Energy-Efficient Transceiver Design for Cache-Enabled Millimeter-Wave Systems

Shiwen He\textsuperscript{\textregistered}, Member, IEEE, Jiaheng Wang\textsuperscript{\textregistered}, Senior Member, IEEE, Wei Huang\textsuperscript{\textregistered}, Member, IEEE, Yongming Huang\textsuperscript{\textregistered}, Senior Member, IEEE, Ming Xiao\textsuperscript{\textregistered}, Senior Member, IEEE, and Yaoxue Zhang\textsuperscript{\textregistered}, Senior Member, IEEE

Abstract—In recent years, network densification and edge caching become effective approaches to reduce the burden on the fronthaul links and the content delivery latency for wireless communication systems. However, maximizing system spectral efficiency cannot directly provide any insight on their energy requirements/efficiency for cache-enabled millimeter-wave (mmWave) radio access networks (RANs). In this paper, we study the design of energy-efficient transceiver, consisting of analog and digital precoder/combiner, for the delivery phase of the downlink of cache-enabled mmWave RANs. Due to the non-convexity of the delivery rate and objective, the coupling between the digital and analog precoders/combiners, and the constant module constraint on the elements of analog precoders/combiners, the problem of interest is non-convex and hard to obtain the global optimal solution, even the local optimal solution. To this end, we first overcome these challenges one-by-one and then transform the original problem into tractable one. Finally, an algorithmic framework that converges to the Karush-Kuhn-Tucker solution with provable is developed to achieve the design of energy-efficient transceiver. Numerical results are provided to evaluate the performance of the proposed algorithm, where fully digital precoding is used as benchmark.

Index Terms—Energy-efficient communication, hybrid transceiver, cache-enabled millimeter-wave communication.

I. INTRODUCTION

A CCOMPANYING with the development of the electronic and communication technologies, various new applications have emerged, such as augmented reality, virtual reality, and autonomous vehicles. Consequently, the fifth generation (5G) and beyond 5G are anticipated to support massive wireless connectivity for both human centric and machine type services with high spectral/energy efficiency and ultra-reliable low-latency delivery [1]. In general, the augmented and virtual realities require gigabit-per-second data rate and the typical emerging Internet of Things (IoT) applications require a latency from 0.25 ms to 10 ms and an outage probability (or packet loss rate) in the order of $10^{-3}$ to $10^{-9}$ [2]. To cater for these new demands on wireless communication networks, millimeter-wave (mmWave) communication and caching popular contents at the network edge stand out as two powerful and promising ways to realize multiple-gigabits data rate and millisecond end-to-end delivery latency [3]–[5].

In the past few years, due to the abundant spectrum resource in the range of 30 – 300 gigahertz (GHz) [6], [7], mmWave communication has been regarded as a promising candidate for providing order of magnitude improvement in the achievable data rate, which has attracted extensive attention from both industry and academia. To balance the tradeoff between the hardware cost/complexity and the system performance of mmWave systems, the digital-analog hybrid antenna architecture has been extensively considered as a cost-efficient architecture in which the number of radio frequency (RF) chains is smaller than that of antennas [8]–[16]. The works in [8]–[11] consider the design of hybrid precoders for point-to-point mmWave communication systems. To release the coupling between the digital and analog precoding matrices, the orthogonal matching pursuit (OMP) [8]–[10] and matrix decomposition [11] methods are used to obtain the analog and digital precoders from fully digital (sub)optimal solutions. A codebook based hybrid precoding design was investigated for mmWave downlink multiuser communication system [12]. Recently, the penalty dual decomposition (PDD)
method is utilized to optimize the hybrid precoding for point-to-point mmWave communication systems [13], [14]. The physical propagation delay of electromagnetic waves on transmitted wideband signal was analyzed for mmWave large-scale multiple-input multiple-output (MIMO) systems [15]. In addition, joint channel estimation and hybrid precoding was investigated by designing a novel hierarchical multi-resolution codebook to construct training analog beamforming vector for mmWave communication systems [16]. Note that the aforementioned literature only focuses on the conventional non-cache-enabled point-to-point or multiuser mmWave communication systems.

In conventional cellular network or cloud radio access network (C-RAN), since the data center is far away from the end-devices, only increasing bandwidth is still difficult to achieve the goal of millisecond level end-to-end delivery latency. Recently, to alleviate the fronthaul capacity requirement and reduce the user-perceived latency in C-RANs, an evolved cache-enabled RAN was proposed in which the remote radio heads (RRHs), referred as enhanced RRHs (eRRHs), have the abilities to store popular contents and perform baseband signal processing [3], [4]. During the off-peak traffic periods, some most frequently requested files can be pre-fetched to the local cache at the network edge to reduce delivery latency and increase spectral efficiency [17]–[19]. A cache placement strategy with limited fronthaul and cache capacity was studied to minimize the average download delay of user requests [17]. The authors of [18] investigated how to cache the files at different caching units with the objective to maximize the average requested data rate subject to a finite service latency. Successful transmission probabilities were analyzed for a two-tier large-scale cache-enabled wireless multicasting network [19]. Joint optimization of cloud-edge precoders was also investigated with aiming to maximize or minimize the system performance criterion under certain constraints for cache-enabled RANs [20]–[22]. A novel partial caching pipelined transmission scheme was proposed to reduce both the burden on fronthaul links and delivery latency for cache-enabled RANs [23], [24]. To further reduce signalling overhead, distributed fog computing was used to investigate the batched sparse code for industrial control [25]. Their studies have shown that caching the popular contents at the network edge has contributed to the reduction of delivery latency and of the burden on the fronthaul links for wireless communication. However, these mentioned literature focused on cache-enabled transmission optimization for sub-6 GHz communication systems. More recently, cache-enabled mmWave precoding design was investigated to maximize the minimum user rate to guarantee the user fairness [26]. The digital-analog hybrid antenna architecture makes the design of transmission scheme more challenging in cache-enabled mmWave communication system. The developed algorithms in [26], which aims only to optimize the hybrid precoding and adopts the Taylor series expansion to optimize the analog precoder, can only guarantee convergence, but cannot be proved to converge to the global optimum or the Karush-Kuhn-Tucker (KKT) solution.

On the other hand, due to the environmental concerns and sustainable growth considerations, green communication is rapidly gaining extensive attention for communication systems [27], [28]. Indeed, information and communication technologies (ICT) consume about 2% of the entire world energy consumption, and the situation is likely to reach the point wherein ICT equipments in large cities will require more energy than it is actually available [29]. To save transmit power consumption, one simple approach is to consider the optimization of the transmit power subject to the requirement of quality of services. However, minimizing transmit power does not mean maximizing spectral efficiency. To achieve high data rate with limited energy consumption, recently, energy efficiency, which is measured in bit/Joule, becomes a new performance optimization criterion for wireless communication, especially for energy efficiency priority mmWave communication systems [30]. Note that the aforementioned references [27]–[30] focus on the design of energy-efficient transmission for non-cache-enabled wireless communication systems.

Compared to the conventional delivery rate maximization or cost minimization problem for cache-enabled RANs [20]–[22], due to the coupling between digital and analog precoders with additional constant modulus constraints, optimizing energy-efficient hybrid transceiver is more challenging and difficult to obtain the specific optimal solution for cache-enabled mmWave RANs. Meanwhile, the introduction of the constraints on the delivery rate and the fronthaul link makes the energy-efficient design of transceivers different from the conventional mmWave communication system [12], [13]. Furthermore, to the best of the authors’ knowledge, the energy efficiency problem of cache-enabled mmWave RANs is still opening problem. Motivated by these observations, in this paper, we investigate the design of energy-efficient transceiver for the delivery phase of the downlink of cache-enabled mmWave RANs. The main contributions of this paper can be summarized as follows:

- We jointly optimize the analog combiner/precoder and digital combiner/precoder to maximize the system energy efficiency for cache-enabled mmWave RANs employing soft fronthaul information transfer;
- To solve the non-convex energy efficiency optimization problem, we first transform the original problem into a tractable form via fractional programming theory and penalty dual decomposition (PDD) method [13], [14]. In particular, the fractional programming theory is used to transform the fractional form into a parameterized subtract one. Then, the PDD method is exploited to put the coupling constraint into the objective;
- An effective optimization algorithm is developed to solve the formulated optimization problem via the so-called alternating direction of multipliers method (ADMM) [31] and block successive upper-bound minimization (BSUM) [32]. Furthermore, we further prove that the presented algorithm converges to the KKT solution of the considered problem;
- Numerical experiments are carried out to evaluate the performance of the presented algorithm and reveal that hybrid antenna architecture is not necessarily the best scheme for cache-enabled mmWave RANs from the perspective of system energy efficiency and delivery rate.
But, compared to the non-caching mmWave RANs, caching popular files at the network edge helps to improve the system performance.

The remainder of this paper is organized as follows. The system model is described in Section II. In Section III, the design of optimization method is investigated for the considered optimization problem. In Section IV, the performance of the developed algorithms is evaluated via simulation. Conclusions are drawn in Section V.

Notations: Bold lower case and upper case letters represent column vectors and matrices, respectively. The superscript $\cdot^H$ by $\text{CN}$ summarize in Table I.

The system model is described in Section II. In Section III, the design of optimization method is investigated for the considered optimization problem. In Section IV, the performance of the developed algorithms is evaluated via simulation. Conclusions are drawn in Section V.

![Fig. 1. Illustration of downlink of cache-enabled millimeter wave RAN.](image)

### Table I

| Symbol         | Meaning                                      | Symbol         | Meaning                                      |
|----------------|----------------------------------------------|----------------|----------------------------------------------|
| $N_{\text{RF}}$, $N_{\text{RF}}$ | Numbers of RF chains of user and eRRH | $N_{\text{RF}}$, $N_{\text{RF}}$ | Numbers of antennas of user and eRRH |
| $N_{\text{RF}}$, $N_{\text{RF}}$ | Sets of antennas and RF chains at eRRH | $N_{\text{RF}}$, $N_{\text{RF}}$ | Sets of antennas and RF chains at eRRH |
| $F$ | Maximum transmit power of eRRH | $C_i$ | Capacity of frontal link to eRRH |
| $B_i$ | Normalized cache size of eRRH | $P$ | Number of files in the library |
| $S_{\text{req}}$ | Sets of all files and all requested files | $S$ | Normalized size of file |
| $f_k$ | File requested by user $k$ | $c_{f, i}$ | Binary cache variable of file $f$ at eRRH $i$ |
| $r_f$ | Data delivery rate of file $f$ | $R_k$ | Achievable rate of user $k$ |
| $y_i$ | Signal received by user $k$ | $x_i$ | Signal transmitted by eRRH $i$ |
| $H_{i, j}$ | Channel matrix from eRRH $i$ to user $k$ | $\Omega_i$ | Quantization noise covariance matrix to eRRH $i$ |
| $F_{\text{RF}, i}$ | Analog precoding matrix at eRRH $i$ | $G_{f, i}$ | Digital precoding matrix for file $f$ at eRRH $i$ |
| $U_{\text{RF}, k}$ | Analog precoding matrix at user $k$ | $U_{\text{RF}, i}$ | Digital precoding matrix at user $k$ |
| $g_{\text{RF}}$ | Set of digital precoding matrices at eRRHs | $U_{\text{RF}}$ | Set of analog combining matrices at users |
| $g_{\text{RF}}$ | Auxiliary variables at eRRH $i$ | $\Phi_{f, i}$ | Lagrange multiplier matrices at eRRH $i$ |
| $d_{\text{RF}}$ | Set of digital combining matrices at users | $\Theta_{f, i}$ | Lagrange multiplier matrices at eRRH $i$ |
| $\Sigma_{\text{RF}}$ | Weight matrix at user $k$ | $Z$ | Set of weight matrices |
| $\mathcal{R}$ | $\mathcal{R} = \{r_f\}_f \in S_{\text{req}}$ | $\mathcal{O}$ | Set of quantization noise covariance matrices |

II. SYSTEM MODEL

In this work, we investigate the downlink transmission of cache-enabled mmWave RANs. There are $K_U$ multi-antenna users establishing wireless communication with $K_R$ eRRHs, as illustrated in Fig. 1, where LPF is the abbreviation of low pass filter. Each eRRH is equipped with an $N_t$-antennas uniform linear array (ULA). Let $K_R = \{1, \cdots, K_R\}$ and $K_U = \{1, \cdots, K_U\}$ be the set of eRRHs and users, respectively. For simplicity, we assume that all eRRHs are equipped with the same number $N_{\text{RF}}$ of RF chains, each connected to $N_t$ transmit antennas via $N_t$ phase shifters. eRRH $i$ is connected to the BBU through an error-free frontal link of capacity $C_i$ nats/Hz/s, $i \in K_R$. Each user is equipped with $N_{\text{RF}}$ RF chains and $N_r$ receive antennas. Each RF chain is connected to the $N_r$ receive antennas via $N_r$ phase shifters.

A. Cache Model

Without loss of generality, we assume that all files in library $\mathcal{F}$ at the BBU have a size of $nS$ nats, where $S$ is the normalized file size and $n$ is the number of symbols of each downlink coded transmission interval. The assumption of equal file-sizes is standard and can be justified, as in practice what are cached and requested by users are chunks of contents, e.g., fragments of a given duration, which may be safely pre-stored in local cache to its local cache. Each eRRH $i \in K_R$ is equipped with a cache of size $nB_i > 0$ nats, where $B_i$ is the normalized cache size. To provide the spatial file diversity (which can improve the performance of dense wireless networks) [19], each eRRH $i \in K_R$ randomly pre-stores $nB_i$ nats of $nFS$ nats in the library to its local cache according to certain preferences [20]–[22]. In other words, each eRRH $i \in K_R$ randomly selects $\lfloor B_i/S \rfloor$ files from library.

![Fig. 1. Illustration of downlink of cache-enabled millimeter wave RAN.](image)
\( \mathcal{F} \) at the BBU to store at its local cache. The cache status of file \( f \) at eRRH \( i \) can be modeled by defining binary variables \( c_{f,i} \), \( f \in \mathcal{F} \), \( i \in \mathcal{K}_R \), as

\[
c_{f,i} = \begin{cases} 
1, & \text{if subfile } f \text{ is cached by eRRH } i, \\
0, & \text{otherwise.}
\end{cases}
\]  

(1)

Let \( \gamma_{f,i} \) denote the complement \( 1 - c_{f,i} \) of binary variable \( c_{f,i} \). In this work, we do not focus on the research on the content popularity and specific caching strategies, the details please refer to other literatures, e.g., [17]–[21]. Therefore, for simplicity, in the following, we assume that the cache state information \( c_{f,i}, f \in \mathcal{F}, i \in \mathcal{K}_R \), is predetermined.

**B. Channel Model**

A narrowband clustered channel model with \( N_c \) non-line of sight (NLoS) scatterers and one line of sight (LoS) path based on the extended Saleh-Valenzuela model is adopted in this work [6]. Each scatterer is further assumed to contribute a single propagation path to the channel between a user and an eRRH [7]. Thus, channel matrix \( \mathbf{H}_{k,i} \) can be modeled as [35, 36]

\[
\mathbf{H}_{k,i} = \sqrt{\frac{N_t N_c}{N_t}} \sum_{l=1}^{N_c} \alpha_{l,k,i} \sqrt{\rho_{l,k,i}} \mathbf{a}_u(\theta_{l,k,i}) \mathbf{a}_l^H(\phi_{l,k,i}) \\
+ \mathbf{I}(\rho_{0,k,i}) \sqrt{\frac{N_t}{N_t}} \rho_{0,k,i} \mathbf{a}_u(\theta_{0,k,i}) \mathbf{a}_l^H(\phi_{0,k,i}),
\]

\( \forall k \in \mathcal{K}_C, i \in \mathcal{K}_R, \)

(2)

where \( \theta_{l,k,i} \in [0, 2\pi) \) and \( \phi_{l,k,i} \in [0, 2\pi) \) are the angles of arrival and departure (AoAs/AoDs) of the \( l \)-th NLoS path between eRRH \( i \) and user \( k \), respectively, \( \theta_{0,k,i} \in [0, 2\pi) \) and \( \phi_{0,k,i} \in [0, 2\pi) \) are the AoA and AoD of the LoS link between eRRH \( i \) and user \( k \), respectively. \( \alpha_{l,k,i} \sim CN(0, \sigma^2_{l,k,i}) \) and \( \rho_{l,k,i} \) denote the complex gain and the attenuation associated to the \( l \)-th path between eRRH \( i \) and user \( k \), respectively, \( p \in \mathcal{K}_C = \{1, \cdots, N_c\} \), \( \rho_{0,k,i} \) denotes the attenuation of the LoS link between eRRH \( i \) and user \( k \). \( \mathbf{I}(\rho_{0,k,i}) \) is a random variate indicating if a LoS link exists between eRRH \( i \) and user \( k \). Denoting by \( p \) the probability that \( \mathbf{I}(\rho_{0,k,i}) = 1 \), i.e., a LOS link exists. The value of \( p \) can be determined according to the Formula given in [37, Eq.(2) and Eq. (3)]. In particular, for an \( N_t \)-element ULA, the array response vector is given by [38]

\[
\mathbf{a}_u(\phi_{l,k,i}) = \sqrt{\frac{1}{N_t}} \left[ 1, e^{j \lambda_d \sin(\phi_{l,k,i})}, \cdots, e^{j (N_t-1) \lambda_d \sin(\phi_{l,k,i})} \right]^T,
\]

(3)

where \( \lambda_d \) and \( d_a \) are the wavelength of carrier frequency and the antenna spacing, respectively. The array response vector \( \mathbf{a}_u(\theta_{l,k,i}) \) at user \( i \) can be written in a similar fashion. For simplicity, we assume that the perfect channel state information (CSI) \( \mathbf{H}_{k,i} \) can be obtained at the BBU. The CSI can be acquired by the eRRHs and reported back to the BBU via the fronthaul links, e.g., [20]–[22]. The analysis of the effect of imperfect CSI is an interesting topic for future work [39, 40].

**C. Signal Model**

According to the caching state of files, the contents of files requested by the users can be retrieved from the local cache or the library located at the BBU. In general, there are two kinds of transfer methods for fetching the uncached requested files from the BBU, i.e., hard- and soft- fronthaul information transfer. For the hard fronthaul information transfer, the hard information of the uncached files is transferred to the eRRHs via fronthaul links. While, for the soft fronthaul information transfer, the fronthaul links are used to transfer a quantized version of the precoded signals of the missing files [20], [26]. In this paper, we adopt the soft fronthaul information transfer to acquire the uncached requested files from the BBU to the eRRHs which does not cached the requested files. Consequently, the signal \( \mathbf{x}_i \in \mathbb{C}^{N_t \times 1} \) transmitted by eRRH \( i \) including the information of the cached and uncached requested files is given by

\[
\mathbf{x}_i = \mathbf{F}_{RF,i} \left( \sum_{f \in \mathcal{F}_{req}} \mathbf{G}_{f,i} \mathbf{s}_f + \mathbf{z}_i \right),
\]

(4)

where \( \mathbf{G}_{f,i} = c_{f,i} \mathbf{G}_{f,i} + \gamma_{f,i} \mathbf{G}_{f,i} \in \mathbb{C}^{N_t \times d_f} \) with \( \mathbf{G}_{f,i} \) and \( \mathbf{U}_{f,i} \) being the digital precoding matrix at eRRH \( i \) for baseband signal \( s_f \) cached at the local cache and not available at the local cache, respectively. \( \mathbf{s}_f \in \mathbb{C}^{d_f \times 1} \) represents the normalized data symbols of requested file \( f \), i.e., \( \mathbb{E}[\mathbf{s}_f^H \mathbf{s}_f] = \mathbf{I}_{d_f \times d_f} \). \( d_f \) denotes the number of spatial data streams of requested file \( f \). \( \mathcal{F}_{req} \) denotes the set of all requested files. \( \mathbf{F}_{RF,i} \in \mathbb{C}^{N_t \times N_t} \) is the analog precoding matrix at eRRH \( i \) with \( \| \mathbf{F}_{RF,i} \|_2^2 = 1 \), \( m_i \in \mathcal{N}_R = \{1, \cdots, N_t\} \), and \( n_i \in \mathcal{N}_C = \{1, \cdots, N_t \} \). The quantization noise \( \mathbf{z}_i \in \mathbb{C}^{N_t \times 1} \) is assumed to be independent of the information of the uncached requested files and distributed as \( \mathbf{z}_i \sim CN(0, \mathbf{\Omega}_i) \). Note if all requested files in \( \mathcal{F}_{req} \) have been stored at the local cache, then the quantization noise \( \mathbf{z}_i \) is a zero vector. We further assume that the quantization noise \( \mathbf{z}_i \) is independent across the eRRHs, i.e., the signals intended for different eRRHs are quantized independently. Consequently, the signal received at user \( k \) can be expressed as

\[
\mathbf{y}_k = \mathbf{H}_{k} \mathbf{F}_{RF} \sum_{f \in \mathcal{F}_{req}} \mathbf{G}_{f} \mathbf{s}_f + \mathbf{H}_{k} \mathbf{F}_{RF} \mathbf{z} + \mathbf{n}_k,
\]

(5)

where \( \mathbf{H}_{k} \triangleq [\mathbf{H}_{k,1}, \cdots, \mathbf{H}_{k,K_R}] \in \mathbb{C}^{N_t \times N_t K_R} \) with \( \mathbf{H}_{k,i} \in \mathbb{C}^{N_t \times N_t} \) being the channel matrix between user \( k \) and eRRH \( i \). \( \mathbf{G}_{f} \triangleq [\mathbf{G}_{f,1}, \cdots, \mathbf{G}_{f,K_R}]^H \in \mathbb{C}^{N_t K_R \times d_f} \) and \( \mathbf{z} \triangleq [\mathbf{z}_1^H, \cdots, \mathbf{z}_{K_R}^H]^H \in \mathbb{C}^{N_t K_R \times 1} \). \( \mathbf{F}_{RF} \) is the analog precoding super-matrix defined as

\[
\mathbf{F}_{RF} = \left[ (\mathbf{F}_{RF,1} \mathbf{P}_{RF,1}^H)^H, \cdots, (\mathbf{F}_{RF,K_R} \mathbf{P}_{RF,K_R}^H)^H \right]^H = \text{diag} \left( \mathbf{F}_{RF,1}, \cdots, \mathbf{F}_{RF,K_R} \right),
\]

(6)

where permutation matrix \( \mathbf{P}_i \) is defined as

\[
\mathbf{P}_i = [0_{N_t \times (i-1) N_t}, \mathbf{I}_{N_t \times N_t}, 0_{N_t \times (K_R-i) N_t}]^T.
\]

(7)
In (5), \( n_k \) is the additive white Gaussian noise (AWGN) with circularly symmetric complex Gaussian 
\( \mathcal{CN}(0, \sigma^2_{n_k} I_{N_{\text{RF}} \times N_{\text{RF}}}) \).

For simplicity, we assume that all users adopt linear receiver to recover the desired signal from the received signals. Let \( U_{\text{RF},k} \in \mathbb{C}^{N_{\text{RF}} \times M_{\text{RF}}} \) be the RF combining matrix and \( U_{\text{BB},k} \in \mathbb{C}^{N_{\text{RF}} \times d_{\text{f}}} \) be the baseband combining matrix. Similar to RF precoding matrix \( F_{\text{RF},i} \), \( U_{\text{RF},k} \) is also implemented via phase shifter network with 
\[
\left| U_{\text{RF},k} \right|_{m_k, n_k}^2 = 1,
\]
where \( m_k \in N_{\text{RF}} = \{1, \ldots, M_{\text{RF}}\} \) and \( n_k \in N_{\text{RF}} = \{1, \ldots, N_{\text{RF}}\} \). Consequently, the recovered baseband signal is expressed as
\[
\mathbf{s}_{f_k} = U_{\text{BB},k}^H U_{\text{RF},k}^H \mathbf{y}_k.
\]

Different from the conventional MIMO communication system, in which each antenna has a dedicated RF chain, the received signal at each RF chain is a combination of received signal of all antennas via a phase shifter network for mmWave communication system. Thus, when Gaussian symbols are transmitted over the mmWave channel, the achievable data rate at user \( k \), in units of nats/Hz, can be defined as [14]
\[
R_k \triangleq \ln \det \left( I_{d_{f,k} \times d_{f,k}} + U_{\text{RF},k}^H \mathbf{F}_{\text{RF}} \mathbf{G}_k \mathbf{G}_k^H U_{\text{RF},k}^H \bar{\Pi}_k^{-1} \right),
\]
where \( \bar{\Pi}_k \) is given by (10), at the bottom of this page. With \( \Omega = \text{diag} (\Omega_1, \ldots, \Omega_{K_{\text{RF}}}) \). In the following subsection, the optimization objective will be formulated to obtain the corresponding digital and analog transceiver strategies.

\[\text{D. Optimization Objective}\]

In this paper, the objective is to maximize the system energy efficiency under certain constraints on the fronthaul transmission rate, eRRH transmit power, and analog precoder together with combiner. In general, energy efficiency is defined as the number of communicated nats per unit power consumption [27]–[30]. Accordingly, the resulting optimization problem is formulated as
\[
\text{maximize} \quad \sum_{f \in \mathcal{F}_{\text{req}}} r_f \quad \frac{\sum_{i \in K_{\text{RF}}} p_i \left( F_{\text{RF},i}, \mathcal{G}_k, \Omega_i \right) + p_{\text{ac}}}{\eta}, \tag{11a}
\]
subject to \( r_f \leq R_k, \quad \forall f \in K_{\text{RF}}, \tag{11b} \)
\( r_f \leq S, \quad \forall f \in \mathcal{F}_{\text{req}}, \tag{11c} \)
\( g_i (\mathcal{G}_k, \Omega_i) \leq C_i, \quad \forall i \in K_{\text{RF}}, \tag{11d} \)
\( p_i (F_{\text{RF},i}, \mathcal{G}_k, \Omega_i) \leq P_i, \quad \forall i \in K_{\text{RF}}, \tag{11e} \)
\( \left| F_{\text{RF},i} \right|_{m_i, n_i}^2 = 1, m_i \in N_{\text{RF}}, n_i \in N_{\text{RF}}, \quad \forall i \in K_{\text{RF}}, \tag{11f} \)
\( \left| U_{\text{RF},k} \right|_{m_k, n_k}^2 = 1, \quad m_k \in N_{\text{RF}}, n_k \in N_{\text{RF}}, \quad \forall k \in K_{\text{RF}}, \tag{11g} \)

where \( r_f \) denotes the delivery rate for requested file \( f \), \( \forall f \in \mathcal{F}_{\text{req}}, \mathcal{G}_k \triangleq \{ \mathcal{G}_{f,i} \}_{f \in \mathcal{F}_{\text{req}}, i \in K_{\text{RF}}}, \quad U_{\text{RF},k} \triangleq \{ U_{\text{RF},k} \}_{k \in K_{\text{RF}}}, \quad \mathcal{R}_i \triangleq \{ r_f \}_{f \in \mathcal{F}_{\text{req}}}, \quad \text{and} \quad \Omega_i \triangleq \{ \Omega_f \}_{f \in \mathcal{F}_{\text{req}}}, \quad \forall i \geq 1 \) is a constant accounting for the inefficiency of power amplifier [13], dynamic power consumption \( p_i (F_{\text{RF},i}, \mathcal{G}_k, \Omega_i) \) is given by
\[
\begin{align*}
p_i (F_{\text{RF},i}, \mathcal{G}_k, \Omega_i) & \triangleq \sum_{f \in \mathcal{F}_{\text{req}}} \text{tr} (F_{\text{RF},i} \mathcal{G}_{f,i} \mathcal{G}_{f,i}^H F_{\text{RF},i}^H ) + \text{tr} (F_{\text{RF},i} \mathcal{G}_k \mathcal{G}_k^H F_{\text{RF},i}^H ), \tag{12}
\end{align*}
\]
and constant circuit power consumption \( p_{\text{ac}} \) is calculated as
\[
\begin{align*}
p_{\text{ac}} \triangleq & \quad P_{\text{RF}} + P_{\text{RF}}, P_{\text{PS}} + P_{\text{LNA}} + 2P_{\text{LO}} + 2P_{\text{BB}}, \tag{13a}
\end{align*}
\]
\[
\begin{align*}
P_{\text{RF}} \triangleq & \quad K_{\text{RF}} N_{\text{RF}}\text{DAC} + P_{\text{mixer}} + P_{\text{LPF}}, \tag{13b}
P_{\text{RF}} \triangleq & \quad K_{\text{RF}} N_{\text{RF}}\text{ADC} + P_{\text{mixer}} + P_{\text{LPF}}, \tag{13c}
P_{\text{PS}} \triangleq & \quad (K_{\text{RF}} N_{\text{RF}}\text{DAC} + K_{\text{RF}} N_{\text{RF}}\text{ADC}) P_{\text{PS}}, \tag{13d}
P_{\text{LNA}} \triangleq & \quad K_{\text{RF}} N_{\text{RF}}(N_{\text{RF}} + 1) P_{\text{LNA}}, \tag{13e}
\end{align*}
\]
In (13), \( P_{\text{RF}} \) and \( P_{\text{RFC}} \) represent the power consumed by all RF chains at eRRHs and users, respectively. \( P_{\text{PS}} \) represents the power consumed by all phase shifters at eRRHs and users. \( P_{\text{LNA}} \) denotes the power consumption of all low noise amplifiers. \( P_{\text{LO}} \) and \( P_{\text{BB}} \) denote respectively the power consumed by the local oscillator and the basic power consumption. \( P_{\text{DAC}} \) and \( P_{\text{ADC}} \) denote the power consumption of the digital-to-analog converter and the analog-to-digital converter, respectively. \( P_{\text{mixer}} \) and \( P_{\text{LPF}} \) denote respectively the power consumed by the mixer and the LPF, respectively. \( P_{\text{PS}} \) stands for the consumption of the phase shifter. \( P_{\text{LNA}} \) represents the power consumption of low noise amplifier [9], [13]. In constraint (11d), fronthaul rate function \( g_i (\mathcal{G}_k, \Omega) \) is defined as
\[
g_i (\mathcal{G}_k, \Omega) \triangleq \ln \det \left( \sum_{f \in \mathcal{F}_{\text{req}}} \mathcal{T}_{f,i} \mathcal{G}_{f,i}^H + \mathbf{1} \right) - \ln \det (\Omega_i) . \tag{14}
\]
According to [41, Ch. 3], to reliably recover signal \( \hat{x}_f \) at eRRH \( i \), constraint \( g_i (\mathcal{U}, \Omega) \leq C_i \) has to be satisfied.

In problem (11), constraint (11b) implies that the data delivery rate for user \( k \) cannot exceed the achievable rate of user \( k \). Constraint (11c) ensures that the data delivery rate of file \( l \) does not exceed the normalized file size \( S \). The rate on each fronthaul link is constrained by the limited fronthaul link capacity \( C_i \) in constraint (11d). Constraint (11e) limits the maximum allowable transmit power of eRRH \( i \). Constraints (11f) and (11g) are the constant modulus constraints on each element of analog precoding matrix due to the implementation of analog phase shifter. Note that there are six main obstacles for solving problem (11), i.e., the non-convex objective (11a), the non-convex achievable rate \( R_k \) in the right hand of constraint (11b), the non-convex fronthaul capacity constraint (11d) and power constraint (11e), and

\[
\bar{\Pi}_k = U_{\text{RF},k}^H \left( \sum_{f \in \mathcal{F}_{\text{req}} \setminus \{ f_k \} } H_k F_{\text{RF},i} G_{f,i} F_{\text{RF},i}^H H_k^H + H_k F_{\text{RF},i} \Omega_i F_{\text{RF},i}^H H_k^H + \sigma_n^2 I_{N_{\text{RF}} \times N_{\text{RF}}} \right) U_{\text{RF},k} \quad (10)
\]
the non-convex constant modulus constraints (11f) and (11g). Therefore, problem (11) is non-convex fractional programming problem and therefore difficult to obtain the global optimal solution.

In addition, if eRRH \( i \) has cached all requested files, i.e., file \( f \), \( \forall f \in \mathcal{F}_{\text{req}} \) is pre-stored at the local cache at the network edge, the corresponding fronthaul link capacity constraint given in (11d) is redundant and can be removed. Furthermore, the dynamic power consumption of eRRH \( i \) is replaced with \( p_i(\mathbf{F}_{\text{RF},i},\mathbf{G}_i,\Omega_i) \triangleq \sum_{f \in \mathcal{F}_{\text{req}}} \mathbf{H} f,i + \mathbf{P}^{F}_{\text{RF},i} \) in constraint (11e). To provide performance comparison, we also here formulate the maximization of delivery rate as follows

\[
\begin{align}
\text{maximize} & \quad \sum_{f \in \mathcal{F}_{\text{req}}} r_f, \\
\text{subject to} & \quad (11b), (11c), (11d), (11e), (11f), (11g).
\end{align}
\]

Compared to problem (11), the objective function in problem (15) is simple, but problem (15) is still non-convex and hard to obtain the global optimal solution. In what follows, we focus on designing an efficient and effective heuristic optimization algorithm to solve problem (11) and (15), respectively.

### III. Design of Optimization Method

For the design of hybrid transceiver of interest, one of the main difficulties is the cascading relation between the analog precoders/combiners and digital precoders/combiners. In the existing literature, many optimization methods are used to overcome this difficulty, such as the OMP method [8]–[10], Lagrange convex approximation method [11] and codecook based analog precoding design method [12] is used to overcome this difficulty. On the other hand, by treating the product of analog precoder/combiner and digital precoder/combiner as a whole, problem (11) can be transformed into another form that is easier to solve compared to the original one [13], [14]. In what follows, we explore the solution of problem (11) in detail.

#### A. Algorithmic Architecture

For ease of presentation, let \( \mathbf{M}_{f,i} \triangleq \mathbf{F}_{\text{RF},i} \mathbf{G}_{f,i}, \mathbf{M}_f \triangleq \left[ \mathbf{M}_{f,1}^H, \cdots, \mathbf{M}_{f,K_{R}}^H \right] \in \mathbb{C}^{N_{K_{R}} \times d_f}, \mathbf{N} = \text{diag} (\mathbf{N}_1, \cdots, \mathbf{N}_{K_{R}}) \), and \( \mathbf{N}_i \triangleq \mathbf{F}_{\text{RF},i} \tilde{\Omega} i, \) with \( \tilde{\Omega} i = \tilde{\Omega}_1 i \mathbf{I}_i^H \). Thus, problem (11) can be equivalently rewritten as the following form

\[
\begin{align}
\text{maximize} & \quad \sum_{f \in \mathcal{F}_{\text{req}}} r_f, \\
\text{subject to} & \quad (11b), (11c), (11d), (11f), (11g), \quad \mathcal{P}_i (\mathcal{M}, \mathcal{N}) \leq P_i, \quad \forall i \in \mathcal{K}_R, \quad (16e)
\end{align}
\]

where \( \mathcal{M} \triangleq \{ \mathbf{M}_{f,i} \} \in \mathcal{F}_{\text{req}} \subseteq \mathcal{K}_R \), \( \mathcal{N} \triangleq \{ \mathbf{N}_i \} \subseteq \mathcal{K}_R \), and \( \mathcal{P}_i (\mathcal{M}, \mathcal{N}) \) is defined as

\[
\mathcal{P}_i (\mathcal{M}, \mathcal{N}) \triangleq \sum_{f \in \mathcal{F}_{\text{req}}} \left( \text{tr} (\mathbf{M}_{f,i}) + \text{tr} (\mathbf{N}_i \mathbf{N}_i^H) \right).
\]

The non-convexity of objective (16a) is mainly incurred with the fractional form. To overcome this difficulty, a classic technique was proposed to transform the single ratio fractional form into a parameterized linear form by introducing an additional auxiliary variable in [42]. Thus, the objective function can be transformed into an linear objective one. This implies that problem (16) can be rewritten as (18), at the bottom of the next page, where \( \lambda \) is an auxiliary tunable parameter. As a result, an iterative algorithm (known as the Dinkelbach method [43]) can be used to solve problem (16). This algorithm is summarized as Algorithm 1, where \( \epsilon \) is a given stopping threshold. The convergence to the optimal energy efficiency is guaranteed [27].

#### B. Inner Optimization Designs

In this subsection, we focus on investigating the solution of problem (18) by using the PDD method [13], [14], ADMM method [31], BSUM method [32] and the relation between the user rate and minimum mean square error (MMSE) [44].

1) Problem Transformation: In problem (18), equality constraints (16d) and (16e) are two main obstacles for solving it. Fortunately, the researches on equality constraint optimization in [13], [14] have shown that problem (17) can be further reformulated as (19), at the bottom of the next page. In (19a), matrix \( \mathbf{Z}_k \) is calculated as (20), at the bottom of the next page, where \( \Phi_{f,i} \) and \( \Psi_i \) are Lagrange Multiplier matrices associated with equality constraints (16d) and (16e), respectively. \( \rho \) and \( \gamma \) are two penalty parameters. Though objective function (19a) is formulated as a linear form by using the fractional programming theory and PDD method, problem (19) is still non-convex and therefore difficult to obtain the global optimal solution. The main difficulties of solving problem (19) are three-fold. The first is the non-convexity of achievable rate \( R_k \) of user \( k \). The second is the coupling of analog and digital precoding matrices. The third is the constant modulus constraint on the elements of analog precoding and combining matrices.

In the sequel, we focus on overcoming the non-convex achievable rate \( R_k \) of user \( k \) by using the weighted MMSE transformation. First, we define the mean square error (MSE)
as follows

\[
\begin{align*}
E_k &= \mathbb{E} \left[ \left( U_{BB,k}^H U_{RF,k}^H y_k - s_{f_k} \right) \left( U_{BB,k}^H U_{RF,k}^H y_k - s_{f_k} \right)^H \right] \\
&= \left( I_{d_{f_k}} \times d_{f_k} - U_{BB,k}^H U_{RF,k}^H H_k M_{f_k} \right) \\
&\quad \times \left( I_{d_{f_k}} \times d_{f_k} - U_{BB,k}^H U_{RF,k}^H H_k M_{f_k} \right)^H \\
&\quad + U_{BB,k}^H \Pi_k U_{BB,k},
\end{align*}
\]

(21)

where \( \Pi_k \) is given by (22), at the bottom of this page. The authors of [44] have shown that the achievable data rate \( R_k \) of user \( k \) can be expressed as

\[
R_k = \maximizeminimize (\ln \det (\Sigma_k) - \text{Tr} (\Sigma_k E_k) + d_{f_k}), \quad \Sigma_k, U_{BB,k}
\]

\[
\Sigma_k \succ 0_{d_{f_k} \times d_{f_k}} \quad \text{and} \quad U_{BB,k} \triangleq \{ U_{BB,k} \}_{k \in K_U}
\]

Note that the objective value of problem (18) is an upper boundary of that of problem (23). This is because the feasible region of problem (23) is a subset of that of problem (18). Similarly, problem (15) can be approximated as the following low boundary problem

\[
\begin{align*}
\text{minimize} & \quad \Phi_{BB,\mathcal{U}B,\mathcal{G},\mathcal{O},\mathcal{R},\mathcal{M},\mathcal{N},Z} \sum_{f \in \mathcal{F}_{req}} r_f - \Xi_s, \\
\text{subject to} & \quad (11c), (11d), (11f), (11g), (16c), (23b).
\end{align*}
\]

(24b)

Though problems (23) and (24) have more tractable forms compared to their original ones, they are still impossible to obtain the global optimal solution via the conventional optimization method. In what follows, we are going to develop an iterative procedure to obtain the solution of problem (23) based on the ADMM and BSUM, which are known for their good behaviors in several non-convex optimization problems [31], [32]. In other words, we maximize the objective (23a) by sequentially fixing partial variables and updating the others.

2) Optimization of Baseband and RF Combiners: In this subsection, we focus on optimizing variables \( U_{BB,k}, \Sigma_k, \) and \( U_{RF,k}, \forall k \in K_U \) with given other variables. Note that these variables appear only in the right side of constraint (23b). Therefore, maximizing the goal of problem (23a) is equivalent to maximizing the right side of constraint (23b) as much as possible, i.e., maximizing the feasible region of problem (23).

\[
\begin{align*}
\text{maximize} & \quad \text{Tr} (\Sigma_k E_k), \\
\text{subject to} & \quad \left| U_{RF,k} \right|_{m_k,n_k}^2 = 1, \quad m_k \in \mathcal{N}_{uR}, \quad n_k \in \mathcal{N}_{uC}.
\end{align*}
\]

(25)

Substituting the analytical expression of \( E_k \) given by (21) into problem (25), we have the following equivalent problem (26), at the bottom of the next page. In (26), matrix \( \Upsilon_k \) is given by

\[
\Upsilon_k = \sum_{f \in \mathcal{F}_{req}} H_k M_f M_f^H H_k^H + H_k N N^H H_k^H + \sigma_k^2 I_{N_{RF} \times N_{RF}}.
\]

(27)

\[
\begin{align*}
\text{maximize} & \quad \sum_{f \in \mathcal{F}_{req}} r_f - \lambda \left( \eta \sum_{i \in K_R} \Phi_i (\mathcal{M}, \mathcal{N}) + p_c \right), \\
\text{subject to} & \quad (11b), (11c), (11d), (11f), (11g), (16c), (16d), (16e).
\end{align*}
\]

(18a)

\[
\begin{align*}
\text{maximize} & \quad \sum_{f \in \mathcal{F}_{req}} r_f - \lambda \left( \eta \sum_{i \in K_R} \Phi_i (\mathcal{M}, \mathcal{N}) + p_c \right) - \Xi_s, \\
\text{subject to} & \quad (11b), (11c), (11d), (11f), (11g), (16c).
\end{align*}
\]

(19a)

\[
\Xi_s \triangleq \sum_{i \in K_R} \Xi_{s,i} \triangleq \sum_{i \in K_R} \left( \sum_{f \in \mathcal{F}_{req}} \frac{1}{2\rho} \left| M_{f,i} - F_{RF,i} G_{f,i} + \rho \Phi_{f,i} \right|^2 + \frac{1}{2\rho} \left| N_{i} - \Phi_{RF,i} \Omega_i + \rho \Psi \right|^2 \right)
\]

(20)

\[
\Pi_k = U_{RF,k}^H \left( \sum_{f \in \mathcal{F}_{req}} H_k M_f M_f^H H_k^H + H_k N N^H H_k^H + \sigma_k^2 I_{N_{RF} \times N_{RF}} \right) U_{RF,k}
\]

(22)

\[
\begin{align*}
\text{maximize} & \quad \sum_{f \in \mathcal{F}_{req}} r_f - \lambda \left( \eta \sum_{i \in K_R} \Phi_i (\mathcal{M}, \mathcal{N}) + p_c \right) - \Xi_s, \\
\text{subject to} & \quad r_{f_k} \leq \ln \det (\Sigma_k) - \text{Tr} (\Sigma_k E_k) + d_{f_k}, \quad \forall k \in K_U, \\
& \quad (11c), (11d), (11f), (11g), (16c).
\end{align*}
\]

(23a)
It is not difficult to find that the objective of problem (26) have the form $\phi(X) \triangleq \text{Tr}(X^HAXC) - 2R(\text{Tr}(X^HB))$, where $A$ and $C$ are positive semidefinite. First, we can express $\phi(X)$ as a quadratic function of $X(i,j)$ in the form of $\phi(X(i,j)) \triangleq a|X(i,j)|^2 - 2R(b^HX(i,j))$. Then, by using the constant module constraint, the derivation of $\phi(X(i,j))$ with respect to $X(i,j)$, and the derivation of $\phi(X)$ with respect to $X$, a block coordinated ascent optimization is designed to solve problem (26) with complexity $O\left(N_{\text{RF}}^2N_{\text{RF}}^2\right)$. The details of the specific optimization algorithm can be referred to [14, Algorithm 4 listed in Table V, pp. 466].

In addition, when other variables are given, the solutions of $U_{\text{RF},k}$ and $\Sigma_k$, $\forall k \in K_{\text{U}}$, can be updated via solving the following problem

$$\text{maximize } U_{\text{RF},k} \ln \text{det}(\Sigma_k) - \text{Tr}(\Sigma_k E_k) + d_{f_k}. \quad (28)$$

Problem (28) is an unconstrained optimization and easy to solve via its stationary points. Since each matrix $\Sigma_k$ is positive definite, the optimal solutions of $U_{\text{RF},k}$ and $\Sigma_k$, $\forall k \in K_{\text{U}}$, can be analytically given respectively by

$$U_{\text{RF},k}(o) = (U_{\text{RF},k}^H \Upsilon_k U_{\text{RF},k})^H \Upsilon_k H_k M_{f_k}, \quad (29)$$

and

$$\Sigma_k(o) = E_k^H = \left(I_{d_{f_k} \times d_{f_k}} - U_{\text{RF},k}^H \Upsilon_k H_k M_{f_k}\right)^H. \quad (30)$$

The complexity of computing $U_{\text{RF},k}(o)$ and $\Sigma_k(o)$ is $O(\varphi_1)$ with $\varphi_1 = N_{\text{RF}}^2N_{\text{RF}} + N_{\text{RF}}d_{f_k} + N_{\text{RF}}^2N_{\text{RF}} + N_{\text{RF}}d_{f_k}$ and $d_m = \max_{k \in K_{\text{U}}} d_{f_k}$ [45].

3) Optimization of Auxiliary Variable, Baseband Precooding and Noise Covariance: In this subsection, we focus on updating the solutions of variables $R$, $M$, $N$, $G$, and $O$ with fixed other variables. When other variables are given, the updates of variables $R$, $M$, $N$, $G$, and $O$ can be achieved via solving problem (31), at the bottom of this page. In problem (32), when optimization variables $F_{\text{RF}}$, $U_{\text{BB},k}$, and $Z$ are given, and constraints (31a), (31b) and (31c) are convex. However, due to the existing of non-convex fronthaul link rate $g_t(G, O)$, i.e., non-convex constraint $g_t(G, O) \leq C_i$, problem (32) is non-convex and difficult to obtain directly the optimal solution. To address problem (32), first, by substituting (21) into (31a), problem (31) can be reformulated as follows

$$\begin{align*}
\text{maximize}_{R,M,N,G,O} & \quad \sum_{f \in F_{\text{req}}} r_f - \lambda \left(\eta \sum_{i \in K_{\text{R}}} \bar{p}_i(M,N) + p_c\right) - \Xi_s, \\
\text{subject to} & \quad a \left|U_{\text{RF},k}\right|^2 = 1, m_k \in N_{\text{RF}}, n_k \in N_{\text{RF}} \\
& \quad \sum_{f \in F_{\text{req}}} r_f - \lambda \left(\eta \sum_{i \in K_{\text{R}}} \bar{p}_i(M,N) + p_c\right) - \Xi_s, \\
& \quad r_f \leq S, \forall f \in F_{\text{req}} \\
& \quad \bar{p}_i(M,N) \leq P_i, \forall i \in K_{\text{R}}, \\
& \quad g_t(G, O) \leq C_i, \Omega_i \geq 0
\end{align*} \quad (31a)$$

Therefore, in the sequel, we replace constraints (31e) with $\bar{p}_i(M,N) \leq P_i$, $\forall i \in K_{\text{R}}$. Thus, problem (32) is reformulated as

$$\begin{align*}
\text{maximize}_{R,M,N,G,O,L} & \quad \sum_{f \in F_{\text{req}}} r_f - \lambda \left(\eta \sum_{i \in K_{\text{R}}} \bar{p}_i(M,N) + p_c\right) - \Xi_s, \\
\text{subject to} & \quad \left|U_{\text{RF},k}\right|_{m_k,n_k}^2 = 1, m_k \in N_{\text{RF}}, n_k \in N_{\text{RF}} \\
& \quad \sum_{f \in F_{\text{req}}} r_f - \lambda \left(\eta \sum_{i \in K_{\text{R}}} \bar{p}_i(M,N) + p_c\right) - \Xi_s, \\
& \quad r_f \leq S, \forall f \in F_{\text{req}} \\
& \quad \bar{p}_i(M,N) \leq P_i, \forall i \in K_{\text{R}}, \\
& \quad g_t(G, O) \leq C_i, \Omega_i \geq 0
\end{align*} \quad (31a)$$
Algorithm 2 Solution of Problem (30)

1: Let \( t = 0 \), initialize \( G^{(t)} \) and \( O^{(t)} \).
2: Compute auxiliary variables \( L_i, \forall i \in K_R \), via (36).
3: Let \( t = t + 1 \). Solve problem (37) for a given \( L_i, \forall i \in K_R \), then obtain \( R^{(t)}, M^{(t)}, N^{(t)}, G^{(t)}, \) and \( O^{(t)} \).
4: Repeat Steps 2 and 3 until convergence.

subject to \( a_k \leq b_k, \forall k \in K_U \), \( r_f \leq S, \forall f \in \mathcal{F}_{\text{req}} \), \( \mathcal{P}_i(M, N) \leq P_i, \forall i \in K_R \), \( \mathcal{G}(G, O, L) \leq C_i, \Omega_i \geq 0 \).

Problem (37) is still non-convex and difficult to obtain directly the global optimal solution.

Note that \( g_i(G, O, L) \) is not jointly convex w.r.t. variables \( G \), \( O \), and \( L \), however, it is jointly convex w.r.t. \( G \) and \( O \) for a given \( L \). When \( L \) is fixed, problem (37) can be easily solved via classical convex optimization methods, such as the primal-dual interior-point methods [46], [47]. While the solution of variable \( L_i \) has an analytical expression, i.e., (36). It implies that problem (37) can be solved via alternative optimization method. In particular, the update of variables \( R \), \( M \), \( N \), \( G \), and \( O \) can be achieved by solving problem (37) for a given \( L \). Then, the update of variable \( L \) can be finished by equality (36) for given \( R \), \( M \), \( N \), \( G \), and \( O \). The detailed steps are summarized in Algorithm 2. After each update at Step 3 of Algorithm 2, this algorithm generates a nondecreasing sequence of the objective (37a). Furthermore, the objective (37a) has an upper boundary due to the limited transmit power in practical communication system. Therefore, the convergence of Algorithm 2 can be guaranteed by the monotonic boundary theorem [48]. The computational complexity of Algorithm 2 is \( O \left((\varphi_2 + \varphi_3)^{1.5} + \varphi_4\right) \), where \( \varphi_2 = K_R |\mathcal{F}_{\text{req}}| N_{d,m} \), \( \varphi_3 = N_{\text{HF}} (N_{\text{HR}} + K_R |\mathcal{F}_{\text{req}}| d_f + N_{\text{HR}} K_R) \), and \( \varphi_4 = |\mathcal{F}_{\text{req}}| N_{\text{HR}}^2 d_f + N_{\text{HR}}^3 \).

4) Optimization of RF Precoders: In this subsection, we pay our attention to the solution of \( F_{\text{RF},i}, \forall i \in K_R \). Similarly, we note that variables \( F_{\text{RF},i}, \forall i \in K_R \) only appear in the third item of objective function (23a). As a result, the solution of \( F_{\text{RF},i}, \forall i \in K_R \), can be obtained by solving the following problem

\[
\begin{align*}
\text{minimize} & \quad \Xi_{m_i, n_i} \\
\text{subject to} & \quad |F_{\text{RF},i}|_{m_i, n_i}^2 = 1, \quad m_i \in N_{I_R}, \quad n_i \in N_{I_C}.
\end{align*}
\]

(38)

By appropriate rearrange, problem (38) can be equivalently formulated as follows

\[
\begin{align*}
\text{minimize} & \quad \text{Tr} \left(F_{\text{RF},i}^H T_{\text{RF},i} C_i \right) - 2R \left(\text{Tr} \left(F_{\text{RF},i}^H C_i \right) \right) \\
\text{subject to} & \quad |F_{\text{RF},i}|_{m_i, n_i}^2 = 1, \quad m_i \in N_{I_R}, \quad n_i \in N_{I_C}.
\end{align*}
\]

(39)

where \( C_i = \sum_{f \in \mathcal{F}_{\text{req}}} \frac{1}{2p} G_{f,i} G_{f,i}^H + \frac{1}{2g} \Omega_i \) and \( T_{\text{RF},i} = \sum_{f \in \mathcal{F}_{\text{req}}} \frac{1}{2p} (M_{f,i} + \rho \Phi_{f,i}) G_{f,i}^H + \frac{1}{2g} (N_i + g \Psi_i) \Omega_i^H \). Similar to problem (26), problem (39) can also be solved with complexity \( O \left(N_{\text{HF}}^2 N_{\text{HR}}^2 \right) \) [14, Algorithm 4 listed in Table V, pp. 466].

5) Summarization of Optimization Algorithm: In this subsection, we summarize the details of our proposed optimization algorithm that is used to solve problem (18). The step-by-step description for solving problem (18) is given in Algorithm 3, where \( l \) and \( t \) denote the number of iterations, respectively. \( \epsilon \) and \( \varepsilon \) denote respectively a stopping threshold and an approximation stopping threshold. \( \omega \) is a control parameter. \( \zeta^{(l)} \) denotes the objective value of problem (23) at the \( t \)-th iteration. In particular, we firstly solve problem (19) with given Lagrange Multiplier matrices \( \Phi_{f,i} \) and \( \Psi_i \) and penalty parameters \( \rho \) and \( \varrho \) using the block coordinate descent method, i.e., Steps 3 to 8. Then we update the Lagrange Multiplier matrices \( \Phi_{f,i} \) and \( \Psi_i \) or the penalty parameters \( \rho \) and \( \varrho \) according to the constraint violation condition \( h(F_{\text{RF},i}^H G_{f,i}, O^{(l)}, M^{(l)}, N^{(l)}) \leq \zeta^{(l)} \) where

\[
\begin{align*}
& h(F_{\text{RF},i}^H G_{f,i}, O^{(l)}, M^{(l)}, N^{(l)}) \leq \zeta^{(l)} \\
& \Delta = \sum_{f \in \mathcal{F}_{\text{req}}} \sum_{i \in K_R} \|M_{f,i} - F_{\text{RF},i} G_{f,i}\|_F \\
& \quad + \sum_{i \in K_R} \|N_i - F_{\text{RF},i} \Xi_i\|_F.
\end{align*}
\]

That is, if the constraint violation condition is satisfied, we update the Lagrange Multiplier matrices \( \Phi_{f,i} \) and \( \Psi_i \), otherwise, we update the penalty parameters \( \rho \) and \( \varrho \). Note that problem (24) can also be solved with a similar procedure used in Algorithm 3. The only difference is that we need to remove the second item in the objective (32a) for updating variables \( R \), \( M \), \( N \), \( G \), and \( O \).

In Algorithm 3, in Steps 4 to 8, each optimization tries to maximize problem (23). It implies that the iterative optimization from Steps 4 to 8 generates a non-decreasing sequence, i.e., \( \zeta^{(1)} \leq \zeta^{(2)} \leq \cdots \leq \zeta^{(t-1)} \leq \zeta^{(t)} \leq \cdots \).

Meanwhile, recalling the limited allowable transmit power in practical communication system, the objective of problem (23) has an upper bound. Therefore, the inner loops can converge to the KKT solution of problem according to the research results in [49, Theorem 1]. Further, following the similar arguments as those in [14, Corollary 3.1], Algorithm 3 converges to a KKT solution of problem (18). Therefore, Algorithm 3 can be guaranteed to converge to a KKT solution of problem (16) [42], [43]. In addition, the total computational complexity of Algorithm 3 is \( O \left(K_{\text{RF}} \left(N_{\text{HF}}^2 N_{\text{HR}}^2 + N_{\text{RF}}^2 N_{\text{HR}}^2 + \varphi_1 + (\varphi_2 + \varphi_3)^{3.5} + \varphi_4\right) \right) \), where \( K_{\text{RF}} \) denotes the average number of the operations of Steps 4 to 8 in each execution of Step 3 and \( \kappa_{\text{RF}} \) represents the number of the operations of Step 3.

C. Discussion

In the aforementioned two subsections, an alternative optimization method is designed to solve problem (11). A few sub-algorithms proposed for solving each subproblem are iterative
algorithms, except for the optimization of baseband combiners and weight matrices. While these sub-algorithms require a number of iterations for achieving convergence. Therefore, to some extent, the proposed algorithm is a time-consuming optimization method. It implies that this algorithm incurs a certain computation delay and consume a huge computational burden, especially for massive mmWave MIMO multiuser system. On the other hand, in this paper, we assume that the BBU can obtain the perfect CSI for each optimization. Once the channel condition changes and the scheduled user group changes, we need to rerun this optimization method to obtain the corresponding transmission mechanism. This means that the proposed algorithm has no the ability to adapt the fast time-varying communication environments. Therefore, how to make full use of the cache at the network edge is still challenging problem for reducing really the burden on the fronthaul links and the delivery latency for cache-enabled wireless communication system.

### IV. Numerical Results

In this section, we present numerical results to evaluate the performance of the proposed transceiver optimization algorithm for cache-enabled energy-efficient mmWave RANs, in which the positions of the eRRHs and the users are uniformly distributed within a circular cell of radius 500 m. For easy of illustration, we assume that all eRRHs have the same maximum transmit power and fronthaul capacity, i.e., $P_t = P$ and $C_i = C$, $\forall i \in K_R$. The average path-loss power is $\rho_{p,k,i} = \frac{1}{1 + (d_{p,k,i}/d_0)}$ with $d_0$, $d_{p,k,i}$, and $\alpha$ being the reference distance, the propagation distance of the $p$-th path between eRRH $i$ and user $k$ in meters ($p = 0, \cdots, N_c$), and the pathloss exponent, respectively. For simplicity, we assume that the LoS link always exists between eRRHs and users. The AoDs/AoAs are assumed to take continuous values and are uniformly distributed in $[0, 2\pi]$. All users have the same noise variance, i.e., $\sigma_k^2 = \sigma^2$, $k \in K_U$. The eRRHs are equipped with caches of equal size, i.e., $B_i = B = \xi SF$, $i \in K_R$, where $\xi$ denotes the fractional caching capacity. The cache states $c_{f,i}$, $f \in F$, $i \in K_R$, are randomly set subject to $\sum_{f \in F} c_{f,i} = \lfloor \xi LF \rfloor$. If not stated otherwise, the values of the system parameters are as in Table II. The set $\mathcal{F}_{req}$ of requested files is randomly generated for each channel realization.

For comparison, we also simulate the minimum user rate for fully digital precoding for downlink cache-enabled RANs as performance benchmark. The power consumption of a complete antenna array with one RF chain per antenna is calculated as $[9]$

$$p_{dc} = P_{RFC} + P_{RFC} + P_{LNAC} + 2P_{LO} + 2P_{BB}, \quad (42a)$$
$$P_{RFC} = K_R N_c (P_{DAC} + P_{mixer} + P_{LPF}), \quad (42b)$$
$$P_{RFC} = K_U N_t (P_{ADC} + P_{mixer} + P_{LPF}), \quad (42c)$$
$$P_{LNAC} = K_U N_t P_{LANA}. \quad (42d)$$

In our simulation, the initialization of RF precoding matrix $\mathbf{F}_{RF,i}$ is randomly generated. The quantization covariance matrix $\mathbf{Q}_i$ is set to be $v I_{N_{RF}}$ with $0 < v < 0.02$. The digital precoding matrix $\mathbf{G}_{f,i}$ is initialized as $\mathbf{G}_{f,i} = \sqrt{\beta} \mathbf{f}_{f,i}$, where the columns of $\mathbf{G}_{f,i}$ are mutual orthogonal unit norm column vectors. The coefficient $\beta$ is defined to satisfy the power constraint and the fronthaul capacity constraint and is given by (43), at the bottom of the next page, where $0 < \kappa < 0.02$. In the legend of the figures, “Hybrid-Precoding-EEMax” and “Digital-Precoding-EEMax” denote the proposed hybrid precoding solution and fully digital precoding solution for maximizing the energy efficiency for partial caching at the network edge, respectively. While, “Hybrid-Precoding-DRMax” and “Digital-Precoding-DRMax” denote the proposed hybrid precoding solution and fully digital precoding solution for maximizing the delivery rate for partial caching at the network edge, respectively. For the case of partial/full/no caching at the network edge, the depicted curves have the term “with

1In the future work, we would like to further compare the performance of our proposed method with that of the exhaustive search method for cache-enabled mmWave communication systems with different configuration of RF chains and antennas.
TABLE II
VALUES OF SIMULATION PARAMETERS

| Symbol | Value | Symbol | Value | Symbol | Value | Symbol | Value | Symbol | Value |
|--------|-------|--------|-------|--------|-------|--------|-------|--------|-------|
| $K_U$  | 5     | $K_R$  | 3     | $N_{RF}$ | 2     | $N_{RF}$ | 1     | $N_c/N_r$ | 8/4   |
| $P$    | 10    | $\eta$, $\sigma^2$ | 1     | $d_f$ | 1     | $\xi$, $\sigma_f^2$ | 1     |
| $d_0$  | 50 m  | $\alpha$, $\beta$ | 3     | $c_i^{(0)}$ | 10$^{-3}$ | $\epsilon$ | 10$^{-3}$ |
| $P_{LNA}$ | 20 mW | $P_{PS}$ | 30 mW | $P_{ADC}$ | 200 mW | $P_{ADD}$ | 19 mW | $P_{DBG}$ | 14 mW |

Fig. 2. Energy efficiency versus fronthaul capacity $C$ for different algorithms.

Fig. 3. Delivery rate versus fronthaul capacity $C$ for different algorithms.

Fig. 4. Performance comparison versus file size $S$ for different algorithms.

PC", “with FC” and “with NC” in the corresponding labels for distinction.

Fig. 2 and Fig. 3 illustrate the energy efficiency and delivery rate comparison versus the fronthaul capacity $C$ for different algorithms with different caching proportions, $S = 1.2$ nats and $P = 20$ dB. Numerical results reveal that when the system performance is limited by other factors rather than the fronthaul capacity $C$, the system energy efficiency and delivery rate obtained by various algorithms increases with the fronthaul capacity $C$. When the fronthaul capacity $C$ is sufficient large, the system energy efficiency and delivery rate obtained by various algorithms keep unchange. Though the hybrid antenna architecture has the ability to reduce the cost of RF chains, from the perspective of system energy efficiency and delivery rate, the hybrid antenna architecture is not an optimum scheme for mmWave communication. This is because the directional transmission of mmWave communication may reduce the spatial/multipath diversity gain and cause additional power consumption and hardware cost with the phase shifter networks. In addition, numerical results also illustrate that caching popular files at the network edge helps to improve the system energy efficiency and delivery rate of cache-enabled mmWave RANs.

Fig. 4 illustrates the energy efficiency and delivery rate comparison versus the file size $S$ for different algorithms with $C = 1.2$ nats/Hz/s and $P = 20$ dB. Numerical results show that when the system performance is limited by the file size $S$, Digital-Precoding-EEMax achieves the best energy efficiency performance, while Hybrid-Precoding-DRMax achieves the worst energy efficiency performance with $p_{ac} = 5.813$ W and $p_{dc} = 11.062$ W, respectively. Digital-Precoding-DRMax and Digital-Precoding-EEMax obtain respectively the optimal and worst system performance in terms of the delivery rate. The results illustrated in Fig. 4 further show that the optimal solution of maximizing system energy efficiency does not necessary maximizing system delivery rate for cache-enabled mmWave RANs with limited file size $S$. The hybrid antenna architecture achieves the worst energy efficiency and an

$$
\beta = \begin{cases} 
\kappa \minimize \left( P \sum_{f \in F_{req}} \text{tr} \left( \mathbf{F}_{R,i} \mathbf{G}_{f,i}^H \mathbf{F}_{RF,i}^H \right) + \text{tr} \left( \mathbf{F}_{R,i} \Omega_i \mathbf{F}_{RF,i}^H \right) \right), & \sum_{f \in F_{req}} c_{f,i} \neq 0, \\
\frac{P}{\kappa} \sum_{f \in F_{req}} \text{tr} \left( \mathbf{F}_{R,i} \mathbf{G}_{f,i}^H \mathbf{F}_{RF,i}^H \right) + \text{tr} \left( \mathbf{F}_{R,i} \Omega_i \mathbf{F}_{RF,i}^H \right), & \sum_{f \in F_{req}} c_{f,i} = 0 
\end{cases}
$$

(43)
eclectic delivery rate performance. This is because the hybrid antenna architecture reduce the hardware cost while increasing the circuit power consumption with an increasing number of phase shifters and reducing the diversity gain with directional transmission. In addition, fully digital antenna architecture has more degree of freedom to optimize the transceivers such that a better performance may be realized.

To further evaluate the system performance with different number of receiving and transmitting antennas, Table III lists the energy efficiency and delivery rate performance for different algorithms with $C = 2$ nats/Hz/s, $S = 1.2$ nats and $P = 20$ dB. Numerical results show that increasing the number of RF chains can improve the delivery rate and energy efficiency performance of hybrid architecture mmWave communication systems. However, for the fixed number of RF chains, increasing the number of antennas can only improve the delivery rate performance, while, the energy efficiency performance becomes worse for hybrid architecture mmWave communication systems.

### V. Conclusion

In this paper, we have studied the design of energy-efficient transceivers for cache-enabled mmWave RANs where each edge node is equipped not only with the functionalities of standard RRHs, but also with local cache and baseband processing capabilities. First, we use the fractional programming theory, PDD, and the relation between user rate and minimum mean square errors to transform the original problem of interest into a tractable form. Then, an effective optimization algorithm based on ADMM with provable convergence was developed to solve the transformed problem. Numerical results were provided to validate the effectiveness of the proposed algorithm and revealed that the hybrid antenna architecture is not necessarily the best scheme for cache-enabled mmWave RANs from the system performance perspective. But, it is a tradeoff between the system performance and the hardware cost and power consumption. In addition, caching requested frequently files at the network edge helps to improve the system performance of wireless communications.

### REFERENCES

[1] A. Gupta and R. K. Jha, “A survey of 5G network: Architecture and emerging technologies,” IEEE Access, vol. 3, pp. 1206–1232, 2015.

[2] P. Schultz et al., “Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture,” IEEE Commun. Mag., vol. 55, no. 2, pp. 70–78, Feb. 2017.

[3] M. Peng, S. Yan, K. Zhang, and C. Wang, “Fog-computing-based radio access networks: Issues and challenges,” IEEE Netw., vol. 30, no. 4, pp. 46–53, Jul. 2016.

[4] H. Liu, F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, and Y. Zhang, “Mobile edge cloud system: Architectures, challenges, and approaches,” IEEE Syst. J., vol. 12, no. 3, pp. 2495–2508, Feb. 2017.

[5] S. He, W. Huang, J. Wang, J. Ren, Y. Huang, and Y. Zhang, “Cache-enabled coordinated mobile edge network: Opportunities and challenges,” IEEE Wireless Commun., pp. 1–8, 2020.

[6] T. S. Rappaport et al., “Millimeter wave mobile communications for 5G cellular: It will work!” IEEE Access, vol. 1, pp. 335–349, 2013.

[7] R. W. Heath, Jr., N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” IEEE J. Sel. Topics Signal Process., vol. 10, no. 3, pp. 436–453, Apr. 2016.

[8] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., “Spatially sparse precoding in millimeter wave MIMO systems,” IEEE Trans. Wireless Commun., vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[9] R. Mendez-Rial, C. Rusu, N. Gonzalez-Prelcic, A. Alkhateeb, and R. W. Heath, Jr., “Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?” IEEE Access, vol. 4, pp. 247–267, 2016.

[10] S. Park, A. Alkhateeb, and R. W. Heath, Jr., “Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems,” IEEE Trans. Wireless Commun., vol. 16, no. 5, pp. 2907–2920, May 2017.

[11] W. Ni, X. Dong, and W.-S. Lu, “Near-optimal hybrid processing for massive MIMO systems via matrix decomposition,” IEEE Trans. Signal Process., vol. 65, no. 15, pp. 3922–3933, Aug. 2017.

[12] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, “Codebook-based hybrid precoding for millimeter wave multiuser systems,” IEEE Trans. Signal Process., vol. 65, no. 20, pp. 5289–5304, Oct. 2017.

[13] C. G. Tsinos, S. Maleki, S. Chatzinotas, and B. Ottersten, “On the energy-efficiency of hybrid analog–digital transceivers for single- and multi-carrier large antenna array systems,” IEEE J. Sel. Areas Commun., vol. 35, no. 9, pp. 1980–1995, Sep. 2017.

[14] Q. Shi and M. Hong, “Spectral efficiency optimization for millimeter wave multiuser MIMO systems,” IEEE J. Sel. Topics Signal Process., vol. 12, no. 3, pp. 455–468, Jun. 2018.

[15] B. Wang, F. Guo, S. Jin, H. Lin, and G. Y. Li, “Spatial- and frequency-wideband effects in millimeter-wave massive MIMO systems,” IEEE Trans. Signal Process., vol. 66, no. 13, pp. 3393–3406, Jul. 2018.

[16] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., “Channel estimation and hybrid precoding for millimeter wave cellular systems,” IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 831–846, Oct. 2014.

[17] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, “Backhaul-aware caching placement for wireless networks,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Dec. 2014, pp. 1–6.

[18] J. Kwak, Y. Kim, L. B. Le, and S. Chong, “Hybrid content caching in 5G wireless networks: Cloud versus edge caching,” IEEE Trans. Wireless Commun., vol. 17, no. 5, pp. 3030–3045, May 2018.

[19] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, “Random caching based cooperative transmission in heterogeneous wireless networks,” IEEE Trans. Commun., vol. 66, no. 7, pp. 2809–2825, Jul. 2018.

| Configurations | Algorithms          | Energy Efficiency | Delivery Rate |
|----------------|---------------------|-------------------|---------------|
| $N_{RF} = 4$, $N_{uRF} = 1$, $N_t = 8$, $N_r = 4$ | Hybrid-Preceding-EEMax with PC | 0.0167          | 4.7235        |
|                 | Digital-Preceding-EEMax with PC | 0.1436          | 2.6324        |
|                 | Hybrid-Preceding-DRMax with PC | 0.0167          | 4.7228        |
|                 | Digital-Preceding-DRMax with PC | 0.0193          | 4.7351        |
| $N_{RF} = 4$, $N_{RF} = 1$, $N_t = 16$, $N_r = 4$ | Hybrid-Preceding-EEMax with PC | 0.0162          | 4.7319        |
|                 | Digital-Preceding-EEMax with PC | 0.0183          | 4.7739        |
|                 | Hybrid-Preceding-DRMax with PC | 0.1336          | 3.3788        |
|                 | Digital-Preceding-DRMax with PC | 0.0223          | 4.790         |
| $N_{RF} = 4$, $N_{RF} = 1$, $N_t = 16$, $N_r = 4$ | Hybrid-Preceding-EEMax with PC | 0.0159          | 4.7738        |
|                 | Digital-Preceding-EEMax with PC | 0.1159          | 4.6689        |
|                 | Hybrid-Preceding-DRMax with PC | 0.0155          | 4.7682        |
|                 | Digital-Preceding-DRMax with PC | 0.0150          | 4.8000        |
Wei Huang (Member, IEEE) received the B.S. and M.S. degrees from the Anhui University of Science and Technology, Huainan, China, in 2010 and 2013, respectively, and the Ph.D. degree in information and communication engineering from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2018. Since December 2018, he has been a Faculty with the School of Computer Science and Information Engineering, Hefei University of Technology. He is currently involved in wireless caching, fog ran, mobile edge computation, and metasurface. His research interests are in the wide range of massive MIMO, millimeter wave MIMO wireless communication, underwater acoustic communication, blind signal processing, sparse signal processing, and optimization theory. He was a Reviewer for various journals, including IEEE Transactions on Wireless Communications, IEEE Transactions on Vehicular Technology, IEEE Wireless Communications Letters, EURASIP Journal on Wireless Communications and Networking, and Circuits, Systems, and Signal Processing.

Yongming Huang (Senior Member, IEEE) received the B.S. and M.S. degrees from Nanjing University, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, China, in 2007. From 2008 to 2009, he was visiting the Signal Processing Lab, Electrical Engineering Department, Royal Institute of Technology (KTH), Stockholm, Sweden. Since March 2007, he has been a faculty with the School of Information Science and Engineering, Southeast University, where he is currently a Full Professor. He has published over 200 peer-reviewed papers and holds over 60 invention patents. He submitted around 20 technical contributions to IEEE standards and was awarded a certificate of appreciation for outstanding contribution to the development of IEEE standard 802.11aj. His current research interests include MIMO wireless communications, cooperative wireless communications, and millimeter wave wireless communications. He has served as an Associate Editor for IEEE Transactions on Signal Processing, IEEE Wireless Communications Letters, EURASIP Journal on Advances in Signal Processing, and EURASIP Journal on Wireless Communications and Networking, and a Guest Editor for the IEEE Journal Selected Areas in Communications.

Ming Xiao (Senior Member, IEEE) received the bachelor’s and master’s degrees in engineering from the University of Electronic Science and Technology of China, Chengdu, in 1997 and 2002, respectively, and the Ph.D. degree from the Chalmers University of Technology, Sweden, in November 2007. From 1997 to 1999, he worked as a Network and Software Engineer at China Telecom. From 2000 to 2002, he held a position at the Sichuan Communications Administration. Since November 2007, he has been with the Department of Information Science and Engineering, School of Electrical Engineering and Computer Science, Royal Institute of Technology, Sweden, where he is currently an Associate Professor. He was an Editor of IEEE Transactions on Communications from 2012 to 2017 and IEEE Wireless Communications Letters from 2012 to 2016, and has been a Senior Editor of IEEE Communications Letters since January 2015 and an Editor of IEEE Transactions on Wireless Communications since 2018. He was the Lead Guest Editor of IEEE JSAC special issue on Millimeter Wave Communications for Future Mobile Networks in 2017. He has served as an Area Editor for the IEEE Open Journal of the Communication Society since 2019.

Yaoxue Zhang (Senior Member, IEEE) received the B.Sc. degree from the Northwest Institute of Telecommunication Engineering, China, in 1982, and the Ph.D. degree in computer networking from Tohoku University, Japan, in 1989. He is currently a Professor with the School of Computer, Central South University, China, and also a Professor with the Department of Computer Science and Technology, Tsinghua University, China. He has published over 200 technical papers in international journals and conferences, as well as 9 monographs and textbooks. His research interests include computer networking, operating systems, ubiquitous/pervasive computing, transparent computing, and big data. He is a fellow of the Chinese Academy of Engineering. He is currently serving as the Editor-in-Chief for Chinese Journal of Electronics.