NOMA-Enabled Cooperative Unicast–Multicast: Design and Outage Analysis

Long Yang, Jian Chen, *Member, IEEE*, Qiang Ni, *Senior Member, IEEE*, Jia Shi, *Member, IEEE*, and Xuan Xue

***Abstract*— This paper designs a novel non-orthogonal multiple access (NOMA) unicast–multicast system, where a number of unicast users (those who require different messages) and a group of multicast users (those who require identical message) share the same time/space/frequency resource. For the designed NOMA unicast–multicast system, an efficient two-phase cooperation strategy is proposed to improve the reliability of all users. In the first phase, the base station (BS) broadcasts a superposed message consisting of all users' information. In the second phase, a multicast user is selected to forward the information intended by unsuccessfully decoded unicast and/or multicast users. Moreover, the multicast user selection is investigated under two different power allocation (PA) approaches: 1) fixed PA (FPA), in which the PA coefficients for both the phases are predetermined, and 2) dynamic PA (DPA), in which the PA coefficients for the first phase are predetermined, while the PA coefficients for the second phase are dynamically determined based on instantaneous channel information. Under the FPA approach, a best user selection (BUS) scheme (called F-BUS) is proposed to minimize the outage probability. Under the DPA approach, the local optimal PA coefficients for the second phase are derived in closed form first. Based on the derived PA coefficients, a BUS scheme (called D-BUS) is then proposed for outage probability minimization. To verify the reliability of the proposed cooperation strategy with employing the BUS schemes, we theoretically analyze the outage probability as well as diversity orders. It is shown that the proposed cooperation strategy achieves diversity orders equal to the number of multicast users, indicating that the inherent diversity orders offered by the multicast users are fully exploited. Finally, simulation results are presented to validate the theoretical results and demonstrate the advantages of the proposed cooperation strategy and the BUS schemes.**

***Index Terms*— Cooperative unicast-multicast, non-orthogonal multiple access (NOMA), outage probability, user selection.**

Manuscript received December 1, 2016; revised April 13, 2017 and July 16, 2017; accepted September 4, 2017. Date of publication September 22, 2017; date of current version December 8, 2017. This work is supported in part by the National Natural Science Foundation of China under Grant 61601347, in part by the EU FP7 CROWN Project under Grant PIRSES-GA-2013-610524, in part by the China Postdoctoral Science Foundation under Grant 2016M600767, in part by the Natural Science Basic Research Plan in the Shaanxi Province of China under Grant 2017JQ6055, in part by the Fundamental Research Funds for the Central Universities under Grant JB160116, and in part by the 111 Project of China under Grant B08038. The associate editor coordinating the review of this paper and approving it for publication was C. Li. (*Corresponding Author: Qiang Ni.*)

L. Yang, J. Chen, and X. Xue are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: lyang@xidian.edu.cn; jianchen@mail.xidian.edu.cn; xuexuan601@gmail.com).

Q. Ni and J. Shi are with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K. (e-mail: q.ni@lancaster.ac.uk; j.shi@lancaster.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2017.2754261

I. Introduction

NON-ORTHOGONAL multiple access (NOMA) has been recognized as a promising multiple access technique to improve spectral efficiency and user fairness in the fifth generation (5G) mobile networks [1]–[7]. In contrast to conventional orthogonal multiple access (OMA) techniques (e.g., time/frequency/code division multiple access), NOMA technique allows multiple users to share same time/frequency/code domain resource with different power levels.

Most recently, cooperative relaying [8]–[10] has been introduced into NOMA to improve its reliability/capacity, where the information is forwarded by dedicated relays or successfully decoded users, termed as *relay-aided cooperative NOMA* or *user-aided cooperative NOMA*. The work in [11]–[14] has studied the relay-aided cooperative NOMA. The ergodic capacity of half/full-duplex relay-aided cooperative NOMA has been studied in [11] and [12]. It has been shown in [13] and [14] that the full diversity orders can be achieved by opportunistically selecting the *best* transmit-antenna/relay. On the other hand, user-aided cooperative NOMA has been investigated in [15]–[17]. Three distance-based user pairing schemes have been proposed in [15]. The work in [16] has derived the outage probability of full-duplex user-aided cooperative NOMA. Moreover, a full-diversity cooperation strategy has been designed in [17], where each user sequentially forwards superposed message.

However, all the aforementioned research efforts on cooperative NOMA are limited in *wireless unicast*, where data is transmitted in the point-to-point mode. In these cooperative NOMA unicast strategies, if some users have the common interest (e.g., Internet protocol television (IPTV), video/audio multicast streaming, live sports streaming, video conferencing, etc.), the same data hence has to be sent separately, leading to low spectral efficiency and inefficient link utilization. On the other hand, by exploiting the broadcasting feature of wireless channel, *wireless multicast* can deliver the same information to each user simultaneously, yielding bandwidth saving and efficient link utilization. Therefore, in NOMA systems, when some users have the same interest, integrating wireless multicast into cooperative NOMA can combine their advantages, thus improving system capacity/reliability, boosting the spectral efficiency and ameliorating link utilization simultaneously.

So far, there is no existing work on the integration of cooperative NOMA and wireless multicast. Even though, the most recent work [18], [19] has provided good understandings of NOMA multicast. In [18], the source sends both data streams of high priority and low priority to two users via NOMA, where the high priority one is intended by both users while low priority one is intended by only one user. A novel beamforming scheme has been proposed to minimize the transmit power under quality-of-service (QoS) constraints of both users. This work focuses on two-user case of NOMA multicast, whereas, in practical networks, a large number of users may intend the same data due to the common interest. Unfortunately, since wireless channels are error-prone and dynamically changing, it is challenging to guarantee good reception quality at a large number of users, as each user may distribute in different locations, experiencing different wireless channel conditions [20]–[26]. The work in [19] has designed a beamforming-based NOMA unicast-multicast strategy for multi-antenna systems. This strategy has employed beamforming and power allocation to improve the reliability of a unicast user while maintaining the reception quality of a group of multicast users. However, the strategy [19] is limited in two aspects: 1) the inherent diversity orders provided by the principles of NOMA are not exploited to improve the reliability, 2) this strategy cannot work properly if more than one unicast users are served together with multicast users.

Motivated by the observations above, we design a novel paradigm named *NOMA-enabled cooperative unicast-multicast system* in this paper. The designed system has the following advantages:

- With the aid of NOMA, a number of *unicast users* (those who have individually different interests) are allowed to share the same wireless resource with a group of *multicast users* (those who have the common interest) in this system to achieve the boost in both spectral efficiency and link utilization.
- With the increased number of multicast users, the designed system is able to achieve increased reliability and spectral utilization simultaneously, as more multicast users are effectively used in our system to improve the spectral utilization and offer more potential "relays".

Further, for the designed system, we propose a novel cooperation strategy and study its outage performance. The main contributions of this paper are summarized as follows.

- For the designed NOMA unicast-multicast system, we propose an efficient cooperation strategy which is motivated to maximize the reliability of all users. The cooperation strategy consists of 1) Direct transmission (DT) phase: the base-station (BS) broadcasts a superposed message to all users; 2) Cooperative relaying (CR) phase: a selected multicast user forwards the messages intended by unsuccessfully decoded multicast and/or unicast users. The proposed strategy is able to achieve high spectral efficiency and link utilization.
- We investigate the multicast user selection of the proposed cooperation strategy under two different power allocation (PA) approaches: 1) fixed power allocation (FPA), in which the PA coefficients for both phases are predetermined, 2) dynamic power allocation (DPA), in which the PA coefficients for DT phase are predetermined, while the PA coefficients for CR phase are dynamically determined based on instantaneous channel information. Under FPA approach, a best user selection (BUS) scheme called F-BUS is proposed to minimize the outage probability. Under DPA approach, we first derive the local optimal PA coefficients for CR phase in closed form. Then, based on derived PA coefficients, a BUS scheme called D-BUS is proposed for outage probability minimization.
- We theoretically analyze the outage performance of the proposed cooperation strategy with employing BUS schemes. When F-BUS scheme is adopted, we derive the exact outage probability of the proposed cooperation strategy in a closed-form expression. When D-BUS scheme is employed, we derive a tight approximation for the outage probability under a practically meaningful special-case. Further, we also show that the proposed cooperation strategy achieves diversity orders that equal to the number of multicast users, regardless of which BUS scheme being employed.

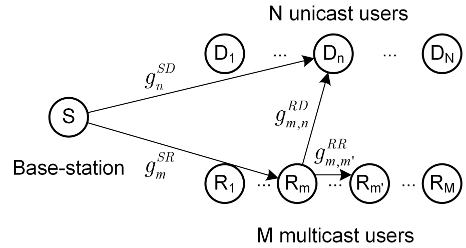


Fig. 1. NOMA-enabled unicast-multicast scenario.

The rest of this paper is organized as follows. Section II describes the considered system model of NOMA unicast-multicast and the proposed cooperation strategy. Section III details the proposed F-BUS and D-BUS schemes. Section IV derives the expressions for outage probability and then analyzes the achieved diversity orders. Section V presents simulation results, followed by Section VI that concludes this paper.

II. System Model and Cooperation Strategy

A. System Model

As depicted in Fig. 1, a BS denoted by S communicates with a number N of unicast users denoted by $D_1, ..., D_N$ who require different messages $x_1, ..., x_N$ and a group of multicast users denoted by $R_1, ..., R_M$ who require the identical message x_{N+1}. The system employs NOMA to serve all users in the same time/space/frequency domain simultaneously. All nodes are equipped with single antenna and operate in half-duplex mode. Denote channel gains (square of channel coefficient magnitude) of links $S-D_n$, $S-R_m$, R_m-D_n, $R_m-R_{m'}$ as g_n^{SD}, g_m^{SR}, $g_{m,n}^{RD}$, and $g_{m,m'}^{RR}$, respectively, where $m, m' \in \mathcal{M} \triangleq \{1, ..., M\}, m \neq m'$ and $n \in \mathcal{N} \triangleq \{1, ..., N\}$. All channels experience Rayleigh block fading, and thus, the channel gains g_n^{SD}, g_m^{SR}, $g_{m,n}^{RD}$, and $g_{m,m'}^{RR}$ follow exponential distribution with parameters λ_{SD_n}, λ_{SR_m}, $\lambda_{R_m D_n}$, $\lambda_{R_m R'_m}$ and remain unchanged within each transmission block (but change among different blocks independently). The noise at each receiver is modeled as additive white Gaussian noise with variance σ^2.

To realize downlink NOMA transmission, it is assumed that average downlink channel information is available at the

TABLE I
TABLE OF MAIN NOTATIONS IN THE PROPOSED COOPERATION STRATEGY

Notation	Description
S	Base-station (BS)
$D_1, ..., D_N$	Unicast users
$R_1, ...R_M$	Multicast users
$x_1, ..., x_N$	Messages intended by unicast users
x_{N+1}	Message intended by multicast users
r_n	Target rate of message x_n
Γ_n	SINR/SNR threshold for correctly decoding message x_n
P_s	Transmit power in DT phase
P_r	Retransmit power in CR phase
α_n	Power allocation coefficient for x_n at BS
$\beta_{m,n}$	Power allocation coefficient for x_n at multicast user R_m
$\mathcal{S}_R$	Indices set of successful multicast users
$\bar{\mathcal{S}}_R$	Indices set of unsuccessful multicast users
$\mathcal{S}_D$	Indices set of successful unicast users
$\bar{\mathcal{S}}_D$	Indices set of unsuccessful unicast users
$\|\bullet\|$	Cardinality of set $\bullet$
$\emptyset$	Null set

BS [3], [27]. Without loss of generality, we assume that the unicast users are ordered as $\mathbb{E}[g_1^{SD}] < \mathbb{E}[g_2^{SD}] < ... < \mathbb{E}[g_N^{SD}]$, implying that, for unicast users, the successive detection with successive interference cancellation (SIC) should follow the order $x_1 \to \cdots \to x_N$ whilst treating x_{N+1} (the message intended by multicast users) as interference. From a statistical viewpoint, this detection order is optimal for unicast users, since it is expected that $g_1^{SD} < g_2^{SD} < ... < g_N^{SD}$ holds in general. It is pertinent to mention here that, if the BS owns the instantaneous downlink channel information, the BS will always be able to optimally determine the detection order for unicast users, and thus, better performance is expected to be achieved. Nevertheless, as pointed in [3] and [27], the lack of instantaneous channel information is a non-trivial case for NOMA system design with practical interest. This case may happen when feedback links are limited in capacity or for other reasons [27]. On the other hand, since all multicast users are randomly distributed, the average downlink channel gain of each multicast user should be individually different. If the rationale for determining detection order at unicast users is applied here, the detection order should be individually different for each multicast user, resulting in a quite complicated design of detection order for multicast users. To simplify the design, it is assumed that successive detection with SIC follows the same order $x_1 \to \cdots \to x_N \to x_{N+1}$ at each multicast user. Note that, such an identical detection order also facilitates fully exploiting the inherent diversity orders offered by multicast users, since each multicast user may be able to serve as a relay for any unicast/multicast user. Therefore, although the detection order for multicast users may not be statistically optimal at each multicast user, the performance loss can be compensated by exploiting the inherent diversity orders.

The worst-case imperfect SIC [28] is considered in successive detection: if a message has been correctly decoded, it can be removed by SIC in the subsequent detection, otherwise, the subsequent detection will inevitably fail. Based on this assumption, a user is declared to be *successful* if its successive detection is successfully performed until its desired message being correctly decoded, or it is declared to be *unsuccessful* otherwise. Further, an *outage* is declared if any user is still unsuccessful at the end of transmission. Note that, this outage declaration is the worst-case setup, since some users may still be successful even when outage happens.

B. Proposed Cooperation Strategy

The proposed NOMA-enabled cooperation strategy is carried out in two phases: direct transmission (DT) phase and cooperative relaying (CR) phase. In DT phase, the BS broadcasts a superposed message consisting of messages $x_1, ..., x_{N+1}$ to each user. In CR phase, a *best* successful multicast user is selected to forward messages intended by unsuccessful unicast users and/or by unsuccessful multicast users.[1] For clarity, Table I summarizes the main notations used in the proposed cooperation strategy. The details of each phase are described as follows.

1) DT Phase: During this phase, BS sends a superposed message $\sum_{n=1}^{N+1} \sqrt{P_s \alpha_n} x_n$ with $\sum_{n=1}^{N+1} \alpha_n = 1$, where P_s denotes the transmit power of BS, α_n denotes the PA coefficient for x_n. Then, the successive detection with SIC is performed at all users. If multicast user R_m has correctly decoded the prior messages $x_1, ..., x_{n-1}$, the signal-to-interference-plus-noise ratio (SINR) for R_m to detect x_n (intended by D_n) is given by

$$\tilde{\gamma}_{m,n}^{\mathrm{DT}} = \frac{P_s g_m^{SR} \alpha_n}{P_s g_m^{SR} \sum_{j=n+1}^{N+1} \alpha_j + \sigma^2}. \quad (1)$$

[1] Inspired by [22], to stimulate the cooperation in CR phase, a trusted agent is assumed to charge fees from unsuccessful users who become successful at the end of CR phase, and forward the total payment to the selected multicast user.

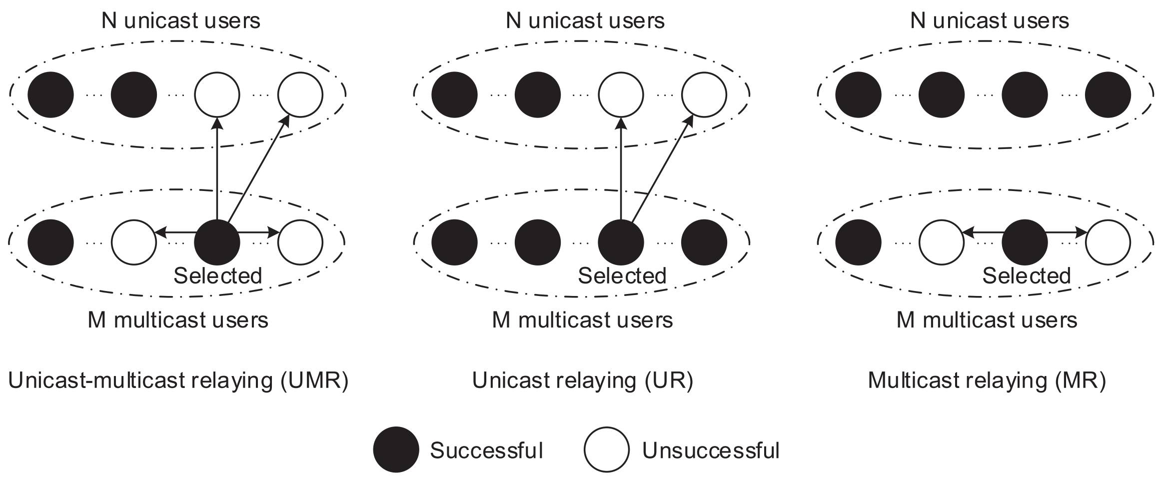


Fig. 2. Three cases of CR phase in the proposed cooperation strategy.

When messages $x_1, ..., x_N$ have been correctly decoded at multicast user R_m, this user then decodes its own message x_{N+1} with the signal-to-noise ratio (SNR) shown as follow

$$\gamma_m^{\mathrm{DT}} = P_s g_m^{SR} \alpha_{N+1}/\sigma^2. \tag{2}$$

After successive detection, a multicast user is successful if it has correctly decoded all messages $x_1, ..., x_{N+1}$. Denoting the target rate of message x_n as r_n for $n = 1, ..., N+1$, the indices set of successful multicast users can be defined as $\mathcal{S}_R \triangleq \{m | \cap_{n=1}^{N} \tilde{\gamma}_{m,n}^{\mathrm{DT}} \geq \Gamma_n, \gamma_m^{\mathrm{DT}} \geq \Gamma_{N+1}\}$, where $\Gamma_n \triangleq 2^{2r_n} - 1$ is the SINR/SNR threshold for decoding x_n. Accordingly, the indices set of unsuccessful multicast users can be expressed as $\bar{\mathcal{S}}_R = \mathcal{M} \setminus \mathcal{S}_R$.

On the other hand, supposing that unicast user D_n has correctly decoded messages $x_1, ...x_{n'-1}$ ($n' < n$), the SINR for D_n to decode the message $x_{n'}$ (intended by $D_{n'}$) can be expressed as

$$\tilde{\gamma}_{n,n'}^{\mathrm{DT}} = \frac{P_s g_n^{SD} \alpha_{n'}}{P_s g_n^{SD} \sum_{j=n'+1}^{N+1} \alpha_j + \sigma^2}. \tag{3}$$

If messages $x_1, ..., x_{n-1}$ have been correctly decoded at D_n, the SINR for D_n to decode its own message is given by

$$\gamma_n^{\mathrm{DT}} = \frac{P_s g_n^{SD} \alpha_n}{P_s g_n^{SD} \sum_{j=n+1}^{N+1} \alpha_j + \sigma^2}. \tag{4}$$

As a result, unicast user D_n is successful only if it has correctly decoded messages $x_1, ..., x_n$. Accordingly, the indices set of successful unicast users can be defined as $\mathcal{S}_D \triangleq \{n | \cap_{n'=1}^{n-1} \tilde{\gamma}_{n,n'}^{\mathrm{DT}} \geq \Gamma_{n'}, \gamma_n^{\mathrm{DT}} \geq \Gamma_n\}$, while the indices set of unsuccessful unicast users is given by $\bar{\mathcal{S}}_D \triangleq \mathcal{N} \setminus \mathcal{S}_D$. In the rest of this paper, if $\bar{\mathcal{S}}_D \neq \emptyset$, we further denote $\bar{\mathcal{S}}_D = \{n_1, ..., n_u\}$ with $1 \leq n_1 < n_2 < ... < n_u \leq N$.

If no multicast user is successful (i.e., $\mathcal{S}_R = \emptyset$) or all users are successful (i.e., $\mathcal{S}_R = \mathcal{M}$ and $\bar{\mathcal{S}}_D = \emptyset$), the CR phase will be cancelled, whilst all users waiting for the next transmission from BS.[2] Otherwise, a channel gains acquisition procedure similar to [26] will be performed to obtain instantaneous channel gains of potential relaying links.[3] Through this procedure, each successful multicast user obtains the channel gains between itself and all unsuccessful users, in order to realize user selection via BUS schemes proposed in Section III.

2) CR Phase: According to the condition of successful users at the end of DT phase, CR phase is performed in one of three cases, namely, unicast-multicast relaying (UMR), unicast relaying (UR), multicast relaying (MR), as shown in Fig. 2. In the following, these three cases are described in detail.

a) UMR case: If at least one unicast user is unsuccessful (i.e., $|\bar{\mathcal{S}}_D| \geq 1$) and not all multicast users are successful (i.e., $1 \leq |\mathcal{S}_R| \leq M-1$), the CR phase will be performed in UMR case, in which a successful multicast user is selected to forward messages intended by both unsuccessful unicast users and unsuccessful multicast users. Recall that, if $\bar{\mathcal{S}}_D \neq \emptyset$, we denote $\bar{\mathcal{S}}_D = \{n_1, ..., n_u\}$ with $1 \leq n_1 < n_2 < ... < n_u \leq N$. Supposing $R_m (m \in \mathcal{S}_R)$ is selected, it sends $\sum_{i=1}^{u} \sqrt{P_r \beta_{m,n_i}} x_{n_i} + \sqrt{P_r \beta_{m,N+1}} x_{N+1}$ with $\sum_{i=1}^{u} \beta_{m,n_i} + \beta_{m,N+1} = 1$, where P_r is the retransmit power, β_{m,n_i} and $\beta_{m,N+1}$ denote the PA coefficients for messages x_{n_i} and x_{N+1} at R_m. Upon receiving the forwarded messages, successive detection with SIC is performed at each unsuccessful user. Assuming that unsuccessful multicast user $R_{m'}$ has correctly decoded messages $x_{n_1}, ..., x_{n_{i-1}}$, the received SINR for $R_{m'}$ to decode message x_{n_i} (intended by unsuccessful unicast user D_{n_i}) is given by

$$\tilde{\gamma}_{m',n_i}^{\mathrm{UMR}}(m) = \frac{P_r g_{m,m'}^{RR} \beta_{m,n_i}}{P_r g_{m,m'}^{RR} (\sum_{j=i+1}^{u} \beta_{m,n_j} + \beta_{m,N+1}) + \sigma^2}. \tag{5}$$

[2]When no multicast user is successful, no multicast user is able to serve as a relay. When all users are successful already, the system does not need the retransmission in CR phase.

[3]In this procedure, each user sequentially announces its presence by broadcasting a flag `Success` or `Failure` with a reference power, according to whether it is successful or not. If the received flag is `Failure`, each successful multicast user measures its reception to estimate the corresponding channel gain (between itself and this unsuccessful user). Finally, each successful multicast user knows the channel gains between itself and all unsuccessful users. Since the two flags `Success` and `Failure` can be encoded into 1 bit codeword with strong channel coding, it is reasonable to assume the transmission of flags is error-free. Further, as each user sends only one flag, the communications overhead of such a procedure is $M+N$ codewords in total.

If messages $x_{n_1}, ..., x_{n_u}$ have been correctly decoded at $R_{m'}$, the SNR for $R_{m'}$ to decode its desired message is expressed as

$$\gamma_{m'}^{\mathrm{UMR}}(m) = P_r g_{m,m'}^{RR} \beta_{m,N+1}/\sigma^2. \quad (6)$$

On the other hand, when messages $x_{n_1}, ..., x_{n_{i'-1}}$ ($n_{i'} < n_i$) have been correctly decoded at unsuccessful unicast user D_{n_i}, the SINR for D_{n_i} to decode message $x_{n_{i'}}$ (intended by unsuccessful unicast user $D_{n_{i'}}$) can be expressed as

$$\tilde{\gamma}_{n_i,n_{i'}}^{\mathrm{UMR}}(m) = \frac{P_r g_{m,n_i}^{RD} \beta_{m,n_{i'}}}{P_r g_{m,n_i}^{RD} (\sum_{j=i'+1}^{u} \beta_{m,n_j} + \beta_{m,N+1}) + \sigma^2}. \quad (7)$$

If messages $x_{n_1}, ..., x_{n_{i-1}}$ have been correctly decoded at D_{n_i}, the SINR for D_{n_i} to decode its own message is given by

$$\gamma_{n_i}^{\mathrm{UMR}}(m) = \frac{P_r g_{m,n_i}^{RD} \beta_{m,n_i}}{P_r g_{m,n_i}^{RD} (\sum_{j=i+1}^{u} \beta_{m,n_j} + \beta_{m,N+1}) + \sigma^2}. \quad (8)$$

b) UR case: If at least one unicast user is unsuccessful (i.e., $|\bar{\mathcal{S}}_D| \geq 1$) and all multicast users are successful (i.e., $|\mathcal{S}_R| = M$), CR phase will be performed in UR case, in which a successful multicast user is selected to forward messages intended by all unsuccessful unicast users. If $R_m (m \in \mathcal{S}_R)$ is selected, it sends $\sum_{i=1}^{u} \sqrt{P_r \beta_{m,n_i}} x_{n_i}$ with $\sum_{i=1}^{u} \beta_{m,n_i} = 1$, where $1 \leq n_1 < n_2 < ... < n_u \leq N$ and $\{n_1, ..., n_u\} = \bar{\mathcal{S}}_D$. Subsequently, successive detection with SIC is performed at each unsuccessful unicast user. Supposing unsuccessful unicast user D_{n_i} has correctly decoded messages $x_{n_1}, ..., x_{n_{i'-1}}$ ($n_{i'} < n_i$), the SINR for D_{n_i} to decode message $x_{n_{i'}}$ (intended by unsuccessful unicast $D_{n_{i'}}$) is given by

$$\tilde{\gamma}_{n_i,n_{i'}}^{\mathrm{UR}}(m) = \frac{P_r g_{m,n_i}^{RD} \beta_{m,n_{i'}}}{P_r g_{m,n_i}^{RD} \sum_{j=i'+1}^{u} \beta_{m,n_j} + \sigma^2}. \quad (9)$$

If unsuccessful unicast user D_{n_i} has correctly decoded messages $x_{n_1}, ..., x_{n_{i-1}}$, D_{n_i} then decodes its own message with the SINR given by

$$\gamma_{n_i}^{\mathrm{UR}}(m) = \frac{P_r g_{m,n_i}^{RD} \beta_{m,n_i}}{P_r g_{m,n_i}^{RD} \sum_{j=i+1}^{u} \beta_{m,n_j} + \sigma^2}. \quad (10)$$

c) MR case: If all unicast users are successful (i.e., $\bar{\mathcal{S}}_D = \emptyset$) and not all multicast users are successful (i.e., $1 \leq |\mathcal{S}_R| \leq M-1$), CR phase will be performed in MR case, in which a successful multicast user is selected to forward only message x_{N+1}. If $R_m (m \in \mathcal{S}_R)$ is selected, it sends $\sqrt{P_r} x_{N+1}$, and then, each unsuccessful multicast user, say $R_{m'}$, directly decodes received message with SNR being $\gamma_{m'}^{\mathrm{MR}}(m) = P_r g_{m,m'}^{RR}/\sigma^2$.

It should be pointed out that the PA coefficients for both phases should be carefully determined to guarantee the decodability of each message. In DT phase, it can be observed from (1) and (4) that, when the transmit power P_s goes to infinity, the received SINR for decoding message x_n reaches its upper bound given by $\alpha_n / \sum_{j=n+1}^{N+1} \alpha_j$. However, if the SINR threshold for decoding x_n is no less than this upper bound, i.e., $\Gamma_n \geq \alpha_n / \sum_{j=n+1}^{N+1} \alpha_j$, it is impossible for any user to decode message x_n. This fact indicates that the PA coefficients for DT phase should satisfy following constraints

$$\alpha_n / \sum_{j=n+1}^{N+1} \alpha_j > \Gamma_n, \quad n = 1, ..., N. \quad (11)$$

On the other hand, it is known from (5) and (8) that, in UMR case of CR phase, the received SINR for decoding message x_{n_i} is upper bounded by $\beta_{m,n_i} / (\sum_{j=i+1}^{u} \beta_{m,n_j} + \beta_{m,N+1})$ for $i = 1, ..., u$. Therefore, the PA coefficients should satisfy following constraints

$$\beta_{m,n_i} / (\sum_{j=i+1}^{u} \beta_{m,n_j} + \beta_{m,N+1}) > \Gamma_{n_i}, \quad i = 1, ..., u. \quad (12)$$

Likewise, when CR phase is performed in UR case, the constraints for PA coefficients are given by

$$\beta_{m,n_i} / \sum_{j=i+1}^{u} \beta_{m,n_j} > \Gamma_{n_i}, \quad i = 1, ..., u-1. \quad (13)$$

III. Proposed BUS Schemes for Outage Probability Minimization

In this section, user selection for CR phase is investigated for minimizing outage probability. To guarantee the decodability of each message, we first design a fixed power allocation (FPA) approach, in which the PA coefficients for both phases are predetermined in order to strictly meet the PA constraints in (11), (12) and (13). Under FPA approach, we propose a BUS scheme (termed as F-BUS) that minimizes outage probability. On the other hand, we also design a dynamic power allocation (DPA) approach, in which the fixed PA coefficients are still employed in DT phase, but the PA coefficients for CR phase are dynamically determined based on instantaneous channel gains of relaying links. The local optimal PA coefficients for CR phase are derived in closed-form, with which a BUS scheme (termed as D-BUS) is then proposed to minimize outage probability.

It should be noted that, when CR phase is performed in MR case, the proposed cooperation strategy is actually reduced to the conventional cooperative multicast in [26]. In this case, the best successful multicast user can be selected by the max-min selection scheme [26], which is given by [4]

$$m^{\dagger} = \arg\max_{m \in \mathcal{S}_R} \min_{m' \in \bar{\mathcal{S}}_R} g_{m,m'}. \quad (14)$$

Therefore, in the rest of this section, we focus on the user selection for UMR and UR cases under FPA/DPA approaches.

[4]In conventional multicast, the outage performance is bottlenecked by the multicast user with worst reception quality. Therefore, the best successful multicast user that minimizes the outage probability is the one whose worst relaying channel is better than that of other successful multicast users, indicating the max-min selection scheme proposed in [26] is optimal for outage probability minimization.

A. FPA Approach

For presentation clarity, we here take the PA for DT phase as an example to describe the idea of FPA approach. Since $\sum_{n=1}^{N+1} \alpha_n = 1$ holds, we have $\sum_{j=n+1}^{N+1} \alpha_j = 1 - \sum_{j=1}^{n-1} \alpha_j - \alpha_n$, for $n = 1, ..., N$. Applying this result into (11) with some algebraic manipulations, we have the following equivalent constraints

$$\alpha_n > \frac{\Gamma_n (1 - \sum_{j=1}^{n-1} \alpha_j)}{1 + \Gamma_n}, \quad \forall n = 1, ..., N. \tag{15}$$

On the other hand, it can be known that $\alpha_n < 1 - \sum_{j=1}^{n-1} \alpha_j$ holds for $n = 1, ..., N$ due to $\alpha_n + \sum_{j=1}^{n-1} \alpha_j < \sum_{j=1}^{N+1} \alpha_j = 1$. Combining this inequality with (15), we have $\alpha_n \in \left(\frac{\Gamma_n}{1+\Gamma_n}(1 - \sum_{j=1}^{n-1} \alpha_j), 1 - \sum_{j=1}^{n-1} \alpha_j\right)$ for $n = 1, ..., N$, which inspires us to design the FPA approach for DT phase as follows. First, we predetermine a number N of *FPA parameters* $\Delta_1, ..., \Delta_N$, which satisfy $\frac{\Gamma_n}{1+\Gamma_n} < \Delta_n < 1$ for $n = 1, ..., N$. Then, the PA coefficients for messages $x_1, ..., x_{N+1}$ can be sequentially determined based on FPA parameters as follows

$$\begin{cases} \alpha_n^{\mathrm{F}} = \Delta_n (1 - \sum_{j=1}^{n-1} \alpha_j^{\mathrm{F}}), & \forall n = 1, ..., N, \\ \alpha_{N+1}^{\mathrm{F}} = 1 - \sum_{j=1}^{N} \alpha_j^{\mathrm{F}}, \end{cases} \tag{16}$$

where the superscript "F" represents FPA. For instance, considering $N = 3$, we first predetermine FPA parameters $\Delta_1, \Delta_2, \Delta_3$ that satisfy $\Delta_n \in (\frac{\Gamma_n}{1+\Gamma_n}, 1)$ for $n = 1, 2, 3$. Then, based on (16), the PA coefficients are sequentially obtained as $\alpha_1^{\mathrm{F}} = \Delta_1$, $\alpha_2^{\mathrm{F}} = \Delta_2(1 - \alpha_1^{\mathrm{F}})$, $\alpha_3^{\mathrm{F}} = \Delta_3(1 - \alpha_1^{\mathrm{F}} - \alpha_2^{\mathrm{F}})$ and $\alpha_4^{\mathrm{F}} = 1 - \alpha_1^{\mathrm{F}} - \alpha_2^{\mathrm{F}} - \alpha_3^{\mathrm{F}}$.

With the same rationale, the FPA approach for the UMR and UR cases of CR phase can be designed as follows. When CR phase is performed in UMR case with $R_m (m \in \mathcal{S}_R)$ being selected, the PA coefficients $\beta_{m,n_1}^{\mathrm{F}}, ..., \beta_{m,n_u}^{\mathrm{F}}, \beta_{m,N+1}^{\mathrm{F}}$ can be sequentially obtained as

$$\begin{cases} \beta_{m,n_i}^{\mathrm{F}} = \Delta_{n_i} (1 - \sum_{j=1}^{i-1} \beta_{m,n_j}^{\mathrm{F}}), & \forall i = 1, ..., u, \\ \beta_{m,N+1}^{\mathrm{F}} = 1 - \sum_{j=1}^{u} \beta_{m,n_j}^{\mathrm{F}}, \end{cases} \tag{17}$$

where $\{n_1, ..., n_u\} = \bar{\mathcal{S}}_D$ and $1 \le n_1 < ... < n_u \le N$. Similarly, when CR phase is performed in UR case with $R_m (m \in \mathcal{S}_R)$ being selected, the PA coefficients can be sequentially derived as

$$\begin{cases} \beta_{m,n_i}^{\mathrm{F}} = \Delta_{n_i} (1 - \sum_{j=1}^{i-1} \beta_{m,n_j}^{\mathrm{F}}), & \forall i = 1, ..., u-1, \\ \beta_{m,n_u}^{\mathrm{F}} = 1 - \sum_{j=1}^{u-1} \beta_{m,n_j}^{\mathrm{F}}, \end{cases} \tag{18}$$

Finally, it is necessary to check whether the derived PA coefficients in (16), (17) and (18) satisfy constraints (11), (12) and (13). For DT phase, since $\Delta_n > \frac{\Gamma_n}{1+\Gamma_n}$ and $1 - \sum_{j=1}^{n-1} \alpha_j^{\mathrm{F}} = \alpha_n^{\mathrm{F}} + \sum_{j=n+1}^{N+1} \alpha_j^{\mathrm{F}}$ hold for $n = 1, ..., N$, we have $\alpha_n^{\mathrm{F}} = \Delta_n (1 - \sum_{j=1}^{n-1} \alpha_j^{\mathrm{F}}) > \frac{\Gamma_n}{1+\Gamma_n} (\alpha_n^{\mathrm{F}} + \sum_{j=n+1}^{N+1} \alpha_j^{\mathrm{F}}), \forall n = 1, ..., N$. With some algebraic manipulations, this inequality can be written as $\alpha_n^{\mathrm{F}} > \Gamma_n \sum_{j=n+1}^{N+1} \alpha_j^{\mathrm{F}}, \forall n = 1, ..., N$, indicating the derived PA coefficients for DT phase satisfy the constraint (11). Following the similar procedure, it can also be shown that the derived PA coefficients in (17) and (18) satisfy constraints (12) and (13).

B. F-BUS Scheme

Recall the indices set of unsuccessful unicast users is denoted as $\bar{\mathcal{S}}_D = \{n_1, ..., n_u\}$ with $1 \le n_1 < ... < n_u \le N$. In UMR case, if the FPA coefficients in (17) is employed, the best successful multicast user $R_{m^\dagger}$ is selected as

$$m^\dagger = \arg\max_{m \in \mathcal{S}_R} \xi_m, \tag{19}$$

with

$$\xi_m = \min\left[\min_{i=1,...u} g_{m,n_i}^{RD} \zeta_{m,i}^{\mathrm{F}}, \min\left(\zeta_{m,u}^{\mathrm{F}}, \beta_{m,N+1}^{\mathrm{F}} / \Gamma_{N+1}\right) \min_{m' \in \bar{\mathcal{S}}_R} g_{m,m'}^{RR}\right], \tag{20}$$

where $\zeta_{m,i}^{\mathrm{F}}$ is defined as

$$\zeta_{m,i}^{\mathrm{F}} \triangleq \min_{i'=1,...i} \left(\beta_{m,n_{i'}}^{\mathrm{F}} / \Gamma_{n_{i'}} - \sum_{j=i'+1}^{u} \beta_{m,n_j}^{\mathrm{F}} - \beta_{m,N+1}^{\mathrm{F}}\right). \tag{21}$$

In UR case, if FPA coefficients in (18) is employed, the best successful multicast user $R_{m^\dagger}$ is selected as

$$m^\dagger = \arg\max_{m \in \mathcal{S}_R} \varepsilon_m, \tag{22}$$

with

$$\varepsilon_m = \min_{i=1,...,u} g_{m,n_i}^{RD} \zeta_{m,i}^{\prime\mathrm{F}}, \tag{23}$$

where $\zeta_{m,n}^{\prime\mathrm{F}}$ is defined as

$$\zeta_{m,i}^{\prime\mathrm{F}} \triangleq \min_{i'=1,...i} \left(\beta_{m,n_{i'}}^{\mathrm{F}} / \Gamma_{n_{i'}} - \sum_{j=i'+1}^{u} \beta_{m,n_j}^{\mathrm{F}}\right). \tag{24}$$

It is noted that, the objective values ξ_m and ε_m are only related to the channel gains associated with successful multicast user R_m, meaning that each successful multicast user can determine its own objective value based on its acquired channel gains. Therefore, the proposed F-BUS scheme can be realized in a distributed manner as follows. First, each successful multicast user maintains a timer with an initial value inversely proportional to its own objective value. Then, the one with the maximal objective value will expire first and serve as a relay.

The optimality of this scheme is shown in the following theorem.

Theorem 1: Under FPA approach, the proposed F-BUS scheme given in (19) and (22) is optimal for outage probability minimization.

Proof: Please refer to Appendix A for details. ■

C. DPA Approach

Since the BS only owns the average downlink channel information, the PA coefficients may not be dynamically determined for the DT phase. Even though, the BS can still utilize the average downlink channel information to optimize its PA coefficients for improving long-term reliability. This is a challenging problem that will be studied in our future work. On the other hand, each successful multicast user owns the instantaneous channel gains of its potential relaying channels, which can be exploited to dynamically determine the PA coefficients in UMR and UR cases of CR phase. In the sequel, we consider that the FPA approach for DT phase (given in (16)) is still employed, and we mainly focus on dynamically optimizing the PA coefficients in UMR and UR cases, aiming at outage probability minimization.

1) DPA Approach for UMR Case: Since outage is the event that at least one user is not eventually successful, its complementary event is that all users are eventually successful. In UMR case, all users become eventually successful if the following events happen at the end of CR phase: 1) message x_{n_i} ($\forall i = 1, ..., u$) has been correctly decoded by not only D_{n_i} (as desired message) but also $\{D_{n_{i'}}\}_{i'=i+1,...,u}$ and $\{R_{m'}\}_{m'\in\bar{\mathcal{S}}_R}$ (as other user's message), 2) message x_{N+1} has been correctly decoded by $\{R_{m'}\}_{m'\in\bar{\mathcal{S}}_R}$ (as desired message). Therefore, in UMR case, the outage probability conditioned on $R_m (m \in \mathcal{S}_R)$ being selected is given by

$$P^{\text{UMR}}_{\text{out}|R_m} = 1 - \Pr\Bigg(\cap_{i=1}^{u} \Big[\gamma^{\text{UMR}}_{n_i}(m) \geq \Gamma_{n_i}, \cap_{i'=i+1}^{u} \tilde{\gamma}^{\text{UMR}}_{n_{i'},n_i}(m) \geq \Gamma_{n_i}, \cap_{m'\in\bar{\mathcal{S}}_R} \tilde{\gamma}^{\text{UMR}}_{m',n_i}(m) \geq \Gamma_{n_i}\Big], \cap_{m'\in\bar{\mathcal{S}}_R} \gamma^{\text{UMR}}_{m'}(m) \geq \Gamma_{N+1}\Bigg). \quad (25)$$

Substituting (5), (6), (7) and (8) into (25) with some algebraic manipulations, this conditional outage probability can be further expressed as

$$\begin{aligned} P^{\text{UMR}}_{\text{out}|R_m} &= 1 - \Pr\Bigg(\cap_{i=1}^{u} \Big[\phi_{m,i} g^{RD}_{m,n_i} \geq \tfrac{\sigma^2}{P_r}, \phi_{m,i} \min_{i'=i+1,...,u} g^{RD}_{m,n_{i'}} \geq \tfrac{\sigma^2}{P_r}, \phi_{m,i} \min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'} \geq \tfrac{\sigma^2}{P_r}\Big], \phi_{u+1} \min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'} \geq \tfrac{\sigma^2}{P_r}\Bigg) \\ &= \Pr\left(\min_{i=1,...,u+1} \phi_{m,i} G_{m,i} < \tfrac{\sigma^2}{P_r}\right), \end{aligned} \quad (26)$$

with the terms $\phi_{m,i}$ and $G_{m,i}$ being defined as

$$\phi_{m,i} \triangleq \begin{cases} \beta_{m,n_i}/\Gamma_{n_i} \\ -\sum_{j=i+1}^{u} \beta_{m,n_j} - \beta_{m,N+1}, & i = 1, ..., u, \\ \beta_{m,N+1}/\Gamma_{N+1}, & i = u+1, \end{cases} \quad (27)$$

$$G_{m,i} \triangleq \begin{cases} \min\left(\min_{i'=i,...,u} g^{RD}_{m,n_{i'}}, \min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'}\right), & i = 1, ..., u, \\ \min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'}, & i = u+1. \end{cases} \quad (28)$$

As known from (26), the minimization of $P^{\text{UMR}}_{\text{out}|R_m}$ is equivalent to the maximization of $\min_{i=1,...,u+1} \phi_{m,i} G_{m,i}$. Therefore, the PA problem for minimizing outage probability can be formulated as follow

$$\begin{aligned} &\underset{\beta_{m,n_1},...,\beta_{m,n_u},\beta_{m,N+1}}{\textbf{maximize}} \quad \min_{i=1,...,u+1} \phi_{m,i} G_{m,i}, \\ &\textbf{s.t.} \quad (12) \text{ and } \sum_{i=1}^{u} \beta_{m,n_i} + \beta_{m,N+1} = 1. \end{aligned} \quad (29)$$

However, due to the presence of the constraints shown in (12), it is challenging to derive the optimal solution to the PA problem above. Instead, we first derive the optimal solution to a relaxed PA problem without constraints (12), and then, prove that the derived optimal solution to the relaxed PA problem is also a feasible and local optimal solution to original PA problem (29).

If constraints (12) are removed, the original PA problem (29) will be reduced to the following relaxed PA problem

$$\begin{aligned} &\underset{\beta_{m,n_1},...,\beta_{m,n_u},\beta_{m,N+1}}{\textbf{maximize}} \quad \min_{i=1,...,u+1} \phi_{m,i} G_{m,i} \\ &\textbf{s.t.} \quad \sum_{i=1}^{u} \beta_{m,n_i} + \beta_{m,N+1} = 1, \end{aligned} \quad (30)$$

where the optimal solution is given in the following lemma.

Lemma 1: The optimal solution to relaxed PA problem (30) can be expressed as

$$\begin{cases} \beta^*_{m,N+1} = \dfrac{E_m \Gamma_{N+1}}{G_{m,u+1}}, \\ \beta^*_{m,n_u} = E_m \left(\dfrac{\Gamma_{n_u}}{G_{m,u}} + \frac{\Gamma_{n_u}\Gamma_{N+1}}{G_{m,u+1}} \right), \\ \beta^*_{m,n_{u-t}} = E_m \Big[\sum_{l=u-t+1}^{u} A_{m,l} \prod_{r=u-t+1}^{l-1} B_r + \left(\dfrac{\Gamma_{n_u}}{G_{m,u}} + \dfrac{\Gamma_{n_u}\Gamma_{N+1}}{G_{m,u+1}} \right) \times \prod_{r=u-t+1}^{u} B_r \Big], \quad \text{for } t = 1, ..., u-1, \end{cases} \quad (31)$$

with

$$E_m = \left[\sum_{t=1}^{u-1} \sum_{l=u-t+1}^{u} A_{m,l} \prod_{r=u-t+1}^{l-1} B_r + \left(1 + \sum_{t=1}^{u-1} \prod_{r=u-t+1}^{u} B_r \right) \times \left(\frac{\Gamma_{n_u}}{G_{m,u}} + \frac{\Gamma_{n_u}\Gamma_{N+1}}{G_{m,u+1}} \right) + \frac{\Gamma_{N+1}}{G_{m,u+1}} \right]^{-1}, \quad (32)$$

where $A_{m,l} \triangleq \Gamma_{n_{l-1}} \left(\frac{1}{G_{m,l-1}} - \frac{1}{G_{m,l}} \right)$ and $B_r \triangleq \Gamma_{n_{r-1}} \left(1 + \frac{1}{\Gamma_{n_r}} \right)$.

Proof: Please refer to Appendix B for details. ■

Lemma 2: The derived solution $\{\beta^*_{m,n_1}, ..., \beta^*_{m,n_u}, \beta^*_{m,N+1}\}$ given in (31) is a feasible and local optimal solution to original PA problem (29).

Proof: Here, we first prove that $\{\beta^*_{m,n_1}, ..., \beta^*_{m,n_u}, \beta^*_{m,N+1}\}$ is a feasible solution to original PA problem (29). It can be seen from (28) that $0 < G_{m,1} \leq G_{m,2} \leq ... \leq G_{m,u+1}$ holds, indicating $A_{m,l} = \Gamma_{n_{l-1}} \left(1/G_{m,l-1} - 1/G_{m,l} \right) \geq 0$ for $l = 1, ..., u$. Then, applying these results into (32), it is known $E_m > 0$ holds. Subsequently, according to the first equation of (B.1), the inequality $\frac{\beta^*_{m,n_i}}{\Gamma_{n_i}} - \sum_{i'=i+1}^{u} \beta^*_{m,i'} - \beta^*_{m,N+1} = \frac{E_m}{G_{m,n}} > 0$ holds for $i = 1, ..., u$, indicating the derived optimal solution to relaxed PA problem (30) also satisfies (12). Therefore, $\{\beta^*_{m,n_1}, ..., \beta^*_{m,n_u}, \beta^*_{m,N+1}\}$ is a feasible solution to original PA problem (29).

Next, we use the proof by contradiction to show that $\{\beta^*_{m,n_1}, ..., \beta^*_{m,n_u}, \beta^*_{m,N+1}\}$ is a local optimal solution to original problem (29).

The claim is negated to assume that $\{\beta^*_{m,n_1}, ..., \beta^*_{m,n_u}, \beta^*_{m,N+1}\}$ is not a local optimal solution to original problem (29), or, equivalently, there must exist another feasible solution to original PA problem (29) that achieves an objective value larger than E_m (achieved by $\{\beta^*_{m,n_1},...,\beta^*_{m,n_u},\beta^*_{m,N+1}\}$). For notational convenience, we denote such a feasible solution as $\{\beta^{\ddagger}_{m,n_1},...,\beta^{\ddagger}_{m,n_u},\beta^{\ddagger}_{m,N+1}\}$ and denote the corresponding objective value as $E^{\ddagger}_m$. Further, as constraints (12) are removed in relaxed PA problem (30), $\{\beta^{\ddagger}_{m,n_1},...,\beta^{\ddagger}_{m,n_u},\beta^{\ddagger}_{m,N+1}\}$ should also be a feasible solution to relax PA problem (30). Therefore, we know that $\{\beta^{\ddagger}_{m,n_1},...,\beta^{\ddagger}_{m,n_u},\beta^{\ddagger}_{m,N+1}\}$ is a feasible solution to relax PA problem (30) and achieves an objective value larger than E_m, i.e.,

$$E^{\ddagger}_m > E_m. \tag{33}$$

On the other hand, it is known from Lemma 1 that $\{\beta^*_{m,n_1},...,\beta^*_{m,n_u},\beta^*_{m,N+1}\}$ is the optimal solution to relaxed PA problem (30), indicating that the objective value E_m achieved by $\{\beta^*_{m,n_1},...,\beta^*_{m,n_u},\beta^*_{m,N+1}\}$ should be no less than $E^{\ddagger}_m$, i.e.,

$$E^{\ddagger}_m \le E_m. \tag{34}$$

Apparently, the results in (33) and (34) are mutually exclusive. As the Lemma 1 has been proved to be true, we have inequality (34) should hold whereas inequality (33) should be false, meaning that the negated claim does not hold. Therefore, $\{\beta^*_{m,n_1},...,\beta^*_{m,n_u},\beta^*_{m,N+1}\}$ should be a local optimal solution to original PA problem (29). This completes the proof. ■

2) DPA Approach for UR Case: When CR phase is performed in UR case, all users are eventually successful if message x_{n_i} is correctly decoded by not only D_{n_i} but also $\{D_{n_{i'}}\}_{i'=i+1,...,u}$ for $i = 1,...,u$. Thus, when $R_m (m \in \mathcal{S}_R)$ is selected, the outage probability is expressed as

$$P^{\text{UR}}_{\text{out}|R_m} = 1 - \Pr\Big[\cap_{i=1}^{u}\Big(\gamma^{\text{UR}}_{n_i}(m) \ge \Gamma_{n_i}, \cap_{i'=i+1}^{u}\gamma^{\text{UR}}_{n_{i'},n_i}(m) \ge \Gamma_{n_i}\Big)\Big]. \tag{35}$$

Applying (9) and (10) into (35) with some algebraic manipulations, we have

$$\begin{aligned}P^{\text{UR}}_{\text{out}|R_m} &= 1 - \Pr\Big[\cap_{i=1}^{u}\Big(\phi'_{m,i} g^{RD}_{m,n_i} \ge \tfrac{\sigma^2}{P_r}, \phi'_{m,i}\min_{i'=i+1,...,u} g^{RD}_{m,n_{i'}} \ge \tfrac{\sigma^2}{P_r}\Big)\Big]\\ &= \Pr\Big(\min_{i=1,...,u}\phi'_{m,i}G'_{m,i} < \tfrac{\sigma^2}{P_r}\Big),\end{aligned} \tag{36}$$

where $\phi'_{m,i}$ and $G'_{m,i}$ are defined as

$$\begin{cases}\phi'_{m,i} \triangleq \dfrac{\beta_{m,n_i}}{\Gamma_{n_i}} - \sum_{j=i+1}^{u}\beta_{m,n_j},\\ G'_{m,i} \triangleq \min_{i'=i,...,u} g^{RD}_{m,n_{i'}}.\end{cases} \tag{37}$$

As known from (36), $P^{\text{UR}}_{\text{out}|R_m}$ monotonically decreases with $\min_{i=1,..u}\phi'_{m,i}G'_{m,i}$, resulting in the following PA problem for minimizing outage probability

$$\begin{aligned}&\underset{\beta_{m,n_1},...,\beta_{m,n_u}}{\textbf{maximize}} \quad \min_{i=1,...,u}\phi'_{m,i}G'_{m,i}\\ &\textbf{s.t.} \quad (13) \text{ and } \sum_{i=1}^{u}\beta_{m,n_i} = 1.\end{aligned} \tag{38}$$

Likewise, although deriving the optimal solution to PA problem (38) is challenging, we can still obtain a feasible and local optimal solution to this problem from a relaxed PA problem without constraints (13).

If constraints (13) are removed, the original PA problem (38) will be simplified to a relaxed PA problem given by

$$\begin{aligned}&\underset{\beta_{m,n_1},...,\beta_{m,n_u}}{\textbf{maximize}} \quad \min_{i=1,...,u}\phi'_{m,i}G'_{m,i}\\ &\textbf{s.t.} \quad \sum_{i=1}^{u}\beta_{m,n_i} = 1.\end{aligned} \tag{39}$$

In the following two lemmas, we provide the optimal solution to the relaxed PA problem above, and then, prove that the optimal solution to relaxed PA problem (39) is also a feasible and local optimal solution to original PA problem (38).

Lemma 3: The optimal solution to relaxed PA problem (39) can be expressed as

$$\begin{cases}\beta^*_{m,n_u} = \dfrac{E'_m\Gamma_{n_u}}{G'_{m,u}},\\ \beta^*_{m,n_{u-t}} = E'_m\Big(\sum_{l=u-t+1}^{u}A'_{m,l}\prod_{r=u-t+1}^{l-1}B_r + \dfrac{\Gamma_{n_u}}{G'_{m,u}}\prod_{r=u-t+1}^{u}B_r\Big), \quad \text{for } t = 1,...,u-1,\end{cases} \tag{40}$$

with

$$E'_m = \left[\sum_{t=1}^{u-1}\sum_{l=u-t+1}^{u}A'_{m,l}\prod_{r=u-t+1}^{l-1}B_r + \left(1+\sum_{t=1}^{u-1}\prod_{r=u-t+1}^{u}B_r\right)\frac{\Gamma_{n_u}}{G'_{m,u}}\right]^{-1}, \tag{41}$$

where $A'_{m,l} \triangleq \Gamma_{n_{l-1}}\left(\frac{1}{G'_{m,l-1}} - \frac{1}{G'_{m,l}}\right)$.

Proof: Following the proof of Lemma 1 in Appendix B, this lemma is obtained. ■

Lemma 4: The solution $\{\beta^*_{m,n_1},...,\beta^*_{m,n_u}\}$ given in (40) is a feasible and local optimal solution to original PA problem (38).

Proof: Following the proof of Lemma 2, this lemma is obtained. ■

D. D-BUS Scheme

When the DPA coefficients are employed, the best successful multicast user $R_{m^\dagger}$ is selected as

$$\begin{cases}m^\dagger = \arg\max_{m\in\mathcal{S}_R} E_m, & \text{UMR case},\\ m^\dagger = \arg\max_{m\in\mathcal{S}_R} E'_m, & \text{UR case},\end{cases} \tag{42}$$

where the terms E_m and E'_m have been defined in (32) and (41). Similar to F-BUS scheme, computing the objective values E_m and E'_m only requires the channel gains associated with R_m, implying the proposed D-BUS scheme can also be realized in a distributed way, as discussed in Section III-B.

Theorem 2: When the DPA coefficients are employed, the proposed D-BUS scheme in (42) can select the successful multicast user that achieves the minimal outage probability.

Proof: The proof is presented in Appendix C. ■

It is noteworthy that, since the DPA coefficients are obtained from the local optimal solutions to original PA problems (29)

$$\begin{aligned}
P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u} &= \Pr\left(\xi_{m^\dagger}=\max_{m\in\mathcal{S}_R}\xi_m<\frac{\sigma^2}{P_r}\middle|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u\right)\\
&\overset{\text{(i)}}{=}\Pr\left(\max_{m\in\mathcal{A}_k}\left[\min\left(\min_{i=1,\ldots u}g^{RD}_{m,n_i}\zeta^{\mathrm{F}}_{m,i},\min\left(\zeta^{\mathrm{F}}_{m,u},\frac{\beta^{\mathrm{F}}_{m,N+1}}{\Gamma_{N+1}}\right)\min_{m'\in\mathcal{M}\setminus\mathcal{A}_k}g^{RR}_{m,m'}\right)\right]<\frac{\sigma^2}{P_r}\right)\\
&=\prod_{m\in\mathcal{A}_k}\left[1-\prod_{i=1}^{u}\Pr\left(g^{RD}_{m,n_i}\geq\frac{\sigma^2}{P_r\zeta^{\mathrm{F}}_{m,i}}\right)\prod_{m'\in\mathcal{M}\setminus\mathcal{A}_k}\Pr\left(g^{RR}_{m,m'}\geq\frac{\sigma^2}{P_r\min\left(\zeta^{\mathrm{F}}_{m,u},\beta^{\mathrm{F}}_{m,N+1}/\Gamma_{N+1}\right)}\right)\right]\\
&=\prod_{m\in\mathcal{A}_k}\left[1-\exp\left(-\frac{\sigma^2}{P_r}\left(\sum_{i=1}^{u}\frac{\lambda_{R_mD_{n_i}}}{\zeta^{\mathrm{F}}_{m,i}}+\frac{c_{m,\mathcal{A}_k}}{\min\left(\zeta^{\mathrm{F}}_{m,u},\beta^{\mathrm{F}}_{m,N+1}/\Gamma_{N+1}\right)}\right)\right)\right],
\end{aligned}\tag{43}$$

and (38), the proposed D-BUS scheme with employing DPA coefficients is expected to perform well and achieve near-optimal outage performance.

IV. Performance Evaluation

In this section, the outage probability of the proposed cooperation strategy is analyzed in Rayleigh fading channels.[5] When F-BUS scheme is adopted, we derive a closed-form expression for outage probability. When D-BUS scheme is employed, we derive a tight approximation for outage probability under a special case which is practically meaningful. Then, based on asymptotic analysis in high SNR regime, we also show that the achievable diversity orders of the proposed cooperation strategy are equal to the number of multicast users.

A. Outage Probability Analysis

1) Outage Probability With Employing F-BUS Scheme: In the proposed cooperation strategy, the outage happens in following conditions: 1) CR phase is performed, but some users remain unsuccessful after CR phase, 2) CR phase is cancelled due to $\mathcal{S}_R=\emptyset$. In the following, we first analyze the outage performance when CR phase is performed in each case or cancelled, and then, we further derive the overall outage probability of the proposed cooperation strategy with employing F-BUS. For notational convenience, if $\mathcal{S}_R\neq\emptyset$ and $\mathcal{S}_R\neq\mathcal{M}$, we denote $\mathcal{S}_R=\{m_1,\ldots,m_k\}\triangleq\mathcal{A}_k$ with $1\leq m_1<\ldots<m_k\leq M$ and $1\leq k\leq M-1$ whereas, if $\bar{\mathcal{S}}_D\neq\emptyset$, we denote $\bar{\mathcal{S}}_D=\{n_1,\ldots,n_u\}\triangleq\mathcal{B}_u$ with $1\leq n_1<\ldots<n_u\leq N$ and $1\leq u\leq N$.

If $\mathcal{S}_R=\mathcal{A}_k$ and $\bar{\mathcal{S}}_D=\mathcal{B}_u$, CR phase will be performed in UMR case. Following the derivations in (A.2), the corresponding conditional outage probability can be obtained as (43), shown at the top of this page, where step (i) uses the expression for ξ_m given in (20), $c_{m,\mathcal{A}_k}\triangleq\sum_{m'\in\mathcal{M}\setminus\mathcal{A}_k}\lambda_{R_mR_{m'}}$. On the other hand, denoting the probability for multicast user R_m being successful after DT phase as Ψ^R_m and the probability for unicast user D_n being successful after DT phase as Ψ^D_n, the probability for condition $\{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u\}$ can be expressed as

$$\begin{aligned}
P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u} &= \prod_{m\in\mathcal{A}_k}\Psi^R_m\prod_{m'\in\mathcal{M}\setminus\mathcal{A}_k}(1-\Psi^R_{m'})\\
&\quad\times\prod_{i=1}^{u}(1-\Psi^D_{n_i})\prod_{n'\in\mathcal{N}\setminus\mathcal{B}_u}\Psi^D_{n'},
\end{aligned}\tag{44}$$

with Ψ^R_m and Ψ^D_n being obtained as

$$\begin{aligned}
\Psi^R_m &= \Pr\left(\cap_{n=1}^{N}\tilde{\gamma}^{\mathrm{DT}}_{m,n}\geq\Gamma_{n'},\gamma^{\mathrm{DT}}_m\geq\Gamma_{N+1}\right)\\
&=\Pr\left(g^{SR}_m\psi_{N+1}\geq\frac{\sigma^2}{P_s}\right)\\
&=\exp\left(-\frac{\sigma^2\lambda_{SR_m}}{P_s\psi_{N+1}}\right),
\end{aligned}\tag{45}$$

$$\begin{aligned}
\Psi^D_n &= \Pr\left(\cap_{n'=1}^{n-1}\tilde{\gamma}^{\mathrm{DT}}_{n,n'}\geq\Gamma_{n'},\gamma^{\mathrm{DT}}_n\geq\Gamma_n\right)\\
&=\Pr\left(g^{SD}_n\psi_n\geq\frac{\sigma^2}{P_s}\right)\\
&=\exp\left(-\frac{\sigma^2\lambda_{SD_n}}{P_s\psi_n}\right),
\end{aligned}\tag{46}$$

where ψ_n is defined as

$$\psi_n=\begin{cases}\min_{n'=1,\ldots,n}\left(\alpha^{\mathrm{F}}_{n'}/\Gamma_{n'}\right.\\ \left.-\sum_{j=n'+1}^{N+1}\alpha^{\mathrm{F}}_j\right), & n=1,\ldots,N,\\ \min\left(\min_{n'=1,\ldots,N}\left(\alpha^{\mathrm{F}}_{n'}/\Gamma_{n'}\right.\right.\\ \left.\left.-\sum_{j=n'+1}^{N+1}\alpha^{\mathrm{F}}_j\right),\frac{\alpha^{\mathrm{F}}_{N+1}}{\Gamma_{N+1}}\right), & n=N+1.\end{cases}\tag{47}$$

If $\mathcal{S}_R=\mathcal{M}$ and $\bar{\mathcal{S}}_D=\mathcal{B}_u$, CR phase will be performed in UR case. Accordingly, the conditional outage probability can be obtained based on (A.4), which is given by

$$\begin{aligned}
&P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u}\\
&=\Pr\left(\varepsilon_{m^\dagger}=\max_{m\in\mathcal{S}_R}\varepsilon_m<\frac{\sigma^2}{P_r}\middle|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u\right)\\
&\overset{\text{(ii)}}{=}\Pr\left(\max_{m\in\mathcal{M}}\min_{i=1,\ldots,u}g^{RD}_{m,n_i}\zeta'^{\mathrm{F}}_{m,i}<\frac{\sigma^2}{P_r}\right)\\
&=\prod_{m\in\mathcal{M}}\left[1-\prod_{i=1}^{u}\Pr\left(g^{RD}_{m,n_i}\geq\frac{\sigma^2}{P_r\zeta'^{\mathrm{F}}_{m,i}}\right)\right]\\
&=\prod_{m\in\mathcal{M}}\left[1-\exp\left(-\frac{\sigma^2}{P_r}\sum_{i=1}^{u}\frac{\lambda_{R_mD_{n_i}}}{\zeta'^{\mathrm{F}}_{m,i}}\right)\right],
\end{aligned}\tag{48}$$

[5]By following the derivations in this section, the analytical results in other fading models can also be obtained. Further, when all users are randomly deployed, the analytical results can be further obtained with the help of stochastic geometry tools [32].

where step (ii) comes from the expression for ε_m given in (23). Moreover, based on (45) and (46), the probability for condition $\{\mathcal{S}_R = \mathcal{M}, \bar{\mathcal{S}}_D = \mathcal{B}_u\}$ can be expressed as

$$P_{\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u} = \prod_{m=1}^{M} \Psi_m^R \prod_{i=1}^{u}(1-\Psi_{n_i}^D) \prod_{n'\in\mathcal{N}\setminus\mathcal{B}_u} \Psi_{n'}^D. \quad (49)$$

If $\mathcal{S}_R = \mathcal{A}_k$ and $\bar{\mathcal{S}}_D = \emptyset$, CR phase will be performed in MR case. Recalling that the max-min selection (14) is employed in MR case, the conditional outage probability can be expressed as

$$\begin{aligned} P^{\mathrm{MR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset} &= \Pr\left(\min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m^\dagger,m'} = \max_{m\in\mathcal{S}_R}\min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'} \right.\\ &\qquad\left. < \frac{\sigma^2\Gamma_{N+1}}{P_r}\Big|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset\right)\\ &= \prod_{m\in\mathcal{A}_k}\left[1-\prod_{m'\in\mathcal{M}\setminus\mathcal{A}_k}\Pr\left(g^{RR}_{m,m'}\geq\frac{\sigma^2\Gamma_{N+1}}{P_r}\right)\right]\\ &= \prod_{m\in\mathcal{A}_k}\left[1-\exp\left(-\frac{\sigma^2\Gamma_{N+1}c_{m,\mathcal{A}_k}}{P_r}\right)\right]. \quad (50)\end{aligned}$$

Likewise, the probability for condition $\{\mathcal{S}_R = \mathcal{A}_k, \bar{\mathcal{S}}_D = \emptyset\}$ can also be obtained based on (45) and (46) as

$$P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset} = \prod_{m\in\mathcal{A}_k}\Psi_m^R \prod_{m'\in\mathcal{M}\setminus\mathcal{A}_k}(1-\Psi_{m'}^R)\prod_{n=1}^{N}\Psi_n^D. \quad (51)$$

Additionally, when CR phase is cancelled due to $\mathcal{S}_R = \emptyset$, outage inevitably happens. Based on (45), the probability for such a condition is $P_{\mathcal{S}_R=\emptyset} = \prod_{m=1}^{M}(1-\Psi_m^R)$.

Then, with the help of Total Probability Theorem [33, Sec.(3.3.8)], the overall outage probability of proposed F-BUS scheme can be expressed as

$$\begin{aligned} P^{\mathrm{F-BUS}}_{\mathrm{out}} &= P_{\mathcal{S}_R=\emptyset} + \widehat{\sum_{k,M-1,\mathcal{A}_k}}\ \widehat{\sum_{u,N,\mathcal{B}_u}} P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u}\\ &\quad\times P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u}\\ &\quad+\widehat{\sum_{u,N,\mathcal{B}_u}} P_{\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u} P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u}\\ &\quad+\widehat{\sum_{k,M-1,\mathcal{A}_k}} P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset} P^{\mathrm{MR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset}, \quad (52)\end{aligned}$$

where $\widehat{\sum_{k,M-1,\mathcal{A}_k}}$ is the short-hand-notation for $\sum_{k=1}^{M-1}\sum_{m_1=1}^{M-k+1}\sum_{m_2=m_1+1}^{M-k+2}\dots\sum_{m_k=m_{k-1}+1}^{M}$, $\widehat{\sum_{u,N,\mathcal{B}_u}}$ is the short-hand-notation for $\sum_{u=1}^{N}\sum_{n_1=1}^{N-u+1}\sum_{n_2=n_1+1}^{N-u+2}\cdots\sum_{n_u=n_{u-1}+1}^{N}$. Finally, summarizing foregoing results, a closed-form expression for overall outage probability is derived.

2) Outage Probability With Employing D-BUS Scheme: When D-BUS scheme is adopted for user selection, we first analyze the outage performance of the proposed cooperation strategy under the special case $N = 1$, i.e., only one unicast user shares the same resource with multicast users. Note that, the outage performance evaluation for this special case is practically meaningful, because two-message NOMA transmission is not only the current research focus (see, e.g., [11]–[15], [18], [34]), but also a simple and effective scenario for realizing NOMA in practical networks.[6]

When the number of unicast users is reduced to $N = 1$, the overall outage probability can be expressed as

$$\begin{aligned} P^{\mathrm{D-BUS}}_{\mathrm{out},N=1} &= P_{\mathcal{S}_R=\emptyset} + \widehat{\sum_{k,M-1,\mathcal{A}_k}} P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}\\ &\quad\times P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}\\ &\quad+P_{\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}} P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}}\\ &\quad+\widehat{\sum_{k,M-1,\mathcal{A}_k}} P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset} P^{\mathrm{MR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset}, \quad (53)\end{aligned}$$

where $P_{\mathcal{S}_R=\emptyset} = \prod_{m=1}^{M}(1-\Psi_m^R)$, $P^{\mathrm{MR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset}$ has been obtained in (50). Moreover, using (44), (49) and (51) with letting $\mathcal{N} = \{1\}$ and $\mathcal{B}_u = \{1\}$, the expressions for probabilities $P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}$, $P_{\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}}$ and $P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset}$ can also be obtained in closed form. Therefore, we only need to further evaluate the conditional outage probabilities $P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}$ and $P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}}$.

When $N = 1$, we have $\bar{\mathcal{S}}_D = \{1\}$ for UMR and UR cases. Consequently, the D-BUS scheme in (42) is simplified as follow

$$m^\dagger = \begin{cases} \arg\max\limits_{m\in\mathcal{S}_R}\left(\dfrac{\Gamma_1}{G_{m,1}}+\dfrac{\Gamma_1\Gamma_2+\Gamma_2}{G_{m,2}}\right)^{-1}, \\ \quad \text{UMR case, } N=1 \\ \arg\max\limits_{m\in\mathcal{S}_R} g^{RD}_{m,1}, \\ \quad \text{UR case, } N=1 \end{cases} \quad (54)$$

where $G_{m,1} = \min(g^{RD}_{m,1}, \min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'})$ and $G_{m,2} = \min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'}$. Further, since the UR case with $N = 1$ is actually reduced to conventional unicast relaying, the corresponding conditional outage probability can be obtained as

$$\begin{aligned} P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}} &= \Pr\left(\frac{P_s}{\sigma^2}g^{RD}_{m^\dagger,1} = \frac{P_r}{\sigma^2}\max_{m=1,\dots,M} g^{RD}_{m,1} < \Gamma_1\right)\\ &= \prod_{m=1}^{M}\left[1-\exp\left(-\frac{\sigma^2\Gamma_1\lambda_{R_mD_1}}{P_r}\right)\right]. \quad (55)\end{aligned}$$

In the following, we focus on deriving the conditional outage probability $P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}$, according to the simplified D-BUS scheme for UMR case given in (54).

Lemma 5: The conditional outage probability $P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}$ can be expressed as

$$\begin{aligned} &P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}\\ &= \prod_{m\in\mathcal{A}_k}\Bigg(1-\underbrace{\int_0^{\frac{P_r}{\sigma^2 a}}\int_0^{\frac{P_r}{\sigma^2\Gamma_1}-bx} f_{X_{m,\mathcal{A}_k}}(x)f_{Y_m}(y)\mathrm{d}y\mathrm{d}x}_{\triangleq\Omega_{m,\mathcal{A}_k}}\Bigg), \end{aligned} \quad (56)$$

[6] As pointed by [35], it may not be realistic to perform NOMA among all users due to strong co-channel interference. A more promising alternative is to partition users into multiple groups and then perform NOMA in each group.

where $X_{m,\mathcal{A}_k} \triangleq 1/\min_{m' \in \mathcal{M}\setminus\mathcal{A}_k} g_{m,m'}^{RR}$ and $Y_m \triangleq 1/g_{m,1}^{RD}$, $a \triangleq \Gamma_1\Gamma_2 + \Gamma_1 + \Gamma_2$, $b \triangleq \Gamma_2 + \Gamma_2/\Gamma_1$.

Proof: Please refer to Appendix D for details. ■

Since $g_{m,m'}^{RR}$ and $g_{m,1}^{RD}$ follow exponential distributions with parameters being $\lambda_{R_m R_{m'}}$ and $\lambda_{R_m D_1}$, the probability density functions (PDFs) of $X_{m,\mathcal{A}_k}$ and Y_m can be obtained as follows

$$f_{X_{m,\mathcal{A}_k}}(x) = \frac{c_{m,\mathcal{A}_k}}{x^2}\exp\left(-\frac{c_{m,\mathcal{A}_k}}{x}\right), \tag{57}$$

$$f_{Y_m}(y) = \frac{\lambda_{R_m D_1}}{y^2}\exp\left(-\frac{\lambda_{R_m D_1}}{y}\right), \tag{58}$$

where $c_{m,\mathcal{A}_k} = \sum_{m' \in \mathcal{M}\setminus\mathcal{A}_k}\lambda_{R_m R_{m'}}$. Using the derived PDFs, the term $\boldsymbol{\Omega}_{m,\mathcal{A}_k}$ can be further expressed as

$$\boldsymbol{\Omega}_{m,\mathcal{A}_k} = \int_0^{\frac{P_r}{\sigma^2 a}} \frac{c_{m,\mathcal{A}_k}}{x^2}\exp\left[-\left(\frac{c_{m,\mathcal{A}_k}}{x} + \frac{\lambda_{R_m D_1}}{\frac{P_r}{\sigma^2\Gamma_1} - bx}\right)\right]\mathrm{d}x. \tag{59}$$

However, it is extremely difficult to derive a closed-form expression for the integral in (59) due to the special form of the integrand. Instead, we here propose a tight approximation for $\boldsymbol{\Omega}_{m,\mathcal{A}_k}$ in the following lemma.

Lemma 6: The integral $\boldsymbol{\Omega}_{m,\mathcal{A}_k}$ can be approximately expressed as follows

$$\begin{aligned}\boldsymbol{\Omega}_{m,\mathcal{A}_k} &\approx \frac{1}{2}\exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r}\right)\left[\exp\left(-\frac{\sigma^2\lambda_{R_m D_1}\Gamma_1}{P_r}\right)\right.\\ &\quad\left. + \exp\left(-\frac{\sigma^2\lambda_{R_m D_1}}{P_r\left(1/\Gamma_1 - b/(Qa)\right)}\right)\right]\\ &\quad + \frac{1}{2}\sum_{q=2}^{Q}\left[\exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r q}\right) - \exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r(q-1)}\right)\right]\\ &\quad \times\left[\exp\left(-\frac{\sigma^2\lambda_{R_m D_1}}{P_r\left(1/\Gamma_1 - [b(q-1)]/(Qa)\right)}\right)\right.\\ &\quad \left. + \exp\left(-\frac{\sigma^2\lambda_{R_m D_1}}{P_r\left(1/\Gamma_1 - (bq)/(Qa)\right)}\right)\right],\end{aligned} \tag{60}$$

where Q is a positive integer.

Proof: Please refer to Appendix E. ■

The accuracy of the approximation given in (60) is verified by simulation in Fig. 3. It can be seen that, when Q increases from 1 to 5, the curve "Approximate" gets closer to the curve "Actual", as expected. Further, if Q is no less than 3, the gap between curves "Approximate" and "Actual" is very small and does not vary with the value of P_r/σ^2, which means the expression in (60) can provide a tight approximation for $\boldsymbol{\Omega}_{m,\mathcal{A}_k}$ given in (56).

Combining Lemmas 5 and 6, a tight approximation is obtained for $P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}}$. Then, applying foregoing results in (53), a tight approximation for overall outage probability in the special case $N = 1$ is derived.

On the other hand, as observed from (31) and (40), the derived DPA coefficients for CR phase are highly correlated for $N \geq 2$. This fact indicates that, when D-BUS scheme is adopted, it is difficult to theoretically derive the overall outage probability for $N \geq 2$. Nevertheless, we numerically evaluate the outage probability of proposed strategy with employing D-BUS scheme in simulation, as given in Section V.

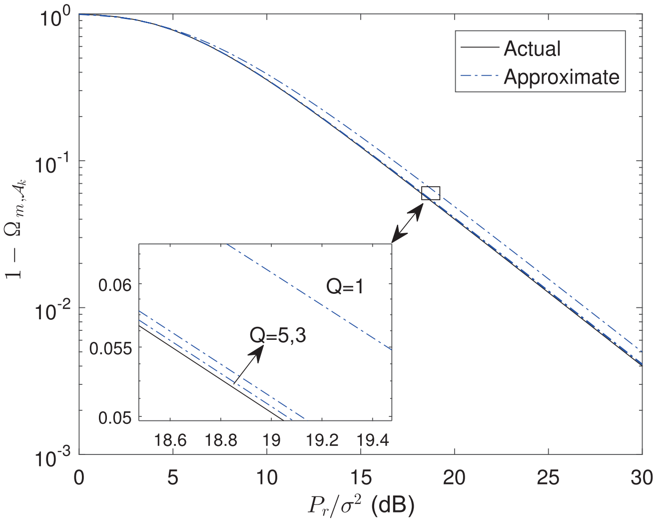


Fig. 3. The accuracy of approximation in (60) with $Q = 1, 3, 5$, where the parameters are $c_{m,\mathcal{A}_k} = \lambda_{R_m D_1} = 1$, $r_1 = r_2 = 1$bps/Hz. Here, the curve "Approximate" is obtained based on the approximate $\boldsymbol{\Omega}_{m,\mathcal{A}_k}$ given in (60), while the curve "Actual" is obtained by numerically computing the integral in (59).

B. Diversity Analysis

1) Diversity Orders of F-BUS Scheme: Define $\rho \triangleq \frac{P_s}{\sigma^2} = \frac{\mu P_r}{\sigma^2}$ as the *system SNR*, where $\mu = \frac{P_s}{P_r}$ is a positive constant.

Based on (43), when system SNR $\rho \to \infty$, the conditional outage probability $P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u}$ can be asymptotically expressed as

$$\begin{aligned}&P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u}\\ &\overset{\rho\to\infty}{\simeq}\left(\frac{\mu}{\rho}\right)^k\prod_{m\in\mathcal{A}_k}\left(\sum_{i=1}^{u}\frac{\lambda_{R_m D_{n_i}}}{\zeta^{\mathrm{F}}_{m,i}} + \frac{c_{m,\mathcal{A}_k}}{\min\left(\zeta^{\mathrm{F}}_{m,u},\beta^{\mathrm{F}}_{m,N+1}/\Gamma_{N+1}\right)}\right)\\ &\propto \quad \rho^{-k}.\end{aligned} \tag{61}$$

According to (45) and (46), we have $\Psi_m^R \overset{\rho\to\infty}{\simeq} 1 - \frac{\lambda_{SR_m}}{\rho\psi_{N+1}}$ and $\Psi_n^D \overset{\rho\to\infty}{\simeq} 1 - \frac{\lambda_{SD_n}}{\rho\psi_n}$ hold for $\rho \to \infty$. Then, based on (44), we have

$$\begin{aligned}P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\mathcal{B}_u} &\overset{\rho\to\infty}{\simeq} \frac{\prod_{m\in\mathcal{M}\setminus\mathcal{A}_k}\lambda_{SR_m}\prod_{i=1}^{u}\lambda_{SD_{n_i}}/\psi_{n_i}}{\rho^{M-k+u}\,\psi_{N+1}^{M-k}}\\ &\propto \quad \rho^{-(M-k+u)}.\end{aligned} \tag{62}$$

Combining (61) and (62) we know that, when CR phase is performed in UMR case, the proposed cooperation strategy achieves diversity orders of $M + u$.

Subsequently, based on (48), the high-SNR asymptotic expression for $P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u}$ can be obtained as

$$P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u} \overset{\rho\to\infty}{\simeq} \left(\frac{\mu}{\rho}\right)^M\prod_{m\in\mathcal{M}}\sum_{i=1}^{u}\frac{\lambda_{R_m D_{n_i}}}{\zeta'^{\mathrm{F}}_{m,i}} \propto \rho^{-M}. \tag{63}$$

Further, when system SNR $\rho \to \infty$, from (49) we also have

$$P_{\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\mathcal{B}_u} \overset{\rho\to\infty}{\simeq} \rho^{-u}\prod_{i=1}^{u}\frac{\lambda_{SD_{n_i}}}{\psi_{n_i}} \propto \rho^{-u}. \tag{64}$$

Therefore, the proposed cooperation strategy also achieves diversity orders of $M + u$ when CR phase is performed in UR case.

Similarly, based on (50) and (51), when CR phase is performed in MR case, we have following high-SNR asymptotic expressions

$$P^{\mathrm{MR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset} \overset{\rho\to\infty}{\simeq} \left(\frac{\mu}{\rho}\right)^k \Gamma^k_{N+1} \prod_{m\in\mathcal{A}_k} c_{m,\mathcal{A}_k} \propto \rho^{-k}, \quad (65)$$

$$P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\emptyset} \overset{\rho\to\infty}{\simeq} \rho^{-(M-k)} \psi^{-(M-k)}_{N+1} \prod_{m\in\mathcal{M}\setminus\mathcal{A}_k} \lambda_{SR_m} \propto \rho^{-(M-k)}, \quad (66)$$

which demonstrate that the cooperation strategy achieves diversity orders of M when CR phase is performed in MR case.

Additionally, in high-SNR regime, the probability that the event $\mathcal{S}_R = \emptyset$ happens can be asymptotically expressed as $P_{\mathcal{S}_R=\emptyset} \overset{\rho\to\infty}{\simeq} \rho^{-M}\psi^{-M}_{N+1}\prod_{m=1}^{M}\lambda_{SR_m} \propto \rho^{-M}$. This fact indicates, even when CR phase is cancelled, the cooperation strategy still achieves diversity orders of M. As a summary, when F-BUS scheme is adopted in the proposed cooperation strategy, diversity orders of M are achieved at least.

2) Diversity Orders of D-BUS Scheme: Here, we first analyze the diversity orders of D-BUS scheme under the special case $N = 1$. When system SNR $\rho \to \infty$, the approximate expression for $\Omega_{m,\mathcal{A}_k}$ given in (60) can be asymptotically expressed as

$$\begin{aligned}\Omega_{m,\mathcal{A}_k} \overset{\rho\to\infty}{\approx}\ & \frac{1}{2}\left(1-\frac{\mu c_{m,\mathcal{A}_k}Qa}{\rho}\right) \\ & \times\left(2-\frac{\mu\lambda_{R_mD_1}\Gamma_1}{\rho}-\frac{\mu\lambda_{R_mD_1}}{\rho(1/\Gamma_1-b/(Qa))}\right) \\ & +\frac{\mu c_{m,\mathcal{A}_k}Qa}{\rho}\sum_{q=2}^{Q}\left(\frac{1}{q-1}-\frac{1}{q}\right) \\ =\ & 1-\frac{\mu}{\rho}\left[c_{m,\mathcal{A}_k}a+\frac{\lambda_{R_mD_1}\Gamma_1}{2}\right. \\ & \left.\times\left(1+\frac{1}{1-(b\Gamma_1)/(Qa)}\right)\right]. \quad (67)\end{aligned}$$

Applying this asymptotic expression into (56), we have $P^{\mathrm{UMR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}} \overset{\rho\to\infty}{\propto} \rho^{-k}$. It is known from (62) that, by letting $\mathcal{B}_u = \{1\}$, we have $P_{\mathcal{S}_R=\mathcal{A}_k,\bar{\mathcal{S}}_D=\{1\}} \overset{\rho\to\infty}{\propto} \rho^{-(M-k+1)}$. By summarizing the results above, we know that, when D-BUS is employed in UMR case with $N = 1$, the achievable diversity orders are equal to $M + 1$. When system SNR $\rho \to \infty$, the conditional outage probability $P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}}$ in (55) can be asymptotically expressed as

$$P^{\mathrm{UR}}_{\mathrm{out}|\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}} \overset{\rho\to\infty}{\simeq} \left(\frac{\mu}{\rho}\right)^M \Gamma^M_1 \prod_{m=1}^{M}\lambda_{R_mD_1} \propto \rho^{-M}. \quad (68)$$

Further, using (64) with letting $\mathcal{B}_u = \{1\}$, we have $P_{\mathcal{S}_R=\mathcal{M},\bar{\mathcal{S}}_D=\{1\}} \propto \rho^{-1}$. Thus, when D-BUS scheme is employed in UR case with $N = 1$, the achievable diversity orders are also equal to $M + 1$. Additionally, as shown in Section IV-B.1, diversity orders of M are achieved when CR phase is cancelled or performed in MR case. Therefore, the overall diversity orders of D-BUS scheme under $N = 1$ are also not less than M.

On the other hand, as mentioned in Section IV-A.2, when D-BUS scheme is adopted under $N \geq 2$, deriving the closed-form outage probability is challenging due to the dependence among the derived DPA coefficients for UMR and UR cases. Thus, it may be impossible to theoretically evaluate the diversity orders based on asymptotic analysis in high-SNR regime. Nevertheless, we can still intuitively evaluate its diversity orders based on the diversity orders achieved by F-BUS scheme. Recalling that the DPA coefficients given (31) and (40) are employed in D-BUS scheme, it is expected that D-BUS scheme outperforms F-BUS scheme in term of outage probability. As proved in Section IV-B.1, when F-BUS scheme is employed, the cooperation strategy achieves diversity orders of M at least. Consequently, it can be deduced that, when D-BUS is employed, the achievable diversity orders should not be smaller than M.

Overall, it can be observed from the above analysis that, regardless of which BUS scheme being employed, the proposed cooperation strategy achieves at least diversity orders of M, which are equal to the number of multicast users. This observation provides following insights: 1) the proposed cooperation strategy with employing either one of F-BUS and D-BUS schemes can fully exploit the inherent diversity provided by all multicast users, 2) having more multicast users can significantly improve the outage performance due to increased diversity orders.

Further, recall that an outage is declared if any user is not eventually successful. Intuitively, the derived outage probability and diversity orders correspond to the worst-case reliability of the proposed cooperation strategy, and thus, the actual diversity orders achieved at each user should be no less than M. On the other hand, when more unicast users are introduced, the outage performance will become worse, since the transmit power for each message will be reduced. However, this outage performance loss can be compensated by introducing more multicast users to increase diversity orders. This observation suggests that, by properly scheduling the unicast users and multicast users, a large number of users can share the same resource without outage performance degradation.

V. Simulation Results

In this section, we use simulations to verify our theoretical results and demonstrate the advantages of the proposed cooperation strategy and BUS schemes. We consider an independently and non-identically distributed Rayleigh fading scenario, where the BS locates at coordinate $(0, 0)$ and all users randomly locate within a circle centered at BS and with a radius being 100 meters. The pathloss model in [36, Sec.2.6] is adopted in our simulation. Specifically, denoting the distance between nodes i and j as d_{ij}, the corresponding pathloss between nodes i and j is $1/\lambda_{ij} = (\frac{d_{ij}}{d_0})^{-\kappa}$, where d_0 is the reference distance and κ is the pathloss exponent. As this model is invalid for $d_{ij} < d_0$ [36],

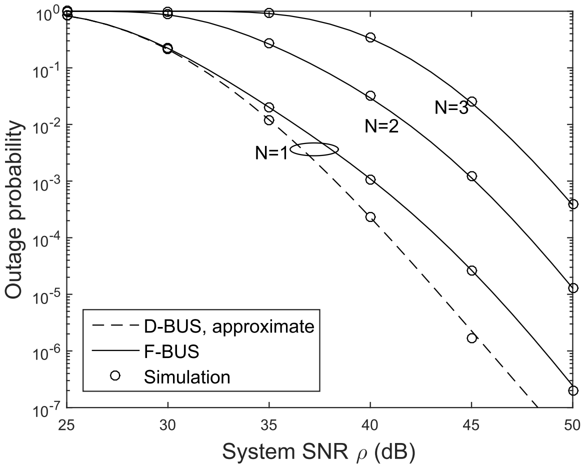


Fig. 4. Theoretical and simulated outage probabilities of the proposed cooperation strategy with employing BUS schemes, where $M = 4$, $\mathbf{r} =$ 1bps/Hz and $\mu = P_s/P_r = 100$.

we do not consider the impact of pathloss if $d_{ij} < d_0$, i.e., $1/\lambda_{ij} = 1$ for $d_{i,j} < d_0$. In our simulation, we set $d_0 = 20$ meters and $\kappa = 3$, which are standard values for urban cellular networks [36]. Without loss of generality, we set the noise power $\sigma^2 = 1$, and thus, we have $P_s = \rho$ and $P_r = \rho/\mu$. Moreover, an identical target rate $\mathbf{r}$ is assumed for all messages transmission, i.e., $\mathbf{r} = r_1 = r_2 = ... = r_{N+1}$ bps/Hz and the FPA parameters are predetermined as $\Delta_n = 0.95 \times \frac{\Gamma_n}{1+\Gamma_n} + 0.05$ for $n = 1, ..., N$. The simulated values are averaged over 1×10^8 independent numerical results.

A. Verification for Derived Theoretical Results

Fig. 4 plots the theoretical and simulated outage probabilities of the proposed cooperation strategy with employing BUS schemes, where $M = 4$, $\mathbf{r} = $ 1bps/Hz, $\mu = P_s/P_r = 100$. It is observed that, when F-BUS scheme is adopted with $N = 1, 2, 3$, the theoretical outage probabilities perfectly coincide with the simulated ones, verifying the derived closed-form outage probability for F-BUS scheme. Recall that, when D-BUS scheme is employed, a tight approximate outage probability is derived under the special case $N = 1$. As seen from this figure, when D-BUS is employed under $N = 1$, the derived approximate outage probability also tightly matches the simulated outage probability, showing the tightness of the derived approximate outage probability.

B. Comparison Between the Proposed Cooperation Strategy and Other Strategies

This subsection compares our proposed cooperation strategy with the *BS retransmission strategy*, the *distributed cooperation strategy* [20] and the *non-cooperation strategy*, which are schematically drawn in Fig. 5 and described as follows. Similar to the proposed strategy, the BS retransmission strategy and the distributed cooperation strategy also perform within two successive phases. In the first phase, the BS broadcasts to all users in both strategies. In the second phase, the BS retransmits the messages intended by the unsuccessful unicast/multicast users in the BS retransmission strategy, whereas, in the distributed cooperation strategy, all successful multicast users

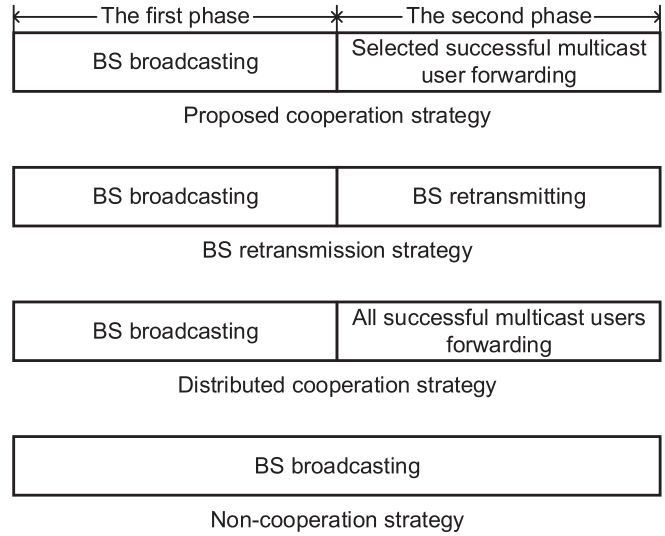


Fig. 5. The schematic diagram of the proposed cooperation strategy, the BS retransmission strategy, the distributed cooperation strategy and the non-cooperation strategy.

simultaneously forward messages intended by the unsuccessful unicast/multicast users. In the non-cooperation strategy, the BS directly broadcasts to all users via NOMA without any cooperation/retransmission. To make the comparison more convincing, we assume that the BS owns instantaneous downlink channel information to optimally determine the detection order for unicast users in these three strategies above.

Fig. 6 compares the proposed cooperation strategy and three strategies described above in terms of outage probability under $M = 5$ and $N = 3$. Since the average downlink channel information is assumed in the proposed cooperation strategy, its outage probability should be lower bounded by its counterpart with the instantaneous downlink channel information. To demonstrate the outage performance gap between the proposed cooperation strategy and its counterpart with instantaneous downlink channel information, we also provide the simulated outage probabilities of its counterpart in Fig. 6 (a) (denoted by "lower bound"), as the lower-bound outage probability of our proposed strategy.

It is observed from Fig. 6(a) that, in high SNR regime ($>$ 50 dB), the outage probabilities of proposed strategy and its lower bounds decay at the rate same as the benchmark $1 \times 10^{19} \times \rho^{-5}$, indicating the proposed strategy achieves diversity orders equal to the number of multicast users $M = 5$, regardless of the availability of instantaneous downlink channel information. However, the outage probability of other strategies decays more slowly than the proposed one in high SNR regime, implying their achievable diversity orders are lower than that of proposed one. Further, by comparing the outage probability of proposed cooperation strategy and its counterpart with instantaneous downlink channel information, we know that owning the instantaneous downlink channel information can achieve around $1 \sim 3$ dB outage probability reduction, since the BS can always optimally determine the detection order for unicast users according to instantaneous downlink channel information, as discussed in Section II-A.

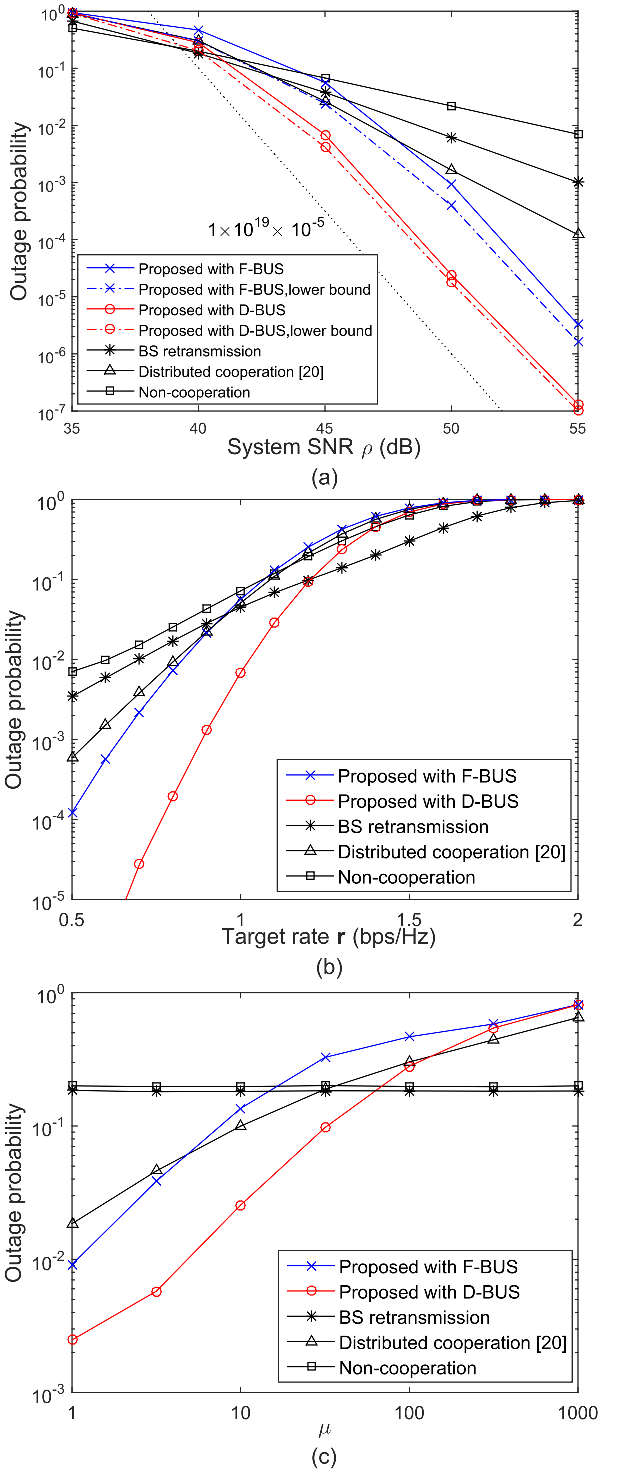


Fig. 6. Comparison among the proposed cooperation strategy, the BS retransmission strategy, the distributed cooperation strategy and the non-cooperation strategy, in terms of outage probability with (a) varying system SNR ρ under $\mathbf{r} = 1$ bps/Hz, $\mu = P_s/P_r = 100$, and with (b) varying target rate $\mathbf{r}$ under $\rho = 45$ dB, $\mu = P_s/P_r = 100$ and with (c) varying μ under $\rho = 40$ dB, $\mathbf{r} = 1$bps/Hz. Here, we set $M = 5$ and $N = 3$.

From Fig. 6(b) we see that, when $\mathbf{r} < 0.9$ bps/Hz, the proposed cooperation strategy and distributed cooperation strategy achieve lower outage probability than the BS retransmission strategy and the non-cooperation strategy, whereas, when $\mathbf{r} > 1.7$ bps/Hz, the outage probability of the proposed cooperation strategy and the distributed cooperation strategy converge to 1. This observation can be intuitively explained as follows. When the value of $\mathbf{r}$ is small, some multicast users may be successful after the first phase, and thus, the cooperation in the second phase can take effect. However, when the value of $\mathbf{r}$ further increases higher than a certain level, almost no multicast user is successful, leading to no available relays for cooperation. On the other hand, when $\mathbf{r} \geq 1.3$ bps/Hz, the BS retransmission strategy outperforms all other strategies, since the BS owns much higher transmit power than the successful multicast users.

Fig. 6 (c) demonstrates the impact of transmit power of successful multicast user on outage probability by varying parameter $\mu = P_s/P_r$ under $P_s = \rho = 40$ dB. When the value of μ increases, the outage performance of proposed cooperation strategy and the distributed cooperation strategy becomes worse, since less transmit power is available at each successful multicast user. It is also seen that the outage probabilities of the BS retransmission strategy and the non-cooperation strategy is independent from the value of μ, since these two strategies do not recruit any user to forward messages. Further, when μ is less than 10, our proposed strategy and distributed cooperation strategy outperform the BS retransmission strategy and non-cooperation strategy. However, when μ further increases higher than 100, the BS retransmission strategy and non-cooperation strategy achieve much better outage performance than our proposed strategy, since each successful multicast user does not own sufficient transmit power for information forwarding.

C. Comparison Between Proposed FPA/DPA Schemes and Existing PA Schemes

Considering that the proposed cooperation strategy is employed, we compare our proposed FPA/DPA approaches with three existing PA approaches in literature: quality-of-service based PA (QoS-PA) [37], channel-state-information based PA (CSI-PA) [37] and fair NOMA PA (F-NOMA-PA) [38]. For making a fair comparison, we here adopt the following simulation configuration: 1) the number of unicast users is reduced to $N = 1$, since the aforementioned existing PA approaches are designed for two-message NOMA; 2) different PA approaches are employed in only CR phase, as the existing PA approaches are designed for one-hop NOMA transmission; 3) the PA coefficients for DT phase is obtained from (16), in order to guarantee the same decoding results after DT phase; 4) the transmit power of each successful multicast user is equal to that of the BS, i.e., $\mu = P_s/P_r = 1$, in order to highlight the impact of different power allocation approaches on outage probability/throughput; 5) for user selection, the proposed F-BUS scheme is adopted under all PA approaches to guarantee the fairness.

Fig. 7 shows the outage probability and outage throughput under $M = 5$, $N = 1$ and $\mathbf{r} = 1$ bps/Hz. Here, the outage throughput is defined as the product of sum target rate and the probability of no outage happening (i.e., $1 - P_{\text{out}}$). It is shown that, our proposed FPA/DPA approaches significantly outperform other three existing PA approaches in both outage probability and outage throughput. Further, as observed

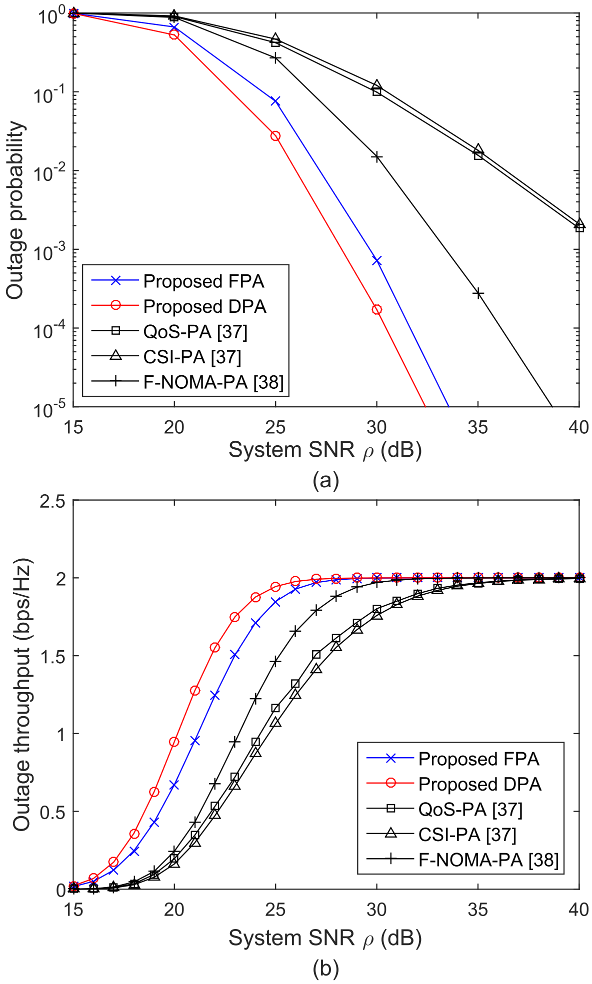


Fig. 7. Comparison among the proposed FPA/DPA approaches, the QoS-PA approach [37], the CSI-PA approach [37] and the F-NOMA-PA approach [38], in terms of (a) outage probability and (b) outage throughput, where $M = 5$, $N = 1$, $\mathbf{r} = 1$ bps/Hz and $\mu = P_s/P_r = 1$.

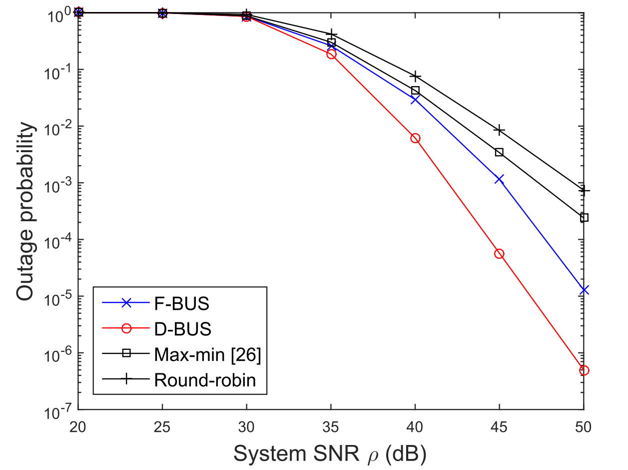


Fig. 8. Comparison between the proposed BUS schemes, the max-min selection scheme [26], and the round-robin selection scheme with $M = 4$, $N = 2$, $\mathbf{r} = 1$ bps/Hz and $\mu = P_s/P_r = 100$.

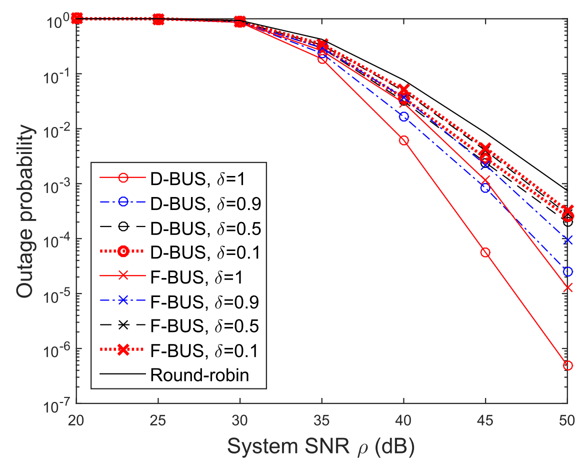


Fig. 9. Outage probability of proposed BUS schemes under imperfect channel information, where $M = 4$, $N = 2$, $\mathbf{r} = 1$ bps/Hz and $\mu = P_s/P_r = 100$.

from Fig. 7(a), as system SNR ρ increases, the outage probability achieved by proposed FPA/DPA approaches decays faster than other approaches, implying that the proposed cooperation strategy cannot achieve diversity orders of M under existing PA approaches in [37] and [38]. On the other hand, as shown in Fig. 7(b), when system SNR $\rho < 30$ dB, the proposed FPA/DPA approaches yield the much higher outage throughput over existing PA approaches due to the higher diversity orders. When system SNR further increases, outage throughput stemmed from each PA approach converges to the ceiling 2 bps/Hz, since the outage probability approaches to zero in high SNR regime.

D. Performance Evaluation for Proposed F-BUS/D-BUS Schemes

It is also expected to compare our proposed BUS schemes with existing user selection schemes in literature. However, since there is no previous work on cooperative NOMA unicast-multicast, we compare them with the max-min selection scheme for conventional cooperative multicast [26], in which the best successful multicast user is selected as $m^\dagger = \arg\max_{m\in\mathcal{S}_R} \min_{m'\in\bar{\mathcal{S}}_R} g^{RR}_{m,m'}$. Moreover, we also provide the simulation results of round-robin selection as a benchmark for comparison. As shown in Fig. 8, the proposed BUS schemes significantly outperform the max-min selection scheme and the round-robin selection scheme. The reason for this observation is twofold: 1) the max-min selection scheme only maximizes the worst reception quality among unsuccessful multicast users, without considering the reception quality of unsuccessful unicast users; 2) the max-min selection scheme is reduced to the round-robin selection scheme in UR case due to $\bar{\mathcal{S}}_R = \emptyset$; 3) the round-robin selection scheme does not utilize any channel information and thus attains the worst outage performance.

Next, we further evaluate the performance of our proposed BUS schemes under imperfect channel information. Let h denote the actual channel coefficient and $\hat{h}$ denote the observed channel coefficient, the relation between them can be formulated as $\hat{h} = \delta h + \sqrt{1-\delta^2}\varpi$, where δ is the correlation coefficient between $\hat{h}$ and h, ϖ conforms the same distribution with h. In Fig. 9, we consider that all potential relaying channels are imperfectly estimated with the same correlation coefficient being $\delta = 0.9, 0.5, 0.1$. Further, Fig. 9 also provides the outage probability of the proposed BUS schemes under perfect channel information (i.e., $\delta = 1$) and outage probability of round-robin selection as upper and lower bounds of outage performance. It is shown that the outage probability of the proposed BUS schemes increases as the correlation coefficient δ decreases. Further, it can also be observed that, the proposed D-BUS scheme always achieves better outage probability than the proposed F-BUS scheme, regardless of the varying correlation coefficient δ.

VI. Conclusion

In this paper, we have designed a novel cooperation strategy for NOMA unicast-multicast, where a multicast user is selected to forward a superposed message to unsuccessfully decoded unicast and/or multicast users. Under FPA and DPA approaches, two BUS schemes (called F-BUS and D-BUS) have been proposed to properly select a multicast user for outage probability minimization. To evaluate the reliability of proposed BUS schemes, we have theoretically derived their outage performance as well as diversity orders. It has been shown that the proposed BUS schemes all achieve diversity orders equal to the number of multicast users, indicating the two proposed BUS schemes all fully utilize the diversity orders offered by multicast users. Simulation results have verified our theoretical results and demonstrated the advantages of our proposed cooperation strategy and BUS schemes.

Appendix A
Proof of Theorem 1

Here, we first prove the optimality of F-BUS scheme in UMR case. In UMR case, the *complementary event* of outage happening is that unicast users $\{D_{n_i}\}_{i=1,\ldots,u}$ and multicast users $\{R_{m'}\}_{m'\in\bar{\mathcal{S}}_R}$ are successful at the end of CR phase. Thus, if $R_m (m \in \mathcal{S}_R)$ is selected, the outage probability can be expressed as

$$\begin{aligned}P^{\mathrm{UMR}}_{\mathrm{out}|R_m} &= 1 - \Pr\Big[\cap_{i=1}^{u} \Big(\gamma^{\mathrm{UMR}}_{n_i}(m) \ge \Gamma_{n_i},\ \cap_{i'=1}^{i-1} \tilde{\gamma}^{\mathrm{UMR}}_{n_i,n_{i'}}(m) \ge \Gamma_{n_{i'}}\Big),\\ &\quad \cap_{m'\in\bar{\mathcal{S}}_R}\Big(\gamma^{\mathrm{UMR}}_{m'}(m) \ge \Gamma_{N+1},\ \cap_{i'=1}^{u} \tilde{\gamma}^{\mathrm{UMR}}_{m,n_{i'}}(m) \ge \Gamma_{n_{i'}}\Big)\Big]. \end{aligned} \tag{A.1}$$

Applying (5), (6) (7) and (8) into (A.1) with some algebraic manipulations, $P^{\mathrm{UMR}}_{\mathrm{out}|R_m}$ can be further expressed as (A.2), shown at the top of the next page, where the equality (a.1) comes from (20). From (A.2) it is known that $P^{\mathrm{UMR}}_{\mathrm{out}|R_m}$ monotonically decreases with ξ_m. Thus, to minimize the outage probability in UMR case, the best successful multicast user should be selected as $m^\dagger = \arg\max_{m\in\mathcal{S}_R} \xi_m$.

Likewise, if $R_m (m \in \mathcal{S}_R)$ is selected in UR case, the outage probability can be expressed as

$$P^{\mathrm{UR}}_{\mathrm{out}|R_m} = 1 - \Pr\Big[\cap_{i=1}^{u} \Big(\gamma^{\mathrm{UR}}_{n_i}(m) \ge \Gamma_{n_i},\ \cap_{i'=1}^{i-1} \tilde{\gamma}^{\mathrm{UR}}_{n_i,n_{i'}}(m) \ge \Gamma_{n_{i'}}\Big)\Big]. \tag{A.3}$$

Substituting (9) and (10) into (A.3) with some algebraic manipulations, $P^{\mathrm{UR}}_{\mathrm{out}|R_m}$ can be further expressed as (A.4), also given at the top of the next page, where equality (a.2) comes from (23). As shown in (A.4), $P^{\mathrm{UR}}_{\mathrm{out}|R_m}$ monotonically decreases with ε_m, indicating it is optimal to select the best successful multicast user $R_{m^\dagger}$ as $m^\dagger = \arg\max_{m\in\mathcal{S}_R} \varepsilon_m$ for outage probability minimization. This completes the proof.

Appendix B
Proof of Lemma 1

If constraints (12) are relaxed, the original problem (29) is reduced to a traditional max-min problem shown in (30), whose maximal objective value is attained when each term inside the "min" operation (i.e., $\phi_{m,i} G_{m,i}$) is equal to the maximal objective value [29]–[31]. Denoting the optimal solution of relaxed problem (30) as $\{\beta^*_{m,n_i}, \ldots, \beta^*_{m,n_u}, \beta^*_{m,N+1}\}$ and the corresponding maximal objective value as E_m, we have the following equations hold based on (27)

$$G_{m,i}\phi_{m,i} = \begin{cases} G_{m,i}\left(\dfrac{\beta^*_{m,n_i}}{\Gamma_{n_i}} - \sum_{i'=i+1}^{u} \beta^*_{m,n_{i'}} - \beta^*_{m,N+1}\right) = E_m, \\ \quad \forall i = 1, \ldots, u, \\ G_{m,u+1}\dfrac{\beta^*_{m,N+1}}{\Gamma_{N+1}} = E_m. \end{cases} \tag{B.1}$$

It is known from the second equation of (B.1) that $\beta^*_{m,N+1}$ can be expressed as

$$\beta^*_{m,N+1} = \frac{E_m \Gamma_{N+1}}{G_{m,u+1}}. \tag{B.2}$$

On the other hand, the first equation of (B.1) can be rewritten as

$$\frac{\beta^*_{m,n_i}}{\Gamma_{n_i}} = \frac{E_m}{G_{m,i}} + \sum_{i'=i+1}^{u} \beta^*_{m,n_{i'}} + \beta^*_{m,N+1}, \quad i = 1, \ldots, u, \tag{B.3}$$

Let $i = \upsilon$ and $i = \upsilon - 1$, $\forall \upsilon \in \{2, \ldots, u\}$, we have

$$\begin{cases} \dfrac{\beta^*_{m,n_\upsilon}}{\Gamma_{n_\upsilon}} = \dfrac{E_m}{G_{m,\upsilon}} + \sum_{i'=\upsilon+1}^{u} \beta^*_{m,n_{i'}} + \beta^*_{m,N+1} \\ \dfrac{\beta^*_{m,n_{\upsilon-1}}}{\Gamma_{n_{\upsilon-1}}} = \dfrac{E_m}{G_{m,\upsilon-1}} + \sum_{i'=\upsilon}^{u} \beta^*_{m,n_{i'}} + \beta^*_{m,N+1} \end{cases} \tag{B.4}$$

Combining these two equations in (B.4) with minus operation on both sides, the relation between β^*_{m,n_υ} and $\beta^*_{m,n_{\upsilon-1}}$ ($\upsilon \in \{2, \ldots, u\}$) is obtained as

$$\beta^*_{m,n_{\upsilon-1}} = E_m \underbrace{\Gamma_{n_{\upsilon-1}}\left(\frac{1}{G_{m,\upsilon-1}} - \frac{1}{G_{m,\upsilon}}\right)}_{=A_{m,\upsilon}} + \underbrace{\Gamma_{n_{\upsilon-1}}\left(1 + \frac{1}{\Gamma_{n_\upsilon}}\right)}_{=B_\upsilon} \beta^*_{m,n_\upsilon}. \tag{B.5}$$

or equivalently written as

$$\begin{cases} \beta^*_{m,n_{u-1}} &= E_m A_{m,u} + B_u \beta^*_{m,n_u}, \\ \beta^*_{m,n_{u-2}} &= E_m A_{m,u-1} + B_{u-1} \beta^*_{m,n_{u-1}}, \\ & \ldots \\ \beta^*_{m,n_1} &= E_m A_{m,2} + B_2 \beta^*_{m,n_2}. \end{cases} \tag{B.6}$$

$$P_{\mathrm{out}|R_m}^{\mathrm{UMR}} = 1-\Pr\Bigg(\min_{i=1,\dots,u} g_{m,n_i}^{RD} \underbrace{\min_{i'=1,\dots,i}\left(\frac{\beta_{m,n_{i'}}^{\mathrm{F}}}{\Gamma_{n_{i'}}}-\sum_{j=i'+1}^{u}\beta_{m,n_j}^{\mathrm{F}}-\beta_{m,N+1}^{\mathrm{F}}\right)}_{=\zeta_{m,i}^{\mathrm{F}}} \geq \frac{\sigma^2}{P_r},$$

$$\min_{m'\in\bar{\mathcal{S}}_R} g_{m,m'}^{RR}\min\Bigg[\underbrace{\min_{i'=1,\dots,u}\left(\frac{\beta_{m,n_{i'}}^{\mathrm{F}}}{\Gamma_{n_{i'}}}-\sum_{j=i'+1}^{u}\beta_{m,n_j}^{\mathrm{F}}-\beta_{m,N+1}^{\mathrm{F}}\right)}_{=\zeta_{m,u}^{\mathrm{F}}},\frac{\beta_{m,N+1}^{\mathrm{F}}}{\Gamma_{N+1}}\Bigg]\geq\frac{\sigma^2}{P_r}\Bigg)$$

$$\overset{(a.1)}{=}\Pr\left(\xi_m<\frac{\sigma^2}{P_r}\right) \tag{A.2}$$

$$P_{\mathrm{out}|R_m}^{\mathrm{UR}} = 1-\Pr\Bigg(\min_{i=1,\dots,u} g_{m,n_i}^{RD} \underbrace{\min_{i'=1,\dots,i}\left(\frac{\beta_{m,n_{i'}}^{\mathrm{F}}}{\Gamma_{n_{i'}}}-\sum_{j=i'+1}^{u}\beta_{m,n_j}^{\mathrm{F}}\right)}_{=\zeta_{m,i}^{\prime\mathrm{F}}}\geq\frac{\sigma^2}{P_r}\Bigg)\overset{(a.2)}{=}\Pr\left(\varepsilon_m<\frac{\sigma^2}{P_r}\right), \tag{A.4}$$

By sequentially applying each equation in (B.6) to the next, the relation between $\beta_{m,n_{u-t}}^{*}$ $(t=1,\dots,u-1)$ and β_{m,n_u}^{*} is derived as follow

$$\beta_{m,n_{u-t}}^{*}=E_m\sum_{l=u-t+1}^{u}A_{m,l}\prod_{r=u-t+1}^{l-1}B_r+\beta_{m,n_u}^{*}\prod_{r=u-t+1}^{u}B_r \tag{B.7}$$

Further, from the first equation of (B.1) with $i=u$, it is known that

$$\beta_{m,n_u}^{*} = \frac{E_m\Gamma_{n_u}}{G_{m,u}}+\Gamma_{n_u}\beta_{m,N+1}^{*}$$
$$\overset{(b.1)}{=} E_m\left(\frac{\Gamma_{n_u}}{G_{m,u}}+\frac{\Gamma_{n_u}\Gamma_{N+1}}{G_{m,u+1}}\right) \tag{B.8}$$

where the equality (b.1) comes from (B.2). Substituting (B.8) into (B.7), we have

$$\beta_{m,n_{u-t}}^{*}=E_m\Bigg[\sum_{l=u-t+1}^{u}A_{m,l}\prod_{r=u-t+1}^{l-1}B_r + \left(\frac{\Gamma_{n_u}}{G_{m,u}}+\frac{\Gamma_{n_u}\Gamma_{N+1}}{G_{m,u+1}}\right)\prod_{r=u-t+1}^{u}B_r\Bigg], \tag{B.9}$$

where $t=1,\dots,u-1$.

Combining (B.2), (B.8) and (B.9) with $\sum_{i=1}^{u}\beta_{m,n_i}^{*}+\beta_{m,N+1}^{*}=1$, the maximal objective value E_m can be expressed as shown in (32). Further, applying (32) into (B.2), (B.8) and (B.9), the optimal solution to relaxed PA problem (30) is derived into closed form, as shown in (31). This completes the proof.

Appendix C
Proof of Theorem 2

As seen from (26) that, when $R_m(m\in\mathcal{S}_R)$ is selected in UMR case, the conditional outage probability can be written as

$$P_{\mathrm{out}|R_m}^{\mathrm{UMR}}=\Pr\left(\min_{i=1,\dots,u+1}\phi_{m,i}G_{m,i}<\frac{\sigma^2}{P_r}\right) \tag{C.1}$$

On the other hand, it is known from (B.1) that, when DPA coefficients in (31) are employed, we have $\phi_{m,i}G_{m,i}=E_m$ for $i=1,\dots,u+1$, indicating $\min_{i=1,\dots,u+1}G_{m,i}\phi_{m,i}=E_m$ holds. Consequently, we further have

$$P_{\mathrm{out}|R_m}^{\mathrm{UMR}}=\Pr\left(E_m<\frac{\sigma^2}{P_r}\right) \tag{C.2}$$

From this expression, we see that $P_{\mathrm{out}|R_m}^{\mathrm{UMR}}$ monotonically decreases with E_m. Therefore, the successful multicast user $R_{m^\dagger}$ that achieves minimal outage probability should be selected as $m^\dagger=\arg\max_{m\in\mathcal{S}_R}E_m$.

Similarly, when the DPA coefficients in (40) are employed in UR case, the outage probability conditioned on $R_m(m\in\mathcal{S}_R)$ being selected can be expressed as

$$P_{\mathrm{out}|R_m}^{\mathrm{UR}}=\Pr\left(\min_{i=1,\dots,u}\phi'_{m,i}G'_{m,i}<\frac{\sigma^2}{P_r}\right)$$
$$=\Pr\left(E'_m<\frac{\sigma^2}{P_r}\right) \tag{C.3}$$

showing that $P_{\mathrm{out}|R_m}^{\mathrm{UR}}$ monotonically decreases with E'_m. Therefore, the successful multicast user that minimizes outage probability should be selected as $m^\dagger=\arg\max_{m\in\mathcal{S}_R}E'_m$. This completes the proof.

Appendix D
Proof of Lemma 5

As known from (C.2), when DPA coefficients in (31) are employed, the outage probability conditioned on $R_m(m\in\mathcal{S}_R)$ being selected can be expressed as $P_{\mathrm{out}|R_m}^{\mathrm{UMR}}=\Pr(E_m<\frac{\sigma^2}{P_r})$. Therefore, when D-BUS scheme is employed in UMR

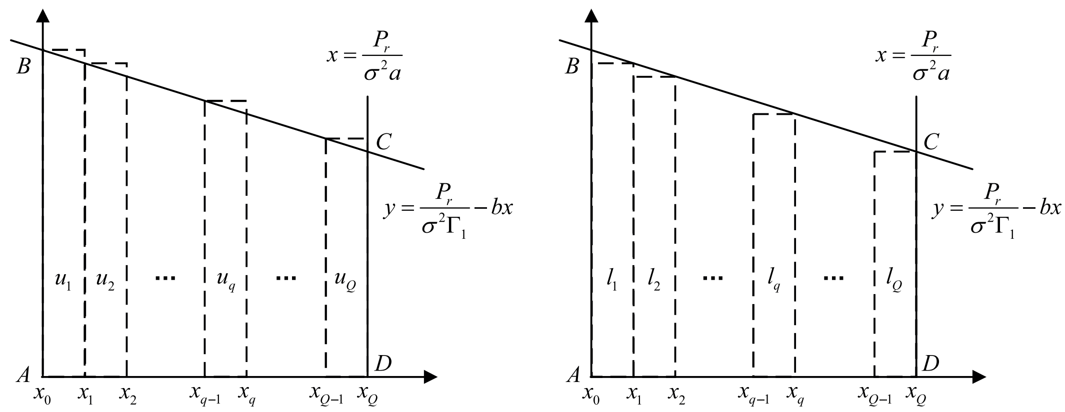


Fig. 10. Integration region of $\Omega_{m,\mathcal{A}_k}$ defined in (56) and its upper and lower bounds.

case, the conditional outage probability $P^{\text{UMR}}_{\text{out}|\mathcal{A}_k,\{1\}}$ can be expressed as

$$\begin{aligned}
&P^{\text{UMR}}_{\text{out}|\mathcal{A}_k,\{1\}}\\
&= \Pr\left(E_{m^\dagger} < \frac{\sigma^2}{P_r}\Big|\mathcal{S}_R=\mathcal{A}_k, \bar{\mathcal{S}}_D=\{1\}\right)\\
&\overset{(d.1)}{=} \Pr\Bigg[\max_{m\in\mathcal{A}_k}\Bigg(\frac{\Gamma_1}{\min(g^{RD}_{m,1}, \min_{m'\in\mathcal{M}\setminus\mathcal{A}_k} g^{RR}_{m,m'})}\\
&\qquad + \frac{\Gamma_1\Gamma_2+\Gamma_2}{\min_{m'\in\mathcal{M}\setminus\mathcal{A}_k} g^{RR}_{m,m'}}\Bigg)^{-1} < \frac{\sigma^2}{P_r}\Bigg]\\
&\overset{(d.2)}{=} \prod_{m\in\mathcal{A}_k}\Pr\Bigg(\Gamma_1\max(X_{m,\mathcal{A}_k}, Y_m)\\
&\qquad +(\Gamma_1\Gamma_2+\Gamma_2)X_{m,\mathcal{A}_k} > \frac{P_r}{\sigma^2}\Bigg), \qquad \text{(D.1)}
\end{aligned}$$

where the equality (d.1) comes from the simplified D-BUS scheme for UMR case in (54), the equality (d.2) uses variable substitutions $X_{m,\mathcal{A}_k} = 1/\min_{m'\in\mathcal{M}\setminus\mathcal{A}_k} g^{RR}_{m,m'}$ and $Y_m = 1/g^{RD}_{m,1}$. Then, based on the Total Probability Theorem [33, Sec.(3.3.8)], we further have

$$\begin{aligned}
&\Pr\left(\Gamma_1\max(X_{m,\mathcal{A}_k}, Y_m) + (\Gamma_1\Gamma_2+\Gamma_2)X_{m,\mathcal{A}_k} > \frac{P_r}{\sigma^2}\right)\\
&= \Pr\left(X_{m,\mathcal{A}_k} \ge Y_m, aX_{m,\mathcal{A}_k} > \frac{P_r}{\sigma^2}\right)\\
&\quad + \Pr\left(X_{m,\mathcal{A}_k} < Y_m, Y_m + bX_{m,\mathcal{A}_k} > \frac{P_r}{\sigma^2\Gamma_1}\right)\\
&= \int\int_{x\ge y, x>\frac{P_r}{\sigma^2 a}} f_{X_{m,\mathcal{A}_k}}(x) f_{Y_m}(y)\mathrm{d}y\mathrm{d}x\\
&\quad + \int\int_{x<y, y+bx>\frac{P_r}{\sigma^2\Gamma_1}} f_{X_{m,\mathcal{A}_k}}(x) f_{Y_m}(y)\mathrm{d}y\mathrm{d}x\\
&= 1 - \underbrace{\int_0^{\frac{P_r}{\sigma^2 a}}\int_0^{\frac{P_r}{\sigma^2\Gamma_1}-bx} f_{X_{m,\mathcal{A}_k}}(x) f_{Y_m}(y)\mathrm{d}y\mathrm{d}x}_{\Omega_{m,\mathcal{A}_k}}, \qquad \text{(D.2)}
\end{aligned}$$

where $a = \Gamma_1\Gamma_2 + \Gamma_1 + \Gamma_2$, $b = \Gamma_2 + \Gamma_2/\Gamma_1$. Substituting (D.2) into (D.1), this lemma is obtained. This completes the proof.

APPENDIX E
PROOF OF LEMMA 6

As plotted in Fig. 10, the integration region of $\Omega_{m,\mathcal{A}_k}$ is the trapezoid $ABCD$, which is upper bounded by a number Q of rectangles $u_1, ..., u_Q$ and is lower bounded by a number Q of rectangles $l_1, ..., l_Q$. Here, the bottom edges of all rectangles are same in length, i.e., $x_q - x_{q-1} = \Delta x \triangleq \frac{P_r}{Q\sigma^2 a}$ for $q = 1, ..., Q$. This observation shows that, the integral $\Omega_{m,\mathcal{A}_k}$ is upper and lower bounded as follow

$$\begin{aligned}
\Omega^{LB}_{m,\mathcal{A}_k} &\triangleq \sum_{q=1}^{Q}\int\int_{(x,y)\in l_q} f_{X_{m,\mathcal{A}_k}}(x) f_{Y_m}(y)\mathrm{d}y\mathrm{d}x < \Omega_{m,\mathcal{A}_k}\\
&< \Omega^{UB}_{m,\mathcal{A}_k} \triangleq \sum_{q=1}^{Q}\int\int_{(x,y)\in u_q} f_{X_{m,\mathcal{A}_k}}(x) f_{Y_m}(y)\mathrm{d}y\mathrm{d}x. \qquad \text{(E.1)}
\end{aligned}$$

Based on the integration region $ABCD$ of $\Omega_{m,\mathcal{A}_k}$, the regions covered by u_q and l_q can be expressed as $\{(q-1)\Delta x \le x < q\Delta x, 0 \le y < \frac{P_r}{\sigma^2\Gamma_1} - b(q-1)\Delta x\}$ and $\{(q-1)\Delta x \le x < q\Delta x, 0 \le y < \frac{P_r}{\sigma^2\Gamma_1} - bq\Delta x\}$. After some straightforward calculations, we have

$$\begin{aligned}
\Omega^{UB}_{m,\mathcal{A}_k} &= \exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r} - \frac{\sigma^2\lambda_{R_m D_1}\Gamma_1}{P_r}\right)\\
&\quad + \sum_{q=2}^{Q}\left[\exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r q}\right)\right.\\
&\quad \left. - \exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r(q-1)}\right)\right]\\
&\quad \times \exp\left(-\frac{\sigma^2\lambda_{R_m D_1}}{P_r\left(1/\Gamma_1 - [b(q-1)]/(Qa)\right)}\right), \qquad \text{(E.2)}\\
\Omega^{LB}_{m,\mathcal{A}_k} &= \exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r} - \frac{\sigma^2\lambda_{R_m D_1}}{P_r\left(1/\Gamma_1 - b/(Qa)\right)}\right)\\
&\quad + \sum_{q=2}^{Q}\left[\exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r q}\right)\right.\\
&\quad \left. - \exp\left(-\frac{\sigma^2 c_{m,\mathcal{A}_k} Qa}{P_r(q-1)}\right)\right]\\
&\quad \exp\left(-\frac{\sigma^2\lambda_{R_m D_1}}{P_r\left(1/\Gamma_1 - (bq)/(Qa)\right)}\right). \qquad \text{(E.3)}
\end{aligned}$$

With the results in (E.2) and (E.3), the integral $\Omega_{m,\mathcal{A}_k}$ can be approximated by averaging its upper and lower bounds, which yields (60). This completes the proof.

References

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

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

[3] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[4] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[5] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and A. Nallanathan, "Non-orthogonal multiple access in massive MIMO aided heterogeneous networks," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[6] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[7] J. Choi, "On the power allocation for a practical multiuser superposition scheme in NOMA systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 438–441, Mar. 2016.

[8] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[9] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—Part I: System description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[10] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—Part II: Implementation aspects and performance analysis," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1939–1948, Nov. 2003.

[11] J.-B. Kim and I.-H. Lee, "Non-orthogonal multiple access in coordinated direct and relay transmission," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 2037–2040, Nov. 2015.

[12] C. Zhong and Z. Zhang, "Non-orthogonal multiple access with cooperative full-duplex relaying," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2478–2481, Dec. 2016.

[13] J. Men and J. Ge, "Non-orthogonal multiple access for multiple-antenna relaying networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1686–1689, Oct. 2015.

[14] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Aug. 2016.

[15] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[16] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-duplex device-to-device-aided cooperative nonorthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4467–4471, May 2017.

[17] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[18] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2016.

[19] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3151–3163, Jul. 2017.

[20] H. V. Zhao and W. Su, "Cooperative wireless multicast: Performance analysis and power/location optimization," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 2088–2100, Jun. 2010.

[21] B. Niu, H. Jiang, and H. V. Zhao, "A cooperative multicast strategy in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 3136–3143, Jul. 2010.

[22] B. Hu, H. V. Zhao, and H. Jiang, "Wireless multicast using relays: Incentive mechanism and analysis," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2204–2219, Jun. 2013.

[23] I.-H. Lee, H. Lee, and H.-H. Choi, "Exact outage probability of relay selection in decode-and-forward based cooperative multicast systems," *IEEE Commun. Lett.*, vol. 17, no. 3, pp. 483–486, Mar. 2013.

[24] L. Yang, J. Chen, Y. Kuo, and H. Zhang, "Outage performance of DF-based cooperative multicast in spectrum-sharing cognitive relay networks," *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1250–1253, Jul. 2014.

[25] P. Yang, Q. Zhang, and J. Qin, "Exact outage probability of N^{th}-best multicast relay networks with co-channel interference," *IEEE Wireless Commun. Lett.*, vol. 2, no. 6, pp. 595–598, Dec. 2013.

[26] L. Yang, J. Chen, H. Zhang, H. Jiang, S. A. Vorobyov, and D. T. Ngo, "Cooperative wireless multicast: Performance analysis and time allocation," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5810–5819, Jul. 2016.

[27] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4718–4731, Jul. 2016.

[28] H. Tabassum, E. Hossain, and M. J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using Poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

[29] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2381–2401, Sep. 2003.

[30] T. Liu, C. Yang, and L. L. Yang, "A low-complexity subcarrier-power allocation scheme for frequency-division multiple-access systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1564–1570, May 2010.

[31] Y. Liang, V. V. Veeravalli, and H. V. Poor, "Resource allocation for wireless fading relay channels: Max-min solution," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3432–3453, Oct. 2007.

[32] M. Haenggi, *Stochastic Geometry for Wireless Networks*, Cambridge, U.K.: Cambridge Univ. Press, 2013.

[33] D. Zwillinger and S. Kokoska, *CRC Standard Probability and Statistics Tables and Formulae*, Boca Raton, FL, USA: CRC Press, 1999.

[34] L. Lv, J. Chen, Q. Ni and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.

[35] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

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

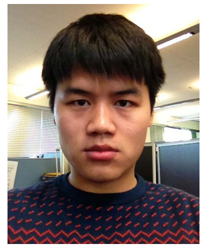
[37] M. M. El-Sayed, A. S. Ibrahim, and M. M. Khairy, "Power allocation strategies for non-orthogonal multiple access," in *Proc. MoWNeT*, Cairo, Egypt, Apr. 2016, pp. 1–6.

[38] J. A. Oviedo and H. R. Sadjadpour, "A new NOMA approach for fair power allocation," in *Proc. IEEE INFOCOM WKSHPS*, San Francisco, CA, USA, Apr. 2016, pp. 843–847.

Long Yang received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2010 and 2015, respectively. He was a visiting Ph.D. student with the University of Alberta, Canada, from 2014 to 2015, and a visiting scholar with Lancaster University, U.K., in 2016. In 2015, he joined the School of Telecommunications Engineering, Xidian University, where he is currently a Lecturer. He is also a Post-Doctoral Research Fellow with the University of Alberta, Canada. His current research interests include physical-layer security, cooperative relaying, and non-orthogonal multiple access.

Jian Chen (M'14) received the B.S. degree from Xi'an Jiaotong University, China, in 1989, the M.S. degree from Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, in 1992, and the Ph.D. degree in telecommunications engineering from Xidian University, China, in 2005. From 2007 to 2008, he was a visitor scholar with The University of Manchester. He is currently a Professor with the School of Telecommunications Engineering, Xidian University. He is also with University of Alberta, as a senior visiting scholar sponsored by the China Scholarship Council. His research interests include cognitive radio, OFDM, wireless sensor networks, and non-orthogonal multiple access.

Qiang Ni (M'04–SM'08) received the B.Sc., M.Sc., and Ph.D. degrees from the Huazhong University of Science and Technology, China, all in engineering. He is currently a Professor and the Head of Communication Systems Research Group, InfoLab21, School of Computing and Communications, Lancaster University, U.K. His research interests focus on the area of future generation communications and networking, including green communications, 5G, non-orthogonal multiple access, energy harvesting, cognitive radio, ultra-dense networks, SDN, IoTs, and vehicular networks. He has authored over 180 papers in these areas. He is a voting member of the IEEE 1932.1 standard. He was an IEEE 802.11 Wireless Standard Working Group voting member and a contributor to various IEEE Wireless Standards.



Jia Shi received the M.Sc. and Ph.D. degrees from the University of Southampton, U.K., in 2010 and 2015, respectively. He was a Research Associate with Lancaster University, U.K., from 2015 to 2017. Since 2017, He has been a Research Fellow with the University of Surrey, U.K. His current research interests include mmwave communications, non-orthogonal multiple access, resource allocation in multicarrier systems, and cooperative communications.

Xuan Xue received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2010 and 2017, respectively. From 2013 to 2015, she was a visiting Ph.D. student with Western University, Canada. Her research interests include massive MIMO, millimeterwave systems, and cooperative communications.