IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 22, 2020, accepted February 10, 2020, date of publication February 18, 2020, date of current version March 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974755

Energy-Efficient Adaptive Modulation and Data Schedule for Delay-Sensitive Wireless Communications

QIAN CHEN[1,2,3], XIAOJING CHEN[1,2,3], (Member, IEEE), SHUNQING ZHANG[1,2,3], (Senior Member, IEEE), SHUGONG XU[1,2,3], (Fellow, IEEE), AND JIE TANG[4], (Senior Member, IEEE)

[1]Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[2]Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, Shanghai 200444, China
[3]Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai University, Shanghai 200444, China
[4]School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510006, China

Corresponding author: Xiaojing Chen (jodiechen@shu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) Grant under Grant 61701293, Grant 61871262, and Grant 61901251, in part by the National Science and Technology Major Project under Grant 2018ZX03001009, in part by the National Key Research and Development Program of China under Grant 2017YFE0121400, and in part by the Research Funds from the Shanghai Institute for Advanced Communication and Data Science (SICS).

ABSTRACT In this paper, targeting at improving the energy efficiency (EE) for Quality-of-Service (QoS)-guaranteed wireless communications, we develop new adaptive modulation and data scheduling algorithms for delay-sensitive bursty data. Assuming a-priori knowledge on data arrivals and latency requirements, the problem is formulated as a mix-integer programming that minimizes the total energy consumption at the transmitter with a non-linear Doherty power amplifier (PA) and non-negligible circuit power. According to the different properties of the PA in different output power regions, we decouple the formulated problem and solve it in two stages. In the first stage, assuming the PA has a linear efficiency, we develop an optimal modulation and data scheduling scheme (MDS) relying on convex relaxation and the resultant optimality conditions. The MDS is able to reveal the specific structure of the optimal policy in a computationally efficient and graphical manner. On top of that, a heuristic MDS scheme (HMDS) is proposed to adjust the MDS when the PA works in the non-linear region in the second stage, where a quadratic function is obtained to approximate the non-linear PA model. The offline HMDS algorithm is further extended to practical online scenarios in a well-structured way, where the modulation and data scheduling policy is produced on-the-fly. Simulation corroborates that the proposed offline algorithm can achieve the exactly same performance as the standard CVX solver, while requiring only 0.69% of its computational time.

INDEX TERMS Energy efficiency (EE), adaptive modulation, quality of service (QoS), circuit power, non-linear power amplifier (PA).

I. INTRODUCTION

Energy efficiency (EE) has been raised as an important issue in the design of wireless communications for economic and ecological concerns [1]. Especially for small battery-powered wireless (e.g., sensor) networks, improving EE is a key solution to prolong the operating lifetime [2]. In addition, Quality-of-Service (QoS), e.g., latency requirement and bit error rate (BER), is extremely important to many applications, including delay-sensitive sensory data in bushfire monitoring and security surveillance [3], [4].

Due to the inherent tradeoff between energy consumption and QoS, challenges arise in improving EE for QoS-guaranteed wireless transmissions [5], especially for short-range wireless networks where the circuit energy consumption due to, e.g., signal processing and filtering, is non-negligible. Adaptive modulation, considered as an effective way to improve system EE, has been extensively studied in the literature [6]–[14].

It was shown in [6] that the system EE can be improved to the greatest extent by jointly optimizing the modulation

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

order at the physical layer and the backoff probability at the Medium Access Control (MAC) layer. An energy-efficient adaptive modulation and power control scheme was proposed for wireless sensor networks in [7], where the sensor node changes its transmit power and modulation scheme in adaptation to the signal-to-noise-ratio (SNR) and target BER. In [8], the EE of a point-to-point link was improved by proposing a dynamic feedback-based adaptive modulation scheme, where the channel state information (CSI) is learned from the receiver feedback per time slot.

In [6]–[8], the circuit power was assumed to be negligible. Capturing non-negligible circuit power in analysis, an energy-efficient data rate for a target BER was obtained in a closed form regarding to the constellation size, distance and bandwidth [9]. The authors in [10] analyzed the effects of bandwidth, power and modulation scheme on the system EE under different channel conditions. A non-convex combinatorial EE maximization problem was solved by obtaining an equivalent one-dimensional problem, and proposing a greedy modulation and power control algorithm [11]. The works [6]–[11] all focused on delay-tolerant data, and thus cannot guarantee QoS for practical delay-sensitive traffics. Besides, in these works, it was assumed that there are always data available in the buffer for transmission; in a more general scenario, data arrivals can be bursty over time. Reference [12] proposed an energy-efficient cross-layer design framework for transmitting Markov modulated Poisson process (MMPP) traffic with delay requirements, where adaptive modulation and coding scheme is performed at the physical layer. Using the notion of energy-delay tradeoff, [13] compared adaptive modulation and coding (AMC) and hybrid automatic repeat request (HARQ) schemes with given equivalent QoS constraints. Yet, [12] and [13] are still not applicable to general data arrival processes.

In addition to adaptive modulation based schemes, increasing hardware efficiency, e.g., adopting a high efficiency Doherty power amplifier (PA), is a straightforward way to improve the system EE [15]. Most works assume that PA has a linear efficiency. For a practical Doherty PA, the power efficiency is non-linear in high output power region. A few recent works captured the non-linearity of PA efficiency in increasing EE for wireless systems [14], [16]. An adaptive polarization-quadrature amplitude modulation (QAM) scheme was developed for OFDM systems in [14], where QAM and polarization modulation are used in the linear and non-linear regions of the PA, respectively. A dynamic carrier allocation strategy was proposed to map carriers into multi-carrier power amplifiers [16], and a comparison of two methods (convex relaxation and deep learning) was provided.

In this paper, to address the challenge of improving EE for QoS-guaranteed wireless communications, we develop new adaptive modulation and data scheduling algorithms for delay-sensitive data arriving in bursts. The problem is formulated as a mix-integer programming that minimizes the total energy consumption at the transmitter with a non-linear Doherty PA and non-negligible circuit power. According to the different properties of the PA in different output power regions, we proceed to solve the formulated problem in two stages. In the first stage, when the PA has a linear efficiency, we develop an optimal modulation and data scheduling scheme (MDS) relying on convex relaxation and the resultant optimality conditions. On top of that, a heuristic MDS scheme (HMDS) is proposed to adjust the MDS when the PA works in the non-linear region in the second stage, where a quadratic function is obtained to approximate the non-linear PA model.

The contributions of this paper can be summarized as follows:

1) The formulated modulation and data scheduling problem is a complex mix-integer programming, especially with non-linear PA efficiency and non-negligible circuit power consumption. By decoupling it and solving it in two stages, a new optimal MDS algorithm is first developed to generate the optimal ‘‘on-off’’ transmission policy in a graphical manner with a low complexity. The MDS, with proven optimality, is insightful by revealing the specific structure of the optimal policy.
2) A HMDS algorithm, which follows the procedure of the MDS, is proposed to address the non-linearity of the PA. A quadratic function is obtained based on Taylor expansion to approximate the non-linear PA model.
3) We further extend the offline HMDS algorithm to practical online scenarios in a well-structured way, where only causal information of data arrivals and latency requirements is available. Extension for online implementations in time-varying channels is also discussed.
4) Extensive numerical results corroborate that the proposed offline algorithm can achieve similar performance as the standard CVX solver, while requiring only 0.69% of its computational time.

The rest part of this paper is organized as follows. Section II introduces the system models including the arrival process of delay-sensitive data and the energy consumption with a non-linear PA model. In Section III, we formulate the energy minimization problem. The optimal MDS algorithm and the HMDS algorithm are proposed in Section IV and Section V, respectively. In Section VI, online extension of the HMDS algorithm is presented. The experimental results are shown in Section VII, followed by the conclusion in Section VIII.

II. SYSTEM MODELS

A. ARRIVAL PROCESS OF DELAY-SENSITIVE DATA

Consider a point-to-point wireless link. We focus on a time period $[0, T]$ without loss of generality. The entire period is partitioned into N epochs, which are defined as the intervals between two adjacent time instants. The length of the ith epoch is $T_i = s_i - s_{i-1}$, $i = 1, \cdots, N$, where $0 = s_0 < s_1 < s_2 < \cdots < s_N = T$ denote the $(N+1)$ time instants.

The data arrive in the burst in amount $\mathcal{A} := \{A_0, A_1, A_2, \cdots, A_{N-1}, 0\}$ at time instant $\mathcal{T} := \{s_0, s_1, s_2, \cdots, s_{N-1}, s_N\}$. A_0 is the amount of initial data in the buffer of the

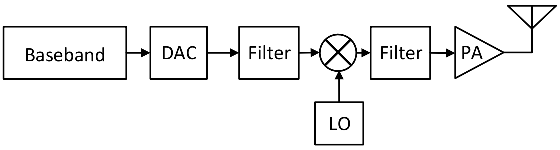


FIGURE 1. An illustration of the transmitter circuit.

transmitter at s_0. The amounts of data with deadlines due are collected in sequence $\mathcal{D} := \{0, D_1, D_2, \cdots, D_{N-1}, D_N\}$, where D_i is the amount of data must be delivered by s_i. It is worth noting that since the data generally are of different traffic types, we consider heterogeneous services here with different latency requirements. The data in the buffer should be re-shuffled once new data arrive, so that those with more stringent latency requirements are always placed head-of-line. It is obvious that we have $\sum_{i=0}^{N-1} A_i = \sum_{i=1}^{N} D_i$, that is, the total amount of data to be delivered is equal to that of data collected over the entire time period $[0, T]$.

B. ENERGY CONSUMPTION WITH A NON-LINEAR PA MODEL

The transmitter comprises a number of radio frequency (RF) components, e.g., the digital-to-analog converter (DAC), filters, local-oscillator (LO), mixer and PA; see Fig. 1. The DAC first converts the baseband signal to an analog signal, which is then filtered and modulated by the filters and mixer. The signal is finally amplified by the PA and delivered to the wireless channel.

To achieve a maximum energy reduction, the transmitter can switch into "sleep" (off) mode when there is no data to transmit to save circuit energy consumption. Denote P_{on} and P_{slp} as the power consumption when the transmitter is in "on" and "off" mode, respectively. The total power consumption of the transmitter when it is on P_{on} consists of three parts: the power consumed by the baseband for signal processing (including coding, digital modulation, etc.) P_{BB}, the total circuit power consumed by the RF components except the PA P_{RF}, and the power consumption of the PA P_{PA}, as given by

$$P_{\text{on}} = P_{\text{BB}} + P_{\text{RF}} + P_{\text{PA}}. \tag{1}$$

Here, the baseband power consumption $P_{\text{BB}} = P_{k1} r + P_{k2} n_c$ increases in proportion to the data rate r and the number of used subcarriers n_c [17], [18]. P_{k1} and P_{k2} are the constant coefficients. The RF chain power consumption P_{RF} is also set to be a constant [19].

Consider the PA with advanced Doherty technology [20], whose power efficiency in the high output power region increases linearly in dB scale [21], as show in Fig. 2. The corresponding power consumption can be approximately modelled as [16]:

$$P_{\text{PA}}(P_t) = \begin{cases} \dfrac{P_t}{\beta \cdot 10 \lg P_t + \delta}, & \text{if } P_{\text{th}} < P_t \le P_{\max}, \\ \eta \cdot P_t, & \text{if } 0 \le P_t \le P_{\text{th}}, \end{cases} \tag{2}$$

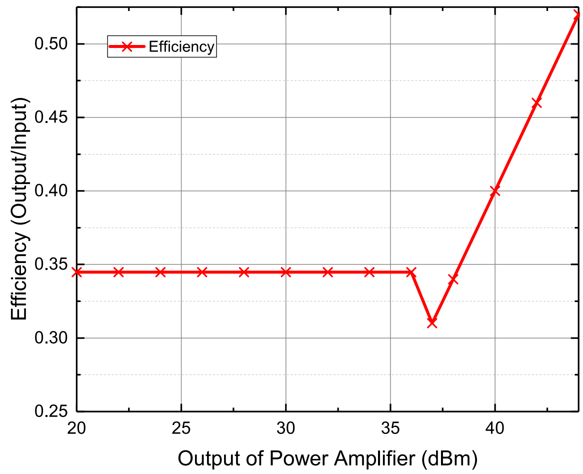


FIGURE 2. The efficiency of the modeled Doherty PA in equation (2).

where P_t is the transmit power, $\eta > 1$ is the inverse of PA's efficiency, β is a constant coefficient, and δ is a biasing factor. P_{th} and $P_{\max}$ are the threshold power and the maximum output power of the PA, respectively.

We consider M-QAM modulation in this paper, with M denoting the constellation size and $b = \log_2 M$ denoting the constellation order (that is, the number of bits per symbol). We have $b \in \mathbf{Z}^+$, where $\mathbf{Z}^+$ denotes the set of positive integers. The transmit power can be modelled as

$$P_t(b) = P_r(b) G, \tag{3}$$

where

$$P_r(b) = N_0 N_f b B \gamma(b) \tag{4}$$

is the received power, and

$$G = M_l G_l d^k \tag{5}$$

is the path-loss component of distance. In (4), N_0 presents the noise power spectral density, N_f denotes the noise figure of the receiver, $\gamma(b)$ is the per-bit SNR, and B is the system bandwidth. In (5), M_l is the link margin, G_l is the gain factor per unit distance, and d and k denote the transmission distance and path loss factor, respectively. Assuming that the symbol rate equals to the bandwidth B, we then have $r = Bb$ (in bit per second).

Given a BER P_e, the SNR for coherently detected M-QAM over additive white Gaussian noise (AWGN) channels can be approximated as [9]

$$\gamma(b) \approx \frac{2^b - 1}{b} \ln\left(\frac{3.10}{P_e}\right). \tag{6}$$

Consequently, the total energy consumption of the transmitter during epoch i is

$$E_{\text{total},i} = P_{\text{on},i} t_i + P_{\text{slp},i}(T_i - t_i), \tag{7}$$

where $P_{\text{on},i}$ and $P_{\text{slp},i}$ are the power consumed by the transmitter in "on" and "off" mode over epoch i, respectively, and $0 \le t_i \le T_i$ is the length of the "on" period in epoch i.

III. PROBLEM FORMULATION

Let $\boldsymbol{b} := \{b_1, \ldots, b_N\}$ collect the constellation orders selected for each epoch, and let $\boldsymbol{t} := \{t_1, \ldots, t_N\}$ collect the lengths of "on" periods in each epoch. The problem of interest is to determine the optimal set of $\{\boldsymbol{b}, \boldsymbol{t}\}$ such that the total energy consumed for delivering delay-sensitive data with a target BER $\sum_{i=1}^{N} E_{\text{total},i}$ is minimized. The energy consumption minimization problem can then be formulated as

$$\min_{\boldsymbol{b},\boldsymbol{t}} \sum_{i=1}^{N}[P_{k1}Bb_i + C_k + P_{\text{PA}}(b_i)]t_i + P_{\text{slp},i}T_i \quad (8a)$$
$$\text{s.t. } b_i \in \mathbf{Z}^+, \; 0 \le t_i \le T_i, \quad (8b)$$
$$\sum_{i=1}^{n}(Bb_it_i) \le \sum_{i=0}^{n-1} A_i, \quad (8c)$$
$$\sum_{i=1}^{n}(Bb_it_i) \ge \sum_{i=1}^{n} D_i, \; n = 1, \cdots, N, \quad (8d)$$
$$0 \le P_{t,i}(b_i) \le P_{\max}. \quad (8e)$$

Here, $P_{k1}Bb_i$ is the power used for channel coding and modulation mapping by the baseband, and $C_k := P_{k2}n_c + P_{\text{RF}} - P_{\text{slp},i}$ is a constant. (8c) is called the data causality constraints: the amount of data delivered $\sum_{i=1}^{n}(Bb_it_i)$ cannot be greater than that collected in the buffer $\sum_{i=0}^{n-1} A_i$ by any time instant t_n; (8d) presents the constraints of latency requirements: the amount of data delivered $\sum_{i=1}^{n}(Bb_it_i)$ must be no less than the data due to be transmitted to meet their deadlines, i.e., $\sum_{i=1}^{n} D_i$; and (8e) indicates that the transmit power cannot exceed the maximum power $P_{\max}$.

It can be observed that problem (8) is a mixed-integer programming problem, which is general difficult to solve. For example, as the number of epochs can be very large, and b_i may vary between epochs, it is complexity-prohibitive to solve the problem by exhaustive searching. We then relax $b_i \in \mathbf{Z}$ to $b_i \ge 0$ for tractability. Since $P_{\text{slp}}T_i$ is a constant, we remove it from the objective function. It is still challenging to solve this problem since the power consumption of the PA $P_{\text{PA}}(b_i)$ is non-continuous and non-differentiable [22].

In accordance with the inconsistent power efficiencies of the PA in different output power regions, we proceed to solve problem (8) in two stages. In the first stage, we solve the energy minimization problem when the PA has a linear efficiency, i.e., $0 \le P_t \le P_{\text{th}}$, and develop an optimal MDS scheme relying on convex relaxation and the Karush-Kuhn-Tucker (KKT) optimality conditions. On top of that, a heuristic HMDS scheme is proposed to adjust the MDS for $P_{\text{th}} < P_t \le P_{\max}$ in the second stage, where a quadratic function is obtained to approximate the non-linear PA model.

IV. PROPOSED OPTIMAL MODULATION AND DATA SCHEDULING ALGORITHM

A. CONVEX REFORMULATION AND OPTIMALITY CONDITIONS

Consider the PA has a linear efficiency. Problem (8) turns to:

$$\min_{\boldsymbol{b},\boldsymbol{t}} \sum_{i=1}^{N}[P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1)]t_i \quad (9a)$$
$$\text{s.t. } b_i \ge 0, \; 0 \le t_i \le T_i, \quad (9b)$$
$$\sum_{i=1}^{n}(Bb_it_i) \le \sum_{i=0}^{n-1} A_i, \quad (9c)$$
$$\sum_{i=1}^{n}(Bb_it_i) \ge \sum_{i=1}^{n} D_i, \; n = 1, \cdots, N, \quad (9d)$$
$$0 \le P_{t,i}(b_i) \le P_{\text{th}}, \quad (9e)$$

where $C_o := \eta N_0 N_f M_l G_l d^k B \ln(\frac{3.10}{P_e})$ is a constant. (9e) is equivalent to $0 \le b_i \le b_{\text{th}}$, where $b_{\text{th}} = \log_2(\frac{\eta P_{\text{th}}}{C_o} + 1)$.

In the relaxed problem (9), neither of b_it_i and $2^{b_i}t_i$ is standard convex or concave form in regard to (b_i, t_i). Nevertheless, problem (9) can be converted to standard convex programming through variables substitution. Define $\phi_i := b_it_i$ and $\boldsymbol{\phi} := \{\phi_1, \cdots, \phi_N\}$. Problem (9) can be rewritten into

$$\min_{\boldsymbol{\phi},\boldsymbol{t}} \sum_{i=1}^{N}[P_{k1}B\frac{\phi_i}{t_i} + C_k + C_o(2^{\frac{\phi_i}{t_i}} - 1)]t_i \quad (10a)$$
$$\text{s.t. } \phi_i \ge 0, \; 0 \le t_i \le T_i, \quad (10b)$$
$$\sum_{i=1}^{n}(B\phi_i) \le \sum_{i=0}^{n-1} A_i, \quad (10c)$$
$$\sum_{i=1}^{n}(B\phi_i) \ge \sum_{i=1}^{n} D_i, \; n = 1, \cdots, N. \quad (10d)$$

where $2^{\frac{\phi_i}{t_i}}t_i = 0$ if $t_i = 0$. For convex function 2^{b_i}, the term $2^{\frac{\phi_i}{t_i}}t_i$ is called its perspective, and is convex of (ϕ_i, t_i) [23]. Consequently, (10) is a convex problem. Note that we drop constraint (9e) here.

Let $\Lambda = \{\lambda_n, \mu_n, \forall n = 1, \cdots, N\}$, where λ_n and μ_n denote the Lagrange multipliers associated with the constraints of data causality (8c) and latency requirements (8d), respectively. The partial Lagrangian function of (10) is given by

$$L(\boldsymbol{\phi}, \boldsymbol{t}, \Lambda) = \sum_{i=1}^{N}[P_{k1}B\frac{\phi_i}{t_i} + C_k + C_o(2^{\frac{\phi_i}{t_i}} - 1)]t_i + \sum_{n=1}^{N}\lambda_n[\sum_{i=1}^{n}(B\phi_i) - \sum_{i=0}^{n-1}A_i] + \sum_{n=1}^{N}\mu_n[\sum_{i=1}^{n}D_i - \sum_{i=1}^{n}(B\phi_i)]$$

$$= C(\Lambda) + \sum_{i=1}^{N}\{[P_{k1}B\frac{\phi_i}{t_i} + C_k + C_o(2^{\frac{\phi_i}{t_i}} - 1)]t_i + B\phi_i \sum_{n=i}^{N}(\lambda_n - \mu_n)\} \quad (11)$$

where $C(\Lambda) := -\sum_{n=1}^{N}\lambda_n(\sum_{i=0}^{n-1}A_i) + \sum_{n=1}^{N}\mu_n(\sum_{i=1}^{n}D_i)$ for notation simplicity.

Let $(\boldsymbol{\phi}^*, \boldsymbol{t}^*)$ denote the optimal solution for (10), and let Λ^* collect the optimal Lagrange multipliers for the dual problem of (10). Define

$$w_i := \sum_{n=i}^{N}\{(\mu_n)^* - (\lambda_n)^*\}. \quad (12)$$

Resorting to the sufficient and necessary KKT optimality conditions [24], we have: $\forall i$,

$$(\phi_i^*, t_i^*) = \arg\min_{\phi_i \geq 0, 0 \leq t_i \leq T_n}\{[P_{k1}B\frac{\phi_i}{t_i} + C_k + C_o(2^{\frac{\phi_i}{t_i}} - 1)]t_i - B\phi_i w_i\}. \quad (13)$$

The complementary slackness conditions indicate that: $\forall n$,

$$\begin{cases}(\lambda_n)^* = 0, & \text{if } \sum_{i=1}^{n}(B\phi_i) < \sum_{i=0}^{n-1}A_i \\ \sum_{i=1}^{n}(B\phi_i) = \sum_{i=0}^{n-1}A_i, & \text{if } (\lambda_n)^* > 0.\end{cases} \quad (14)$$

$$\begin{cases}(\mu_n)^* = 0, & \text{if } \sum_{i=1}^{n}D_i < \sum_{i=1}^{n}(B\phi_i) \\ \sum_{i=1}^{n}D_i = \sum_{i=1}^{n}(B\phi_i), & \text{if } (\mu_n)^* > 0.\end{cases} \quad (15)$$

Let $b_i^* = \frac{\phi_i^*}{t_i^*}$ if $t_i^* > 0$, and let $b_i^* = 0$ if $t_i^* = 0$. Clearly $(\boldsymbol{b}^*, \boldsymbol{t}^*)$ is the optimal solution to (8).

Based on (13)–(15), we can obtain the sufficient and necessary optimality conditions for problem (9):

$$(b_i^*, t_i^*) = \arg\min_{b_i \geq 0, 0 \leq t_i \leq T_n}[P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1) - Bb_i w_i]t_i \quad (16)$$

$$\begin{cases}(\lambda_n)^* = 0, & \text{if } \sum_{i=1}^{n}(Bb_i t_i) < \sum_{i=0}^{n-1}A_i \\ \sum_{i=1}^{n}(Bb_i t_i) = \sum_{i=0}^{n-1}A_i, & \text{if } (\lambda_n)^* > 0.\end{cases} \quad (17)$$

$$\begin{cases}(\mu_n)^* = 0, & \text{if } \sum_{i=1}^{n}D_i < \sum_{i=1}^{n}(Bb_i t_i) \\ \sum_{i=1}^{n}D_i = \sum_{i=1}^{n}(Bb_i t_i), & \text{if } (\mu_n)^* > 0.\end{cases} \quad (18)$$

Given a positive t_i, the optimal constellation order b_i^* can be derived from (16), i.e.,

$$b_i^* = \arg\min_{b_i \geq 0}[P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1) - Bb_i w_i], \quad (19)$$

which is equivalent to: $P_w(b_i) := P_{k1} + \frac{C_o 2^{b_i^*}\ln 2}{B} = w_i$, or

$$b_i^* = \log_2 \frac{B(w_i - P_{k1})}{C_o \ln 2}. \quad (20)$$

It is obvious that b_i^* increases with w_i. Consequently, the optimal duration for the transmitter in "on" mode over epoch i is:

$$t_i^* = \arg\min_{0 \leq t_i \leq T_n}[C_k + C_o(2^{b_i^*} - 1 - b_i^* 2^{b_i^*}\ln 2)]t_i. \quad (21)$$

Now we introduce a bits-per-Joule EE-maximizing rate Bb_{ee}, where b_{ee} is defined as

$$b_{ee} = \arg\max_{b \geq 0} \frac{Bb}{P_{k1}Bb + C_k + C_o(2^b - 1)}. \quad (22)$$

Since the term on the right-hand side of (22) is concave-over-linear, it is a quasi-concave function and has a unique maximizer [24]; therefore, b_{ee} can be efficiently derived by a bisectional search [25].

According to (19), (21) and (22), we then establish the following two important lemmas.

Lemma 1 (Three Candidate Schemes for the Optimal Policy): The optimal modulation and data scheduling policy for (9) over epoch i can be chosen from one of the following three schemes: (i) "off" mode with $t_i^* = 0$, (ii) "on-off" mode with $b_i^* = b_{ee}$ and $t_i^* \leq T_i$, or (iii) "on" mode with $b_i^* > b_{ee}$ and $t_i^* = T_i$.

Proof: See Appendix A. ■

Lemma 1 indicates that the constellation orders smaller than b_{ee} should never be used in the optimal policy. An "on-off" strategy with $b_i^* = b_{ee}$ should always be considered first as it can consume less energy to transmit a given data amount. Only when the latency requirements are stringent, should we adopt $b_i^* > b_{ee}$ to deliver more data and meet the latency constraints; in such cases, the transmitter is in an "on" mode, i.e., $t_i^* = T_i$, over epoch i.

According to (20), and the complementary slackness conditions (17)–(18), we can obtain the specific structure of the optimal policy, as established in Lemma 2.

Lemma 2 (Specific Structure of the Optimal Policy): In the optimal policy for (9), b_i^* changes only at some s_n when the constraints of data causality and latency requirements are effective with equality. Particularly, b_i^* increases after s_n when $\sum_{i=1}^{n}(Bb_i^* t_i^*) = \sum_{i=0}^{n-1}A_i$, and decreases after s_n when $\sum_{i=1}^{n}(Bb_i^* t_i^*) = \sum_{i=1}^{n}D_i$.

Proof: See Appendix B. ■

Lemma 2 unveils that the optimal constellation order of the transmitter follows an interesting pattern. A constant b_i should always be adopted whenever possible. This is because

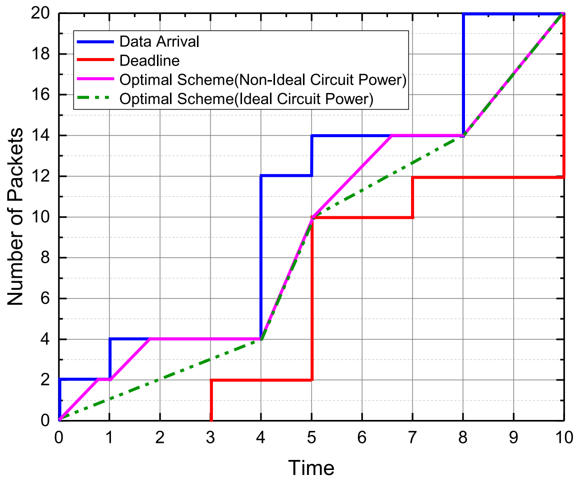


FIGURE 3. An illustration of the proposed MDS scheme.

$P_{\mathrm{on}}(b_i)$ is convex, and then a constant b_i can result in a minimum power consumption. The constellation order changes only when the constraints become active. An effective data causality constraint indicates that the data buffer is emptied at s_n, if the data arriving rate is relatively low; as a consequence, b_i adopted before s_n is smaller than that after. Likewise, an effective latency constraint indicates that the latency requirement is strict at s_n; therefore, b_i adopted before s_n should be larger than that after.

It is noteworthy that this offline schedule could be obtained by standard convex programming solvers. However, standard solvers designed for general convex problems would require a complexity higher than $\mathcal{O}(N^3)$ [24]. Also, the general-purpose solvers cannot unveil the underlying structure of the optimal modulation and data scheduling policy. To this end, the Lagrange multiplier method, in coupling with the KKT optimality conditions, is applied in this paper for a simpler and more insightful solution, which can guide the design of energy-efficient online scheduling as a benchmark.

B. VISUALIZATION OF MDS

We proceed to propose a new MDS algorithm, which generates the optimal modulation and data scheduling policy for delay-sensitive data, given the non-causal information of data arrivals and latency requirements.

FIGURE 3 depicts our proposed MDS procedure, where the data arrival curve $A_d(s)$ represents the accumulative amount of arrived data. The deadline curve $D_{\min}(s)$ represents the latency requirements of the arrived data. Specifically, it depicts the total amount of data that must be transmitted by s. The data arrival curve and thee deadline curve can be expressed as

$$A_d(s) = \sum_{i=0}^{N-1} [A_i u(s - s_i)], \tag{23}$$

$$D_{\min}(s) = \sum_{i=1}^{N} [D_i u(s - s_i)], \tag{24}$$

where $0 \leq s \leq T$ and $u(s)$ is the unit-step function: $u(s) = 1$ if $s \geq 0$, and $u(s) = 0$ otherwise.

A closed feasible solution region is presented. The data arrival curve $A_d(s)$ specifies the upper boundary of the feasible solution region, while the deadline curve $D_{\min}(s)$ specifies its lower boundary, so that the optimal transmission schedule can satisfy both data causality and latency requirements.

We can then specify the optimal data transmission curve $D^*(s)$ within the solution region. The slope of $D^*(s)$ denotes the optimal transmit rates $r^* = Bb_i^*$. The procedure is as follows.

1) Pass a string through the origin (0, 0) and the intersection of the upper and lower boundaries (i.e., $A_d(s)$ and $D_{\min}(s)$) at T, and then tauten the string between the boundaries until it bends only at some corners.
2) Compare the slope of each straight segment of the string $B\tilde{b}$ with Bb_{ee}.
   a) If the slope is larger than Bb_{ee}, set $B\tilde{b}$ as the optimal transmit rate, and $\tilde{b}$ as the optimal constellation order.
   b) If the slope is no larger than Bb_{ee}, set Bb_{ee} as the optimal transmit rate, and b_{ee} as the optimal constellation order.

Procedure 1 follows Lemma 2. Tautening a string tight between the boundaries ensures that the slope of the string increases after the string intersecting with the upper boundary, and decreases after it intersecting with the lower boundary. Note that such a string specifies the optimal transmission schedule in an ideal case [26], where the circuit power consumption is ignored; refer to the green dash line in FIGURE 3.

Procedure 2 follows Lemma 1 that $b_i^* \geq b_{ee}$. We set

$$\begin{cases} b_i^* = b_{ee}, t_i^* = \tilde{b}_i T_i / b_{ee}, & \text{if } \tilde{b}_i < b_{ee}, \\ b_i^* = \tilde{b}_i, t_i^* = T_i, & \text{if } \tilde{b}_i \geq b_{ee}, \end{cases} \tag{25}$$

where $B\tilde{b}_i$ is the slope of the string obtained in Step 1. The procedure in Step 2-a is most energy efficient, as for the epochs with $\phi = b_i^* t_i \geq b_{ee} t_i$, any "on-off" policy (b_i, t_i) with $b_i > b_i^*$ and $b_i t_i = \phi_i$ would cause more energy consumption, since

$$\begin{aligned} &[P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1)]t_i \\ &\quad = \phi_i \frac{P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1)}{b_i} \\ &\quad > \phi_i \frac{P_{k1}Bb_i^* + C_k + C_o(2^{b_i^*} - 1)}{b_i^*}, \end{aligned} \tag{26}$$

where the inequality holds because $\frac{P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1)}{b_i}$ is strictly increasing if $b_i \geq b_{ee}$.

The procedure in Step 2-b is optimal as the energy consumed by transmitting data of amount $B\phi_i = Bb_i t_i$ over epoch i is minimized by an "on-off" transmission with $b_{ee} \geq b_i$, since

$$\begin{aligned}
&[P_{k1}Bb_{ee} + C_k + C_o(2^{b_{ee}} - 1)]t_i^* \\
&= \phi_i \frac{P_{k1}Bb_{ee} + C_k + C_o(2^{b_{ee}} - 1)}{b_{ee}} \\
&= \phi_i \min_{b_i \geq 0} \frac{P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1)}{b_i} \\
&= \min_{b_i t_i = \phi_i} [P_{k1}Bb_i + C_k + C_o(2^{b_i} - 1)]t_i. \qquad (27)
\end{aligned}$$

C. DYNAMIC STRING TAUTENING ALGORITHM

The proposed offline procedure is summarized in Algorithm 1, which is later applied to yield the practical online scheme in Section VI.

Algorithm 1 Proposed MDS Algorithm

1: Input $\mathcal{A}$, $\mathcal{D}$ and $\mathcal{T}$, set $n_{\text{offset}} = 0$, $b_i^* = t_i^* = 0, \forall i$.
2: **while** $n_{\text{offset}} < N$ **do**
3: Calculate b_n^a and b_n^d, $n = n_{\text{offset}} + 1, \ldots, N$;
4: $b^- = 0, b^+ = \infty, \tau^- = \tau^+ = 0$;
5: $\tau = N, \tilde{b} = b_N^a = b_N^d$;
6: **for** $n = n_{\text{offset}} + 1$ to N **do**
7: **if** $b^+ \geq b_n^a$ **then**
8: $b^+ = b_n^a, \tau^+ = n$;
9: **end if**
10: **if** $b^- \leq b_n^d$ **then**
11: $b^- = b_n^d, \tau^- = n$;
12: **end if**
13: **if** $b^- \geq b^+$ **then**
14: **if** $\tau^+ \geq \tau^-$ **then**
15: $\tau = \tau^-, \tilde{b} = b^-$;
16: **else**
17: $\tau = \tau^+, \tilde{b} = b^+$;
18: **end if**
19: **end if**
20: **end for**
21: **for** $i = n_{\text{offset}} + 1$ to τ **do**
22: $b_i^* = \max\{b_{ee}, \tilde{b}\}$;
23: **end for**
24: find a feasible set of $\{t_i^*\}$ satisfying $\sum_{i=n_{\text{offset}}+1}^{\tau} t_i^* = \sum_{i=n_{\text{offset}}+1}^{\tau} \frac{\tilde{b}T_i}{b_i^*}$;
25: update $(\mathcal{A}, \mathcal{D}, \mathcal{T})$;
26: $n_{\text{offset}} = \tau$;
27: **end while**

In Steps 6 to 20, the constellation order changing time τ and the constellation order $\tilde{b}$ tentatively applied before time τ are determined in each iteration. They are specified by comparing and updating b^+ to the minimum rates $b_n^a = \frac{\sum_{i=0}^{n-1} A_i}{\sum_{i=1}^{n} L_i}$ obtained from the upper boundary $A_d(s)$ (from Step 7 to 9), and comparing and updating b^- to the maximum rates $b_n^d = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} L_i}$ obtained from the lower boundary $D_{\min}(s)$ (from Step 10 to 12), from index $n = n_{\text{offset}} + 1$ until $b^- \geq b^+$ (from Step 13 to 20). As a result, Steps 6 to 20 produce the exact string identified in Procedure 1 in Section IV-B.

Steps 21 to 24 implement Procedure 2 by setting the constellation order to be no smaller than b_{ee}. The lengths of the "on" periods of the transmitter are consequently determined, guaranteeing the total amount of transmitted data is unchanged. Note that the optimal policy may not be unique over the "on-off" epochs. We may have multiple feasible sets of $\{t_i^*\}$ to meet $\sum_{i=n_{\text{offset}}+1}^{\tau} t_i^* = \sum_{i=n_{\text{offset}}+1}^{\tau} \frac{\tilde{b}T_i}{b_i^*}$. In some cases, we can even allow $t_i^* = 0$ (i.e., turn off the transmitter) for some epochs, and perform the "on-off" data schedules over the remaining epochs.

After determining the optimal (b_i^*, t_i^*) for epochs $i, i \in [n_{\text{offset}}, \tau]$ in each iteration, we adjust $(\mathcal{A}, \mathcal{D}, \mathcal{T})$ by considering the time offset and the amount of data that have been transmitted. This procedure continues until the entire transmission schedule is derived.

Theorem 1 is readily established to assert the optimality and the efficiency of the proposed Algorithm 1.

Theorem 1: Algorithm 1 can yield the optimal transmission policy for (9) *with a complexity of* $\mathcal{O}(N^2)$.

Proof: See Appendix C. ■

The theorem is achieved by first proving that a Lagrange multiplier vector $\boldsymbol{\Lambda}^*$ exists, which guarantees that $(\boldsymbol{b}^*, \boldsymbol{t}^*)$ satisfies the sufficient and necessary conditions (16)–(18). It is also shown that $(\boldsymbol{b}^*, \boldsymbol{t}^*)$ ensures $t_i^* = T_i$ when $b_i^* > b_{ee}$, and $t_i^* \leq T_i$ when $b_i^* = b_{ee}$. As a result, $(\boldsymbol{b}^*, \boldsymbol{t}^*)$ is global optimal.

For each iteration that determines the optimal (b_i^*, t_i^*) for epochs $i, i \in [n_{\text{offset}}, \tau]$, we need to go through at most $(N - n_{\text{offset}})$ future time instants. Thus, the computational complexity of Algorithm 1 is $\mathcal{O}(N^2)$ in the worst case, where the optimal constellation order changes at every instant $s_i, i = 0, \cdots, N-1$, i.e., we need to calculate N optimal (b_i^*, t_i^*). In general, the optimal constellation order may remain unchanged over many epochs, and much fewer time instants are to be evaluated in the process. Therefore, the complexity of the proposed algorithm is often much lower than $\mathcal{O}(N^2)$. On the contrary, general-purpose convex programming solvers require high-order multiplications and many iterations, leading to slow convergence and a polynomial complexity higher than $\mathcal{O}(N^3)$.

V. HEURISTIC MDS ALGORITHM

Consider now the high output power region with non-linear PA efficiency. To deal with the non-convex function $f(P_{t,i}) := P_{\text{PA}}(P_{t,i})$ in (2) when $P_{\text{th}} < P_{t,i} \leq P_{\max}$, a quadratic function of $P_{t,i}$ is obtained by using Taylor expansion to approximate it at the middle point $P_m = \frac{P_{\text{th}}+P_{\max}}{2}$. We have

$$\begin{aligned}
f(P_{t,i}) &\approx f_a(P_{t,i}) \\
&= f(P_m) + \frac{f'(P_m)}{1!}(P_{t,i} - P_m) \\
&\quad + \frac{f''(P_m)}{2!}(P_{t,i} - P_m)^2, \qquad (28)
\end{aligned}$$

where $f'(\cdot)$ and $f''(\cdot)$ are the first and second derivatives of $f(\cdot)$, respectively. The high order terms $\sum_{i=3}^{+\infty} \frac{f^{(n)}(P_m)}{n!}(P_{t,i} - P_m)^n$ can be ignored, since the co-efficient, $\frac{f^{(n)}(P_m)}{n!}$, converges to zero fast when n goes to infinity [16]. For the quadratic function $f_a(P_{t,i})$, it is easy to find a minimizer $P^*_{t,i}$. If $P^*_{t,i}$ is smaller than P_{th} or larger than $P_{\max}$, we check P_{th} and $P_{\max}$ for the one minimizing $f_a(P_{t,i})$. As $P_{t,i}(b_i) = \frac{C_o(2^{b_i}-1)}{\eta}$ is monotonically increasing with b_i, we can find a unique $\tilde{b}^*_i \in [b_{\text{th}}, b_{\max}]$ that minimizes P_{PA}, where $b_{\max} = \log_2(\frac{\eta P_{\max}}{C_o} + 1)$.

We then propose a heuristic MDS algorithm for problem (8) based on Algorithm 1. When the proposed $b^*_i \in [b_{\text{th}}, b_{\max}]$ in Algorithm 1, let $b^*_i = \max\{b^*_i, \tilde{b}_i^*\}$, and recalculate the required transmission time accordingly. The proposed HMDS algorithm is summarized in Algorithm 2.

Algorithm 2 Proposed HMDS Algorithm

1: **while** there is data to transmit **do**
   Calculate $\tilde{b}^*_i \in [b_{\text{th}}, b_{\max}]$ that minimizes $f_a(P_{t,i})$;
2: run Algorithm 1 and obtain $\{b^*_i, t^*_i\}$ and $\Delta_i = b^*_i t^*_i$, $i = 1, 2, \cdots, N$;
3: **if** $b_{\text{th}} \le b^*_i \le b_{\max}$ **then**
4: $b^*_i = \max\{b^*_i, \tilde{b}^*_i\}$;
5: **end if**
6: **if** $b^*_i > b_{\max}$ **then**
7: error 'infeasible';
8: **end if**
9: $b^*_i = \lceil b^*_i \rceil$, $t^*_i = \frac{\Delta_i}{b^*_i}$;
10: transmit the data with the modulation and data scheduling policy $\{b^*_i, t^*_i\}$, $i = 1, 2, \cdots, N$;
11: **end while**

In Step 9 of Algorithm 2, $\lceil b^*_i \rceil$ denotes the smallest integer no less than b^*_i (a.k.a, the ceiling operator). Note that problem (8) can be infeasible, when the latency requirement is too stringent such that the transmit power exceeds its maximum value. Once the infeasibility happens, the proposed algorithm terminates and outputs the error message 'infeasible'. Clearly the complexity of Algorithm 2 is still $\mathcal{O}(N^2)$.

VI. ONLINE EXTENSION OF THE HMDS ALGORITHM

When developing the HMDS algorithm, we assumed non-causal information about data arrivals. Considering it is impractical to have a-priori knowledge on data arrivals, we proceed to generalize the offline HMDS algorithm to online scenarios where only current data arrival information is available. The main idea is to transmit the arrived data using the HMDS algorithm with current data arrival information, and reschedule the transmission once new data arrive.

When new data arrive, we set the current time instant as s_0, and set the last time instant by which all the buffered data must be delivered as s_N. In this case, we have $\mathcal{A} = \{A_0, 0, \ldots, 0\}$, $\mathcal{D} = \{0, D_1, \ldots, D_N\}$ measured at time $\mathcal{T} = \{s_0, \ldots, s_N\}$, and $\sum_{i=1}^N D_i = A_0$. We then run the proposed HMDS algorithm for this $(\mathcal{A}, \mathcal{D}, \mathcal{T})$ system, and obtain the optimal transmission strategy over time $[s_0, s_N]$. Adopt the optimal strategy for data transmission, until a new data arrival occurs at $s_i < s_N$.

Then we take s_i as the new initial time instant, and update $(\mathcal{A}, \mathcal{D}, \mathcal{T})$ by considering the time offset, remaining data in the buffer and new latency requirements of arrived data. The optimal transmission strategy is also reconsidered for the time instants after s_i by using the proposed HMDS algorithm. This procedure is repeatedly conducted, until there is no more data to deliver. The proposed online scheme is summarized in Algorithm 3.

Algorithm 3 Proposed Online Scheduling based on the HMDS Algorithm

while there is data to transmit **do**
  if a new data arrival occurs at the current instant **then**
    set the current instant as s_0, and the last time instant by which all the buffered data must be delivered as s_N;
    update $(\mathcal{A}, \mathcal{D}, \mathcal{T})$;
    run Algorithm 2 to update the transmission strategy over $[s_0, s_N]$;
  end if
  transmit the data following the updated transmission strategy;
end while

The online scheme may degrade the performance compared to Algorithm 2. When new data arrive at s_i during $[s_0, s_N]$, new transmission strategy is considered for the time instants after s_i. This may cause the violation of Lemma 2, where a specific pattern of the optimal policy is revealed. Note that Lemma 2 is established with a-priori knowledge on data arrivals and latency requirements for the offline scenario. Due to the unavailability of the future knowledge in the online case, it is not possible to develop an online scheme without violating Lemma 2. Nevertheless, when no data arrive during $[s_0, s_N]$, the online scheme can achieve the same performance as Algorithm 2, providing a well-structured way for practical data transmissions.

The proposed HMDS Algorithm can be readily extended for online implementations in time-varying channels, e.g., a flat fading Rayleigh channel. The average SNR in a Rayleigh channel for M-QAM is given by [27]

$$\overline{\gamma(b)} \approx \frac{1}{6P_e}\frac{(2^b - 1)}{b} \tag{29}$$

Substituting $\gamma(b)$ in (6) with $\overline{\gamma(b)}$, we can obtain similar rules as in Algorithms 1 and 2 to generate the modulation and data scheduling policy for Rayleigh fading channels. During each epoch, the transmitter can send data to the receiver with a certain modulation size. The receiver can feed back ACK to confirm the successful reception of the data, and feed back the estimated CSI to the transmitter through a feedback channel. The adaptive modulation and scheduling

TABLE 1. Detailed simulation parameters.

Parameter	Value	Parameter	Value
Channel Model	AWGN	P_e	10^{-4}
B	150 kHz	$P_{\rm slp}$	0 mW
P_{k1}	4×10^{-7}	P_{k2}	0.96
β	0.03	δ	0.1
$P_{\rm RF}$	40 mW	η	2.9
N_0	-204 dBJ	N_f	10 dB
M_l	40 dB	G_l	30 dB
d	30 m	k	3.5
$P_{\rm th}$	6 dB	$P_{\rm max}$	70 dB

controller at the transmitter then determines modulation and scheduling schemes based on the received CSI and current data information. In the online implementation, all steps are inherited from Algorithms 1 and 2. The only difference is that the amount of unsuccessfully delivered data (due to fading channels), which is not confirmed by ACK, needs to be added to the arrived data $\mathcal{A}$ and the deadline-approaching data $\mathcal{D}$ in the new system. The unsuccessfully delivered data can be re-transmitted as part of new and undelivered data. The online algorithm for fading channels is optimal in the case that all messages can be successfully delivered at the first transmission attempts.

VII. EXPERIMENTAL RESULTS

In this section, we carry out simulations to evaluate the proposed algorithms. The detailed parameters used in the simulations are listed in Table 1. The data arrivals are modelled as Poisson processes. The average rate of data arrival is set to 18 kbits per second (kbps), unless otherwise specified. We assume all data have the same latency requirement (that is, the maximum latency allowed is the same). Note that the proposed algorithms are applicable to any stochastic data arrival processes with different latency requirements.

We compare our proposed offline algorithm (i.e., the HMDS Algorithm) and online algorithm with two benchmarks. One is "CVX tool" solving (10) and substituting Algorithm 1 in the HMDS by the standard MATLAB CVX toolbox. The other one is a heuristic offline method stemmed from the "water-level tautening" approach in [26], where the circuit power consumption and non-linear efficiency of PA are overlooked.

FIGURE 4 plots the CPU running time of the proposed offline and online algorithms, the CVX tool and the heuristic method, where the transmission interval T ranges from 10 to 80 seconds. It is obvious that the CPU time required for the algorithms increase with growth of the transmission interval. It is also observed that when T is large, the proposed offline

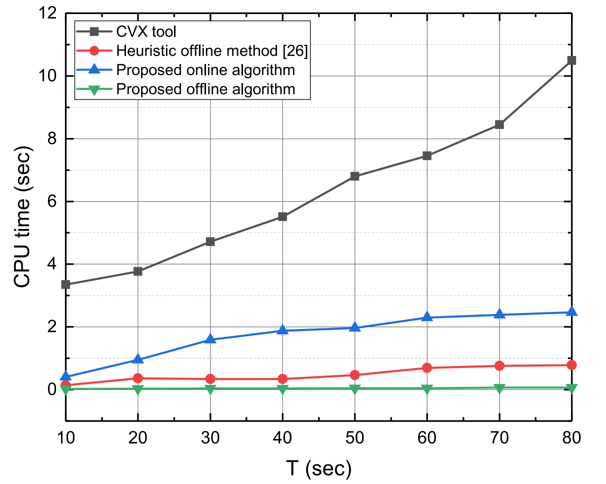


FIGURE 4. CPU running time of different algorithms versus transmission interval T. The average data arrival rate is 18 kbps and the deadline requirement is 2 seconds. Our proposed algorithms are much more computationally efficient than the CVX tool.

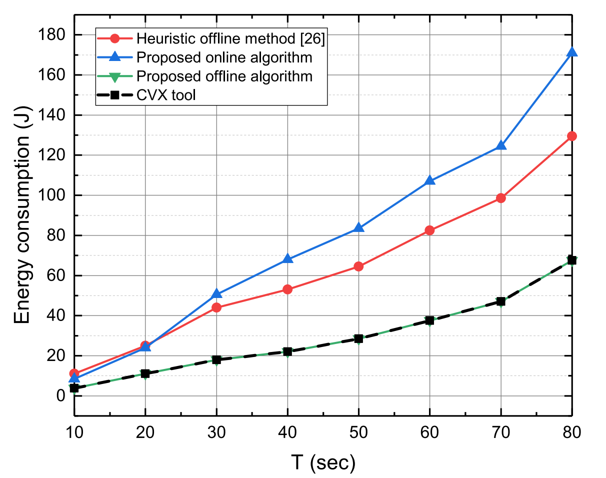


FIGURE 5. Energy consumption of different algorithms versus transmission interval T. The average data arrival rate is 18 kbps and the deadline requirement is 2 seconds.

and online algorithms only require about 0.69% and 27.67% of the CPU time with the CVX tool, respectively. As mentioned in Section IV, the proposed algorithms can produce the optimal schedule directly according to the optimality conditions, and lead to a complexity of $\mathcal{O}(N^2)$ in the worst case. In contrast, the CVX tool uses the interior point methods designed for general convex optimization problems, which has a complexity higher than $\mathcal{O}(N^3)$. It is corroborated that our proposed algorithms are more computationally efficient than the CVX tool.

FIGURE 5 depicts the energy consumption of different algorithms as T increases. As expected, with an average data arrival rate of 18 kbps, the energy consumption of all the four algorithms increases as T becomes larger. We can also see that the energy consumption of the proposed offline algorithm is exactly the same as that of the CVX tool, which validates the optimality of Algorithm 1.

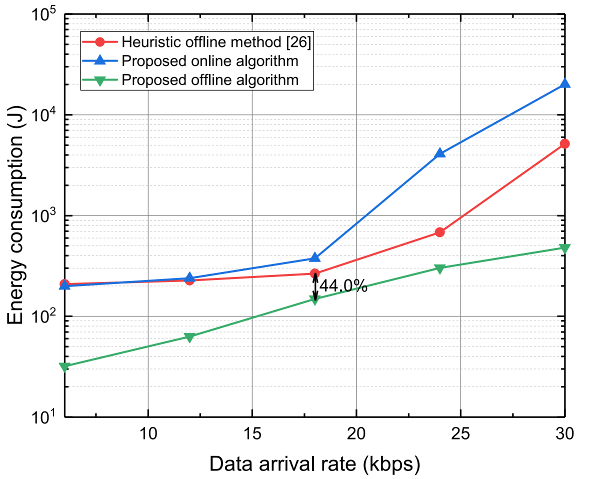


FIGURE 6. **Energy consumption of different algorithms versus data arrival rate. The transmission interval *T* is 200 seconds, and the deadline requirement is 2 seconds.**

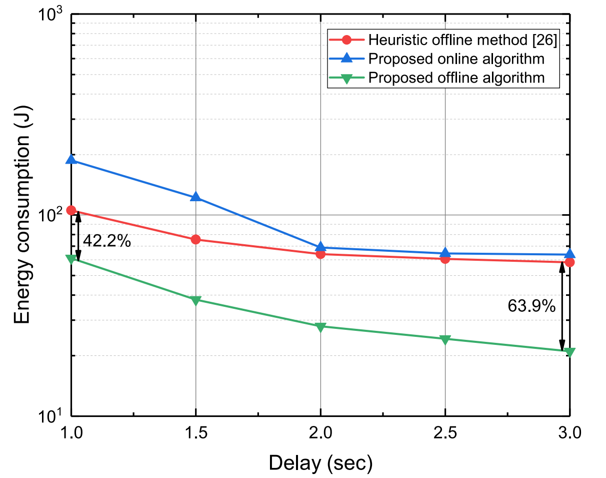


FIGURE 7. **Energy consumption of different algorithms versus latency requirement. The transmission interval *T* is 50 seconds, and the average data arrival rate is 18 kbps.**

As expected, FIGURE 5 shows that the proposed online algorithm can increase the energy consumption compared to its offline counterpart. This is caused by the unavailability of the future information on data arrivals. Consequently, the online constellation order changes more frequently than the offline one, resulting in larger energy consumption. The impact of circuit power consumption on the transmission strategy is also revealed in FIGURE 5. It can be seen that the heuristic offline method consumes more energy than our proposed one, as the former assumes negligible circuit power consumption and linear PA efficiency, and keeps the PA active over the entire transmission interval, leading to a significant energy loss.

The energy consumption of different transmission schemes is compared in FIGURE 6 under different data arrival rates. The transmission interval T is set to 200 seconds. As we can see, the proposed online algorithm consumes the most energy, followed by the heuristic method and the proposed offline algorithm. The proposed offline algorithm can save at least 44% energy consumption compared to the heuristic algorithm. Moreover, as the data arrival rate grows, the energy consumption of all algorithms increases. This is because as the data arrival rate becomes higher, more amount of data need to be transmitted within T, resulting in larger energy consumption; on the other hand, as more data need to be delivered before each deadline, the transmission becomes more urgent and constellation order larger than b_{ee} is to be more frequently used. This further leads to larger energy consumption.

The energy consumption of different transmission schemes is also compared in FIGURE 7, under different latency requirements. It can be observed that, as the latency requirement becomes looser (i.e., the delay becomes larger), the energy consumption of all the three algorithms decreases. This is because with loose latency requirement, our algorithms can, to the most extent, apply the "on-off" strategy with the most energy-efficient constellation order b_{ee} for transmission. It is also observed that our proposed offline algorithm can save about 42.2%-63.9% energy consumption compared to the heuristic one. The advantage of the proposed offline algorithm in terms of energy reduction over the heuristic method become more and more significant, as the delay grows.

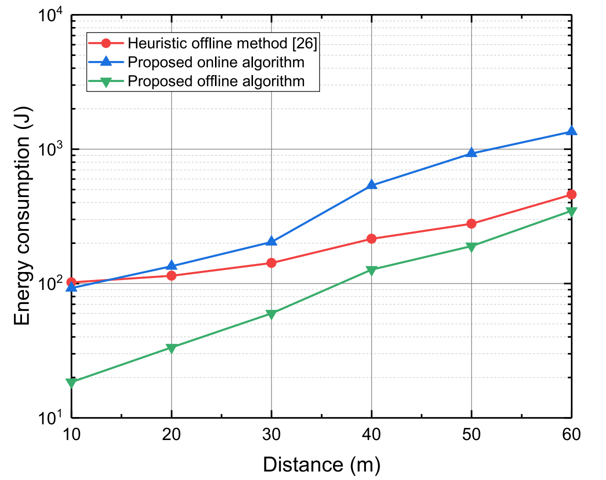


FIGURE 8. **Energy consumption of different algorithms versus transmission distance *d*. The transmission interval *T* is 50 seconds, the average data arrival rate is 18 kbps and the deadline requirement is 2 seconds.**

FIGURE 8 compares the energy consumption of different algorithms under different transmission distances. It is observed that our proposed offline algorithm always outperforms the heuristic one, resulting in a 40.8% energy saving on average. We can also see that the energy consumption of all schemes grow as the transmission distance increases. Given a target BER, the longer the transmission distance, the larger the path loss. A larger path loss leads to higher transmit power, consequently resulting in a larger total energy consumption.

VIII. CONCLUSION

In this paper, we proposed the new two-stage based HMDS algorithm to generate the energy-efficient modulation and data scheduling schemes for delay-sensitive data, where non-negligible circuit power and non-linear PA efficiency were taken into account. The optimal MDS was first developed based on convex relaxation and the resultant optimality conditions, and reveals the specific structure of the optimal policy. The HMDS scheme was further proposed for the PA working in the non-linear region. The offline HMDS algorithm was then extended to practical online scenarios in a well-structured way. Simulation showed that the proposed offline algorithm can achieve the exactly same performance as the standard CVX solver, while requiring only 0.69% of its computational time.

APPENDIXES

APPENDIX A

PROOF OF LEMMA 1

Define $\eta_{ee}(b) := \frac{P_{k1}Bb+C_k+C_o(2^b-1)}{Bb}$. The first derivative of $\eta_{ee}(b)$ is:

$$\frac{d\eta_{ee}(b)}{db} = \frac{C_oB(2^b b\ln 2 - 2^b + 1) - C_kB}{B^2b^2}. \tag{30}$$

As $\eta_{ee}(b)$ is "convex-over-linear", it first decreases and then increases with b, and achieves the minimum at b_{ee}. Therefore,

$$\begin{cases} C_oB(2^b b\ln 2 - 2^b + 1) - C_kB < 0, & \text{if } b < b_{ee}, \\ C_oB(2^b b\ln 2 - 2^b + 1) - C_kB = 0, & \text{if } b = b_{ee}, \\ C_oB(2^b b\ln 2 - 2^b + 1) - C_kB > 0, & \text{if } b > b_{ee}. \end{cases} \tag{31}$$

If there exists a $b_i^* < b_{ee}$ when $t_i^* > 0$, it follows from (31) that $C_oB(2^{b_i^*} b_i^* \ln 2 - 2^{b_i^*} + 1) - C_kB < 0$. But when $C_oB(2^{b_i^*} b_i^* \ln 2 - 2^{b_i^*} + 1) - C_kB < 0$, (21) implies that $t_i^* = 0$, leading to a contradiction. Thus, $b_i^* < b_{ee}$ should never be adopted when $t_i^* > 0$.

If $b_i^* > b_{ee}$, we have $C_oB(2^{b_i^*} b_i^* \ln 2 - 2^{b_i^*} + 1) - C_kB > 0$ based on (31), and (21) implies that $t_i^* = T_i$. If $b_i^* = b_{ee}$, we have $C_oB(2^{b_i^*} b_i^* \ln 2 - 2^{b_i^*} + 1) - C_kB = 0$, and any $t_i^* \in [0, T_i]$ can be selected for the optimal policy.

APPENDIX B

PROOF OF LEMMA 2

It is clear that $b_i^* = \log_2 \frac{B(w_i - P_{k1})}{C_o\ln 2}$ changes only with w_i. For w_i defined in (12), if $(\lambda_n)^*, (\mu_n)^* = 0, \forall n = 1, \ldots, N-1$, a constant $w = (\mu_N)^* - (\lambda_N)^*$ is to be adopted over all epochs. A change of w_i occurs only when a Lagrange multiplier is positive at a time instant $s_n, n \in [1, N-1]$. Based on the complementary slackness conditions (17)-(18), at such a s_n, the constraints of data causality or latency requirements are met with equality.

If the constellation order changes at a certain s_n when $\sum_{i=1}^{n}(Bb_i^*t_i^*) = \sum_{i=0}^{n-1} A_i$, then the corresponding $(\lambda_n)^* > 0$. For epoch n and $n+1$, we have $w_n = \sum_{l=n}^{N}[(\mu_l)^* - (\lambda_l)^*]$, and $w_{n+1} = \sum_{l=n+1}^{N}[(\mu_l)^* - (\lambda_l)^*]$, respectively. Hence, $w_{n+1} - w_n = (\lambda_n)^* > 0$. It can then be concluded that the constellation order increases after such s_n since $\log_2 \frac{B(w_i-P_{k1})}{C_o\ln 2}$ increases with w_i.

If the constellation order changes at a certain s_n when $\sum_{i=1}^{n}(Bb_i^*t_i^*) = \sum_{i=1}^{n} D_i$, then $(\mu_n)^* > 0$. Similarly, we have $w_{n+1} - w_n = -\mu_n^* < 0$, which implies that the constellation order decreases after such s_n.

APPENDIX C

PROOF OF THEOREM 1

Given the procedure in Algorithm 1, it is shown that the changing pattern of the optimal transmission strategy $(\boldsymbol{b}^*, \boldsymbol{t}^*)$ generated by Algorithm 1 is consistent with Lemma 2, i.e., (i) if the constellation order applied is first b and then changed to $\tilde{b}$ at s_τ where $\sum_{i=1}^{\tau}(Bbt_i^*) = \sum_{i=0}^{\tau-1} A_i$, then we have $\tilde{b} > b$; and (ii) if b is changed at s_τ where $\sum_{i=1}^{\tau}(Bbt_i^*) = \sum_{i=1}^{\tau} D_i$, then we have $\tilde{b} < b$.

Suppose that the constellation order changes M times at instants $\{s_{\tau_1}, s_{\tau_2}, \ldots, s_{\tau_M}\}$. We separate the whole transmission policy into $M+1$ phases: constellation order $b_i^* = \check{b}_1$ over epochs $i \in [1, \tau_1]$, $b_i^* = \check{b}_2$ over epochs $i \in [\tau_1 + 1, \tau_2], \ldots, b_i^* = \check{b}_{M+1}$ over epochs $i \in [\tau_M + 1, N]$. Then define a set of Lagrange multipliers $\boldsymbol{\Lambda}^* := \{(\lambda_n)^*, (\mu_n)^*, n = 1, \ldots, N\}$ as follows:

For convenience, let $\Delta_1 := P_w(\check{b}_{m+1}) - P_w(\check{b}_m)$. For a certain $\tau_m, \forall m = 1, \ldots, M$,

1) if $\sum_{i=1}^{\tau_m}(Bb_i^*t_i^*) = \sum_{i=0}^{\tau_m-1} A_i$, then

$$(\lambda_{\tau_m})^* = \Delta_1;$$

2) if $\sum_{i=1}^{\tau_m}(Bb_i^*t_i^*) = \sum_{i=1}^{\tau_m} D_i$, then

$$(\mu_{\tau_m})^* = -\Delta_1.$$

We have proven that the constellation order $\check{b}_{m+1} > \check{b}_m$ if the data causality constraint is tight at s_{τ_m}, and $\check{b}_{m+1} < \check{b}_m$ if the latency requirement constraint is tight at s_{τ_m}. As $P_w(b)$ increases with b, we have $(\lambda_{\tau_m})^* > 0$ or $(\mu_{\tau_m})^* > 0$, when a certain constraint is tight at s_{τ_m}. In addition, let $(\mu_N)^* = P_w(\check{b}_{M+1}) > 0$. Except these $M+1$ positive $(\mu_N)^*$, $(\lambda_{\tau_m})^*$ and $(\mu_{\tau_m})^*$, other elements in $\boldsymbol{\Lambda}^*$ are set to zero.

With such a $\boldsymbol{\Lambda}^*$, the complementary slackness conditions (17)-(18) clearly hold. Using such a $\boldsymbol{\Lambda}^*$ leads to $w_i := \sum_{n=i}^{N}[(\mu_n)^* - (\lambda_n)^*] = P_w(\check{b}_m), \forall i \in [\tau_{m-1}+1, \tau_m]$ (with $\tau_0 := 1$ and $\tau_{M+1} := N$). This implies that $b_i^* = \check{b}_m = \log_2 \frac{B(w_i-P_{k1})}{C_o\ln 2}, \forall i \in [\tau_{m-1}+1, \tau_m]$. Moreover, the construction of the optimal schedule ensures $t_i^* = T_i$ when $b_i^* = \check{b}_m > b_{ee}$, and obtains a feasible set of $t_i^* \le T_i$ when $b_i^* = \check{b}_m = b_{ee}$ in each phase m. This ensures that every (b_i^*, t_i^*) satisfies (16); hence, $(\boldsymbol{b}^*, \boldsymbol{t}^*)$ follows Lemma 1.

REFERENCES

[1] G. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: Tutorial, survey, and open issues," *IEEE Wireless Commun.*, vol. 18, no. 6, pp. 28–35, Dec. 2011.

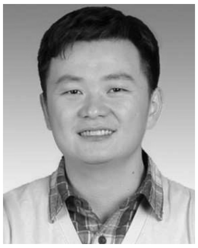
[2] Z. Nan, T. Chen, X. Wang, and W. Ni, "Energy-efficient transmission schedule for delay-limited bursty data arrivals under nonideal circuit power consumption," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6588–6600, Aug. 2016.

[3] X. Chen, X. Wang, and Y. Sun, "Energy-harvesting powered transmissions of bursty data packets with strict deadlines," in *Proc. IEEE Int. Conf. Commun.(ICC)*, Jun. 2014, pp. 4060–4065.

[4] X. Chen, W. Ni, X. Wang, and Y. Sun, "Optimal quality-of-service scheduling for energy-harvesting powered wireless communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3269–3280, May 2016.
[5] X. Chen, W. Ni, X. Wang, and Y. Sun, "Provisioning quality-of-service to energy harvesting wireless communications," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 102–109, Apr. 2015.
[6] T. Shu, H. A. B. Salameh, and M. Krunz, "WSN11-5: Cross-layer optimization of a CSMA protocol with adaptive modulation for improved energy efficiency in wireless sensor networks," in *Proc. IEEE Globecom*, Nov. 2006, pp. 1–5.
[7] C.-E. Weng, J.-M. Zhang, and H.-L. Hung, "An efficient power control scheme and joint adaptive modulation for wireless sensor networks," *Comput. Electr. Eng.*, vol. 40, no. 2, pp. 641–650, Feb. 2014.
[8] P. Mukherjee and S. De, "Dynamic feedback-based adaptive modulation for energy-efficient communication," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 946–949, May 2019.
[9] R. Jaouadi, G. Andrieux, J.-Y. Baudais, and J.-F. Diouris, "Rate optimization for energy efficient system with M-QAM," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Jan. 2017, pp. 975–979.
[10] S. Zhang, Y. Chen, and S. Xu, "Improving energy efficiency through bandwidth, power, and adaptive modulation," in *Proc. IEEE 72nd Veh. Technol. Conf. (Fall)*, Sep. 2010, pp. 1–5.
[11] Q. Sun, L. Li, A. Tolli, M. Juntti, and J. Mao, "Optimal energy efficient bit and power loading for multicarrier systems," *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1194–1197, Jul. 2014.
[12] K. Wang, M. Tao, W. Chen, and Q. Guan, "Delay-aware energy-efficient communications over Nakagami-*m* fading channel with MMPP traffic," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 3008–3020, Aug. 2015.
[13] J. Choi and J. Ha, "On the energy efficiency of AMC and HARQ-IR with QoS constraints," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3261–3270, Sep. 2013.
[14] S. Zhao, Z. Zeng, C. Feng, F. Liu, and Y. Nie, "An adaptive polarization-QAM modulation scheme for improving the power amplifier energy efficiency in OFDM systems," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–6.
[15] S. Zhang, S. Xu, G. Y. Li, and E. Ayanoglu, "First 20 years of green radios," *IEEE Trans. Green Commun. Netw.*, to be published.
[16] S. Zhang, C. Xiang, S. Cao, S. Xu, and J. Zhu, "Dynamic carrier to MCPA allocation for energy efficient communication: Convex relaxation versus deep learning," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 628–640, Sep. 2019.
[17] C. Isheden and G. P. Fettweis, "Energy-efficient multi-carrier link adaptation with sum rate-dependent circuit power," in *Proc. IEEE Global Telecommun. Conf. GLOBECOM*, Dec. 2010, pp. 1–6.
[18] H. S. Kim and B. Daneshrad, "Energy-constrained link adaptation for MIMO OFDM wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2820–2832, Sep. 2010.
[19] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1089–1098, Aug. 2004.
[20] H. Kang, H. Lee, H. Oh, W. Lee, C. S. Park, K. C. Hwang, K. Y. Lee, and Y. Yang, "Symmetric three-way doherty power amplifier for high efficiency and linearity," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 64, no. 8, pp. 862–866, Aug. 2017.
[21] R. Darraji, F. M. Ghannouchi, and O. Hammi, "A dual-input digitally driven doherty amplifier architecture for performance enhancement of doherty transmitters," *IEEE Trans. Microw. Theory Techn.*, vol. 59, no. 5, pp. 1284–1293, May 2011.
[22] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, vol. 3. New York, NY, USA: Springer, 2012.
[23] X. Wang and G. B. Giannakis, "Resource allocation for wireless multiuser OFDM networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4359–4372, Jul. 2011.
[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[25] J. Xu and R. Zhang, "Throughput optimal policies for energy harvesting wireless transmitters with non-ideal circuit power," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 322–332, Feb. 2014.
[26] X. Wang and Z. Li, "Energy-efficient transmissions of bursty data packets with strict deadlines over time-varying wireless channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2533–2543, May 2013.
[27] S. Cui, A. J. Goldsmith, and A. Bahai, "Modulation optimization under energy constraints," in *Proc. ICC*, vol. 4, 2003, pp. 2805–2811.

QIAN CHEN received the B.E. degree from Shanghai University China, in 2017, where she is currently pursuing the master's degree in information and communication engineering. Her research fields include MIMO detection, machine learning, deep learning in the PHY layer, and energy-efficient communication networks.

XIAOJING CHEN (Member, IEEE) received the B.E. degree in communication science and engineering and the Ph.D. degree in electromagnetic field and microwave technology from Fudan University, China, in 2013 and 2018, respectively, and the Ph.D. degree in engineering from Macquarie University, Australia, in 2019. She is currently a Lecturer with Shanghai University, China. Her research interests include wireless communications, energy-efficient communications, stochastic network optimization, and network functions virtualization.



SHUNQING ZHANG (Senior Member, IEEE) received the B.S. degree from the Department of Microelectronics, Fudan University, Shanghai, China, in 2005, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2009.

He was with the Communication Technologies Laboratory, Huawei Technologies, as a Research Engineer, and then a Senior Research Engineer, from 2009 to 2014, and a Senior Research Scientist of the Intel Collaborative Research Institute on Mobile Networking and Computing, Intel Labs, from 2015 to 2017. Since 2017, he has been with the School of Communication and Information Engineering, Shanghai University, Shanghai, China, as a Full Professor. His current research interests include energy efficient 5G/5G+ communication networks, hybrid computing platform, and joint radio frequency and baseband design. He has published over 60 peer-reviewed journal articles and conference papers, and more than 50 granted patents. He has received the National Young 1000-Talents Program and won the Paper Award for Advances in Communications from the IEEE Communications Society, in 2017.

SHUGONG XU (Fellow, IEEE) graduated from Wuhan University, China, in 1990. He received the master's degree in pattern recognition and intelligent control from the Huazhong University of Science and Technology (HUST), China, in 1993, and the Ph.D. degree in EE from HUST, in 1996.

In September 2013, he was the Research Director and a Principal Scientist with the Communication Technologies Laboratory, Huawei Technologies. He was the Center Director and the Intel Principal Investigator of the Intel Collaborative Research Institute for Mobile Networking and Computing (ICRI-MNC), prior to December 2016, when he joined Shanghai University. Among his responsibilities at Huawei, he founded and directed Huawei's green radio research program, Green Radio Excellence in Architecture and Technologies (GREAT). He is currently a Professor with Shanghai University and the Head of the Shanghai Institute for Advanced Communication and Data Science (SICS). His current research interests include wireless communication systems and machine learning.

He was also the Chief Scientist and PI for the China National 863 project on End-to-End Energy Efficient Networks. He was one of the co-founders of the Green Touch consortium together with Bell Labs etc. He has served as the Co-Chair of the Technical Committee for three terms in this international consortium. Prior to joining Huawei, in 2008, he was with Sharp Laboratories of America as a Senior Research Scientist. Before that, he conducted research as a Research Fellow with the City College of New York, Michigan State University, and Tsinghua University. He has published more than 100 peer-reviewed research articles in top international conferences and journals. One of his most referenced articles has more than 1400 Google Scholar citations, in which the findings were among the major triggers for the research and standardization of the IEEE 802.11S. He has more than 20 U.S. patents granted. Some of these technologies have been adopted in international standards, including the IEEE 802.11, 3GPP LTE, and DLNA. He was awarded National Innovation Leadership Talent by Chinese Government, in 2013, was elevated to the IEEE Fellow, in 2015, for contributions to the improvement of wireless networks efficiency. He is also the winner of the 2017 Award for Advances in Communication from the IEEE Communications Society.

JIE TANG (Senior Member, IEEE) received the B.Eng. degree in information engineering from the South China University of Technology, Guangzhou, China, in 2008, the M.Sc. degree (Hons.) in communication systems and signal processing from the University of Bristol, U.K., in 2009, and the Ph.D. degree from Loughborough University, Loughborough, U.K., in 2012.

He held a Postdoctoral Research positions at the School of Electrical and Electronic Engineering, The University of Manchester, U.K. He is currently an Associate Professor with the School of Electronic and Information Engineering, South China University of Technology. His research interests include green communications, NOMA, 5G networks, simultaneous wireless information and power transfer, heterogeneous networks, cognitive radio, and device-to-device communications. He was a co-recipient of the 2018 IEEE ICNC Best Paper Award. He has served as the Track Co-Chair for the IEEE Vehicular Technology Conference, in 2018. He also serves as an Editor for IEEE Access, the *EURASIP Journal on Wireless Communications and Networking*, *Physical Communication*, and *Ad Hoc & Sensor Wireless Networks*.