(\boldsymbol{\zeta},\lambda,\boldsymbol{\nu})$, the inner maximization problem is a convex optimization problem w.r.t. $(\mathbf{p},\mathbf{u})$. Using the standard optimization techniques and the KKT conditions, the optimal power allocation for user k on subcarrier i can be obtained by

$$\widetilde{p}_{k,i}^{*} = u_{k,i}p_{k,i}^{*} = u_{k,i}\left[\frac{1+\nu_k}{(\delta \mathrm{EE}_{\mathrm{Sys}}+\lambda)\ln 2} - \frac{\sigma^2}{|h_{k,i}|^2}\right]^{+}, \forall k,i. \tag{31}$$

The optimal power allocation has the form of multilevel water filling [8], [34], [35]. From (30), we can observe that the Lagrange multiplier ν_k becomes larger when its minimum data-rate requirement becomes stringent, and vice versa. In other words, the Lagrange multiplier ν_k controls the water level of user k and it forces the BS to increase the transmit power to fulfill the minimum data rate requirement $R_{\min}$. In addition, we can observe that the system EE $\mathrm{EE}_{\mathrm{Sys}} > 0$ prevents energy inefficient transmission via clipping the water-level. On the other hand, to obtain the optimal user scheduling, we take the derivative of the Lagrangian function w.r.t. $u_{k,i}$ which yields

$$\frac{\partial \mathcal{L}}{\partial u_{k,i}} = M_{k,i} = (1+\nu_k)\left[\log_2\left(1+\frac{p_{k,i}^{*}|h_{k,i}|^2}{\sigma^2}\right) - \frac{p_{k,i}^{*}|h_{k,i}|^2}{\ln 2\left(p_{k,i}^{*}|h_{k,i}|^2+\sigma^2\right)}\right] - \zeta_i, \forall k,i. \tag{32}$$

The constant derivative in (32) implies that the Lagrangian function grows linearly w.r.t. $u_{k,i}$. In fact, $M_{k,i}$ has the physical meaning of marginal benefit to the system EE when subcarrier i is assigned to user k. We can observe that allocating user k to subcarrier i with a higher ν_k and a higher channel gain $|h_{k,i}|^2$ can provide a higher marginal benefit to the system. Since $M_{k,i}$ is independent of $u_{k,i}$ and $\sum_{k=1}^{K} u_{k,i} \leq 1$, the allocation of subcarrier i at the BS should base on the following winner-take-all criterion:

$$u_{k,i}^{*} = \begin{cases} 1 & \text{if } k = \arg\max\limits_{j} M_{j,i}, \\ 0 & \text{otherwise}, \end{cases} \tag{33}$$

which implies that user k is assigned to subcarrier i if it can provide the maximal marginal benefit to the system. The derived subcarrier allocation solution in (33) demonstrates that although time sharing is assumed for solving the optimization problem, the optimal solution indicates that the maximum system performance is achieved when there is no time sharing on any subcarrier.

On the other hand, the Layer 2 outer minimization problem can be solved by using the gradient methods, which leads to the following iterative Lagrange multiplier update rules:

$$\lambda\left(m+1\right)=\left[\lambda\left(m\right)-\xi_1\left(m\right)\times\left(p_{\max}-\sum_{k=1}^{K}\sum_{i=1}^{N_{\mathrm{F}}}\tilde{p}_{k,i}\right)\right]^{+}\text{ and} \tag{34}$$

$$\nu_k\left(m+1\right)=\left[\nu_k\left(m\right)-\xi_2\left(m\right)\times\left(\tilde{R}_k-R_{\min}\right)\right]^{+}, \tag{35}$$

where $m>0$ is the iteration index and $\xi_1\left(m\right),\xi_2\left(m\right)>0$ are positive step sizes. We note that there is no need to update the Lagrangian multiplier ζ_i as the optimal subcarrier allocation can always be found via (33). Therefore, we ignore ζ_i in each iteration.

Now, given the parameter $\mathrm{EE}_{\mathrm{Sys}}$, alternatively solving the Layer 1 inner problem and Layer 2 outer problem can solve the main loop problem in (28). In particular, based on the Lagrangian multipliers $(\boldsymbol{\zeta},\lambda)$ of the last iteration, the Layer 1 inner problem is solved by (31) and (33). Then, the obtained intermediate resource allocation policies $(\mathbf{p},\mathbf{u})$ are passed to Layer 2 for updating the Lagrangian multipliers for next iteration according to (34) and (35). The procedure repeats until convergence is achieved or the number of maximum iterations is reached. Since the transformed problem in (28) is convex for a given parameter $\mathrm{EE}_{\mathrm{Sys}}$, the iteration between Layer 1 and Layer 2 can converge to the optimal solution of (28) under some mild conditions on the step sizes [40].

D. Simulation Results and Discussions

In this section, we evaluate the performance of the proposed energy-efficient resource allocation design via simulations. Besides, we unveil some interesting insights about energy-efficient resource allocation design. Unless specified otherwise, the simulation setting is given as follows. The number of users is $K=5$. The system bandwidth is 10 MHz and it is divided into $N_{\mathrm{F}}=16$ subcarriers with equal subcarrier bandwidth. The fixed circuit power consumption is $P_{\mathrm{C}}=0.1$ Watt and $\delta=2$. The maximum transmit power of the considered OFDMA system $p_{\max}$ ranges from 5 to 45 dBm and the noise power in each subcarrier bandwidth is $\sigma^2=0.1$ Watt. In addition, the requested minimum data rate $R_{\min}$ ranges from 0

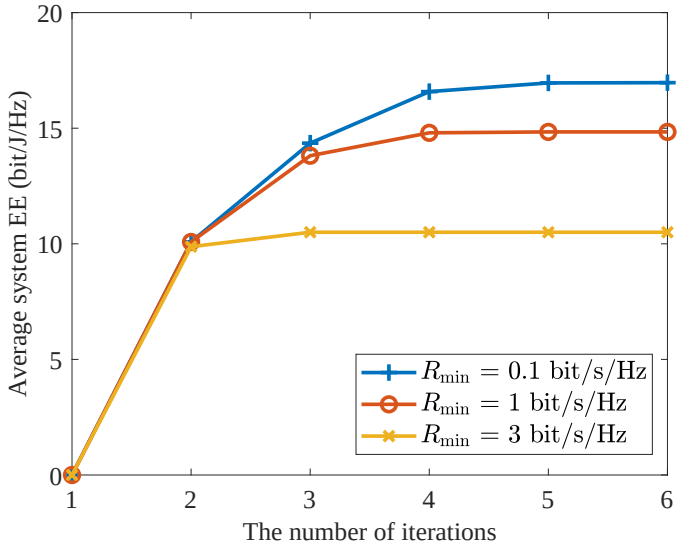


Fig. 3. The convergence of the Dinkelbach's algorithm.

to 3 bit/s/Hz. For simplicity, we only take into account the small scale fading $h_{k,i} \sim \mathcal{CN}(0,1)$ and average the simulation results over 100 channel realizations.

Case I: Convergence of Dinkelbach's Algorithm

We first verify the convergence of Dinkelbach's algorithm in Fig. 3 with $p_{\max} = 30$ dBm and $R_{\min} = [0.1, 1, 3]$ bit/s/Hz. We can observe that the system EE monotonically increases with the number of iterations. Furthermore, the Dinkelbach's algorithm enjoys both quick convergence and low complexity. Besides, we can observe that the larger the required minimum data rate, the smaller the converged system EE. In fact, when increasing $R_{\min}$ in the problem in (25), the feasible solution set becomes smaller and the maximum system EE is reduced.

Case II: Average System EE versus the Maximum Transmit Power

Fig. 4 illustrates the average system EE versus the maximum transmit power $p_{\max}$ with $R_{\min} = [0.1, 1, 3]$ bit/s/Hz. It can be seen that the system EE first increases with $p_{\max}$ and then saturates in the high SNR regime. Indeed, transmitting with the maximum available power is the most energy-efficient option in the low SNR regime. However, with increasing the system transmit power budget, there is a diminishing return in the spectral efficiency when allocating more transmit power. Hence, the energy consumption in the system would outweigh the spectral efficiency gain in the high SNR regime. Therefore, the most energy-efficient operation point in the high SNR regime only utilizes the right enough amount of

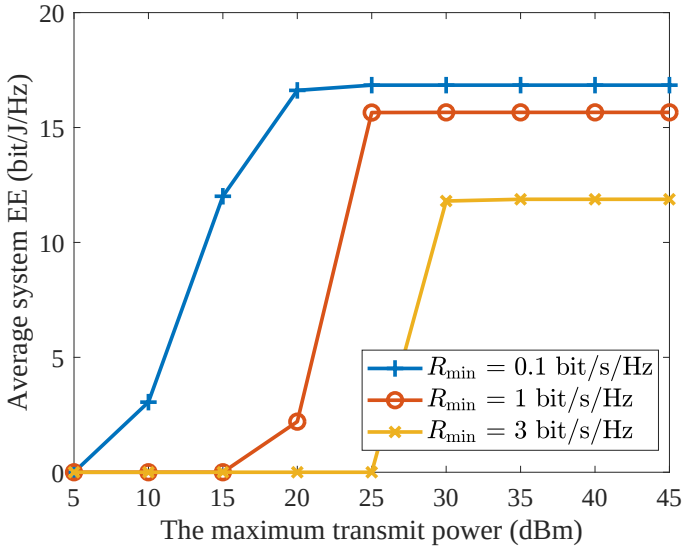


Fig. 4. The average system EE versus the maximum transmit power.

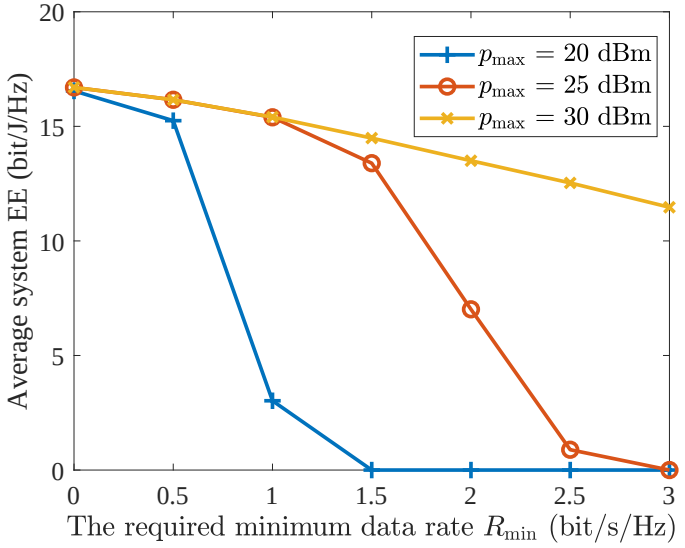


Fig. 5. The average system EE versus the requested minimum data rate.

the total power budget and further increasing $p_{\max}$ cannot increase the system EE.

Case III: Average System EE versus the Requested Minimum Data Rate

Fig. 5 shows the average system EE versus the requested minimum data rate $R_{\min}$ with $p_{\max} = [20, 25, 30]$ dBm. We can observe that the system EE decreases dramatically when the minimum data rate $R_{\min}$ increases, especially with a low transmit power. In fact, a higher transmit power needs to be allocated to the user to satisfying the increased minimum data

rate despite their possibly weak channel conditions. This reduces the flexibility in resource allocation which yields a smaller system sum-rate and limits the system EE.

III. Conclusion and Future Directions

In this article, we have explored the basic ideas and technical challenges in energy-efficient resource management. In particular, two types of commonly adopted definitions of EE metrics, i.e., system-centric EE and user-centric EE, were presented and discussed. The fundamental trade-off between the EE and spectral efficiency was unveiled. Maximizing the system-centric EE can be classified as fractional programming, which can be solved efficiently via the Dinkelbach's algorithm. On the other hand, the user-centric EE is a sum-of-ratios function and maximizing the user-centric EE can be achieved via an iterative parametric resource allocation algorithm. As an illustrative example, we studied the energy-efficient resource allocation design in an OFDMA system. The energy-efficient power allocation and user scheduling were formulated as an optimization problem with taking into account the minimum data rate requirements. After transforming the fractional objective function to an equivalently subtractive form, dual decomposition method has been used to solve the main loop problem in each iteration of Dinkelbach's algorithm. Our simulation results have demonstrated the fast convergence of Dinkelbach's algorithm and the excellent performance of the proposed resource allocation design.

It is clearly expected that the energy-efficient resource management will continue to be a valuable technique for the future 5G and beyond 5G wireless networks. In particular, the following are promising extensions of energy-efficient resource allocation to meet the requirements of future wireless networks.

A. Energy-Efficient NOMA Communications

Recently, NOMA has received considerable attention as a promising multiple access technique for the 5G wireless networks [9], [12], [41]. The basic principle of NOMA is to exploit the power domain for users multiplexing and to employ successive interference cancelation (SIC) at receivers to retrieve the users' messages. In contrast to conventional orthogonal multiple access (OMA) schemes, NOMA is a promising solution to fulfill the demanding requirements of the 5G communication systems [1], [2], such as massive connectivity [4], high spectral efficiency [42], [43] and enhanced user fairness [44]. Green radio design for NOMA systems [45]–[47] has become an important focus in both academia and industry due

to the growing demands of energy consumption and arising environmental concerns around the world. Therefore, it is important to investigate the energy-efficient resource allocation for NOMA communication systems.

B. Energy-Efficient mmWave Communications

MmWave communications have recently triggered and attracted tremendous research interests due to their potential to meet the stringent requirements of ultra-high data rate for the future fifth-generation (5G) wireless networks [48], [49]. However, due to the prohibitively large power consumption of RF chains, the EE of mmWave systems is generally limited and remains to be further improved. Hence, the energy-efficient resource allocation design is a fundamentally important issue to be tackled for realizing mmWave communications in future 5G networks.

C. Energy-Efficient UAV Communications

Owing to the high flexibility and low cost in the deployment of UAVs, UAV-enabled communication offers a promising solution to fulfill the stringent requirements of future wireless networks. In practice, the total energy budget for maintaining a stable flight and communication is limited by the onboard battery capacity. Hence, EE has become an important figure of merit for UAV-based communications. Jointly designing the trajectory and resource allocation for UAV communication systems is critical to improving the system EE.

IV. APPENDIX

A. Proof of Theorem 1

We start the proof by verifying the forward via a similar approach in [37]. Without loss of generality, let $\mathbf{x}^*$ be a solution of the problem in (4), i.e.,

$$\mathrm{EE}_{\mathrm{Sys}}^{*}=\frac{f\left(\mathbf{x}^{*}\right)}{g\left(\mathbf{x}^{*}\right)}\geq\frac{f\left(\mathbf{x}\right)}{g\left(\mathbf{x}\right)},\mathbf{x}\in\mathcal{X}. \tag{36}$$

Hence, we have

$$f\left(\mathbf{x}\right)-\mathrm{EE}_{\mathrm{Sys}}^{*}g\left(\mathbf{x}\right)\leq 0,\mathbf{x}\in\mathcal{X}, \tag{37}$$

$$f\left(\mathbf{x}^{*}\right)-\mathrm{EE}_{\mathrm{Sys}}^{*}g\left(\mathbf{x}^{*}\right)=0. \tag{38}$$

Comparing (37) and (38), we can observe that $\mathbf{x}^*$ is the optimal solution of the problem in (6).

Now, we prove the converse. Let $\mathbf{x}^*$ be a solution of the problem in (6), i.e.,

$$f(\mathbf{x}) - \mathrm{EE}_{\mathrm{Sys}}^* g(\mathbf{x}) \leq f(\mathbf{x}^*) - \mathrm{EE}_{\mathrm{Sys}}^* g(\mathbf{x}^*) = 0, \mathbf{x} \in \mathcal{X}. \tag{39}$$

Hence, we have

$$\mathrm{EE}_{\mathrm{Sys}}^* \geq \frac{f(\mathbf{x})}{g(\mathbf{x})}, \mathbf{x} \in \mathcal{X}, \tag{40}$$

$$\mathrm{EE}_{\mathrm{Sys}}^* = \frac{f(\mathbf{x}^*)}{g(\mathbf{x}^*)}. \tag{41}$$

Comparing (40) and (41), we can observe that $\mathbf{x}^*$ is the optimal solution of the problem in (4) and it results in the optimal value $\mathrm{EE}_{\mathrm{Sys}}^*$. It completes the proof of Theorem 1.

B. Proof of Theorem 2

Let us define the optimal value of the problem in (6) given the parameter $\mathrm{EE}_{\mathrm{Sys}}$ as

$$F(\mathrm{EE}_{\mathrm{Sys}}) = \underset{\mathbf{x}\in\mathcal{X}}{\text{maximize}} \;\; f(\mathbf{x}) - \mathrm{EE}_{\mathrm{Sys}} g(\mathbf{x}). \tag{42}$$

For any feasible solution $\widetilde{\mathbf{x}} \in \mathcal{X}$ and $\widetilde{\mathrm{EE}}_{\mathrm{Sys}} = \frac{f(\widetilde{\mathbf{x}})}{g(\widetilde{\mathbf{x}})}$, we can observe that

$$F\left(\widetilde{\mathrm{EE}}_{\mathrm{Sys}}\right) = \underset{\mathbf{x}\in\mathcal{X}}{\text{maximize}} \;\; f(\mathbf{x}) - \frac{f(\widetilde{\mathbf{x}})}{g(\widetilde{\mathbf{x}})} g(\mathbf{x}) \geq f(\widetilde{\mathbf{x}}) - \frac{f(\widetilde{\mathbf{x}})}{g(\widetilde{\mathbf{x}})} g(\widetilde{\mathbf{x}}) = 0. \tag{43}$$

Therefore, following the updating rule in line 8 of Dinkelbach's algorithm, we have

$$F\left(\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}\right) = f\left(\mathbf{x}^{\mathrm{iter}}\right) - \mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}} g\left(\mathbf{x}^{\mathrm{iter}}\right) = \left(\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}+1} - \mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}\right) g\left(\mathbf{x}^{\mathrm{iter}}\right) \geq 0. \tag{44}$$

Since $g(\mathbf{x}) > 0$, we have $\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}+1} \geq \mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}$. It implies that the system EE $\mathrm{EE}_{\mathrm{Sys}}$ monotonically increases with the number of iterations in Dinkelbach's algorithm. Besides, we can observe that

$$\begin{aligned} F\left(\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}+1}\right) &= f\left(\mathbf{x}^{\mathrm{iter}+1}\right) - \mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}+1} g\left(\mathbf{x}^{\mathrm{iter}+1}\right) \\ &< f\left(\mathbf{x}^{\mathrm{iter}+1}\right) - \mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}} g\left(\mathbf{x}^{\mathrm{iter}+1}\right) \\ &\leq \underset{\mathbf{x}\in\mathcal{X}}{\text{maximize}} \;\; f(\mathbf{x}) - \mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}} g(\mathbf{x}) \\ &= F\left(\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}\right), \end{aligned} \tag{45}$$

which implies the monotonically decreasing feature of $F(\mathrm{EE}_{\mathrm{Sys}})$ with the proceeding of iterations in Dinkelbach's algorithm. It completes the proof of the convergence of Dinkelbach's algorithm.

In the following, we prove the optimality of the Dinkelbach's algorithm by contradiction. Let $\mathrm{EE}_{\mathrm{Sys}}^{*}$ denote the optimal value of the problem in (4). According to Theorem 1, we have

$$F\left(\mathrm{EE}_{\mathrm{Sys}}^{*}\right)=0. \tag{46}$$

Assuming the convergence point $\left\{\mathbf{x}^{\mathrm{iter}},\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}\right\}$ is not the optimal point, i.e., $\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}<\mathrm{EE}_{\mathrm{Sys}}^{*}$ and $\mathbf{x}^{\mathrm{iter}}=\arg\underset{\mathbf{x}\in\mathcal{X}}{\mathrm{maximize}}\quad f\left(\mathbf{x}\right)-\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}g\left(\mathbf{x}\right)$. Thus we have $F\left(\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}\right)=0$. However, according to (45), we have $F\left(\mathrm{EE}_{\mathrm{Sys}}^{*}\right)<F\left(\mathrm{EE}_{\mathrm{Sys}}^{\mathrm{iter}}\right)=0$, which contradicts to (46).

The proof above is based on the condition that the problem in (4) can be solved optimally.

C. Proof of Theorem 3

We follow a similar approach in [38] to prove Theorem 3. The problem in (7) is equivalent to the following problem:

$$\begin{aligned}\underset{\mathbf{x}}{\mathrm{maximize}}\quad & \sum_{k=1}^{K}\beta_k \\ \text{s.t.}\quad & t_n(\mathbf{x})\leq 0,\forall n=1,\ldots,N,\\ & \omega_k\frac{f_k\left(\mathbf{x}\right)}{g_k\left(\mathbf{x}\right)}\geq\beta_k,\forall k,\end{aligned} \tag{47}$$

where $\mathcal{X}=\left\{\mathbf{x}\,|t_n(\mathbf{x})\leq 0,n=1,\ldots,N\right\}$ and $t_n(\mathbf{x})$ is a convex function with respect to (w.r.t.) $\mathbf{x}$. Following the proof outline in [38], we define the following function for the problem in (48):

$$\mathcal{L}\left(\mathbf{x},\varpi,\alpha_k,\beta_k\right)=-\varpi\sum_{k=1}^{K}\beta_k+\sum_{k=1}^{K}\alpha_k\left(\beta_k g_k\left(\mathbf{x}\right)-\omega_k f_k\left(\mathbf{x}\right)\right)+\sum_{n=1}^{N}\nu_n t_n(\mathbf{x}). \tag{48}$$

According to the Fritz-John optimality condition, there exist variables ϖ^{*}, α_k^{*}, β_k^{*} and ν_n^{*} such that they satisfy

$$\frac{\partial\mathcal{L}}{\partial\beta_k}=-\varpi^{*}+\alpha_k^{*}g_k\left(\mathbf{x}\right)=0,\forall k, \tag{49}$$

$$\alpha_k^{*}\frac{\partial\mathcal{L}}{\partial\alpha_k}=\alpha_k^{*}\left(\beta_k^{*}g_k\left(\mathbf{x}\right)-\omega_k f_k\left(\mathbf{x}\right)\right)=0,\forall k, \tag{50}$$

$$\nu_n^{*}\frac{\partial\mathcal{L}}{\partial\nu_n}=\nu_n^{*}t_n(\mathbf{x})=0,\forall n, \tag{51}$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{x}}=\sum_{k=1}^{K}\alpha_k^{*}\left(\beta_k^{*}\nabla g_k\left(\mathbf{x}\right)-\omega_k\nabla f_k\left(\mathbf{x}\right)\right)+\nu_n^{*}\nabla t_n\left(\mathbf{x}\right)=\mathbf{0}, \tag{52}$$

$$\beta_k^{*}g_k\left(\mathbf{x}\right)-\omega_k f_k\left(\mathbf{x}\right)\leq 0,\alpha_k^{*}\geq 0,\forall k, \tag{53}$$

$$t_n(\mathbf{x})\leq 0,\nu_n^{*}\geq 0,\forall k, \tag{54}$$

$$\varpi^{*}\geq 0,\left\{\varpi^{*},\alpha_k^{*},\beta_k^{*}\right\}\neq\left\{0,0,0\right\},\forall k, \tag{55}$$

where $\nabla f(\mathbf{x})$ denotes the derivative of function $f(\mathbf{x})$ w.r.t. $\mathbf{x}$. Suppose that $\varpi^* = 0$. Then, from (49), we have $\alpha_k^* = 0$ since $g_k(\mathbf{x}) > 0$, $\forall k = 1, \ldots, K$ for $\mathbf{x} \in \mathcal{X}$. Following from (51), (52), (54) and (55), we have

$$\sum_{n \in I(\bar{\mathbf{x}})} \nu_n^* \nabla t_n(\bar{\mathbf{x}}) = \mathbf{0}, \tag{56}$$

$$\sum_{n \in I(\bar{\mathbf{x}})} \nu_n^* > 0, \nu_n^* \geq 0, n \in I(\bar{\mathbf{x}}), \tag{57}$$

where $I(\bar{\mathbf{x}}) = \{n \,|t_n(\bar{\mathbf{x}}) = 0, 1 \leq n \leq N\}$. According to the Slaters constraint qualification condition, there exists an inner point $\widetilde{\mathbf{x}}$ such that

$$t_n(\widetilde{\mathbf{x}}) < 0, n = 1, \ldots, N. \tag{58}$$

Since $t_n(\widetilde{\mathbf{x}})$, $n = 1, \ldots, N$ are convex, it follows from (58) that

$$\nabla t_n(\bar{\mathbf{x}})^{\mathrm{T}} (\widetilde{\mathbf{x}} - \bar{\mathbf{x}}) \leq t_n(\widetilde{\mathbf{x}}) - t_n(\bar{\mathbf{x}}) < 0, n \in I(\bar{\mathbf{x}}). \tag{59}$$

Letting $\mathbf{d} = \widetilde{\mathbf{x}} - \bar{\mathbf{x}}$, from (59) and (57), we have $\left(\sum_{n \in I(\bar{\mathbf{x}})} \nu_n^* \nabla t_n(\bar{\mathbf{x}}) \right)^{\mathrm{T}} \mathbf{d} < 0$, which contradicts (56). Thus, we have $\varpi^* > 0$ and $\alpha_k^* > 0$. Redefining $\alpha_k^* = \frac{\alpha_k^*}{\varpi^*}$ and $\nu_n^* = \frac{\nu_n^*}{\varpi^*}$, we can observe that (49) and (50) are equivalent to (10) and (9), respectively.

Given $\alpha_k = \alpha_k^*$ and $\beta_k = \beta_k^*$, (51), (52) and (54) are just the KKT conditions for the problem in (8). Since the problem in (8) is convex programming for given parameters α_k and β_k, the KKT conditions are also the sufficient optimality condition and then $\bar{\mathbf{x}}$ is the solution for the problem in (8) with $\alpha_k = \alpha_k^*$ and $\beta_k = \beta_k^*$.

REFERENCES

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[3] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Sep. 2018.

[4] Z. Sun, Z. Wei, L. Yang, J. Yuan, X. Cheng, and L. Wan, "Joint user identification and channel estimation in massive connectivity with transmission control," in *Proc. IEEE Intern. Sympos. on Turbo Codes Iterative Information Process.*, 2018, pp. 1–5.

[5] Z. Sun, Y. Xie, J. Yuan, and T. Yang, "Coded slotted ALOHA for erasure channels: Design and throughput analysis," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4817–4830, Nov. 2017.

[6] Z. Sun, L. Yang, J. Yuan, and D. W. K. Ng, "Physical-layer network coding based decoding scheme for random access," *IEEE Trans. Veh. Technol.*, pp. 1–1, Jan. 2019.

[7] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[8] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.

[9] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[10] M. Qiu, Y. Huang, J. Yuan, and C. Wang, "Lattice-partition-based downlink non-orthogonal multiple access without SIC for slow fading channels," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1166–1181, Feb. 2019.

[11] M. Qiu, Y. Huang, S. Shieh, and J. Yuan, "A lattice-partition framework of downlink non-orthogonal multiple access without SIC," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2532–2546, Jun. 2018.

[12] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *ZTE Commun.*, vol. 14, no. 4, pp. 17–25, Oct. 2016.

[13] Z. Wei, D. W. K. Ng, J. Yuan, and H. Wang, "Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944–3961, Sep. 2017.

[14] Z. Wei, L. Yang, D. W. K. Ng, and J. Yuan, "On the performance gain of NOMA over OMA in uplink single-cell systems," in *Proc. IEEE Global Commun. Conf.*, 2018, pp. 1–7.

[15] Z. Wei, D. W. K. Ng, and J. Yuan, "Joint pilot and payload power control for uplink MIMO-NOMA with MRC-SIC receivers," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 692–695, Apr. 2018.

[16] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, J. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[17] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmwave systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1705–1719, Feb. 2019.

[18] Z. Wei, D. W. Kwan Ng, and J. Yuan, "NOMA for hybrid mmwave communication systems with beamwidth control," *IEEE J. Select. Topics Signal Process.*, pp. 1–1, Feb. Early access, 2019.

[19] L. Zhao, G. Geraci, T. Yang, D. W. K. Ng, and J. Yuan, "A tone-based AoA estimation and multiuser precoding for millimeter wave massive MIMO," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5209–5225, Dec. 2017.

[20] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE J. Select. Areas Commun.*, vol. 35, no. 7, Jul. 2017.

[21] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: Past, Present, and Future," *IEEE J. Select. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[22] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, May 2014.

[23] J. Guo, W. Yu, and J. Yuan, "Enhancing cellular performance through device-to-device distributed MIMO," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6096–6109, Dec. 2018.

[24] J. Guo, J. Yuan, and J. Zhang, "An achievable throughput scaling law of wireless device-to-device caching networks with distributed MIMO and hierarchical cooperations," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 492–505, Jan. 2018.

[25] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Mar. 2018.

[26] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, "Optimal 3D-trajectory design and resource allocation for solar-powered UAV communication systems," *IEEE Trans. Commun.*, vol. 1, no. 1, pp. 1–1, Feb. 2019.

[27] Y. Cai, Z. Wei, R. Li, D. W. K. Ng, and J. Yuan, "Energy-efficient resource allocation for secure UAV communication systems," in *Proc. IEEE Wireless Commun. and Networking Conf.*, 2019, pp. 1–8.

[28] R. Li, Z. Wei, L. Yang, D. W. K. Ng, N. Yang, J. Yuan, and J. An, "Joint trajectory and resource allocation design for UAV communication systems," in *Proc. IEEE Global Commun. Conf.*, 2018, pp. 1–7.

[29] C. Lin and G. Y. Li, "Energy-efficient design of indoor mmwave and sub-THz systems with antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4660–4672, Mar. 2016.

[30] C. H. News, "Tsunami of data could consume one-fifth of global electricity by 2025," Online: https://www.climatechangenews.com.

[31] "SMARTer2030: ICT solutions for 21st-century challenges," Global e-Sustainability Initiative, Tech. Rep., 2015.

[32] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.

[33] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524–540, Nov. 2011.

[34] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3618–3631, Oct. 2012.

[35] ——, "Energy-efficient resource allocation for secure OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2572–2585, Jul. 2012.

[36] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2082–2085, Dec. 2015.

[37] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[38] Y. Jong, "An efficient global optimization algorithm for nonlinear sum-of-ratios problem," Online: https://www.optimizationonline.org, 2012.

[39] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[40] S. Boyd, "Subgradient methods," Oct. 2013.

[41] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[42] Z. Wei, L. Dai, D. W. K. Ng, and J. Yuan, "Performance analysis of a hybrid downlink-uplink cooperative NOMA scheme," in *Proc. IEEE Veh. Techn. Conf.*, 2017, pp. 1–7.

[43] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *arXiv preprint arXiv:1903.01683*, 2019.

[44] Z. Wei, J. Guo, D. W. K. Ng, and J. Yuan, "Fairness comparison of uplink NOMA and OMA," in *Proc. IEEE Veh. Techn. Conf.*, 2017, pp. 1–6.

[45] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang, and K. K. Wong, "Energy-efficient NOMA enabled heterogeneous cloud radio access networks," *IEEE Network*, vol. 32, no. 2, pp. 152–160, Mar. 2018.

[46] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[47] H. Zhang, F. Fang, J. Cheng, K. Long, W. Wang, and V. C. M. Leung, "Energy-efficient resource allocation in NOMA heterogeneous networks," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 48–53, Apr. 2018.

[48] L. Zhao, Z. Wei, D. W. K. Ng, J. Yuan, and M. C. Reed, "Multi-cell hybrid millimeter wave systems: Pilot contamination and interference mitigation," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5740–5755, Nov. 2018.

[49] L. Zhao, J. Guo, Z. Wei, D. W. K. Ng, and J. Yuan, "A distributed multi-RF chain hybrid mmwave scheme for small-cell systems," in *Proc. IEEE Intern. Commun. Conf.*, May 2019, pp. 1–7.