Performance Analysis of NOMA Multicast Systems Based on Rateless Codes with Delay Constraints

Performance Analysis of NOMA Multicast Systems Based on Rateless Codes with Delay Constraints

Yingmeng Hu, Rongke Liu, *Senior Member, IEEE,* Aryan Kaushik, *Member, IEEE,* and John Thompson, *Fellow, IEEE*

***Abstract*—To achieve an efficient and reliable data transmission in time-varying conditions, a novel non-orthogonal multiple access (NOMA) transmission scheme based on rateless codes (NOMA-RC) is proposed in the multicast system in this paper. Using rateless codes at the packet level, the system can generate enough encoded data packets according to users' requirements to cope with adverse environments. The performance of the NOMA-RC multicast system with delay constraints is analyzed over Rayleigh fading channels. The closed-form expressions for the frame error ratio and the average transmission time are derived for two cases which are a broadcast communication scenario (Scenario 1) and a relay communication scenario (Scenario 2). Under the condition that the quality of service for the edge user is satisfied, an optimization model of power allocation is established to maximize the sum rate. Simulation results show that Scenario 2 can provide better block error ratio performance and exhibit less transmission time than Scenario 1. When compared with orthogonal multiple access (OMA) with rateless codes system, the proposed system can save on the transmission time and improve the system throughput.**

***Index Terms*—Multicast system, rateless codes, NOMA, delay constraints, throughput.**

I. Introduction

MULTIMEDIA services such as video conferencing, online teaching, and interactive games will gradually become mainstream services on the mobile network [1]–[3]. The multimedia broadcast/multicast service (MBMS) scheme has been proposed for use in Long Term Evolution (LTE) [4], which can simultaneously transmit data to multiple users. The MBMS system provides the services for multiple users with fewer resources [5]. However, receivers in a multicast scenario may suffer from different packet loss ratios due to their adverse environments, which results in high feedback and re-transmission costs. As an effective error-control code, rateless code is applied in multicast systems [6]–[8]. The rateless coding system does not depend on channel conditions, it generates an infinite number of encoded packets as needed. The multicast system with rateless codes can cope with more complex conditions and reduce the feedback.

Manuscript received XXX, XX, 2020; revised XXX, XX, 2021. This work was supported by the National key research and development program under Grant 2020YFB1807102.

Y. Hu, and R. Liu are with the School of Electronic and Information Engineering, Beihang University, Beijing, 100191, China (e-mail: { huyingmeng, rongke_liu}@buaa.edu.cn). *(Corresponding authors: Rongke Liu)*

A. Kaushik is with Department of Electronic and Electrical Engineering, University College London (UCL), UK (e-mail: a.kaushik@ucl.ac.uk).

J. Thompson is with the Institute for Digital Communications, The University of Edinburgh, Edinburgh EH3 9JL, UK (e-mail: john.thompson@ed.ac.uk).

However, the traditional systems fail to satisfy the rapidly growing various services of mobile devices in the future due to finite spectrum resources. It also brings some new challenges for the next generation of wireless access networks [9]. Non-orthogonal multiple access (NOMA) technology can carry the signals of multiple users to the same time-frequency resources for transmission [10]–[12]. Thus, the NOMA technology is an effective way to improve the spectral efficiency, and is considered as a promising candidate for multiple access in future mobile communication networks [13]. For downlink multicast scenarios in the paper, the NOMA technology based on rateless codes is proposed to reduce the overhead caused by feedbacks and further improve the channel utilization.

A. Literature Review

NOMA has recently gained widespread attention in industry and academia. Combined with other key technologies such as hybrid automatic repeat request (HARQ), multiple input multiple output (MIMO), NOMA has many potential applications such as relay assisted communications, millimeter wave communications and drone communications [14]–[16]. To decrease the outage probability, a cooperative HARQ-assisted NOMA scheme is presented for the large-scale device-to-device networks in [14]. NOMA transmission technology is applied in a unmanned aerial vehicle aided communication network to serve a large number of mobile users in a hotspot area in [15]. The performance of the MIMO-NOMA systems with multiple users is analyzed in [16], where the MIMO-NOMA scheme outperforms the MIMO-OMA system in terms of sum rate. Power factors play extremely important role on users for NOMA system. Some researchers have conducted some works on power factors [17]–[19]. Taken as a joint optimization problem among the coordinating base stations, a power allocation algorithm is presented for multi-point transmission in multi-cell netorks in [17]. The power allocation strategies are proposed to maximize the minimum user transmission rates and the sum rate in [18] and [19], respectively. These strategies are closely related to the gains of the channels, so they dependent on the real-time feedbacks of the channel conditions heavily. The performance of the system can also deteriorate seriously when the channels change rapidly.

In addition, these works are designed for the unicast systems. With the increasing number of services, as part of the third generation (3GPP) LTE standard [20], multicast systems have attracted a lot of attention. They can transmit multicast data services to multiple users at the same time. Cooperative

multicast mmWave wireless networks are discussed in [21], where users can decode multicast data layers according to their quality of services. The authors demonstrate an approach to obtain the maximum sum multicast rates and analyze the impact of data transmission rate and power allocation on the sum multicast rate. A cooperative non-orthogonal layered multicast multiple access is proposed to improve multicast users' spectrum efficiency and reliability in [22], where a successful user severs as a relay to help others. Both [21] and [22] improve the multicast users performance by layering and introducing relay assistance methods. Also, some researchers try to achieve the goal through feedbacks. The performance comparison of the fixed-block-length and the end-to-end delay systems is discussed when feedback is allowed, in [23], where the fixed-delay systems can achieve better gains for output-symmetric discrete memoryless channels. An radio resource management policy with users' channel feedbacks is proposed to satisfy video quality of the receivers and improve the system spectral efficiency in [24]. Besides, a hybrid unicast-multicast network selection scheme is designed to provide a trade-off between throughput, energy consumption and user satisfaction in dense heterogeneous networks in [25], where a user in signal overlay area of multiple cells can select a unicast or multicast transmission through the feedbacks.

Making use of the feedbacks helps to improve the system performance, but it also increases the complexity and the signaling overhead of the system. Most of the existing literature about NOMA-multicast systems considers the users in the same group, and the throughput performance needs to be further improved. When the channel conditions are relatively stable, the users may obtain good decoding results by using the traditional fixed rate coding schemes such as turbo codes. However, the systems will suffer from performance degradation in fast time-varying channels. Rateless codes, as a special kind of channel coding technique, can automatically adjust the rate in time-varying conditions without knowing the channel state information (CSI) in advance [26]–[29]. According to the decoding requirements, the transmitter can continuously generate many encoded data packets for a user until it successfully decodes its message. Then the user will feedback an acknowledgement message (ACK) to the transmitter. During the communication process, if a packet is lost, the transmitter will generate a new packet instead of re-transmitting the missing packet again with the feedbacks [30]. Luby transform (LT) codes proposed in [31] are the first practical rateless codes, but they suffer from high error floors in additive white Gaussian noise (AWGN) channels. As an improvement of LT codes, raptor codes take the concatenated coding scheme that the input data is first encoded by a high rate LDPC code, then the outputs are delivered to the LT encoder [32]. Raptor codes have a low encoding and decoding complexity and they have been applied in 3GPP and DVB standards [33]. With characteristics of good noise rejection and adaptive transmission ability in time varying conditions, raptor codes have attracted widespread attention since they were proposed, and widely are used in multicast systems. A massive access strategy based on raptor codes is proposed for dense cellular networks in [34], where several machine-type communication devices can be transmitted in the same resource block to improve the system throughput. By opportunistically utilizing the specific conditions of the users, a novel signal superposition transmission scheme based on rateless codes is proposed to enhance the system reliability in [35]. With the development of channel coding technologies, many other rateless codes have been proposed in literature such as stride codes [36] and spinal codes [37]. They have been widely applied in wireless communication systems such as broadcasting systems and relay communication systems. However, most of the existing literature on the analysis of rateless codes uses an ideal transmission scheme without delay constraints. The rateless coding system can continuously generate many different packets for a receiver to decode its own message, so it will complete the decoding with probability close to 1 when there is no limitation on the transmission time. Unfortunately, there is always a maximum tolerance time in a real system. When the receiver cannot complete the decoding of a block within the maximum transmission time, the block will be abandoned. Thus, the maximum transmission delay refers to the maximum tolerance time to deal with a block in this paper.

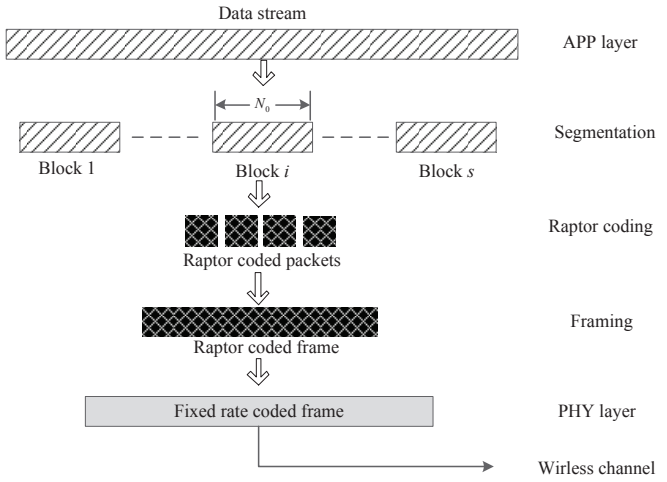

Fig. 1: Cross layer design structure.

B. Contributions

Considering the problems mentioned above, this paper proposes a NOMA downlink transmission scheme based on rateless codes in the multicast system, and analyzes the performance of the system applied in two scenarios with delay constraints over Rayleigh fading channels. Referring to [27]–[29], we present a simple cross layer design scheme, where the data in application (APP) layer is encoded by a raptor encoder at the packet level first. The cross layer design structure is shown in Fig. 1, where data stream from a service in the APP layer is divided into many blocks and each block consists of N_0 short data packets. Many rateless coded packets will be produced by a raptor encoder according to the decoding requirements. Next, a fixed number of packets form a rateless

coded data frame[1], which is used to unify the input length of the fixed rate encoder. As the rateless frame cannot be decoded independently, it will be further encoded by the fixed rate codes in the physical (PHY) layer.

According to the specific scenario environment, the NOMA-RC system can achieve advantageous block error rate (BLER) performance by adjusting the transmission power or the maximum transmission delay. For example, under the condition of limited delay, we can improve the system performance by amplifying the transmission power. Similarly, good performance of the system with power constraints can also be achieved by increasing the maximum transmission time. In addition, power allocation factors are very important parameters, which affect the performance of each user. Forcing the power of a user to increase may improve its performance, however it causes stronger interference to other users. Thus, how to allocate the power to the users is also an important problem. The main contributions of this paper are summarized as follows:

(1) This paper proposes a NOMA transmission scheme based on rateless codes in the multicast systems. According to the specific conditions of the users, the scheme provides different strategies to meet their needs, such as adjusting the transmission power or the maximum transmission time. Many encoded data packets can be generated by the proposed system to cope with the differences in channels between multicast users.

(2) The proposed scheme is applied to two practical application scenarios, where the performance of the systems with delay constraints is investigated including the transmission time, frame error ratio (FER) and BLER. We also derive the closed-form expressions for the FER and the transmission time in different stages. Furthermore, we obtain the lower bound of the transmission time in different scenarios.

(3) Given the total transmission power and the quality of service (QoS) of users, a model of power allocation is established to maximize the sum rate. Besides, the performance of the edge users are discussed in the two scenarios, where the throughputs of the proposed schemes is twice that of the OMA system when the transmission power is greater than 50 dBm.

The organization of the rest of this paper is as follows: Section II introduces the model of the NOMA-RC system. Section III presents the received signal of every node in different multicast scenarios. The performance of the NOMA-RC system in Scenario 1 is analyzed in detail in Section IV. The optimization strategy for power allocation is discussed in Section V. Next, Section VI provides the simulation results. Finally, Section VII summarizes this paper.

II. System Model

The block diagram of the NOMA-RC system is shown in Fig. 2. There are m_0 services need to be transmitted. First, data stream of a service is divided into many blocks. A block will be encoded by a raptor encoder to generate multiple data packets at the packet level. A fixed rate coded frame will be produced by the framing and channel encoding modules. Next, the signals of m_0 channels are respectively multiplied by different power factors adjusted by the power control module (PC). These signals are superimposed to form a composite one. It will be sent out by the transmitter antenna using an orthogonal frequency division multiple (OFDM) module.

There are m_0 groups to receive the superimposed signal. A user $U_{i,j}\,(1 \le i \le m_0, j \ge 1)$ with one antenna receives the signal through its demodulation module, where $U_{i,j}$ denotes the *j*-th user in the *i*-th G_i group. As there is no interference between users in the same group, a user can eliminate the interference from other groups using the successive interference cancellation (SIC) algorithm [38], and finally decodes its own message. This will be further discussed in Section III. The revised signal will be delivered to the channel decoder and if it successfully decodes, the message will be passed to the frame dividing module. Otherwise, the receiver will abandon it and continues to receive the subsequent packets. Thus, the channel between the channel decoding module and the frame dividing module is equivalent to the binary erasure channel (BEC). Next, these successfully decoded packets are forwarded to the raptor decoder. When the ACKs from the users in the same group are all received, the transmitter can re-allocate the power to the remaining users for the next transmission, or continues to send the packets of the next block according to different power allocation strategies. We will further discuss the two methods in Section III. A user with poor channel conditions may fail to accumulate enough packets within the maximum tolerance time. As a result, it cannot decode the message. As the transmission time reaches the maximum tolerance delay T_s, the system has to abandon the block and delete the packets of the block. In the next T_s, it will produce some new packets of the next block.

One PHY frame data is transmitted in one slot. We assume that the channels are frequency-flat, block-fading Rayleigh channels and the feedback channel is ideal and error-free. The channel gain $|h|$ obeys the Rayleigh distribution, so the cumulative distribution function (CDF) of the channel power gain $x{=}|h|^2$ is

$$F\left(x\right) = 1 - \exp\left(-\frac{x}{w}\right) \quad , \quad x \ge 0, \tag{1}$$

where w denotes the average received power. It satisfies [39]

$$w\left(\mathrm{dBm}\right) {=} P_t\left(\mathrm{dBm}\right) + K\left(\mathrm{dB}\right) - 10\Upsilon \lg\left(d_i\right), \tag{2}$$

where K is a constant coefficient related to each antenna element and the average channel loss, P_t is the transmission power, Υ is the path loss exponent and d_i is the distance between a transmitter and a receiver.

A user continuously accumulates the data packets until the number of the received packets satisfies $N \ge N_0\left(1+\nu\right)$, where ν denotes the overhead[2], which is a random variable [28]. Fig. 3 shows the CDF of ν, which is well approximated

[1]If the system transmits one packet at every slot, it may occur a very large number of transmission times when N_0 is large. So the practical systems do not use this method generally. Referring to [37], we take a frame consisted of 100 packets as a basic transmission unit in this paper.

[2]The overhead is defined as the number of output packets that a user needs to collect to recover the message, minus the number of N_0 input packets. We measure the overhead as a multiple of the numbe of N_0 input packets, so an overhead of ν means that $N \ge N_0\left(1+\nu\right)$ output packets need to be collected to ensure successful decoding with high probability [32].

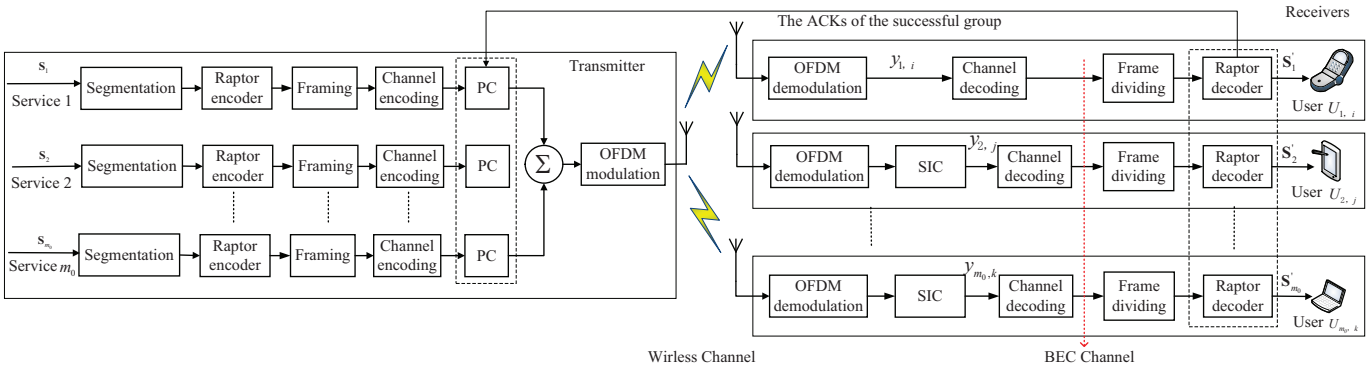

Fig. 2: A block diagram of the downlink NOMA-RC multicast system.

by the Gaussian mixture method. The larger the number of received packets is, the higher the probability of decoding is. Besides, the overhead becomes smaller as the length of a block increases. Using Gaussian mixture fitting method, we can obtain the CDF of ν is given by

$$\Gamma(\nu) = \sum_{i=1}^{y_0} \varsigma_i \exp\left(-\frac{(\nu - b_i)}{c_i}\right)^2, \tag{3}$$

where y_0 denotes the number of Gaussian distribution components, and ς_i, b_i and c_i are fitting parameters. Thus, we can express the CDF for N packets as follows:

$$\Phi(N) = \frac{1}{N_0}\Gamma\left(\frac{N}{N_0} - 1\right), \quad N \in \mathbf{N}^+. \tag{4}$$

where N_0 is the length of a block.

III. Signal Analysis

This section introduces a multicast communication scenario (Scenario 1) and a relay cooperative multicast communication scenario (Scenario 2). Then, the NOMA-RC system is applied to the two scenarios. Next, the received signals are discussed in detail.

A. The multicast communication scenario

As shown in Fig. 4, there are m_0 groups consisted of some users randomly distributed around BS. It provides different services for the groups. It is $R_1 \leq R_2 \leq \cdots \leq R_{m_0-1} \leq R_{m_0}$, where $R_i\,(1 \leq i \leq m_0)$ is the transmission rate of G_i that the BS provides.

The services of the groups are multiplied by power factors and superimposed together to form a composite signal. It will be broadcasted to multiple groups. The signal received by $U_{i,j}$ in the first stage is

$$y_{U_{i,j}}(t) = h_{BU_{i,j}}(t)\left(\sum_{k=1}^{m_0}\sqrt{a_k P_t}\,x_{\mathbf{s}_k}(t)\right) + n_{BU_{i,j}}(t), \tag{5}$$

where $U_{i,j}\,(1 \leq i \leq m_0, j \geq 1)$ denotes the j-th user in G_i, $h_{BU_{i,j}}(t)$ is the channel gain between BS and $U_{i,j}$ at the time t. The parameter $n_{BU_{i,j}}(t)$ is the additive Gaussian white noise and a_i is a power factor of G_i. It is $a_1 > a_2 > \cdots > a_{m_0}$ and $\sum_{k=1}^{m_0} a_k = 1$. The optimization of the power factors will be discussed in detail in Section V. If $U_{i,j}$ tries to decode the service message $x_{\mathbf{s}_i}$, it has to decode $x_{\mathbf{s}_k}\,(1 \leq k \leq i-1)$ first. Next, we take $U_{i,j}$ to decode $x_{\mathbf{s}_i}$ as an example. The SIC structure of the receiver $U_{i,j}$ is shown in Fig. 5. After $U_{i,j}$ decodes the first $i-1$ services, the revised signal will be updated as

$$y'_{U_{i,j}}(t) = h_{BU_{i,j}}(t)\left(\sum_{l=i}^{m_0}\sqrt{a_l P_t}\,x_{\mathbf{s}_l}(t)\right) + n_{BU_{i,j}}(t). \tag{6}$$

The signal to interference and noise ratio (SINR) of $U_{i,j}$ to

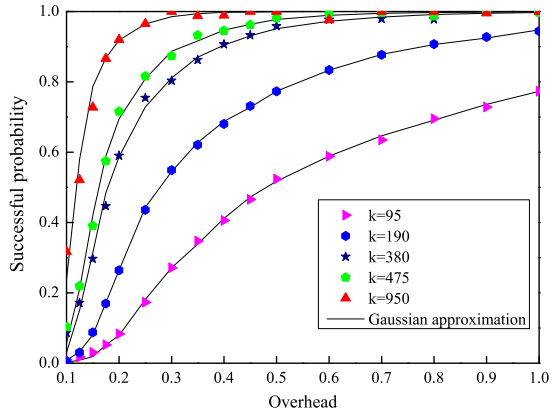

Fig. 3: Gaussian mixture approximation for the CDF of ν for raptor codes. (LDPC code rate is 0.95, and the degree distribution $\Omega(x)$ is referenced as [40].)

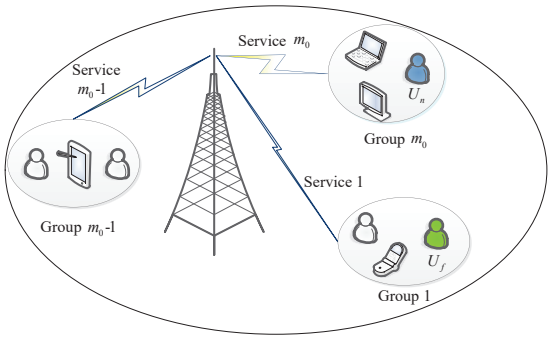

Fig. 4: The multicast system where the users in the same group share the same service.

decode $x_{\mathbf{S}_i}$ can be expressed as

$$\gamma_{U_{i,j}\to x_{\mathbf{S}_i}} = \frac{|h_{i,j}|^2 a_i P_t}{|h_{i,j}|^2 P_t \sum_{l=i+1}^{m_0} a_l + P_{\sigma^2}}, \tag{7}$$

where $1 \le i \le m_0$.

We present two power adjustment strategies such as Scheme 1 and Scheme 2 in this paper. For Scheme 1, according to the number of groups, the decoding process will be divided into multiple stages. When there is a group where all the users have successfully decoded, the system will allocate the power for the remaining groups. The power allocation process will be further discussed in Section VI in detail. Then, the next stage will begin after the system achieves the new power allocation factors. Unlike Scheme 1, Scheme 2 does not adjust the power factors again even the states of the group are changed. For the successful groups, the BS continues to transmit the packets of the next block. However, for the remaining groups, they will continue accumulating the subsequent new packets of the current block. Next, we take Scheme 1 as an example to explain the decoding process. The user $U_{i,j}$ will return an ACK signal when it decodes successfully. After receiving all the ACKs of the users in G_i, BS can re-allocate the power to the remaining users to speed up decoding of the remaining users. Then the signal received by the remaining group $G_g\,(1 \le g \le m_0, g \ne i)$ in the second stage is

$$y_{U_{g,l}}(t) = h_{BU_{g,l}}(t)\left(\sum_{g=1}^{m_0}\sqrt{a'_g P_t}\,x_{\mathbf{S}_g}(t)\right) + n_{BU_{g,l}}(t), \tag{8}$$

where a'_g is the power factor of $U_{g,l}\,(l \ge 1)$ in the next stage. Given the new power allocation factors, the BS continues transmitting data packets to the remaining groups until a new successful one appears.

B. Cooperative multicast communication scenario

As shown in Fig. 6, there are two groups and an amplify-and-forward (AF) relay in the signal coverage of a BS. The two groups can receive signals from both the relay and the BS. In the first phase, the BS broadcasts signals to the relay and the users. The relay will forward the signals to each user in the second phase [41]. Besides, the whole communication process also can be divided into two different stage according to the decoding specific cases.

1) The First Stage: In the first phase, the signals from BS received by the relay, the user $U_{1,i}$ and the user $U_{2,j}$ are, respectively

$$y_R(t) = h_{BR}(t)\left(\sqrt{a_1 P_t}\,x_{\mathbf{s}_1}(t) + \sqrt{a_2 P_t}\,x_{\mathbf{s}_2}(t)\right) + n_{BR}(t), \tag{9}$$

$$y_{BU_{1,i}}(t) = h_{BU_{1,i}}(t)\left(\sqrt{a_1 P_t}\,x_{\mathbf{s}_1}(t) + \sqrt{a_2 P_t}\,x_{\mathbf{s}_2}(t)\right) + n_{BU_{1,i}}(t), \tag{10}$$

$$y_{BU_{2,j}}(t) = h_{BU_{2,j}}(t)\left(\sqrt{a_1 P_t}\,x_{\mathbf{s}_1}(t) + \sqrt{a_2 P_t}\,x_{\mathbf{s}_2}(t)\right) + n_{BU_{2,j}}(t). \tag{11}$$

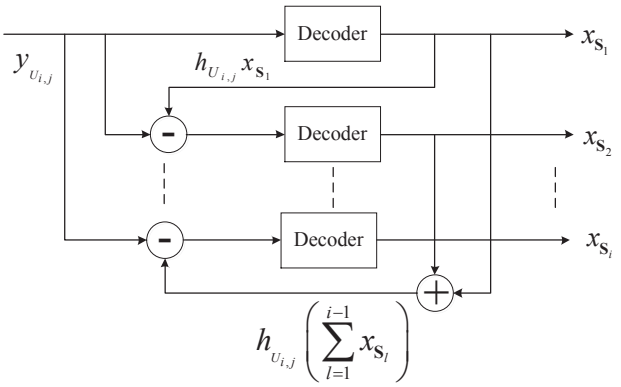

Fig. 5: SIC structure of the receiver $U_{i,j}$.

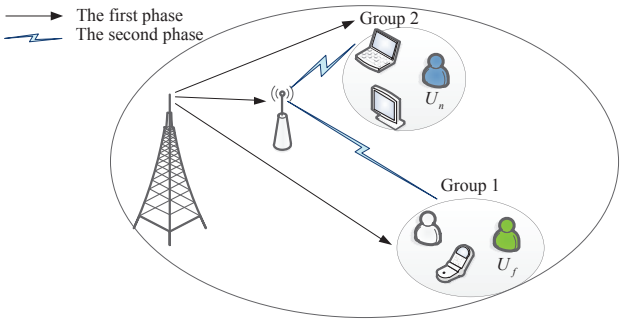

Fig. 6: Cooperative multicasting scheme based on relay-assisted communication.

In the second phase, the signals from the relay received by $U_{1,i}\,(i \ge 1)$ and $U_{2,j}\,(j \ge 1)$ are respectively

$$y_{RU_{1,i}}(t) = \eta h_{RU_{1,i}} y_{BU_{1,i}}(t) + n_{RU_{1,i}}(t), \tag{12}$$

and

$$y_{RU_{2,j}}(t) = \eta h_{RU_{2,j}} y_{BU_{2,j}}(t) + n_{RU_{2,j}}(t), \tag{13}$$

where $\eta = \sqrt{\frac{P_R}{P_t \mathrm{E}\left(|h_{BR}|^2\right) + P_{\sigma^2}}}$ is the amplification gain of the relay and we consider $P_R = P_t$. If there is a group where all the users achieve success, the system will adjust the power. It only needs to transmit the signal of the remaining group in the second stage when Scheme 1 is adopted.

2) The Second Stage: If all the users in $G_i\,(i = 1, 2)$ successfully decode in the first stage, the signal of a remaining user in $G_{\bar{i}}\left(\bar{i} = 2, 1\right)$ received by the relay in the first phase is

$$y_R(t) = h_{BR}(t)\sqrt{P_t}\,x_{S_{\bar{i}}}(t) + n_{BR}(t). \tag{14}$$

Denote: There are two groups in Scenario 2. If it is $i = 1$, so we can get $\bar{i} = 2$.

The signals received by $U_{\bar{i},j}$ in the first phase and the second phase are respectively

$$y_{BU_{\bar{i},j}}(t) = h_{BU_{\bar{i},j}}(t)\sqrt{P_t}\,x_{S_{\bar{i}}}(t) + n_{BU_{\bar{i},j}}(t), \tag{15}$$

and

$$y_{RU_{\bar{i},j}}(t) = \eta h_{RU_{\bar{i},j}} y_{BU_{\bar{i},j}}(t) + n_{RU_{\bar{i},j}}(t). \tag{16}$$

IV. Performance evaluation

There is no interference between users in the same group, but it exists between different groups. As the channel gain of each user obeys an independent and identical distribution, the FER performance of the users in the same group can be discussed in the same way. Thus, we focus on there is only one user in each group. First, this section analyzes the FER performance of each user in the two scenarios. Then, taking Scenario 1 as an example, we discuss the transmission time and the throughput of the system.

A. The analysis of FER performance

1) Scenario 1: After we simplify the model, there are m_0 users in the m_0 groups, and each group has one user. Thus, we can denote $U_{i,\bullet}$ as U_i. The message of each user can be decoded step by step by the SIC algorithm. For user $U_i\,(1 \le i \le m_0)$, it just treats the signals of $U_{i+1} \sim U_{m_0}$ as interference. So the FER of U_i is given by

$$P_{U_i} = 1 - \Pr\left(\begin{array}{l} \gamma_{BU_i \to U_1} > \gamma_{th(1)}, \gamma_{BU_i \to U_2} > \gamma_{th(2)}, \\ \cdots, \gamma_{BU_i \to U_i} > \gamma_{th(i)} \end{array} \right), \tag{17}$$

where $\gamma_{th(i)}{=}2^{2r_i}-1 \ge 0, 1 \le i \le m_0$ and r_i is the minimum target rate of U_i.

Theorem 1: *The closed-form expression for the FER of* $U_i\,(1 \le i \le m_0)$ *is*

$$P_{U_i} = 1 - \exp\left(-\frac{\mu_i}{w_{BU_i}}\right), \tag{18}$$

$$\mu_i = \max\left(z_l\right), l \in [1, i], \tag{19}$$

$$z_l = \frac{\gamma_{th(l)}}{\rho\left(a_l - \gamma_{th(l)} \sum\limits_{k=l+1}^{m_0} a_k\right)}, \tag{20}$$

where $a_l - \gamma_{th(l)} \sum\limits_{k=l+1}^{m_0} a_k \ge 0$ and $\rho{=}\frac{P_t}{P_{\sigma^2}}$.

Proof: According to (7), we obtain that

$$\begin{aligned} P_{U_i} &= 1 - \Pr\left(\begin{array}{l} \gamma_{BU_i \to U_1} > \gamma_{th(1)}, \gamma_{BU_i \to U_2} > \gamma_{th(2)}, \\ \cdots, \gamma_{BU_i \to U_i} > \gamma_{th(i)} \end{array} \right) \\ &= 1 - \Pr\left(\begin{array}{l} \frac{\left|h_{BU_i}\right|^2 a_1 \rho}{\rho\left|h_{BU_i}\right|^2 \sum\limits_{j=2}^{m_0} a_j + 1} \ge \gamma_{th(1)}, \\ \frac{\left|h_{BU_i}\right|^2 a_2 \rho}{\rho\left|h_{BU_i}\right|^2 \sum\limits_{j=3}^{m_0} a_j + 1} \ge \gamma_{th(2)}, \\ \ldots, \left|h_{BU_i}\right|^2 a_i \rho \ge \gamma_{th(i)} \end{array} \right) \\ &= \Pr\left(\left|h_{BU_i}\right|^2 < \max\left(\begin{array}{l} \frac{\gamma_{th(1)}}{\rho\left(a_1 - \gamma_{th(1)} \sum\limits_{j=2}^{m_0} a_j\right)}, \\ \frac{\gamma_{th(2)}}{\rho\left(a_2 - \gamma_{th(2)} \sum\limits_{j=3}^{m_0} a_j\right)}, \\ \cdots, \frac{\gamma_{th(i)}}{a_i \rho} \end{array} \right) \right) \\ &= 1 - \exp\left(-\frac{\mu_i}{w_{BU_i}}\right). \end{aligned} \tag{21}$$

When a user decodes successfully, the system with Scheme 1 will re-allocate the power to the remaining users. The specific power allocation process will be discussed in detail in Algorithm 1 in Section V. Given the power factor and the target rate of each user, the FER of the remaining users in subsequent stages can also be obtained using Theorem 1. Besides, if there are more than one user in a group, we also can get the FER of a user with (21).

2) Scenario 2: There are two groups in Scenario 2. we assume that $U_f = U_{1,f}\,(1 \le f \le g_1)$ is in G_1 and $U_n = U_{2,n}\,(1 \le n \le g_2)$ is in G_2, where g_1 and g_2 are the number of users in G_1 and G_2, respectively. The whole communication process can be divided into two different stages for the Scenario 2.

(1) The First Stage

User U_f can receive the signals from the relay and the BS at the same time, so the FER of U_f is written as [42]

$$P_{U_{f1}} = \Pr\left(\gamma_{BU_f} < \gamma_{th(f)}\right) \Pr\left(\gamma_{RU_f} < \gamma_{th(f)}\right). \tag{22}$$

Theorem 2: *The closed-form expression for the FER of* U_f *in the first stage is*

$$\begin{aligned} P_{U_{f1}} &= \Pr\left(\gamma_{BU_f} < \gamma_{th(f)}\right) \times \Pr\left(\gamma_{RU_f} < \gamma_{th(f)}\right) \\ &= \left(\begin{array}{l} \left(1 - \exp\left(-\frac{\tau}{w_{BU_f}}\right)\right) \times \\ \left(\begin{array}{l} 1 - \frac{2}{w_{RU_f}} \exp\left(-\tau\left(\frac{1}{w_{RU_f}} + \frac{1}{w_{BR}}\right)\right) \times \\ \sqrt{\frac{\tau(1+\tau\rho) w_{RU_f}}{\rho w_{BR}}} \mathrm{K}_1\left(2\sqrt{\frac{\tau(1+\tau\rho)}{\rho w_{BR} w_{RU_f}}}\right) \end{array} \right) \end{array} \right), \end{aligned} \tag{23}$$

where $\tau{=}\frac{\gamma_{th(f)}}{\rho\left(a_f - a_n \gamma_{th(f)}\right)}$, $\mathrm{K}_1(\cdot)$ denotes the modified Bessel function of second kind with one order.[3]

User U_n needs to decode the signal of U_f before it decodes its own message. Besides, U_n also receives the signals from the relay and the BS. Thus, the FER of U_n is

$$\begin{aligned} P_{U_{n1}} = &\left(1 - \Pr\left(\gamma_{BU_n \to U_f} > \gamma_{th(f)}, \gamma_{BU_n} > \gamma_{th(n)}\right)\right) \times \\ &\left(1 - \Pr\left(\gamma_{RU_n \to U_f} > \gamma_{th(f)}, \gamma_{RU_n} > \gamma_{th(n)}\right)\right) \end{aligned}. \tag{24}$$

Theorem 3: *The closed-form expression for the FER of* U_n *in the first stage is*

$$\begin{aligned} P_{U_{n1}} &= \left(\begin{array}{l} \left(1 - \Pr\left(\gamma_{BU_n \to U_f} > \gamma_{th(f)}, \gamma_{BU_n} > \gamma_{th(n)}\right)\right) \\ \times \left(1 - \Pr\left(\gamma_{RU_n \to U_f} > \gamma_{th(f)}, \gamma_{RU_n} > \gamma_{th(n)}\right)\right) \end{array} \right) \\ &= \left(\begin{array}{l} \left(\begin{array}{l} 1 - \frac{2}{w_{RU_n}} \exp\left(-\Omega\left(\frac{1}{w_{RU_n}} + \frac{1}{w_{BR}}\right)\right) \times \\ \sqrt{\frac{\Omega(1+\Omega\rho) w_{RU_n}}{\rho w_{BR}}} \mathrm{K}_1\left(2\sqrt{\frac{\Omega(1+\Omega\rho)}{\rho w_{BR} w_{RU_n}}}\right) \end{array} \right) \\ \times \left(1 - \exp\left(-\frac{\Omega}{w_{BU_n}}\right)\right) \end{array} \right), \end{aligned} \tag{25}$$

where $\chi{=}\frac{\gamma_{th(n)}}{a_n \rho}$ and $\Omega{=}\max(\tau, \chi)$.

(2) The Second Stage

[3]Due to the limitation of the article length, the proof process of Theorem 2, Theorem 3 and Theorem 4 are all omitted in the paper.

There are two users U_n and U_f in Scenario 2. Thus, if it is $i=n$, we can get $\overline{i}=f$. If $U_i\,(i=n,f)$ successfully decodes in the first stage, the BS only needs to transmit the data of $U_{\overline{i}}$ in the second stage. The FER of $U_{\overline{i}}$ is

$$P_{U_{\overline{i}}}=\Pr\left(\gamma_{RU_{\overline{i}}}<\gamma_{th(\overline{i})}\right)\Pr\left(\gamma_{BU_{\overline{i}}}<\gamma_{th(\overline{i})}\right). \tag{26}$$

Theorem 4. *The closed-form expression for the FER of $U_{\overline{i}}$ in the second stage is*

$$\begin{aligned}P_{U_{\overline{i}}}&=\Pr\left(\gamma_{RU_{\overline{i}}}<\gamma_{th(\overline{i})}\right)\Pr\left(\gamma_{BU_{\overline{i}}}<\gamma_{th(\overline{i})}\right)\\&=\left(\begin{array}{l}\left(\begin{array}{l}1-\frac{2}{w_{RU_{\overline{i}}}}\exp\left(-\beta\left(\frac{1}{w_{RU_{\overline{i}}}}+\frac{1}{w_{BR}}\right)\right)\\\times\sqrt{\frac{\beta(1+\beta\rho)w_{RU_{\overline{i}}}}{\rho w_{BR}}}\\\times\mathrm{K}_1\left(2\sqrt{\frac{\beta(1+\beta\rho)}{\rho w_{BR}w_{RU_{\overline{i}}}}}\right)\end{array}\right)\\\times\left(1-\exp\left(-\frac{\beta}{w_{BU_{\overline{i}}}}\right)\right)\end{array}\right),\end{aligned} \tag{27}$$

where $\beta=\frac{\gamma_{th(\overline{i})}}{\rho}$. Thus, we can get the FER of a user in $G_i\,(i=1,2)$ using Theory 2, 3, 4.

B. Transmission time

The time required by the system to deal with a block is defined as transmission time[4]. The system without delay constraints can generate plenty of different packets until all the users decode successfully. However, there is always a maximum tolerance delay in a real system [27], [43]. If a block consisted of N_0 packets cannot decoded within a given time, the system will abandon the current data or send a new batch of packets. Each user tries decode the service message by accumulating the packets continually. The longer the transmission time is, the more packets it accumulates and the greater the successful decoding probability is. Unfortunately, the transmission efficiency of the system gets lower. There are two groups in Scenario 2. We assume that $U_f=U_{1,f}\,(1\le f\le g_1)$ is the last one to achieve the service message $\mathbf{s}_1$ in G_1 and $U_n=U_{2,n}\,(1\le n\le g_2)$ is the last one to obtain the service message $\mathbf{s}_2$ in G_2. If it is $i=n$, we can get $\overline{i}=f$.

1) The First Stage: As shown in Table I, there are four cases for U_n and U_f in the time T_1.

TABLE I: Four different cases for U_n and U_f

Users	Case 1	Case 2	Case 3	Case 4
U_n	Yes	No	Yes	No
U_f	No	Yes	Yes	No

Case 1: U_n succeeded, U_f failed.

[4]In this paper, we evaluate the transmission time of a block by counting the number of the transmitted frames in the system. Compared with the transmission time, the delay caused by rateless decoder and SIC algorithm is negligible. A simple signal amplification relay is applied in Scenario 2, so the delay caused by the relay is also small.

At time T_1, the probability that U_n just succeeded in decoding is

$$f_n=\sum_{z_1=N_0}^{T_1}\Psi_1\left(P_{U_{n1}},z_1\right)P\left(N_n=z_1\right), \tag{28}$$

where

$$\Psi_1\left(P_{U_{n1}},z_1\right)=\left(\begin{array}{c}T_1-1\\z_1-1\end{array}\right)\left(1-P_{U_{n1}}\right)^{z_1}\left(P_{U_{n1}}\right)^{T_1-z_1}. \tag{29}$$

and $P_{U_{i1}}$ is the FER of $U_i, i\in(n,f)$ in the first stage. The expression $P\left(N_n=z_1\right)$ is the probability that U_n needs to accumulate z_1 frames to achieve the success. A frame is consisted of a fixed number of packets. The probability of a failed message decoding for U_f in T_1 is

$$f_{\overline{f}}=\left(\begin{array}{l}\sum\limits_{z_2=N_0}^{T_1}\Psi_2\left(P_{U_{f1}},z_2\right)P\left(N_f=z_2\right)+\\\sum\limits_{k=T_1+1}^{\infty}P\left(N_f=k\right)\end{array}\right), \tag{30}$$

where

$$\Psi_2\left(P_{U_{f1}},z_2\right)=\sum_{k=0}^{z_2-1}\left(\begin{array}{c}T_1\\k\end{array}\right)\left(1-P_{U_{f1}}\right)^{k}\left(P_{U_{f1}}\right)^{T_1-k}. \tag{31}$$

The expression $P\left(N_f=z_2\right)$ is the probability that U_f needs to accumulate z_2 frames to achieve the success. It is $P_{U_{f1}}=P_{U_f},P_{U_{n1}}=P_{U_n}$. We can obtain the PDF of N_n and N_f from (4). So the CDF of Case 1 in T_s is

$$\begin{aligned}&F_{n\overline{f}}\left(T_s\right)\\&=\sum_{T_1=N_0}^{T_s}\left(\begin{array}{l}\left(\sum\limits_{z_1=N_0}^{T_1}\Psi_1\left(P_{U_{n1}},z_1\right)P\left(N_n=z_1\right)\right)\times\\\left(\begin{array}{l}\sum\limits_{z_2=N_0}^{T_1}\Psi_2\left(P_{U_{f1}},z_2\right)P\left(N_f=z_2\right)+\\\sum\limits_{k=z_2+1}^{\infty}P\left(N_f=k\right)\end{array}\right)\end{array}\right).\end{aligned} \tag{32}$$

The average time in T_s is

$$\begin{aligned}&E_{n,\overline{f}}\left(T_s\right)=\\&\sum_{T_1=N_0}^{T_s}\left(\begin{array}{l}\left(\sum\limits_{z_1=N_0}^{T_1}\Psi_1\left(P_{U_{n1}},z_1\right)P\left(N_n=z_1\right)\right)\times T_1\times\\\left(\begin{array}{l}\sum\limits_{z_2=N_0}^{T_1}\Psi_2\left(P_{U_{f1}},z_2\right)P\left(N_f=z_2\right)+\\\sum\limits_{k=z_2+1}^{\infty}P\left(N_f=k\right)\end{array}\right)\end{array}\right).\end{aligned} \tag{33}$$

Case 2: U_f succeeded, U_n failed.

Similarly, we can deduce that the CDF of Case 2 and the average time in T_s are respectively

$$\begin{aligned}&F_{f,\overline{n}}\left(T_s\right)=\\&\sum_{T_1=N_0}^{T_s}\left(\begin{array}{l}\left(\sum\limits_{z_1=N_0}^{T_1}\Psi_1\left(P_{U_{f1}},z_1\right)P\left(N_f=z_1\right)\right)\times\\\left(\begin{array}{l}\sum\limits_{z_2=N_0}^{T_1}\Psi_2\left(P_{U_{n1}},z_2\right)P\left(N_n=z_1\right)+\\\sum\limits_{z_3=T_1+1}^{\infty}P\left(N_n=z_3\right)\end{array}\right)\end{array}\right),\end{aligned} \tag{34}$$

and

$$E_{f,\overline{n}}\left(T_s\right)= \sum_{T_1=N_0}^{T_s}\left(\begin{array}{l}\left(\sum_{z_1=N_0}^{T_1}\Psi_1\left(P_{U_{f1}},z_1\right)P\left(N_f=z_1\right)\right)\times T_1\times \\ \left(\begin{array}{l}\sum_{z_2=N_0}^{T_1}\Psi_2\left(P_{U_{n1}},z_2\right)P\left(N_n=z_1\right)+ \\ \sum_{k=T_1+1}^{\infty}P\left(N_n=k\right)\end{array}\right)\end{array}\right). \tag{35}$$

Case 3: U_n and U_f succeeded simultaneously.

$$F_{n,f}\left(T_s\right)= \sum_{T_1=N_0}^{T_s}\left(\begin{array}{l}\left(\sum_{z_1=N_0}^{T_1}\Psi_1\left(P_{U_{n1}},z_1\right)P\left(N_n=z_1\right)\right)\times \\ \left(\sum_{z_2=N_0}^{T_1}\Psi_1\left(P_{U_{f1}},z_2\right)P\left(N_f=z_2\right)\right)\end{array}\right), \tag{36}$$

and

$$E_{nf}\left(T_s\right)= \sum_{T_1=N_0}^{T_s}\left(\begin{array}{l}\left(\sum_{z_1=N_0}^{T_1}\Psi_1\left(P_{U_{n1}},z_1\right)P\left(N_n=z_1\right)\right)\times T_1\times \\ \left(\sum_{z_2=N_0}^{T_1}\Psi_1\left(P_{U_{f1}},z_2\right)P\left(N_f=z_2\right)\right)\end{array}\right). \tag{37}$$

Case 4: Both U_n and U_f failed.

The CDF and the average time of Case 4 in T_s are respectively

$$F_{\overline{n},\overline{f}}\left(T_s\right)=1-F_{n,f}\left(T_s\right)-F_{f,\overline{n}}\left(T_s\right)-F_{n,\overline{f}}\left(T_s\right), \tag{38}$$

and

$$E_{\overline{n},\overline{f}}\left(T_s\right)=\left(1-F_{n,f}\left(T_s\right)-F_{f,\overline{n}}\left(T_s\right)-F_{n,\overline{f}}\left(T_s\right)\right)T_s. \tag{39}$$

Thus, the transmission time in the first stage can be obtained as follows:

$$E_1\left(T_s\right)=E_{n,\overline{f}}\left(T_s\right)+E_{f,\overline{n}}\left(T_s\right)+E_{n,f}\left(T_s\right)+E_{\overline{n},\overline{f}}\left(T_s\right). \tag{40}$$

2) The Second Stage: The probability of the event that U_i has received (z_1-1) frames within T_1-1 and achieved the z_1-th frame at T_1 time is

$$f_{z_1|T_1}\left(P_{U_{i1}}\right)=\Psi_1\left(P_{U_{i1}},z_1\right). \tag{41}$$

The probability of $U_{\overline{i}}$ receiving v_0 frames in T_1 is

$$f_{v_0|T_1}\left(P_{U_{\overline{i}1}}\right)=\left(\begin{array}{c}T_1\\ v_0\end{array}\right)\left(1-P_{U_{\overline{i}1}}\right)^{v_0}\left(P_{U_{\overline{i}1}}\right)^{T_1-v_0}. \tag{42}$$

In the second stage, the probability of the event that $U_{\overline{i}}$ has received z_2-v_0 frames in ΔT and achieved the z_2-th frame at $T_1+\Delta T$ time is

$$f_{z_2-v_0|k}\left(P_{U_{\overline{i}2}}\right)= \left(\begin{array}{c}k-1\\ z_2-v_0-1\end{array}\right)\left(1-P_{U_{\overline{i}2}}\right)^{z_2-v_0}\left(P_{U_{\overline{i}2}}\right)^{k-z_2+v_0}, \tag{43}$$

where $P_{U_{\overline{i}2}}$ is the FER of $U_{\overline{i}}, \overline{i}\in(f,n)$ in the second stage. So the CDF of the event is expressed in (44) that the two users decode successfully in T_s and $U_i\,(i=n,f)$ is earlier than $U_{\overline{i}}\left(\overline{i}=f,n\right)$ to achieve success. The average time is given in (45). There are also following four cases in the second stage.

(1) U_n succeeds first, and U_f is the second user to achieve success.

Substituting U_i=U_n and $U_{\overline{i}}$=U_f to (44) and (45), we can obtain $F_{f2}\left(T_s\right)$ and $E_{f2}\left(t\right)$.

(2) U_f succeeds first, and U_n is the second user to achieve success.

Similarly, we can obtain $F_{n2}\left(T_s\right)$ and $E_{n2}\left(t\right)$ with U_i=U_f and $U_{\overline{i}}$=U_n from (44) and (45).

(3) Both U_f and U_f succeed simultaneously.

$F_{n,f}\left(T_s\right)$ and $E_{n,f}\left(T_s\right)$ is expressed in (36) and (37).

(4) There are more than one failed user.

The CDF of the event that at least one user fails to decode is

$$F_s\left(T_s\right)=1-F_{n,f}\left(T_s\right)-F_{f2}\left(T_s\right)-F_{n2}\left(T_s\right). \tag{46}$$

The average time is

$$E_s\left(T_s\right)=T_sF_s\left(T_s\right). \tag{47}$$

Finally, combining (37) and (47), we can obtain the average transmission time required in Scenario 1 (Please refer to Section III.A) for the two stages as

$$E_2\left(T_s\right)=E_{n2}\left(T_s\right)+E_{f2}\left(T_s\right)+E_{n,f}\left(T_s\right)+E_s\left(T_s\right). \tag{48}$$

According to (23), (25) and (27), the average transmission time in the Scenario 2 (Please refer to Section III.B) can be obtained by the analysis method above. For G_i with multiple users, only when all the users in the group achieve $\mathbf{s}_i$, a new block will be transmitted. So the average required time of the first stage is $\min\limits_{i=1,\cdots,g_1}\left(\max\limits_{j=1,\cdots,g_2}\left(E_{i,j}\left(T_1\right)\right)\right)$, where $E_{i,j}\left(T_1\right)$ denotes the average time of $U_{i,j}$ in the first stage. Thus, the average time of the system is $E\left(T_2\right)=\max\limits_{i=1,\cdots,g_1,j=1,\cdots,g_2}\left(E_{i,j}\left(T_1\right)+E_{i,j}\left(\Delta T\right)\right)$ in the two stages, where $E_{i,j}\left(\Delta T\right)$ denotes the average time of $U_{i,j}$ in the second stage. Besides, we define the throughput as [28]

$$\eta=\frac{m_0N_0}{\mathrm{E}\left(T_{m_0}\right)}, \tag{49}$$

where T_{m_0} is the time required for m_0 groups to decode successfully.

V. Power allocation strategy

The power allocation factors play an important role in the performance of the users. Under the condition that the total power is limited, we achieve the maximum sum rate of the system at a given FER in this section. As all users in the same group share the same power factor and the same content, also there are no interference between the users in the same group, we can still choose one user to represent the whole group to simplify the analysis process. To guarantee the QoS of each user in a group, we need to discuss the performance of the edge users with the poorest channel conditions in their groups first.

$$
\begin{aligned}
F_{\bar{i}2}\left(T_s\right) &= \sum_{t=N_0}^{T_s}\sum_{k=1}^{T_s-t}\Pr\nolimits_{\bar{i}}\left(\Delta T=k, T_1{=}t\right)\\
&= \sum_{t=N_0}^{T_s}\sum_{k=1}^{T_s-t}\sum_{z_2=N_0}^{t+k}\sum_{z_1=N_0}^{t}\sum_{v_0=0}^{z_2-1}\left(\begin{array}{l} f_{z_2-v_0|k}\left(P_{U_{\bar{i}2}}\right) f_{v_0|T_1}\left(P_{U_{\bar{i}1}}\right) f_{z_1|T_1}\left(P_{U_{i1}}\right)\\ \times P\left(N_i=z_1\right)P\left(N_{\bar{i}}=z_2\right)\end{array}\right)\\
&= \sum_{t=N_0}^{T_s}\sum_{k=1}^{T_s-t}\sum_{z_2=N_0}^{t+k}\sum_{z_1=N_0}^{t}\sum_{v_0=0}^{z_2-1}\left(\begin{array}{l}\left(\begin{array}{c}k-1\\ z_2-v_0-1\end{array}\right)\left(1-P_{U_{\bar{i}2}}\right)^{z_2-v_0}\left(P_{U_{\bar{i}2}}\right)^{k-z_2+v_0}\\ \times\left(\begin{array}{c}T_1\\ v_0\end{array}\right)\left(1-P_{U_{\bar{i}1}}\right)^{v_0}\left(P_{U_{\bar{i}1}}\right)^{T_1-v_0}\\ \times\left(\begin{array}{c}T_1-1\\ z_1-1\end{array}\right)\left(1-P_{U_{i1}}\right)^{z_1}\left(P_{U_{i1}}\right)^{T_1-z_1}\\ \times P\left(N_i=z_1\right)P\left(N_{\bar{i}}=z_2\right)\end{array}\right).
\end{aligned}
\tag{44}
$$

$$
\begin{aligned}
&E_{\bar{i}2}\left(T_s\right)=\\
&\sum_{t=N_0}^{T_s}\sum_{k=1}^{T_s-t}\left(\sum_{z_2=N_0}^{t+k}\sum_{z_1=N_0}^{t}\sum_{v_0=0}^{z_2-1}\left(\begin{array}{l}\left(\begin{array}{c}k-1\\ z_2-v_0-1\end{array}\right)\left(1-P_{U_{\bar{i}2}}\right)^{z_2-v_0}\left(P_{U_{\bar{i}2}}\right)^{k-z_2+v_0}\\ \times\left(\begin{array}{c}T_1\\ v_0\end{array}\right)\left(1-P_{U_{\bar{i}1}}\right)^{v_0}\left(P_{U_{\bar{i}1}}\right)^{T_1-v_0}\\ \times\left(\begin{array}{c}T_1-1\\ z_1-1\end{array}\right)\left(1-P_{U_{i1}}\right)^{z_1}\left(P_{U_{i1}}\right)^{T_1-z_1}\\ \times P\left(N_i=z_1\right)P\left(N_{\bar{i}}=z_2\right)\end{array}\right)\times\left(t+k\right)\right).
\end{aligned}
\tag{45}
$$

A. Problem formulation

To maximize the sum rate of the system, the power allocation problem in the NOMA-RC system can be expressed as:

$$
\begin{aligned}
&\max_{\bar{\mathrm{p}}=\left(p_{1},\ldots,p_{m_0}\right)}\sum_{i=1}^{m_0}\log\left(1+\frac{P_t|h_i|^2a_i}{P_t|h_i|^2\sum\limits_{j=i+1}^{m_0}a_j+P_{\sigma^2}}\right)\\
&\mathrm{C1}: s.t.\sum_{i=1}^{m_0}p_{i}\le P_t, p_{i}\ge 0,\\
&\mathrm{C2}: P_{U_i}\le\varepsilon_{i},\\
&\mathrm{C3}: \left(1-\varepsilon_i\right)^{\Omega_i}\ge\left(1-\varepsilon_{i+1}\right)^{\Omega_{i+1}}, 1\le i\le m_0-1,
\end{aligned}
\tag{50}
$$

where U_i is the edge user of G_i, Ω_i is the received power of U_i, ε_i is the target FER at a given rate $R_i\ge r_i$.

According to (18), C2 can be re-written as

$$
\xi_i\ge\max_{j=1\ldots,i}\left(\frac{\gamma_{th(j)}}{\Delta p_j}\right),
\tag{51}
$$

where $\xi_i{=}-\frac{\Omega_i\ln(1-\varepsilon_i)}{P_{\sigma^2}}$ and $\Delta p_j=p_j-\gamma_{th(j)}\sum\limits_{k=j+1}^{m_0}p_k$.

Theorem 5: *Equation (51) can be re-written as*

$$
\min\left(\xi_i,\xi_{i+1},\ldots\xi_{m_0}\right)\ge\frac{\gamma_{th(i)}}{\Delta p_{i}}.
\tag{52}
$$

Proof: According to (51), it is $\xi_i\ge\frac{\gamma_{th(i)}}{\Delta p_i}$. We can get $\xi_{i+1}\ge\max\limits_{j=1,\ldots,i+1}\left(\frac{\gamma_{th(j)}}{\Delta p_j}\right)$ when $i\leftarrow i+1$. Similarly, we can deduce that $\xi_{i+2}\ge\frac{\gamma_{th(i)}}{\Delta p_i},\cdots,\xi_{m_0}\ge\frac{\gamma_{th(i)}}{\Delta p_i}$. Thus, it is $\min\left(\xi_i,\xi_{i+1},\ldots,\xi_{m_0}\right)\ge\frac{\gamma_{th(i)}}{\Delta p_i}$. According to C3, we can deduce that $\xi_i\le\xi_{i+1}$. Similarly, we can obtain $\xi_1\le\cdots\le\xi_i\le\xi_{i+1}\le\cdots\le\xi_{m_0}$. Thus, $\{\xi_i\}$ is a non-decreasing sequence. Finally, (52) can be re-written as

$$
\xi_i\ge\frac{\gamma_{th(i)}}{\Delta p_{i}}.
\tag{53}
$$

The above optimization problem in (52) can be transformed into the following form

$$
\begin{aligned}
&\max_{\bar{\mathrm{p}}=\left(p_{1},\ldots,p_{m_0}\right)}\sum_{i=1}^{m_0}\log\left(1+\frac{P_t|h_i|^2a_i}{P_t|h_i|^2\sum\limits_{j=i+1}^{m_0}a_j+P_{\sigma^2}}\right),\\
&\mathrm{C4}: s.t.\sum_{i=1}^{m_0}a_{i}\le 1, a_{i}\ge 0,\\
&\mathrm{C5}: \xi_i\ge\tfrac{\gamma_{th(i)}}{\Delta p_i},\\
&\mathrm{C6}: \left(1-\varepsilon_i\right)^{\Omega_i}\ge\left(1-\varepsilon_{i+1}\right)^{\Omega_{i+1}}, 1\le i\le m_0-1.
\end{aligned}
\tag{54}
$$

B. Optimal solution

Considering the references [17] and [19], we notice that the objective function (54) is a concave function. The convex optimal problem can be solved by applying Karush-Kuhn-Tucker (KKT) conditions. The Lagrangian function of the problem can be written as

$$
\begin{aligned}
\Psi\left(\overline{a},\upsilon,\overline{\beta}\right)=\sum_{i=1}^{m_0}&\left\{\begin{array}{l}\log_2\left(P_t|h_i|^2\sum\limits_{j=i}^{m_0}a_j+P_{\sigma^2}\right)-\\ \log_2\left(P_t|h_i|^2\sum\limits_{j=i+1}^{m_0}a_j+P_{\sigma^2}\right)\end{array}\right\}\\
&+\upsilon\left(P_t-\sum_{i=1}^{m_0}p_i\right)+\sum_{i=1}^{m_0}\beta_i\left(\xi_i-\tfrac{rth_i}{\Delta p_i}\right).
\end{aligned}
\tag{55}
$$

The KKT conditions of the problem are

$$\begin{aligned}
&\mathrm{C7}: \tfrac{\partial \Psi}{\partial a_k} = \\
&\tfrac{1}{\ln 2}\sum_{i=1}^{k}\tfrac{|h_i|^2 P_t}{P_t|h_i|^2\sum_{j=i}^{m_0} a_j + P_{\sigma^2}} - \tfrac{1}{\ln 2}\sum_{i=2}^{k}\tfrac{|h_{i-1}|^2 P_t}{P_t|h_{i-1}|^2\sum_{j=i}^{m_0} a_j + P_{\sigma^2}} \\
&-\upsilon - \sum_{i=1}^{k-1}\beta_i\tfrac{\left(\gamma_{th(k)}\right)^2 P_t}{\Delta p_k^2} + \beta_k\tfrac{\gamma_{th(k)}P_t}{\Delta p_k^2} = 0, \\
&\mathrm{C8}: \upsilon\left(P_t - \sum_{i=1}^{m_0} p_i\right) = 0, \upsilon \geq 0, \\
&\mathrm{C9}: \beta_i\left(\xi_i - \tfrac{\gamma_{th(i)}}{\Delta p_i}\right) = 0, 1 \leq i \leq m_0, \beta_i \geq 0.
\end{aligned} \tag{56}$$

Substituting $k = 1$ into C7, we can get $\upsilon > 0$. Besides, we can also obtain

$$\begin{aligned}
&\tfrac{\partial \Psi}{\partial a_k} - \tfrac{\partial \Psi}{\partial a_{k-1}} = \\
&\tfrac{1}{\ln 2}\tfrac{|h_k|^2 P_t}{|h_k|^2 P_t\sum_{i=k}^{m_0} a_i + P_{\sigma^2}} - \tfrac{1}{\ln 2}\tfrac{|h_{k-1}|^2 P_t}{|h_{k-1}|^2 P_t\sum_{i=k}^{m_0} a_i + P_{\sigma^2}} \\
&+P_t\beta_k\tfrac{\gamma_{th(k)}}{\Delta p_k^2} - P_t\beta_{k-1}\tfrac{\gamma_{th(k-1)}}{\Delta p_{k-1}^2}\left(1+\gamma_{th(k-1)}\right) = 0.
\end{aligned} \tag{57}$$

Thus, we have

$$\begin{aligned}
&-\beta_k\tfrac{\gamma_{th(k)}}{\Delta p_k^2} + \beta_{k-1}\tfrac{\gamma_{th(k-1)}}{\Delta p_{k-1}^2}\left(1+\gamma_{th(k-1)}\right) \\
&= \tfrac{1}{\ln 2}\tfrac{|h_k|^2}{|h_k|^2 P_t\sum_{i=k}^{m_0} a_i + P_{\sigma^2}} - \tfrac{1}{\ln 2}\tfrac{|h_{k-1}|^2}{|h_{k-1}|^2 P_t\sum_{i=k}^{m_0} a_i + P_{\sigma^2}}.
\end{aligned} \tag{58}$$

The average gain of the channel satisfies $|h_i|^2 > |h_{i-1}|^2$, so we can get that

$$\tfrac{1}{\ln 2}\tfrac{|h_k|^2}{|h_k|^2 P_t\sum_{i=k}^{m_0} a_i + P_{\sigma^2}} - \tfrac{1}{\ln 2}\tfrac{|h_{k-1}|^2}{|h_{k-1}|^2 P_t\sum_{i=k}^{m_0} a_i + P_{\sigma^2}} > 0, \tag{59}$$

and

$$\beta_{k-1}\tfrac{\gamma_{th(k-1)}}{\Delta p_{k-1}^2}\left(1+\gamma_{th(k-1)}\right) > \beta_k\tfrac{\gamma_{th(k)}}{\Delta p_k^2}. \tag{60}$$

Combining $\gamma_{th(i)} = 2^{2R_i} - 1 \geq 0, 1 \leq i \leq m_0$ and (60), we can obtain $\beta_{k-1} > 0$. It is easy to achieve $\beta_{k-2} > 0$ when $k \leftarrow k-1$. Similarly, we can deduce $\beta_j > 0, 1 \leq j \leq m_0 - 1$. When $\beta_{m_0} > 0$ is established, the optimization problem is converted into linear equations to be solved. Then it is easy to find the optimal solution for all the users. Therefore, the maximum sum rate is the sum of rate of each user with the constraints of the target FER. The BS preferentially satisfies the power demand of the first $m_0 - 1$ users if $\beta_{m_0} = 0$ is established. Then it allocates the remaining power to the nearest user U_{m_0}.

The optimal value of the power factor for $U_i\,(i = 1, 2, \ldots, m_0)$ is

$$a_i = \begin{cases} \left(\tfrac{2^{2R_i}-1}{\xi_i} + \kappa\left(m_0, i\right)\right) \Big/ P_t, 1 \leq i \leq m_0 - 1, \\ 1 - \sum_{j=1}^{m_0-1} a_j, i = m_0. \end{cases} \tag{61}$$

where

$$\begin{aligned}
&\kappa\left(m_0, i\right) = \\
&\left(2^{2R_i} - 1\right)\left(\tfrac{2^{2R_{i+1}}-1}{\xi_{i+1}} + \sum_{n=i+2}^{m_0}\prod_{g=i+1}^{n-1}\tfrac{2^{2R_g}\left(2^{2R_n}-1\right)}{\xi_n}\right).
\end{aligned} \tag{62}$$

Proof: For $\Delta p_i = p_i - \gamma_{th(i)}\sum_{j=i+1}^{m_0} p_j$, we can get

$$p_i = \Delta p_i + \left(2^{2R_i} - 1\right)\sum_{j=i+1}^{m_0} p_j. \tag{63}$$

$$\sum_{j=i+1}^{m_0} p_j = p_{i+1} + \sum_{j=i+2}^{m_0} p_j = \Delta p_{i+1} + 2^{2R_{i+1}}\sum_{j=i+2}^{m_0} p_j. \tag{64}$$

Equation (64) can be further expanded as

$$\begin{aligned}
\sum_{j=i+1}^{m_0} p_j &= p_{i+1} + \sum_{j=i+2}^{m_0} p_j \\
&= \Delta p_{i+1} + 2^{2R_{i+1}}\Delta p_{i+2} + 2^{2R_{i+1}+2R_{i+2}}\sum_{j=i+3}^{m_0} p_j.
\end{aligned} \tag{65}$$

Thus, we can obtain

$$\sum_{j=i+1}^{m_0} p_j = \Delta p_{i+1} + \sum_{n=i+2}^{m_0}\prod_{g=i+1}^{n-1} 2^{2R_g}\Delta p_n. \tag{66}$$

Substituting (66) into (63), we can obtain

$$p_i = \Delta p_i + \left(2^{2R_i} - 1\right)\left(\Delta p_{i+1} + \sum_{n=i+2}^{m_0}\prod_{g=i+1}^{n-1} 2^{2R_g}\Delta p_n\right). \tag{67}$$

It can be observed from (67), the power of U_i is affected by the others users. Increasing the power of a user will inevitably enhance the interference to other users. Thus, it is necessary for the system to take into account the QoS of all the users simultaneously. The power factor of $U_i\,(1 \leq i \leq m_0 - 1)$ is

$$a_i = \frac{1}{P_t}\left(\frac{2^{2R_i}-1}{\xi_i} + \kappa\left(m_0, i\right)\right). \tag{68}$$

So the power factor of U_{m_0} is $a_{m_0} = 1 - \sum_{j=1}^{m_0-1} a_j$. The minimum value of the total power required to m_0 users is

$$P_s = \sum_{i=1}^{m_0-1} 2^{\sum_{j=1}^{i-1} 2R_j}\frac{\left(2^{2R_i}-1\right)}{\xi_i} + \frac{\left(2^{2R_{m_0}}-1\right)}{\xi_{m_0}}. \tag{69}$$

Algorithm 1 presents the allocation power process of the users in the NOMA-RC system. The transmit power cannot guarantee the QoS of all the users when $P_t < P_s$. The system will be forced to abandon some of the worst users in $\mathbf{U} = \left(U_1, U_2, \ldots, U_i, \ldots, U_{m_0}\right), 1 \leq k \leq m_0$ until the sum power of the remaining users is less than or equal to the transmit power, i.e., $P_t \geq P_s$. Then, the system will allocate different power to the remaining users $\left(U_j', \cdots, U_{k-1}'\right)$ according to (68) and (69). The rest of the power is assigned to U_k'. Thus, the power factor will be set to 1 when the channel conditions are very poor or the transmission power is very small. In other words, there is only one user to be served under the extremely poor communication environment. If there is a user who has successfully decoded, the system will remove it from $\mathbf{U}$ and re-order the remaining users according to their target rates. Finally, the service messages for all the users will be decoded iteratively.

Algorithm 1 A dynamic power allocation (DPA) scheme

Initialization: $\mathbf{U}=(U_1,U_2,\ldots,U_i,\ldots,U_{m_0}),k\leftarrow m_0$.
while $k\geq 1$ **do**
  Compute P_s with m_0=k and (69).
  if $P_t<P_s$ **then**
    for $j\leftarrow 2$ to k **do**
      $f(j)=\sum_{i=j}^{k-1}2^{\sum_{j=1}^{i-1}2R_j}\frac{\left(2^{2R_i}-1\right)}{\xi_i}+\frac{\left(2^{2R_k}-1\right)}{\xi_k}$.
      if $f(j)<P_t$ **then**
        break.
      end if
    end for
    Compute R_{m_0} with $P_s=P_t$ and (69).
    Allocate power to the remaining users $\left(U_j',\cdots,U_{k-1}'\right)$ according to (68).
  else
    Allocate power to $U_i'\,(1\leq i\leq k-1)$ according to (68).
    Allocate power to U_k' with $p_k=\left(1-\sum_{j=1}^{k-1}a_j\right)P_t$.
  end if
  Count the number of succeeded users s_0.
  $k\leftarrow k-s_0$
  Update the remaining users $\mathbf{U}=(U_1',U_2',\ldots,U_k')$.
end while

VI. Simulation Results

For the two scenarios mentioned above, several experiments are introduced to demonstrate the performance of the proposed scheme over Rayleigh fading channels. The edge users play an important role in the multicast system, they always affect the performance of the multicast system. So we conduct some experiments to investigate the edge users' performance and verify the results in Section IV and V. Firstly, we introduce an experiment to verify the effectiveness of the proposed power allocation scheme. Then, given the optimal power factors, the two edge users' performance are discussed in detail such as the FER, the transmission time and the BLER. We assume that the edge users are the last one in the groups to get their service information. Some simulation parameters about the two edge users U_n and U_f are as follows: $P_{\sigma^2}=-174$ dBm, $R_f=1$ bps/Hz, $R_n=1.5$ bps/Hz, $d_n=500$ m, $d_f=1000$ m, $K=-38.757$, $\Upsilon=3.71$. Some parameters about raptor codes are as follows: $N_0=950$, LDPC code rate is 0.95, and the degree distribution $\Omega(x)$ is given by [40]

$$\begin{aligned}\Omega(x)=&0.008x+0.494x^2+0.166x^3+0.073x^4+\\&0.083x^5+0.056x^8+0.037x^9+0.056x^{19}+\\&0.025x^{65}+0.003x^{66}\end{aligned}\quad(70)$$

A. Power allocation

Combining (69) and $\xi_i=-\frac{\Omega_i\ln(1-\varepsilon_i)}{P_{\sigma^2}}$, we can get that the system with P_t=50 dBm can just meet the requirements of the two users ($R_f=1,\ R_n=1.5,\ \varepsilon_n=0.0332,\ \varepsilon_f=0.004$). As the transmission power P_t increases, not only can the needs of each user be guaranteed, but also the BS will have more residual power to enhance the performance of the users.

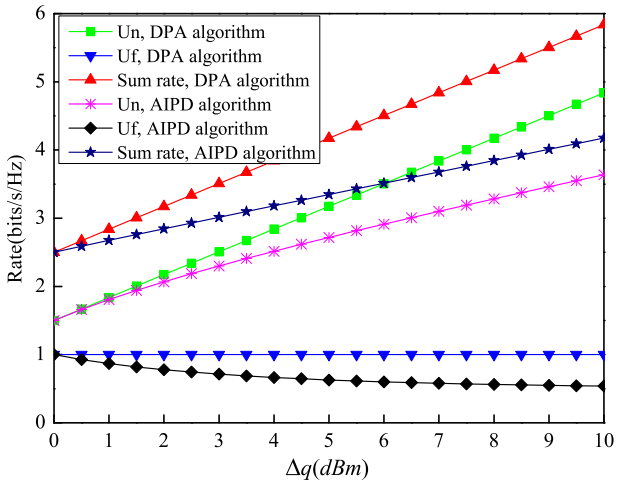

Fig. 7: Incremental power Δq versus rate with $P_t=50$ dBm, where $\varepsilon_n=0.004,\varepsilon_f=0.0332$.

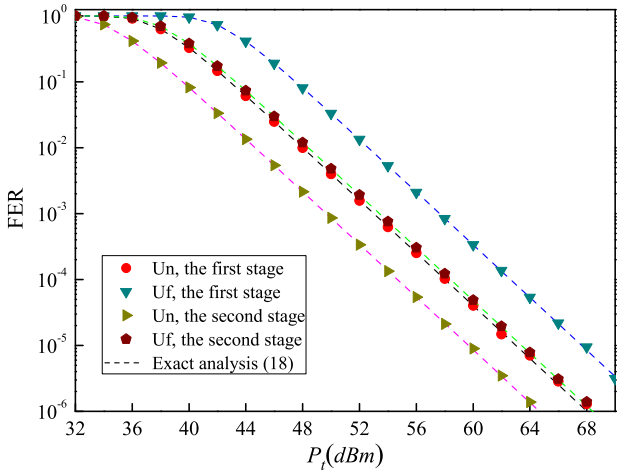

Fig. 8: The FER versus the transmission power in the different stage in Scenario 1.

Next, we will use two methods to distribute the remaining power. The first scheme, called the average incremental power distribution (AIPD) algorithm, divides all the remaining power to each user equally. The second scheme is the dynamic power allocation (DPA) scheme as presented in Algorithm 1. The performance of the DPA algorithm is better than that of the AIPD algorithm as shown in Fig. 7, where the U_n's rate and the sum rate increase with P_t. However, the rate of U_f in the AIPD algorithm reduces as P_t increases. Although the power received by each user increases as P_t increases, the interference between users also gets stronger. We also can deduce this from Theorem 1. Thus, the system has to reduce the rate of U_f to obtain the target FER performance.

B. Performance analysis of the edge users

As shown in Fig. 8, the FER of each user is continuously decreasing at different stages as P_t increases. The curve of U_n drops faster in the two stages as it has a better channel condition. In the second stage, the NOMA-RC system re-adjusts the transmission power. It just allocates the power to the remaining user. So the interference from the succeed user

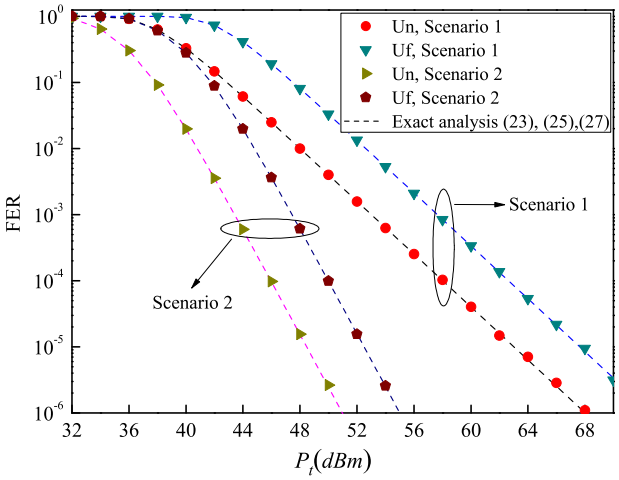

Fig. 9: Transmission time of the two stages in the two scenarios with $T_s = 30$.

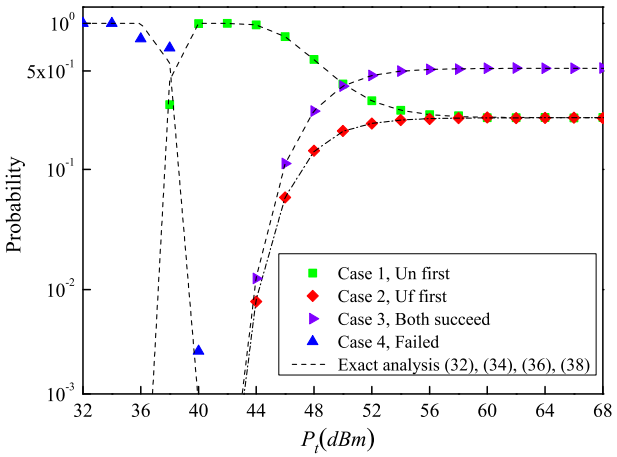

Fig. 10: The decoding probability versus the transmission power in the first stage in Scenario 1 with $T_s = 30$.

can be cancelled. The performance of the remaining user gets better in the second stage. In addition, the theoretical curves are basically consistent with the experimental simulation results, which also verifies (18).

Fig. 9 shows the comparison of the FER in two different scenarios. Due to the relay assistance in Scenario 2, the users can receive the signals from the relay and the BS. Thus, the performance of the users in Scenario 2 is better than that in Scenario 1. Besides, the simulation results match with the theoretical results derived in (23), (25) and (27), respectively.

Fig. 10 shows the decoding probabilities of different cases during the first stage in the Scenario 1. Four cases have been defined in Table I. The probability of Case 4 is greater than 90% when $P_t < 34$ dBm. It indicates that the users hardly receive a frame data. As P_t increases, the curve of Case 4 drops rapidly, while the other curves increase fast. They continue to approach a certain constant. The curves of Case 1 and Case 2 have the same bound, which is lower than that of Case 3. This is because the FER of the two users is very small when P_t is greater than 52 dBm. The order of successful decoding for the two users is normally determined by N_n and N_f. The two random numbers follow the same distribution.

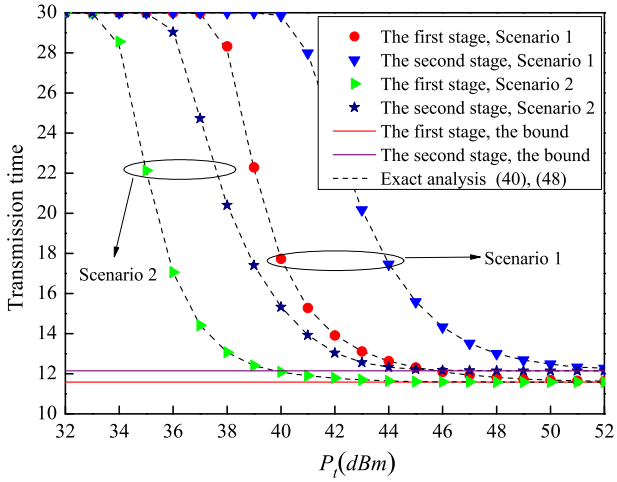

Fig. 11: Transmission time of the two stages in the two scenarios with $T_s = 30$.

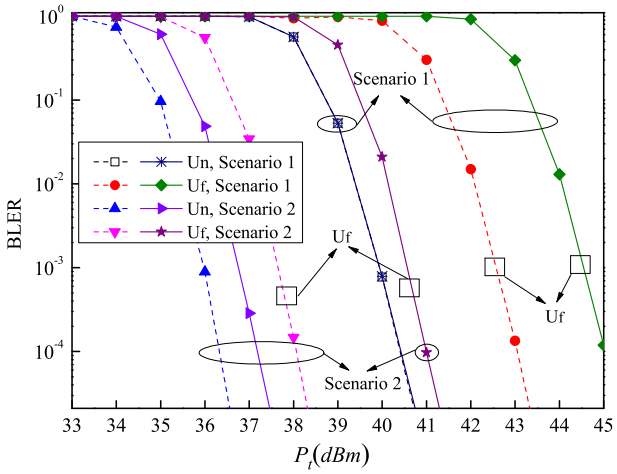

Fig. 12: BLER performance for the two users in the two scenarios with $T_s = 30$, where the dotted line denotes Scheme 1, and the solid one denotes Scheme 2.

Besides, the dashed lines are obtained from (32), (34), (36) and (38). Results show that the theoretical curves are close to the results of experimental simulations.

From Fig. 11, we get that the FER performance of the users is worse when P_t is small, so it will take a long time for $U_i\,(i = n, f)$ to accumulate N_i frames. Fig. 9 shows that the probability of receiving a frame increases as P_t increases, so the transmission time also reduces continually in each stage. When P_t is greater than a certain value, the two users can receive the frames simultaneously. Thus, the number of frames successfully received by the two users is relatively close within the same time. The system only needs to spend a small amount of time to complete the decoding of the remaining user after the first stage is over. With the increase of P_t, the average transmission time in the two stages will tend to two different bounds, $E\,(T_1) = E\,(\min\,(N_n, N_f))$ and $E\,(T_1 + \Delta T) = E\,(\max\,(N_n, N_f))$, respectively. Besides, the simulation results match with the theoretical results derived in (40) and (48).

From Fig. 1, we get that one block consists of N_0 data packets. If the decoder cannot complete decoding within T_s,

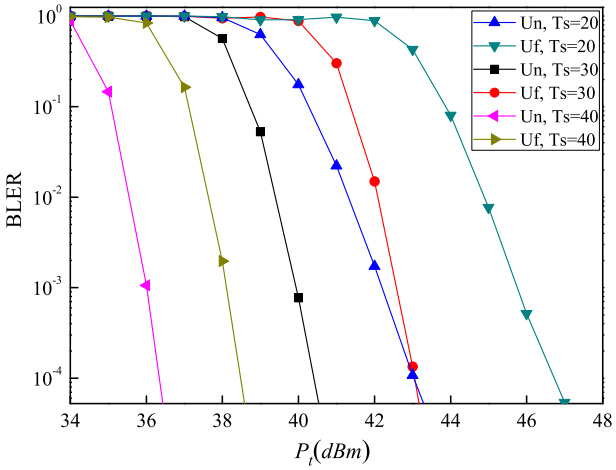

Fig. 13: BLER performance for the two users with different maximum delay in Scenario 1.

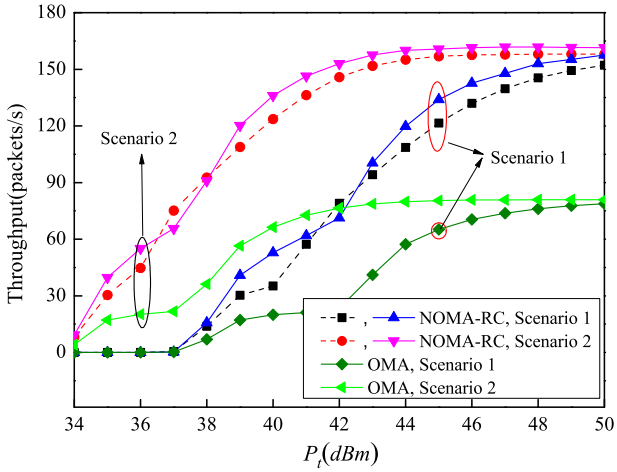

Fig. 14: Throughput performance for the OMA and NOMA-RC systems with $T_s = 30$ and $N_0 = 960$, where the dotted line is Scheme 1, and the solid line is Scheme 2.

the error block will be abandoned. It can be seen from Fig.12 that as P_t increases, the BLER curve of Scenario 2 drops faster than that of Scenario 1. In addition, Scheme 1 is also better than Scheme 2, especially the performance advantage of U_f is more obvious. So Scheme 1 helps to improve the performance of the remaining user. For U_n in Scenario 1, the curves of Scheme 1 and Scheme 2 basically overlap. This is because when P_t is less than 41 dBm, user U_f has a high FER performance as shown in Fig. 8. It is almost impossible for U_f to complete decoding before U_n. Thus, the power factor for U_n hardly needs to be adjusted during the whole decoding process in Scenario 1.

As shown in Fig. 13, the BLER performance of each user will be continuously improved by adjusting the size of T_s. With the increase of P_t, these curves keep dropping. The $T_s = 40$ curves drop more quickly, but the $T_s = 20$ curves decline more slowly. From Fig. 12 and 13, we can get that the performance of the NOMA-RC users can be improved by extending the maximum transmission time or amplifying the transmit power.

C. Throughput

Next, we will discuss the throughput of the multicast systems. There are two groups in the each scenario. Every group has three users. In Scenario 1, the distances between the BS and the users are (600, 800, 1000) meters in G_1 and (300, 400, 500) meters in G_2 respectively. In Scenario 2, the distance between the BS and the relay is 300 meters, and the distances between the relay and the users are (300, 400, 500) meters in G_1 and (60, 80, 100) meters in G_2, respectively. From Fig. 14, as P_t increases, the throughput curves increase and become stable finally. For the OMA scheme, the users in group G_2 are decoded first, the next group is G_1. When P_t is relatively low, there is less time left for G_1's users as the users in G_2 spend a lot of time to decode the service message $\mathbf{s}_2$. As a result, some users in G_1 frequently fail to decode the service message $\mathbf{s}_1$. Thus, there will be a slowly increasing interval for OMA system. As P_t increases further, the throughput performance are improved rapidly. In addition, it seems that Scheme 2 can achieve a slightly better performance than Scheme 1 in Fig. 14, but Scheme 2 fails to guarantee the performance of the edge users in G_1. We also can find this result from Fig. 12. Finally, those curves will tend to different upper bounds. The bounds of Scheme 1, Scheme 2 and OMA are $2N_0/E(T_2)$, $N_0/E(T_{G_1}) + N_0/E(T_{G_2})$ and $0.5N_0/E(T_{G_1}) + 0.5N_0/E(T_{G_2})$ respectively, where T_{G_i} $(i = 1, 2)$ is the average transmission time in the group G_i. Even if the transmission power tends to infinity, a user still needs to accumulate $N_0(1+\nu)$ packets to complete the decoding. This is determined by the structure of rateless codes. When compared with other schemes, Scenario 2 achieves the highest throughput. Especially when P_t is greater than 50 dBm, the throughputs of the proposed schemes are twice that of the OMA system in the two scenarios.

VII. Conclusion

This paper proposes a NOMA multicast transmission technology based on rateless codes and analyzes its performance in two different application scenarios. The NOMA-RC system generates many different encoded data packets for multicast users. Within T_s, the system can continuously transmit different packets to the remaining users in a group. The FER performance and the average transmission time are discussed in detail. We also demonstrate the performance of the NOMA-RC system in different multicast scenarios. The best solution for power allocation is obtained to maximize the sum rate. Compared with the OMA system with rateless codes, the NOMA-RC system helps to reduce the transmission time and improve the throughput of the system. Besides, the proposed scheme can be applied in multimedia communication scenarios, where it can provide the services with different requirements for multimedia users. The better channel condition the user has, the more data packets it will receive within the same time and the better the QoS is. Besides, the edge users can also meet the minimum requirements by adjusting the transmit power or extending the transmission time.

There are some areas worthy of further study from this paper. As the number of groups increases, the complexity of the SIC algorithm in the receivers will increase quickly.

Thus, how to reduce the complexity of the SIC algorithm is a problem to be solved. Besides, an ideal SIC scheme is adopted in this paper. We assume that the signals of other groups can be completely eliminated using SIC algorithm. However, it is difficult to achieve the perfect result in a practical system. Although this paper only discusses the system performance in the single-cell network, it also be extended to the multicell network, where the interference between the cells needs to be considered. As the proposed scheme has the backward compatibility, it also can be combined with MIMO and other technologies to further improve system performance.

REFERENCES

[1] J. Montalban *et al.*, "Multimedia multicast services in 5G networks: Subgrouping and non-orthogonal multiple access techniques," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 91–95, Mar. 2018.

[2] G. Araniti *et al.*, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Network*, vol. 31, no. 2, pp. 80–89, Apr. 2017.

[3] A. de la Fuente, R. P. Leal, and A. G. Armada, "New technologies and trends for next generation mobile broadcasting services," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 217–223, Nov. 2016.

[4] *Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description*, Sophia Antipolis Cedex, France, Mar. 2012.

[5] G. A. andothers, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Netw.*, vol. 31, no. 2, pp. 80–89, Apr. 2017.

[6] M. Sardari *et al.*, "Multilevel diversity coding via rateless codes for reliable and scalable video multicasting," *IEEE Commun. Lett.*, vol. 17, no. 5, pp. 956–959, May 2013.

[7] Y. Sun *et al.*, "Scheduling of multicast and unicast services under limited feedback by using rateless codes," in *Proc. IEEE INFOCOM 2014*, May 2014, pp. 1671–1679.

[8] B. W. Khoueiry and M. R. Soleymani, "A novel machine-to-machine communication strategy using rateless coding for the internet of things," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 937–950, Dec. 2016.

[9] L. Dai *et al.*, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[10] Y. Chen *et al.*, "Toward the standardization of non-orthogonal multiple access for next generation wireless networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 19–27, Mar. 2018.

[11] D. Wan *et al.*, "Non-orthogonal multiple access for cooperative communications: Challenges, opportunities, and trends," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 109–117, Apr. 2018.

[12] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, Secondquarter 2017.

[13] Z. Ding *et al.*, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[14] Z. Shi *et al.*, "Cooperative HARQ-assisted NOMA scheme in large-scale D2D networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4286–4302, Sep. 2018.

[15] N. Rupasinghe *et al.*, "Non-orthogonal multiple access for mmwave drone networks with limited feedback," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 762–777, Jan. 2019.

[16] M. Zeng *et al.*, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.

[17] M. S. Ali *et al.*, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3982–3998, Sep. 2018.

[18] J. Cui, Z. Ding, and P. Fan, "A novel power allocation scheme under outage constraints in NOMA systems," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1226–1230, Sep. 2016.

[19] Z. Yang *et al.*, "On the optimality of power allocation for NOMA downlinks with individual QoS constraints," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649–1652, Jul. 2017.

[20] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[21] Z. Zhang *et al.*, "Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1794–1808, Aug. 2017.

[22] L. Yang *et al.*, "Cooperative non-orthogonal layered multicast multiple access for heterogeneous networks," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1148–1165, Feb. 2019.

[23] A. Sahai, "Why do block length and delay behave differently if feedback is present?" *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1860–1886, May 2008.

[24] M. Condoluci *et al.*, "Multicast resource allocation enhanced by channel state feedbacks for multiple scalable video coding streams in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 2907–2921, May 2016.

[25] G. Araniti *et al.*, "A hybrid unicast-multicast network selection for video deliveries in dense heterogeneous network environments," *IEEE Trans. Broadcast.*, vol. 65, no. 1, pp. 83–93, Mar. 2019.

[26] U. Erez, M. D. Trott, and G. W. Wornell, "Rateless coding for gaussian channels," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 530–547, Feb. 2012.

[27] A. Rajanna and M. Haenggi, "Enhanced cellular coverage and throughput using rateless codes," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 1899–1912, May 2017.

[28] X. Wang, W. Chen, and Z. Cao, "SPARC: Superposition-aided rateless coding in wireless relay systems," *IEEE Trans. Veh. Commun.*, vol. 60, no. 9, pp. 4427–4438, Nov. 2011.

[29] T. A. Courtade and R. D. Wesel, "A cross-layer perspective on rateless coding for wireless channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, June 2009, pp. 1–6.

[30] A. F. Molisch *et al.*, "Performance of fountain codes in collaborative relay networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 11, pp. 4108–4119, Nov. 2007.

[31] M. Luby, "LT codes," in *Proc. 43rd Ann. IEEE Symp. Found. Comput. Sci.*, Nov. 2002, pp. 271 – 280.

[32] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.

[33] T. Mladenov *et al.*, "Efficient incremental raptor decoding over BEC for 3GPP MBMS and DVB IP-Datacast services," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 313–318, Jun. 2011.

[34] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive multiple access based on superposition raptor codes for cellular M2M communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 307–319, Jan. 2017.

[35] Y. Hu *et al.*, "Performance analysis of rateless-coded non-orthogonal multiple access," in *Proc. 2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, Jun. 2019, pp. 397–402.

[36] A. Gudipati and S. Katti, "Strider: automatic rate adaptation and collision handling," in *Proc. ACM SIGCOMM*, vol. 41, no. 4, Jun. 2011, pp. 158–169.

[37] J. Perry *et al.*, "Spinal codes," in *Proc. ACM SIGCOMM*, Sep. 2012, pp. 49–60.

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

[39] A. Goldsmith, *Wireless communications*. New York, NY, USA: Cambridge University Press, 2007.

[40] A. Shokrollahi, "Raptor codes," in *Proc. IEEE Int. Symp. Information Theory*, Jue. 2004, pp. 36–36.

[41] H. Liu *et al.*, "Decode-and-forward relaying for cooperative NOMA systems with direct links," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8077–8093, Dec. 2018.

[42] J. Men and J. Ge, "Performance analysis of non-orthogonal multiple access in downlink cooperative network," *IET Commun.*, vol. 9, no. 18, pp. 2267–2273, Dec. 2015.

[43] A. James *et al.*, "Spectrally efficient packet recovery in delay constrained rateless coded multihop networks," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4462–4474, Nov. 2013.

Yingmeng Hu is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, Beihang University, Beijing, China. He received the M.S. degree in school of Information Engineering Jiangxi University of Science and Technology, Jiangxi, China, in 2016. His research interests include rateless codes, non-orthogonal multiple access technologies, resource allocation and optimization and deep reinforcement learning algorithms.

Rongke Liu (SM'20) is currently a Full Professor with the School of Electronics and Information Engineering, Beihang University. He received the B.S. and Ph.D. degrees from Beihang University in 1996 and 2002, respectively. He was a Visiting Professor with the Florida Institution of Technology, USA, in 2005; The University of Tokyo, Japan, in 2015; and the University of Edinburgh, U.K., in 2018, respectively. He received the support of the New Century Excellent Talents Program from the Minister of Education, China. He has attended many special programs, such as China Terrestrial Digital Broadcast Standard. He has published over 100 papers in international conferences and journals. He has been granted 20 patents. His current research interest covers wireless communication5G/6G, channel coding and satellite internet.

Aryan Kaushik (M'20) is a Research Fellow at the Department of Electronic and Electrical Engineering, University College London (UCL), U.K., from Feb. 2020. He received PhD in Communications Engineering at the School of Engineering, The University of Edinburgh, U. K., in Jan. 2020. He received MSc in Telecommunications from The Hong Kong University of Science and Technology, Hong Kong, in 2015. He has held visiting research appointments at the Wireless Communications and Signal Processing Lab, Imperial College London, U. K., from 2019-20, the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, in 2018, and the School of Electronic and Information Engineering, Beihang University, China, from 2017-19. He has been a TPC Member for the IEEE ICC 2021, Conference Champion at the IEEE PIMRC 2020, and regular reviewer for IEEE journals and conferences. His research interests include signal processing for wireless communications, joint communications and radar transmission, energy efficient communications, millimeter wave and massive multi-antenna communications.

John Thompson (S'94-M'03-SM'13-F'16) is currently a Professor of signal processing and communications and the Director of discipline with The University of Edinburgh, U.K. He is listed by Thomson Reuters as a Highly Cited Scientist from 2015 to 2018. His research interests include millimeter wave wireless communications, signal processing for wireless networks, smart grid concepts for energy efficiency green communications systems and networks, and rapid prototyping of MIMO detection algorithms. He has published over 300 journal and conference papers on these topics. He coordinated the EU Marie Curie International Training Network Advantage on smart grid from 2014 to 2017. He is an Editor of IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING and Communications Magazine Green Series. He is also a Former founding Editor-in-Chief of the IET Signal Processing. He is also the Technical Programme Co-Chair for the IEEE Communication Society ICC 2007 Conference, the Globecom 2010 Conference, the IEEE Vehicular Technology Society VTC Spring 2013 Conference, and the IEEE Smartgridcomm 2018 Conference. He is also the Track Co-Chair for the selected areas in communications topic on Green Communication Systems and Networks at ICC 2014 Conference, the Member at Large of the IEEE Communications Society Board of Governors from 2012 to 2014. He is also the Tutorial Co-Chair for IEEE ICC 2015 Conference and the ICC 2015 Conference. He is also the Local Student Counselor for the IET and the Local Liaison Officer for the U.K. Communications Chapter of the IEEE.