Cooperative Non-Orthogonal Layered Multicast Multiple Access for Heterogeneous Networks

Long Yang, *Member, IEEE*, Qiang Ni, *Senior Member, IEEE*, Lu Lv, Jian Chen, *Member, IEEE*, Xuan Xue, *Member, IEEE*, Hailin Zhang, *Member, IEEE*, and Hai Jiang, *Senior Member, IEEE*

***Abstract*—This paper proposes a novel design of cooperative non-orthogonal layered multicast multiple access in a heterogeneous network, where the information is encoded into the messages of high-priority (HP) and low-priority (LP). Two types of multicast users coexist in the network: 1) regular users (RUs), which are located far away from the base-station (BS) and expect to decode only the HP message (due to the weak channels); 2) advanced users (AUs), which are located close to the BS and expect to decode both HP and LP messages. To improve the reliability of layered multicast, we consider that the successful AUs (those AUs who successfully decode the HP and LP messages) serve as potential relays to assist other AUs/RUs. Based on this idea, two novel cooperation strategies are proposed for different cases of channel information availability. For each proposed strategy, we derive closed-form exact outage probabilities of AUs and RUs, and then further analyze their diversity orders. Moreover, considering that the layered multicast is outage-constrained, we theoretically evaluate the energy consumption of both strategies and demonstrate their energy saving gains over the direct non-orthogonal multiple access for layered multicast. Finally, our theoretical analysis is verified by numerical results, and the advantages of the proposed strategies are also demonstrated.**

***Index Terms*—Cooperative non-orthogonal multiple access, energy saving, layered multicast, outage probability.**

Manuscript received January 2, 2018; revised May 8, 2018 and August 16, 2018; accepted September 25, 2018. The associate editor coordinating the review of this paper and approving it for publication was Prof. A. Nallanathan. Part of the results in this paper has been accepted to be presented in IEEE GLOBECOM 2018 [1]. This work was supported in part by the National Natural Science Foundation of China under Grants 61601347 and 61771366, in part by the Natural Science Basic Research Plan in the Shaanxi Province of China under Grant 2017JQ6055, in part by the Royal Society project under Grant IEC/NSFC/170324, in part by the EPSRC IAA project under Grant CSA7113, in part by the EPSRC project under Grant EP/K011693/1, in part by the EU FP7 CROWN project under Grant PIRSES-GA-2013-610524, in part by Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2018-06307, in part by the International Postdoctoral Exchange Fellowship Program 2017 from the Office of China Postdoctoral Council, and in part by the 111 Project of China under Grant B08038. (*Corresponding Authors: Qiang Ni and Jian Chen*)

L. Yang is with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China, and also with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada (e-mail: lyang@xidian.edu.cn).

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

L. Lv, J. Chen, X. Xue and H. Zhang are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: lulv@stu.xidian.edu.cn; jianchen@mail.xidian.edu.cn; {xuanx, hlzhang}@xidian.edu.cn).

H. Jiang is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada (e-mail: hai1@ualberta.ca).

I. Introduction

The application of non-orthogonal multiple access (NOMA) to the future wireless networks has drawn increased attention from both industrial and academic communities [2]–[6]. By employing superposition coding and successive interference cancellation (SIC) at the sender and receiver sides, respectively, NOMA is able to serve multiple users at the same time/frequency/code domain with distinct power levels. Compared with the conventional orthogonal multiple access (OMA), NOMA can provide better fairness and connectivity for users with poor channel quality, thus being viewed as a potential technique for the fifth generation (5G) wireless networks [3]–[11].

To improve the reliability/capacity of NOMA systems, cooperative relaying technique has been incorporated into NOMA, termed as *cooperative NOMA*, in which the information is forwarded by a successful user (i.e., a user that successfully decodes the information) [12]–[14] or by a dedicated relay [15]–[19]. As the SIC is employed in NOMA, a user with strong channel condition may correctly decode its own message along with other users' messages, thus being able to naturally serve as a relay for other users. Compared to deploying dedicated relays in NOMA systems [15]–[19], recruiting successful users as relays [12]–[14] can save the infrastructure cost for dedicated relays. When NOMA is performed between two users, it is shown in [12] that the user-assisted cooperative NOMA can be realized by recruiting the near user (the one which is closer to the source) to serve as a half-duplex relay for the far user (the one which is further away from the source). Further, if the near user can work as a full-duplex relay, the far user will achieve better performance in terms of outage probability and ergodic rate, as demonstrated in [13] and [14].

The aforementioned efforts on cooperative NOMA consider only *wireless unicast*, i.e., point-to-point data delivery. In practical wireless networks, some users may need the same data in many scenarios, e.g., in Internet protocol television and live streaming of sports games. If cooperative NOMA unicast is employed, the same data will be sent repeatedly, resulting in low spectrum/energy efficiency. Thanks to the broadcast feature of wireless channels, *wireless multicast* can simultaneously send the same data to multiple users, thus being an efficient manner to serve the users with common interest [20]–[22]. When some users want the same data in NOMA systems, integrating wireless multicast into cooperative NOMA, i.e., *cooperative NOMA multicast*, is expected to combine their advantages. Motivated by this fact, two opportunistic cooperative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2874239, IEEE Transactions on Communications

NOMA multicast strategies are designed in most recent work [23] and [24], which opportunistically recruit a successful multicast user that helps forward its received messages to unsuccessful multicast/unicast users. It has been theoretically proved in [23] and [24] that, by appropriately selecting the successful multicast user, full diversity can be achieved at each multicast/unicast user.

However, the cooperative NOMA multicast strategies proposed in [23] and [24] have only focused on the *non-layered multicast*, which does not consider the *heterogeneity* among multicast users. In fact, as multicast users own heterogeneous and time-varying channel conditions, the non-layered multicast which employs fixed data-rate and coding scheme cannot best serve all multicast users simultaneously, especially in video multicast that requires seamless connectivity and low latency [25]. To cope with this problem, *layered multicast* is designed to deliver the same video content to multicast users with different video resolutions. In layered multicast, the original video information is split into a base-layer stream and some enhancement-layer streams, where the base-layer stream provides a basic video quality level and each enhancement-layer stream can further refine video resolution. Consequently, each multicast user can adaptively decode its received layered streams, according to its reception quality of each layered stream. However, the current layered-multicast mechanism is mainly performed over the application layer without cooperative support at the physical layer, thereby limiting the performance of layered multicast in wireless systems. The application of NOMA in layered multicast is investigated in most recent works [26]–[28], showing that the spectrum efficiency and coverage probability of layered multicast can be significantly improved by using NOMA.

In this paper, we propose a novel design called *cooperative non-orthogonal layered multicast multiple access*, which can significantly improve both spectrum efficiency and reliability. More precisely, by combining NOMA with layered multicast, the source can send both base-layer and enhancement-layer streams with the same link, thus achieving high spectrum efficiency and link utilization. Further, by recruiting the users that successfully decode the multicast messages to serve as relays, great spatial diversity offered by cooperative NOMA can be harvested to improve the reliability/connectivity of layered multicast. Therefore, compared to the cooperative NOMA strategies for non-layered multicast in [23] and [24], the design of cooperative NOMA strategy for layered multicast is a more challenging and practically useful issue for video multicast in 5G systems. Specifically, in this novel design, the video information is encoded into data streams of high-priority (HP) and low-priority (LP), corresponding to the base-layer and enhancement-layer streams in the digital video broadcasting (DVB) [29]. Two types of multicast users coexist in this network: 1) regular users (RUs), which are located far away from the base station (BS) and expect to decode only the HP data stream (due to the weak channels) for basic quality-of-service (QoS), and 2) advanced users (AUs), which are located close to the BS and expect to decode both the HP and LP data streams for better QoS. To improve the reception quality of layered multicast, successful AUs (those AUs who successfully decode the HP and LP data streams) can serve as available relays to assist other users.

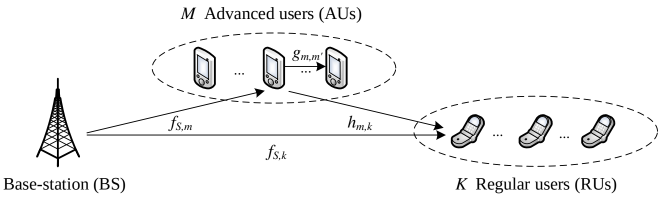


Fig. 1. The considered NOMA network for layered multicast.

The main contributions of this work can be summarized as follows.

- To enhance reliability of layered multicast in a downlink NOMA network, we propose two user-assisted cooperative NOMA multicast strategies, namely, the distributed cooperative NOMA (DC-NOMA) multicast and the opportunistic cooperative NOMA (OC-NOMA) multicast, for different cases of channel state information (CSI) availability.
- For each proposed strategy, the outage probabilities of AUs and RUs are derived in closed form and further asymptotically analyzed in high signal-to-noise ratio (S-NR) regime. The derived theoretical results demonstrate that, the DC-NOMA multicast strategy achieves a diversity order of two, while the OC-NOMA multicast strategy achieves a diversity order not less than M (the number of AUs).
- Considering that the layered multicast is outage-constrained, the energy consumption of proposed cooperation strategies are also theoretically evaluated. It is shown that the energy saving gains of DC-NOMA multicast and OC-NOMA multicast strategies scale with factors $p_o^{-1/2}$ and $p_o^{-(M-1)/M}$ as $p_o \to 0$, respectively, where p_o is the threshold of outage probability. This fact indicates that significant energy saving can be achieved by both proposed strategies.

The rest of this paper is organized as follows. Section II presents the considered system model. Section III discusses the direct NOMA multicast strategy as a benchmark. Sections IV and V detail the proposed DC-NOMA multicast and OC-NOMA multicast strategies, respectively, and derive the theoretical expressions for their performance metrics, including outage probability, diversity order, energy consumption and energy saving gains. Section VI provides simulation results, followed by Section VII that is conclusion of this paper.

II. System Model

Consider the multicast transmission from a BS denoted by S to a group of AUs $A_1, ..., A_M$ ($M \geq 2$) and a group of RUs $R_1, ..., R_K$ ($K \geq 2$), as depicted in Fig. 1. The group of AUs are close to the BS and the group of RUs are far away from the BS. Let x_H and x_L denote the messages of HP and LP, respectively, corresponding to the multi-resolution DVB services [29]. By using NOMA, the BS superimposes both HP and LP messages and then sends the superposition to AUs/RUs. Following the principles of layered multicast [25], it is assumed that the RUs expect to decode only HP message for basic QoS and the AUs expect to decode both HP and LP messages for better QoS, as the channel conditions of AUs are better than those of RUs.[1] It is also assumed that more power is allocated to x_H while the rest of power is allocated to x_L, since message x_H is given high priority [26], [27]. According to the principles of SIC-based detection [32], the information detection should be performed from the message with more power to the message with less power. As more power is allocated to HP message, each AU first detects HP message x_H by treating LP message x_L as interference, and then, performs SIC to further detect LP message x_L. For each RU, only HP message x_H needs to be detected by considering LP message x_L as interference. Throughout this paper, we declare an AU is successful if it has correctly decoded both x_H and x_L, and declare an RU is successful if it has correctly decoded x_H. Otherwise, the AU/RU is declared to be unsuccessful. As the purpose of multicast is to deliver the same information to all users, an outage event is declared for AUs if any AU is eventually unsuccessful while an outage event is declared for RUs if any RU is eventually unsuccessful.

The BS, AUs, and RUs all have a single antenna, operating under a half-duplex fashion. Denote the channel coefficients pertaining to links $S-A_m$, $S-R_k$, A_m-R_k and $A_m-A_{m'}$ as $f_{S,m}$, $f_{S,k}$, $h_{m,k}$ and $g_{m,m'}$, respectively. Assume that, all links experience independent but nonidentically distributed Rayleigh fading, and thus, the channel coefficients follow circularly symmetric complex Gaussian distribution as $f_{S,m} \sim \mathcal{CN}(0,\Omega^f_{S,m})$, $f_{S,k} \sim \mathcal{CN}(0,\Omega^f_{S,k})$, $h_{m,k} \sim \mathcal{CN}(0,\Omega^h_{m,k})$, $g_{m,m'} \sim \mathcal{CN}(0,\Omega^g_{m,m'})$, where $\Omega^f_{S,m} \triangleq \mathbb{E}[|f_{S,m}|^2]$, $\Omega^f_{S,k} \triangleq \mathbb{E}[|f_{S,k}|^2]$, $\Omega^h_{m,k} \triangleq \mathbb{E}[|h_{m,k}|^2]$, and $\Omega^g_{m,m'} \triangleq \mathbb{E}[|g_{m,m'}|^2]$. Here, $\mathbb{E}[\cdot]$ represents the mathematical expectation. Channel reciprocity is also assumed. A block-fading model is considered, and thus, the channel coefficients keep unchanged within every transmission block, but vary independently among different transmission blocks. The duration of each transmission block is T_0. The noise at each user is modeled as additive white Gaussian noise (AWGN) with the identical variance σ^2. For each transmission block, the target information rate for x_H and x_L are denoted as r_H and r_L, respectively.

III. Direct NOMA Multicast

To provide a benchmark for the proposed cooperation strategies, this section presents the direct NOMA multicast strategy that does not employ any cooperation.

By employing NOMA, the BS sends superimposed signal $\sqrt{P_S\alpha_H}x_H+\sqrt{P_S\alpha_L}x_L$ to all users, where P_S is the transmit power of BS, α_H and α_L are the power allocation coefficients for messages x_H and x_L, respectively, with $\alpha_H+\alpha_L=1$ and $\alpha_H>\alpha_L$. Further, to ensure that NOMA can be realized, the power allocation coefficients also satisfy $\alpha_H-\alpha_L(2^{r_H}-1)>0$ [12].

Once receiving the superimposed signals, all AUs successively detect messages x_H and x_L by using SIC. For AU A_m, its received signal can be expressed as $y_{S\to A_m} = \sqrt{P_S\alpha_H}f_{S,m}x_H+\sqrt{P_S\alpha_L}f_{S,m}x_L+n_m$, where n_m represents the AWGN at AU A_m. Defining $\rho \triangleq P_S/\sigma^2$ as the *transmit SNR*, the signal-to-interference-plus-noise ratio (SINR) for AU A_m to decode x_H can be expressed as

$$\gamma_{S\to A_m,H} = \frac{\alpha_H|f_{S,m}|^2}{\alpha_L|f_{S,m}|^2+\rho^{-1}}. \tag{1}$$

If AU A_m correctly decodes x_H, it performs SIC to remove x_H from its observation, and then, decodes x_L with the SNR being

$$\gamma_{S\to A_m,L} = \rho\alpha_L|f_{S,m}|^2. \tag{2}$$

Recalling that an AU is successful if it correctly decodes both messages x_H and x_L, the condition for AU A_m being successful is given by $\{\log(1+\gamma_{S\to A_m,H}) \ge r_H,\ \log(1+\gamma_{S\to A_m,L}) \ge r_L\}$.

On the other hand, all RUs only decode message x_H from their received signals. The received signal at RU R_k can be expressed as $y_{S\to R_k} = \sqrt{P_S\alpha_H}f_{S,k}x_H+\sqrt{P_S\alpha_L}f_{S,k}x_L+n_k$, where n_k represents the AWGN at RU R_k. By treating LP message x_L as interference, RU R_k decodes the HP message x_H with the following SINR

$$\gamma_{S\to R_k,H} = \frac{\alpha_H|f_{S,k}|^2}{\alpha_L|f_{S,k}|^2+\rho^{-1}}. \tag{3}$$

As RUs only need to decode the message x_H, the condition for RU R_k being successful is given by $\{\log(1+\gamma_{S\to R_k,H}) \ge r_H\}$.

As an outage is declared for AUs if any AU is unsuccessful, the complementary event of AU outage happening is that all AUs are successful. Therefore, using (1) and (2), the outage probability of AUs can be expressed as

$$\begin{aligned} P^{\text{AU}}_{\text{out}} =& 1-\Pr\Bigg(\bigcap_{m=1}^{M}\Big\{\gamma_{S\to A_m,H} \ge \tau_H \triangleq 2^{r_H}-1,\\ &\qquad\qquad \gamma_{S\to A_m,L} \ge \tau_L \triangleq 2^{r_L}-1\Big\}\Bigg)\\ =& 1-\prod_{m=1}^{M}\Pr\left(|f_{S,m}|^2 \ge \max(\varepsilon_1,\varepsilon_2)/\rho\right)\\ =& 1-e^{-\frac{\max(\varepsilon_1,\varepsilon_2)}{\rho}\sum_{m=1}^{M}1/\Omega^f_{S,m}}, \end{aligned} \tag{4}$$

[1] In this paper, two priority levels are considered for the NOMA layered multicast. The reason is twofold. First, as pointed by [30], it may not be practical to superimpose more than two messages in NOMA systems because of the possible strong co-channel interference. As an effective alternative, superimposing two messages in power domain is a simple but effective way to realize NOMA in practical systems. Second, the layered multicast with two priority levels has been accepted by the current multimedia multicast standards, such as DVB [29] and ATSC 3.0 [31]. Therefore, considering two priority levels for NOMA layered multicast is practically meaningful.
Note that, although only two priority levels are considered in this paper, the idea of our proposed cooperative NOMA for layered multicast can also be extended to layered multicast with multiple priority levels. For example, one HP message is intended by all users and $N(\ge 2)$ LP messages are intended by only AUs. The extension can be outlined as follows. In each transmission block, the BS superimposes all messages with proper power allocation, and then, sends the superimposed messages to all AUs/RUs. Once receiving the superimposed messages from the BS, all AUs decode each message by using SIC, while all RUs only decode the HP message by treating all N LP messages as interference. Then, some successful AUs (which have correctly decoded all messages) can be recruited to serve as relays, to forward the HP message and all LP messages to unsuccessful AUs/RUs by using NOMA. The performance analysis with multiple priority levels is similar to those in Section IV and Section V in this paper.

where $\Pr(\cdot)$ means probability, $\varepsilon_1 \triangleq (\alpha_H/\tau_H - \alpha_L)^{-1}$, and $\varepsilon_2 \triangleq \tau_L/\alpha_L$. Following the same rationale, the outage probability of RUs can be derived as

$$\begin{aligned}P_{\text{out}}^{\text{RU}} &= 1 - \Pr\left(\bigcap_{k=1}^{K}\{\gamma_{S\to R_k,H} \geq \tau_H\}\right)\\ &= 1 - \prod_{k=1}^{K}\Pr\left(|f_{S,k}|^2 \geq \varepsilon_1/\rho\right)\\ &= 1 - e^{-\frac{\varepsilon_1}{\rho}\sum_{k=1}^{K} 1/\Omega_{S,k}^{f}}. \end{aligned} \tag{5}$$

Based on the derived outage probability above, the diversity order achieved by the direct NOMA multicast strategy is demonstrated in the following theorem.

Theorem 1: In the direct NOMA multicast, both AUs and RUs achieve a unit diversity order.

Proof: Based on the series representation of exponential function [33, eq. (1.211.1)], we have $e^{-c/\rho} \simeq 1 - c/\rho$ holds for $\rho \to \infty$, where c is a positive constant. Therefore, using (4) and (5) with letting $\rho \to \infty$, we have

$$P_{\text{out}}^{\text{AU}} \overset{\rho\to\infty}{\simeq} \frac{\max(\varepsilon_1,\varepsilon_2)}{\rho}\sum_{m=1}^{M} 1/\Omega_{S,m}^{f} \propto \rho^{-1}, \tag{6}$$

$$P_{\text{out}}^{\text{RU}} \overset{\rho\to\infty}{\simeq} \frac{\varepsilon_1}{\rho}\sum_{k=1}^{K} 1/\Omega_{S,k}^{f} \propto \rho^{-1}. \tag{7}$$

As the diversity order is defined as $\lim_{\rho\to\infty} -\frac{\log P_{\text{out}}(\rho)}{\log \rho}$ [34], we know from (6) and (7) that a unit diversity order is achieved by both AUs and RUs. ■

Next, the energy consumption of direct NOMA multicast strategy is shown in the following theorem. Following [20], the energy consumption refers to the energy consumed by transmit power in each transmission block, termed as *energy consumption per block (ECPB)*. More precisely, for the direct NOMA multicast strategy, the ECPB can be expressed as

$$E^{\text{direct}} = P_S T_0 = \rho\sigma^2 T_0, \tag{8}$$

where the second equality uses the fact $\rho = P_S/\sigma^2$.

Theorem 2: Given a target outage probability for AUs denoted as p_o and a target outage probability for RUs denoted as βp_o ($\beta > 0$), the minimal ECPB of direct NOMA multicast that guarantees $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ is given by

$$E_{\text{min}}^{\text{direct}} = \Phi\sigma^2 T_0, \tag{9}$$

with Φ being defined as

$$\begin{aligned}\Phi \triangleq \max\Bigg(&\frac{\max(\varepsilon_1,\varepsilon_2)}{\ln(\frac{1}{1-p_o})}\sum_{m=1}^{M}\frac{1}{\Omega_{S,m}^{f}},\\ &\frac{\varepsilon_1}{\ln(\frac{1}{1-\beta p_o})}\sum_{k=1}^{K}\frac{1}{\Omega_{S,k}^{f}}\Bigg). \end{aligned} \tag{10}$$

Further, if the multicast service is highly reliability-sensitive, i.e., $p_o \to 0$, the minimal ECPB that achieves $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ can be asymptotically expressed as

$$E_{\text{min}}^{\text{direct}} \simeq \frac{\Psi\sigma^2 T_0}{p_o}, \tag{11}$$

where the term Ψ is defined as

$$\Psi \triangleq \max\left(\max(\varepsilon_1,\varepsilon_2)\sum_{m=1}^{M}\frac{1}{\Omega_{S,m}^{f}}, \frac{\varepsilon_1}{\beta}\sum_{k=1}^{K}\frac{1}{\Omega_{S,k}^{f}}\right).$$

Proof: By using (4), the inequality $P_{\text{out}}^{\text{AU}} \leq p_o$ can be equivalently expressed as

$$1 - p_o \leq e^{-\frac{\max(\varepsilon_1,\varepsilon_2)}{\rho}\sum_{m=1}^{M} 1/\Omega_{S,m}^{f}}. \tag{12}$$

Taking logarithm on both sides of (12) and then using some algebraic manipulations, we further have $\rho \geq \frac{\max(\varepsilon_1,\varepsilon_2)}{\ln(\frac{1}{1-p_o})}\sum_{m=1}^{M}\frac{1}{\Omega_{S,m}^{f}}$.

From this inequality we see that the minimal transmit SNR that ensures $P_{\text{out}}^{\text{AU}} \leq p_o$ can be obtained as

$$\rho_{\text{min}}^{\text{AU}} = \frac{\max(\varepsilon_1,\varepsilon_2)}{\ln(\frac{1}{1-p_o})}\sum_{m=1}^{M}\frac{1}{\Omega_{S,m}^{f}}. \tag{13}$$

Applying the same rationale into (5), the minimal transmit SNR that ensures $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ is

$$\rho_{\text{min}}^{\text{RU}} = \frac{\varepsilon_1}{\ln(\frac{1}{1-\beta p_o})}\sum_{k=1}^{K}\frac{1}{\Omega_{S,k}^{f}}. \tag{14}$$

Apparently, to ensure that both $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ hold, the minimal transmit SNR should be $\max(\rho_{\text{min}}^{\text{AU}}, \rho_{\text{min}}^{\text{RU}})$. Therefore, combining (13) and (14), the minimal transmit SNR is obtained as

$$\begin{aligned}&\rho_{\text{min}}^{\text{direct}} = \\ &\underbrace{\max\left(\frac{\max(\varepsilon_1,\varepsilon_2)}{\ln(\frac{1}{1-p_o})}\sum_{m=1}^{M}\frac{1}{\Omega_{S,m}^{f}},\ \frac{\varepsilon_1}{\ln(\frac{1}{1-\beta p_o})}\sum_{k=1}^{K}\frac{1}{\Omega_{S,k}^{f}}\right)}_{=\Phi}. \end{aligned} \tag{15}$$

Then, applying (15) into (8), the minimal ECPB that ensures both $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ hold is derived as $E_{\text{min}}^{\text{direct}} = \Phi\sigma^2 T_0$, as shown in (9).

Moreover, it is known from [33, eq.(1.513.4)] that $\ln(\frac{1}{1-p_o}) \simeq p_o$ and $\ln(\frac{1}{1-\beta p_o}) \simeq \beta p_o$ hold for $p_o \to 0$. Applying this fact into (10), we have $\Phi \simeq \Psi/p_o$ for $p_o \to 0$, indicating that $E_{\text{min}}^{\text{direct}} \simeq \frac{\Psi\sigma^2 T_0}{p_o}$ holds for $p_o \to 0$, as shown in (11). This completes the proof. ■

As seen from the direct NOMA multicast strategy, both HP and LP messages have been correctly decoded by successful AUs. Thus, we can recruit the successful AUs to help those unsuccessful AUs/RUs, in order to improve the reliability of NOMA multicast by achieving cooperative diversity. Motivated by this fact, we propose two cooperative NOMA multicast strategies that recruit successful AU(s) to serve as *relay(s)* in the following Sections IV and V. For operational simplicity, we consider that each participating relay uses the same power allocation coefficients as those used by the BS. Note that extension to different power allocation coefficients at the relay(s) is straightforward. Moreover, each unsuccessful AU/RU is assumed to detect the forwarded signals without combining with the signals from the BS. As combining signals and optimizing power allocation can improve the reception quality, the results in this paper provide a lower-bound performance of the system.

IV. Distributed Cooperative NOMA Multicast

To enhance the reliability of NOMA multicast, this section proposes a distributed cooperative NOMA strategy for layered multicast (termed as *DC-NOMA multicast*), which does not require the CSI from successful AUs to unsuccessful AUs/RUs. Outage performance, diversity order, and energy consumption of the proposed DC-NOMA multicast are also derived in this section.

A. Strategy Description

The proposed DC-NOMA multicast strategy is performed in two phases, where the duration of each phase is $T_0/2$. During the first phase, the BS sends the superimposed signal $\sqrt{P_S\alpha_H}x_H + \sqrt{P_S\alpha_L}x_L$ to all users, which is same as the direct NOMA multicast. Similarly, the power allocation coefficients satisfy $\alpha_H+\alpha_L=1$, $\alpha_H>\alpha_L$ and $\alpha_H-\alpha_L(2^{2r_H}-1)>0$ [12]. At the end of the first phase, the condition for A_m being successful is $\{\frac{1}{2}\log(1+\gamma_{S\to A_m,H}) \ge r_H, \frac{1}{2}\log(1+\gamma_{S\to A_m,L})\ge r_L\}$, while the condition for R_k being successful is $\{\frac{1}{2}\log(1+\gamma_{S\to R_k,H})\ge r_H\}$. Accordingly, the indices sets of successful AUs and RUs can be defined as $\mathcal{S}_A \triangleq \{m|\gamma_{S\to A_m,H}\ge\tilde{\tau}_H\triangleq 2^{2r_H}-1, \gamma_{S\to A_m,L}\ge\tilde{\tau}_L\triangleq 2^{2r_L}-1\}$ and $\mathcal{S}_R\triangleq\{k|\gamma_{S\to R_k,H}\ge\tilde{\tau}_H\}$, respectively. Here, the expressions for terms $\gamma_{S\to A_m,H}$, $\gamma_{S\to A_m,L}$ and $\gamma_{S\to R_k,H}$ have been given in (1), (2), and (3), respectively.

During the second phase, all successful AUs simultaneously forward the superimposed signal $\sqrt{P_A\alpha_H}\,x_H+\sqrt{P_A\alpha_L}x_L$ to the unsuccessful AUs/RUs[2], where $P_A\triangleq\mu P_S(\mu>0)$ is the transmit power at each AU. For unsuccessful AU A_m ($m\in\{1,...,M\}\backslash\mathcal{S}_A$), it observes $y_{\mathcal{S}_A\to A_m}=\sum_{l\in\mathcal{S}_A}\left(\sqrt{P_A\alpha_H}x_H +\sqrt{P_A\alpha_L}x_L\right)g_{l,m}+n_m$. Defining $X_{\mathcal{S}_A,m}\triangleq|\sum_{l\in\mathcal{S}_A}g_{l,m}|^2$, the SINR for A_m to decode x_H is given by

$$\gamma_{\mathcal{S}_A\to A_m,H}=\frac{\alpha_H X_{\mathcal{S}_A,m}}{\alpha_L X_{\mathcal{S}_A,m}+(\mu\rho)^{-1}}. \tag{16}$$

If A_m succeeds in decoding x_H, it removes x_H by SIC and then further detects x_L with SNR being

$$\gamma_{\mathcal{S}_A\to A_m,L}=\mu\rho\alpha_L X_{\mathcal{S}_A,m}. \tag{17}$$

Similarly, at unsuccessful RU R_k ($k\in\{1,...,K\}\setminus\mathcal{S}_R$), the received signal is $y_{\mathcal{S}_A\to R_k}=\sum_{l\in\mathcal{S}_A}\left(\sqrt{P_A\alpha_H}\;x_H +\sqrt{P_A\alpha_L}x_L\right)h_{l,k}+n_k$. Defining $Y_{\mathcal{S}_A,k}\triangleq|\sum_{l\in\mathcal{S}_A}h_{l,k}|^2$, the SINR for RU R_k to decode message x_H is given by

$$\gamma_{\mathcal{S}_A\to R_k,H}=\frac{\alpha_H Y_{\mathcal{S}_A,k}}{\alpha_L Y_{\mathcal{S}_A,k}+(\mu\rho)^{-1}}. \tag{18}$$

B. Outage Analysis

1) Decoding Results after the First Phase: Recall that the condition for A_m being successful after the first phase is $\{\frac{1}{2}\log(1+\gamma_{S\to A_m,H})\ge r_H, \frac{1}{2}\log(1+\gamma_{S\to A_m,L})\ge r_L\}$. Thus, after the first phase, the probability that AU A_m is successful can be expressed as

$$\begin{aligned}\Pr(m\in\mathcal{S}_A)=&\Pr\left(\gamma_{S\to A_m,H}\ge\tilde{\tau}_H,\gamma_{S\to A_m,L}\ge\tilde{\tau}_L\right)\\ =&\Pr\left(|f_{S,m}|^2\ge\max(\varphi_1,\varphi_2)/\rho\right)\\ =&\underbrace{e^{-\max(\varphi_1,\varphi_2)/(\rho\Omega^f_{S,m})}}_{\triangleq\theta_m},\end{aligned} \tag{19}$$

where $\varphi_1\triangleq(\alpha_H/\tilde{\tau}_H-\alpha_L)^{-1}$ and $\varphi_2\triangleq\tilde{\tau}_L/\alpha_L$. Since the channel gains $|f_{S,m}|^2, m=1,...,M$, are independently distributed, the probability for condition $\{\mathcal{S}_A=\mathcal{A}\}$ can thus be expressed as

$$\Pr(\mathcal{S}_A=\mathcal{A})=\prod_{m\in\mathcal{A}}\theta_m\prod_{m\in\{1,...,M\}\backslash\mathcal{A}}(1-\theta_m), \tag{20}$$

where $\mathcal{A}$ is a subset of $\{1,...,M\}$. On the other hand, as the condition for R_k being successful after the first phase is $\{\frac{1}{2}\log(1+\gamma_{S\to R_k,H})\ge r_H\}$, the probability for RU R_k being successful can be expressed as

$$\begin{aligned}\Pr(k\in\mathcal{S}_R)=&\Pr\left(\gamma_{S\to R_k,H}\ge\tilde{\tau}_H\right)\\ =&\Pr\left(|f_{S,k}|^2\ge\varphi_1/\rho\right)\\ =&\underbrace{e^{-\varphi_1/(\rho\Omega^f_{S,k})}}_{\triangleq\zeta_k}.\end{aligned} \tag{21}$$

Let $\mathcal{B}$ represent a subset of $\{1,...,K\}$. Using the independence among channel gains $|f_{S,k}|^2, k=1,...,K$, the probability for $\{\mathcal{S}_R=\mathcal{B}\}$ is expressed as

$$\Pr(\mathcal{S}_R=\mathcal{B})=\prod_{k\in\mathcal{B}}\zeta_k\prod_{k\in\{1,...,K\}\backslash\mathcal{B}}(1-\zeta_k). \tag{22}$$

2) Exact Outage Probability: Since an outage is declared for AUs when any AU is not eventually successful, the AUs experience an outage when either one of the following outage events happens:

- Outage event $\mathbf{O}_1^{\mathrm{AU}}\triangleq\{\mathcal{S}_A=\emptyset\}$, which means that all AUs are unsuccessful after the first phase so that no AU is able to serve as a relay in the second phase;
- Outage event $\mathbf{O}_2^{\mathrm{AU}}\triangleq\{1\le|\mathcal{S}_A|\le M-1, \cup_{m\in\{1,...,M\}\backslash\mathcal{S}_A}\mathbf{o}_{2,m}^{\mathrm{AU}}\}$ with $\mathbf{o}_{2,m}^{\mathrm{AU}}\triangleq\{\gamma_{\mathcal{S}_A\to A_m,H}<\tilde{\tau}_H\}$ $\cup\{\gamma_{\mathcal{S}_A\to A_m,L}<\tilde{\tau}_L\}$, meaning that some AUs remain unsuccessful after receiving the forwarded signals from successful AUs. Here $|\mathcal{S}_A|$ means cardinality of set $\mathcal{S}_A$.

The probabilities for outage events $\mathbf{O}_1^{\mathrm{AU}}$ and $\mathbf{O}_2^{\mathrm{AU}}$ are derived in the following Lemma.

Lemma 1: The probability that outage event $\mathbf{O}_1^{\mathrm{AU}}$ happens can be derived as

$$\Pr(\mathbf{O}_1^{\mathrm{AU}})=\prod_{m=1}^{M}(1-\theta_m). \tag{23}$$

The probability that outage event $\mathbf{O}_2^{\mathrm{AU}}$ happens can be derived as

$$\Pr(\mathbf{O}_2^{\mathrm{AU}})=\sum_{i=1}^{M-1}\sum_{|\mathcal{A}|=i}\left(\prod_{m\in\mathcal{A}}\theta_m\right)$$

[2] For the convenience of practical implementation, all successful AUs serve as relays and forward both messages simultaneously. This can be easily implemented in a distributed manner, where each AU automatically forwards both HP and LP messages if both messages are correctly detected.

$$\times \left[\prod_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} (1-\theta_m) \right] [1-\varpi(\mathcal{A})], \quad (24)$$

with

$$\varpi(\mathcal{A}) \triangleq e^{-\frac{\max(\varphi_1,\varphi_2)}{\mu\rho} \sum_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} \Lambda^g_{\mathcal{A},m}}, \quad (25)$$

where $\Lambda^g_{\mathcal{A},m} \triangleq (\sum_{l \in \mathcal{A}} \Omega^g_{l,m})^{-1}$.

Proof: Please refer to Appendix A. ■

Moreover, as outage events $\mathbf{O}_1^{\text{AU}}$ and $\mathbf{O}_2^{\text{AU}}$ are mutually exclusive, the overall outage probability of AUs can be expressed as $P_{\text{out}}^{\text{AU}} = \Pr(\mathbf{O}_1^{\text{AU}}) + \Pr(\mathbf{O}_2^{\text{AU}})$. By using Lemma 1, a closed-form expression of the overall outage probability of AUs is obtained.

Likewise, an outage is declared for RUs if one of the following outage events happens:

- Outage event $\mathbf{O}_1^{\text{RU}} \triangleq \{|\mathcal{S}_R| < K, \mathcal{S}_A = \emptyset\}$, which represents the event that not all RUs are successful after the first phase and no AU can serve as relay in the second phase;
- Outage event $\mathbf{O}_2^{\text{RU}} \triangleq \{|\mathcal{S}_R| < K, \mathcal{S}_A \neq \emptyset, \cup_{k \in \{1,\ldots,K\} \setminus \mathcal{S}_R} \mathbf{o}_{2,k}^{\text{RU}}\}$ with $\mathbf{o}_{2,k}^{\text{RU}} \triangleq \{\gamma_{\mathcal{S}_A \to R_k, H} < \tilde{\tau}_H\}$, which accounts for the event that some unsuccessful RUs remain unsuccessful after receiving the forwarded signals from successful AUs.

The exact expressions for $\Pr(\mathbf{O}_1^{\text{RU}})$ and $\Pr(\mathbf{O}_2^{\text{RU}})$ are derived in the following Lemma.

***Lemma** 2:* The probability for outage event $\mathbf{O}_1^{\text{RU}}$ can be derived as

$$\Pr(\mathbf{O}_1^{\text{RU}}) = \prod_{m=1}^{M} (1-\theta_m) \left(1 - \prod_{k=1}^{K} \zeta_k\right), \quad (26)$$

while the probability for outage event $\mathbf{O}_2^{\text{RU}}$ can be derived as

$$\Pr(\mathbf{O}_2^{\text{RU}}) = \sum_{i=1}^{M} \sum_{|\mathcal{A}|=i} \sum_{j=0}^{K-1} \sum_{|\mathcal{B}|=j} \left(\prod_{m \in \mathcal{A}} \theta_m \right) \times \left[\prod_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} (1-\theta_m) \right] \left(\prod_{k \in \mathcal{B}} \zeta_k \right) \times \left[\prod_{k \in \{1,\ldots,K\} \setminus \mathcal{B}} (1-\zeta_k) \right] [1-\eta(\mathcal{A},\mathcal{B})], \quad (27)$$

with $\eta(\mathcal{A},\mathcal{B}) = e^{-\frac{\varphi_1}{\mu\rho} \sum_{k \in \{1,\ldots,K\} \setminus \mathcal{B}} \Lambda^h_{\mathcal{A},k}}$, where $\Lambda^h_{\mathcal{A},k} \triangleq 1/\sum_{l \in \mathcal{A}} \Omega^h_{l,k}$.

Proof: Please refer to Appendix B ■

As the outage events $\mathbf{O}_1^{\text{RU}}$ and $\mathbf{O}_2^{\text{RU}}$ are mutually exclusive, the overall outage probability of RUs is $P_{\text{out}}^{\text{RU}} = \Pr(\mathbf{O}_1^{\text{RU}}) + \Pr(\mathbf{O}_2^{\text{RU}})$. Combining this fact with Lemma 2, a closed-form expression of the overall outage probability of RUs can be obtained.

C. Diversity Order Analysis

Based on the derived outage probabilities of AUs and RUs, we further demonstrate the achieved diversity order of DC-NOMA multicast strategy in the following theorem.

***Theorem** 3:* In the DC-NOMA multicast, both AUs and RUs achieve a diversity order of two.

Proof: From the series representation of exponential function [33, eq. (1.211.1)], it is known that $\exp(-c/\rho) \simeq 1 - c/\rho$ holds for $\rho \to \infty$, where c is a positive constant. Thus, as $\rho \to \infty$, we know from (19) that $\theta_m \overset{\rho\to\infty}{\simeq} 1 - \max(\varphi_1,\varphi_2)/(\rho\Omega^f_{S,m})$. Applying this asymptotic expression into (23) and ignoring the high order infinitesimals, we have

$$\Pr(\mathbf{O}_1^{\text{AU}}) \overset{\rho\to\infty}{\simeq} \frac{[\max(\varphi_1,\varphi_2)]^M}{\rho^M \prod_{m=1}^{M} \Omega^f_{S,m}} \propto \rho^{-M}. \quad (28)$$

Further, it can also be known from (25) that $\varpi(\mathcal{A}) \overset{\rho\to\infty}{\simeq} 1 - \max(\varphi_1,\varphi_2) \sum_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} \Lambda^g_{\mathcal{A},m}/(\mu\rho)$. Therefore, the expression in (24) can be asymptotically expressed as

$$\Pr(\mathbf{O}_2^{\text{AU}}) \overset{\rho\to\infty}{\simeq} \sum_{i=1}^{M-1} \rho^{-(M-i+1)} \times \sum_{|\mathcal{A}|=i} \left(\prod_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} \frac{\max(\varphi_1,\varphi_2)}{\Omega^f_{S,m}} \right) \times \left(\frac{\max(\varphi_1,\varphi_2)}{\mu} \sum_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} \Lambda^g_{\mathcal{A},m} \right). \quad (29)$$

Note that, when $\rho \to \infty$, the term $\rho^{-(M-i+1)}$ is high order infinitesimal of ρ^{-2} for $i = 1, \ldots, M-2$. Thus, by ignoring these high order infinitesimals with $i = 1, \ldots, M-2$, expression (29) can be further simplified as

$$\Pr(\mathbf{O}_2^{\text{AU}}) \overset{\rho\to\infty}{\simeq} \rho^{-2} \sum_{m=1}^{M} \frac{[\max(\varphi_1,\varphi_2)]^2 \Lambda^g_{\mathcal{I}_m,m}}{\mu\Omega^f_{S,m}} \propto \rho^{-2}, \quad (30)$$

where $\mathcal{I}_m \triangleq \{1, \ldots, m-1, m+1, \ldots M\}$. Overall, combining (28) and (30) with the fact that $P_{\text{out}}^{\text{AU}} = \Pr(\mathbf{O}_1^{\text{AU}}) + \Pr(\mathbf{O}_2^{\text{AU}})$, the overall outage probability of AUs can be asymptotically expressed as

$$P_{\text{out}}^{\text{AU}} = \Pr(\mathbf{O}_1^{\text{AU}}) + \Pr(\mathbf{O}_2^{\text{AU}}) \overset{\rho\to\infty}{\simeq} \rho^{-2}\Gamma, \quad (31)$$

where Γ is defined as

$$\Gamma \triangleq \begin{cases} [\max(\varphi_1,\varphi_2)]^2 \sum\limits_{m=1}^{M} \frac{\Lambda^g_{\mathcal{I}_m,m}}{\mu\Omega^f_{S,m}}, & M > 2, \\ [\max(\varphi_1,\varphi_2)]^2 \left(\sum\limits_{m=1}^{2} \frac{\Lambda^g_{\mathcal{I}_m,m}}{\mu\Omega^f_{S,m}} + \frac{1}{\prod\limits_{m=1}^{2} \Omega^f_{S,m}} \right), & M = 2. \end{cases} \quad (32)$$

Therefore, the AUs achieve a diversity order of two in DC-NOMA multicast.

With the same rationale, we can also obtain the high-SNR asymptotic expressions for $\Pr(\mathbf{O}_1^{\text{RU}})$ and $\Pr(\mathbf{O}_2^{\text{RU}})$ as follows

$$\Pr(\mathbf{O}_1^{\text{RU}}) \overset{\rho\to\infty}{\simeq} \frac{[\max(\varphi_1,\varphi_2)]^M}{\rho^{M+1} \prod_{m=1}^{M} \Omega^f_{S,m}} \sum_{k=1}^{K} \frac{\varphi_1}{\Omega^f_{S,k}} \propto \rho^{-(M+1)}, \quad (33)$$

$$\Pr(\mathbf{O}_2^{\text{RU}}) \overset{\rho\to\infty}{\simeq} \frac{\varphi_1^2}{\rho^2 \mu} \sum_{k=1}^{K} \frac{\Lambda^h_{\mathcal{J},k}}{\Omega^f_{S,k}} \propto \rho^{-2}, \quad (34)$$

where $\mathcal{J} \triangleq \{1, ..., M\}$. As $\rho^{-(M+1)}$ for $M \geq 2$ is high order infinitesimal of ρ^{-2}, we have $\Pr(\mathbf{O}_1^{\text{RU}}) \ll \Pr(\mathbf{O}_2^{\text{RU}})$ holds in high-SNR regime. Therefore, by ignoring the asymptotic expression of $\Pr(\mathbf{O}_1^{\text{RU}})$, the high-SNR asymptotic expression for the outage probability of RUs is obtained as

$$\begin{aligned} P_{\text{out}}^{\text{RU}} &= \Pr\left(\mathbf{O}_1^{\text{RU}}\right) + \Pr\left(\mathbf{O}_2^{\text{RU}}\right) \\ &\overset{\rho\to\infty}{\simeq} \Pr(\mathbf{O}_2^{\text{RU}}) \overset{\rho\to\infty}{\simeq} \frac{\varphi_1^2}{\rho^2\mu}\sum_{k=1}^{K}\frac{\Lambda_{\mathcal{J},k}^h}{\Omega_{S,k}^f} \propto \rho^{-2}, \end{aligned} \quad (35)$$

which demonstrates that the RUs also achieve a diversity order of two in DC-NOMA multicast. ■

***Remark** 1:* Compared with the direct NOMA multicast, the reliability of AUs and RUs are improved simultaneously by exploiting the diversity provided by the DC-NOMA multicast. However, since each AU may be able to serve as a relay, the maximal achievable diversity order should be no less than the number of AUs, i.e., diversity order of M. This fact implies that the DC-NOMA multicast strategy does not fully exploit the inherent diversity offered by AUs. Intuitively, the reason for this fact is that, when multiple AUs serve as relays to forward information simultaneously, the received signals at every unsuccessful AU/RU may be a destructive combination of the signals from those relays, thus leading to a loss in diversity order.

D. Minimal Energy Consumption for Guaranteed Reliability

For the DC-NOMA multicast strategy, the ECPB refers to the sum of transmission energy consumptions in both phases, which can be expressed as [20]

$$E^{\text{DC}} = \frac{P_S T_0}{2} + \frac{|\mathcal{S}_A| P_A T_0}{2} = \frac{(1+\mu|\mathcal{S}_A|)\rho\sigma^2 T_0}{2}, \quad (36)$$

where the second equality uses the facts $\rho = P_S/\sigma^2$ and $P_A = \mu P_S$.

***Theorem** 4:* Consider that the outage probabilities are constrained by $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ ($\beta > 0$). When multicast services are highly reliability-sensitive, i.e., $p_o \to 0$, the minimal ECPB that ensures $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ is asymptotically upper bounded by

$$E_{\min}^{\text{DC}} \overset{p_o\to 0}{\leq} \Xi\frac{\sigma^2 T_0}{2\sqrt{p_o}}, \quad (37)$$

with

$$\Xi = (1+\mu M)\max\left(\sqrt{\Gamma}, \varphi_1\sqrt{\frac{1}{\mu\beta}\sum_{k=1}^{K}\frac{\Lambda_{\mathcal{J},k}^h}{\Omega_{S,k}^f}}\right).$$

Proof: When $p_o \to 0$, the transmit SNR that achieves $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ should be sufficiently high. Using the high-SNR asymptotic expression in (31), we have $P_{\text{out}}^{\text{AU}} \leq p_o$ holds for $\rho \geq \sqrt{\Gamma/p_o}$. Thus, the minimal transmit SNR that guarantees $P_{\text{out}}^{\text{AU}} \leq p_o$ can be expressed as $\rho_{\min}^{\text{AU}} \overset{p_o\to 0}{\simeq} \sqrt{\Gamma/p_o}$. Using (35) with the same rationale, the minimal transmit SNR that ensures $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ can be expressed as $\rho_{\min}^{\text{RU}} \overset{p_o\to 0}{\simeq} \varphi_1\sqrt{(1/\mu\beta p_o)\sum_{k=1}^{K}\Lambda_{\mathcal{J},k}^h/\Omega_{S,k}^f}$. Therefore, the minimal transmit SNR that ensures both $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ hold is obtained as

$$\begin{aligned} \rho_{\min} &= \max(\rho_{\min}^{\text{AU}}, \rho_{\min}^{\text{RU}}) \\ &\overset{p_o\to 0}{\simeq} \max\left(\sqrt{\frac{\Gamma}{p_o}}, \varphi_1\sqrt{\frac{1}{\mu\beta p_o}\sum_{k=1}^{K}\frac{\Lambda_{\mathcal{J},k}^h}{\Omega_{S,k}^f}}\right). \end{aligned} \quad (38)$$

Substituting (38) into (36), we have

$$\begin{aligned} E_{\min}^{\text{DC}} \overset{p_o\to 0}{\simeq} & \frac{(1+\mu|\mathcal{S}_A|)\sigma^2 T_0}{2\sqrt{p_o}} \\ & \times \max\left(\sqrt{\Gamma}, \varphi_1\sqrt{\frac{1}{\mu\beta}\sum_{k=1}^{K}\frac{\Lambda_{\mathcal{J},k}^h}{\Omega_{S,k}^f}}\right). \end{aligned} \quad (39)$$

Then, applying the fact that $|\mathcal{S}_A| \leq M$ into the right-hand side of (39), the upper bound for minimal ECPB of DC-NOMA multicast strategy is obtained as shown in (37). This completes the proof. ■

***Corollary** 1:* When $p_o \to 0$, the asymptotic energy saving gain achieved by the DC-NOMA multicast over the direct NOMA multicast is lower bounded as $\frac{E_{\min}^{\text{direct}}}{E_{\min}^{\text{DC}}} \overset{p_o\to 0}{\geq} \frac{2}{\sqrt{p_o}} \cdot \frac{\Psi}{\Xi} \propto \frac{1}{\sqrt{p_o}}$.

Proof: Combining Theorems 2 and 4, this corollary is obtained. ■

***Remark** 2:* As known from Corollary 1, as the layered multicast becomes more reliability-sensitive, i.e., $p_o \to 0$, the lower bound of the energy saving gain achieved by the DC-NOMA multicast increases with a scaling factor ${p_o}^{-\frac{1}{2}}$. This observation shows that the DC-NOMA multicast is not only more reliable but also more energy-efficient than the direct NOMA multicast. The reason for this observation can be intuitively explained as follows. As the DC-NOMA multicast achieves a diversity order of two at both AUs and RUs, it can meet the requirement of reliability with much lower transmit SNR than the direct NOMA multicast which achieves only a unit diversity order.

E. Further Discussions

For the RUs in the DC-NOMA multicast strategy, since the forwarded signals may be stronger than the signal from the BS, the RUs may further decode the LP message from the forwarded signals by performing SIC.[3] More specifically, if message x_H has been decoded by RU R_k (either from the BS's signal or from the forwarded signals), the received SNR for R_k to further decode message x_L from the forwarded signals is $\gamma_{\mathcal{S}_A\to R_k,L} = \mu\rho\alpha_L Y_{\mathcal{S}_A,k}$. An *advanced outage* for RUs is declared if at least one RU is unable to decode both HP and LP messages.

All RUs can successfully decode both messages only if the following three events happen simultaneously: 1) some AUs are successful after the first phase, i.e., $\{\mathcal{S}_A \neq \emptyset\}$, 2) all successful RUs successfully decode the LP message in the second phase, i.e., $\cap_{k\in\mathcal{S}_R}\{\gamma_{\mathcal{S}_A\to R_k,L} \geq \tilde{\tau}_L\}$, and 3) all unsuccessful RUs successfully decode both messages in the second phase, i.e., $\cap_{k\in\{1,...,K\}\setminus\mathcal{S}_R}\{\gamma_{\mathcal{S}_A\to R_k,H} \geq \tilde{\tau}_H, \gamma_{\mathcal{S}_A\to R_k,L} \geq \tilde{\tau}_L\}$.

[3] For the RUs, as the signals from the BS are weak, which may not facilitate SIC, the RUs do not try to decode LP message in the first phase.

Therefore, the advanced outage probability of RUs can be expressed as

$$\begin{aligned}P^{\text{RU}}_{\text{adv,out}}=&1-\Pr\Bigg(\mathcal{S}_A\neq\emptyset,\bigcap_{k\in\mathcal{S}_R}\gamma_{\mathcal{S}_A\to R_k,L}\geq\tilde{\tau}_L,\\&\bigcap_{k\in\{1,...,K\}\setminus\mathcal{S}_R}\{\gamma_{\mathcal{S}_A\to R_k,H}\geq\tilde{\tau}_H,\gamma_{\mathcal{S}_A\to R_k,L}\geq\tilde{\tau}_L\}\Bigg)\\=&1-\sum_{i=1}^{M}\sum_{|\mathcal{A}|=i}\sum_{j=0}^{K}\sum_{|\mathcal{B}|=j}\Pr(\mathcal{S}_A=\mathcal{A})\Pr(\mathcal{S}_R=\mathcal{B})\\&\times\prod_{k\in\mathcal{B}}\underbrace{\Pr\left(\gamma_{\mathcal{S}_A\to R_k,L}\geq\tilde{\tau}_L\middle|\mathcal{S}_A=\mathcal{A}\right)}_{\triangleq I_{k,1,\mathcal{A}}}\\&\times\prod_{k\in\{1,...,K\}\setminus\mathcal{B}}I_{k,2,\mathcal{A}},\end{aligned}\tag{40}$$

with

$$I_{k,2,\mathcal{A}}\triangleq\Pr\left(\gamma_{\mathcal{S}_A\to R_k,H}\geq\tilde{\tau}_H,\gamma_{\mathcal{S}_A\to R_k,L}\geq\tilde{\tau}_L\middle|\mathcal{S}_A=\mathcal{A}\right).$$

Here, the conditional probabilities denoted by $I_{k,1,\mathcal{A}}$ and $I_{k,2,\mathcal{A}}$ can be further derived as $I_{k,1,\mathcal{A}}=\Pr\left(Y_{\mathcal{A},k}\geq\varphi_2/(\mu\rho)\right)=e^{-\varphi_2\Lambda^h_{\mathcal{A},k}/(\mu\rho)}$ and $I_{k,2,\mathcal{A}}=\Pr\left(Y_{\mathcal{A},k}\geq\max(\varphi_1,\varphi_2)/(\mu\rho)\right)=e^{-\max(\varphi_1,\varphi_2)\Lambda^h_{\mathcal{A},k}/(\mu\rho)}$. Combining the above results with the help of (20) and (22), a closed-form expression is obtained for the advanced outage probability of RUs.

V. Opportunistic Cooperative NOMA Multicast

Recall that the DC-NOMA multicast suffers a loss in diversity order due to the possible destructive combination in the second phase. To fully exploit the inherent diversity offered by the AUs, this section proposes an opportunistic cooperative NOMA layered multicast strategy, termed as *OC-NOMA multicast*, in which one of successful AUs is selected to forward messages. To evaluate the performance of proposed OC-NOMA multicast strategy, this section also theoretically analyzes the outage probability, diversity orders, and energy consumption. Compared with the DC-NOMA multicast strategy, the OC-NOMA multicast strategy achieves much better reliability and energy-efficiency, but requires that each successful AU has instantaneous CSI to unsuccessful AUs and RUs.

A. Strategy Description

Similar to the DC-NOMA multicast strategy, the OC-NOMA multicast strategy is also performed within two phases. The duration of each phase is $T_0/2$. During the first phase, the BS sends the superimposed signal $\sqrt{P_S\alpha_H}x_H+\sqrt{P_S\alpha_L}x_L$ to all users, where the power allocation coefficients satisfy $\alpha_H+\alpha_L=1$, $\alpha_H>\alpha_L$ and $\alpha_H-\alpha_L(2^{2r_H}-1)>0$ [12]. At the end of the first phase, the indices set of successful AUs is $\mathcal{S}_A\triangleq\{m|\gamma_{S\to A_m,H}\geq\tilde{\tau}_H,\gamma_{S\to A_m,L}\geq\tilde{\tau}_L\}$ while the indices set of successful RUs is $\mathcal{S}_R\triangleq\{k|\gamma_{S\to R_k,H}\geq\tilde{\tau}_H\}$, where the terms $\gamma_{S\to A_m,H}$, $\gamma_{S\to A_m,L}$ and $\gamma_{S\to R_k,H}$ have been shown in (1), (2) and (3).

If all users have been successful after the first phase, the second phase will be cancelled, and the BS will immediately start a new transmission block and proceed to transmit new messages. Otherwise, the cooperative transmission will be performed in the second phase. Prior to the second phase, a relay selection procedure is performed to select the best successful AU to serve as a relay. The detailed selection scheme will be presented in the next subsection.

Without loss of generality, here we assume that successful AU A_m is selected to serve as a relay. Thus, the selected AU A_m sends superimposed signal $\sqrt{P_A\alpha_H}x_H+\sqrt{P_A\alpha_L}x_L$ during the second phase. As a result, the observed signal at unsuccessful AU $A_{m'}(m'\in\{1,...,M\}\setminus\mathcal{S}_A)$ is $y_{A_m\to A_{m'}}=\sqrt{P_A\alpha_H}g_{m,m'}x_H+\sqrt{P_A\alpha_L}g_{m,m'}x_L+n_{m'}$. Accordingly, the SINR for AU $A_{m'}$ to decode x_H is given by

$$\gamma_{A_m\to A_{m'},H}=\frac{\alpha_H|g_{m,m'}|^2}{\alpha_L|g_{m,m'}|^2+(\mu\rho)^{-1}}.\tag{41}$$

Once x_H is correctly decoded, unsuccessful AU $A_{m'}$ performs SIC and then decodes x_L with SNR given by

$$\gamma_{A_m\to A_{m'},L}=\mu\rho\alpha_L|g_{m,m'}|^2.\tag{42}$$

Similarly, the received signals at unsuccessful RU R_k ($k\in\{1,...,K\}\setminus\mathcal{S}_R$) is $y_{A_m\to R_k}=\sqrt{P_A\alpha_H}h_{m,k}x_H+\sqrt{P_A\alpha_L}$ $h_{m,k}x_L+n_k$, and the SINR for R_k to decode x_H is

$$\gamma_{A_m\to R_k,H}=\frac{\alpha_H|h_{m,k}|^2}{\alpha_L|h_{m,k}|^2+(\mu\rho)^{-1}}.\tag{43}$$

B. Two-Step Selection Scheme

From (41), (42) and (43) we know that, the reception quality of unsuccessful AUs/RUs is affected by the selected relay. Therefore, the *best* successful AU should be properly selected in order to maximally enhance the reception quality of all unsuccessful AUs/RUs. To this end, we design a two-step selection scheme described as follows.

First, we pick up a number of successful AUs as *potential relays*. Here, a successful AU is called a potential relay if it can reliably forward information (HP and LP messages) to all unsuccessful AUs. If successful AU A_m is selected, the condition that all unsuccessful AUs become successful after the second phase is $\cap_{m'\in\{1,...,M\}\setminus\mathcal{S}_A}\{\gamma_{A_m\to A_{m'},H}\geq\tilde{\tau}_H,\gamma_{A_m\to A_{m'},L}\geq\tilde{\tau}_L\}$, which can be equivalently expressed as $\min_{m'\in\{1,...,M\}\setminus\mathcal{S}_A}|g_{m,m'}|^2\geq\max(\varphi_1,\varphi_2)/(\mu\rho)$. Therefore, the indices set of potential relays can be expressed as[4]

$$\begin{aligned}&\mathcal{R}_p\triangleq\\&\left\{m\middle|\min_{m'\in\{1,...,M\}\setminus\mathcal{S}_A}|g_{m,m'}|^2\geq\frac{\max(\varphi_1,\varphi_2)}{\mu\rho},m\in\mathcal{S}_A\right\}.\end{aligned}\tag{44}$$

Second, among all potential relays, we select the one, denoted $A_{m^\dagger}$, such that the worst relaying link gain to the

[4] When all AUs have already been successful after the first phase, there is no need to guarantee the reliability of unsuccessful AUs. In this condition, each successful AU can be a potential relay, i.e., $\mathcal{R}_p=\{1,...,M\}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2874239, IEEE Transactions on Communications

unsuccessful RUs is maximized, i.e., we have:

$$m^{\dagger} = \arg\max_{m \in \mathcal{R}_p} \left(\min_{k \in \{1,...,K\} \backslash \mathcal{S}_R} |h_{m,k}|^2 \right). \tag{45}$$

Note that, if $\mathcal{S}_R = \{1, ..., K\}$, the above selection criterion will reduce to random selection, since each potential relay in $\mathcal{R}_p$ can guarantee the reliability of AUs.

Similar to [35], we design the following procedure that consists of $M + K + 3$ minislots denoted as minislot $0, 1, ...,$ and $M + K + 2$, to realize the two-step selection scheme. Note that the minislots are located at the beginning of the second phase of each transmission block. In other words, the procedure is implemented after the BS's broadcasting in the first phase and before the message forwarding in the second phase.

First step: In minislot 0, the BS sends a message `Channel Estimation Request (CER)` to all AUs and RUs. By reception of the CER, an AU, say AU $A_{m'}(m' \in \{1, ..., M\})$, sends a message `AU Success (AU-S)` at minislot m' if it is a successful AU, or sends a message `AU Failure (AU-F)` at minislot m' otherwise. If the message is `AU-F` at minislot m', each successful AU estimates its channel gain to AU $A_{m'}$ based on reception of the message (due to channel reciprocity). Thus, at the end of minislot M, each successful AU has channel gain information to each unsuccessful AU. Then, in minislot $(M + 1)$, each successful AU can decide whether or not it can be a potential relay based on the criterion given in (44). In specific, if $\min_{m' \in \{1,...,M\} \backslash \mathcal{S}_A} |g_{m,m'}|^2 \geq \max(\varphi_1, \varphi_2)/(\mu\rho)$ holds, successful AU $A_m (m \in \mathcal{S}_A)$ can be a potential relay; otherwise, it will become inactive until the end of the transmission block.

Second step: Similarly, in minislot $(M + k + 1)$, $k \in \{1, ..., K\}$, RU R_k sends a message `RU-Success (RU-S)` if it is a successful RU, or sends message `RU-Failure (RU-F)` otherwise. If message `RU-F` is received in minislot $(M + k + 1)$, each potential relay uses its reception of the message to get its channel gain information to RU R_k. Thus, at the end of minislot $(M + K + 1)$, each potential relay has channel gain information to every unsuccessful RU. Then at minislot $(M + K + 2)$, each potential relay, say AU A_m, counts down a timer with initial value equal to $t_m = t_0 \exp(-\min_{k \in \{1,...,K\} \backslash \mathcal{S}_R} |h_{m,k}|^2)$, where $t_0 (\ll T_0)$ is a constant. Consequently, the timer of AU $A_{m^{\dagger}}$ that satisfies (45) expires first and then AU $A_{m^{\dagger}}$ announces that it has been selected by sending a message `Relay Selected`. After getting the message, all other potential relays become inactive until the end of the transmission block.

It can be seen that we have six message types in the above procedure. Thus, three-bit codewords are sufficient to encode those messages, and therefore, we can adopt a strong channel coding for the messages. So it is reasonable to assume that the messages are transmitted in a very short duration[5] and without errors.

C. Outage Analysis

1) Probability for Successful AUs being Potential Relays: Conditioned on AU A_m being successful after the first phase, the probability that AU A_m can further be a potential relay is expressed as

$$\begin{aligned} &\Pr(m \in \mathcal{R}_p | m \in \mathcal{S}_A) \\ =&\Pr\left(\min_{m' \in \{1,...,M\} \backslash \mathcal{S}_A} |g_{m,m'}|^2 \geq \frac{\max(\varphi_1, \varphi_2)}{\mu\rho} \right) \\ =&\underbrace{e^{-\frac{\max(\varphi_1,\varphi_2)}{\mu\rho} \sum_{m' \in \{1,...,M\} \backslash \mathcal{S}_A} 1/\Omega^g_{m,m'}}}_{\triangleq \omega_m(\mathcal{S}_A)}. \end{aligned} \tag{46}$$

Then, letting $\mathcal{C}$ be a subset of $\mathcal{A}$ (recall that $\mathcal{A}$ is a subset of $\{1, ..., M\}$), the probability of $\{\mathcal{R}_p = \mathcal{C}\}$ conditioned on $\{\mathcal{S}_A = \mathcal{A}\}$ can be expressed as

$$\Pr(\mathcal{R}_p = \mathcal{C} | \mathcal{S}_A = \mathcal{A}) = \prod_{m \in \mathcal{C}} \omega_m(\mathcal{A}) \prod_{m \in \mathcal{A} \backslash \mathcal{C}} [1 - \omega_m(\mathcal{A})]. \tag{47}$$

2) Exact Outage Probability: Since each potential relay can guarantee reliable reception at all unsuccessful AUs, the AUs experience an outage only when not all AUs are successful after the first phase and no potential relay exists, i.e., $\{|\mathcal{S}_A| < M, \mathcal{R}_p = \emptyset\}$. Consequently, the outage probability of AUs can be expressed as

$$\begin{aligned} P^{\text{AU}}_{\text{out}} =& \Pr(|\mathcal{S}_A| < M, \mathcal{R}_p = \emptyset) \\ =& \sum_{i=0}^{M-1} \sum_{|\mathcal{A}|=i} \Pr(\mathcal{S}_A = \mathcal{A}) \Pr(\mathcal{R}_p = \emptyset | \mathcal{S}_A = \mathcal{A}). \end{aligned} \tag{48}$$

Using (20) and (47) with letting $\mathcal{C} = \emptyset$, a closed-form expression of $P^{\text{AU}}_{\text{out}}$ is derived as

$$\begin{aligned} P^{\text{AU}}_{\text{out}} = \sum_{i=0}^{M-1} \sum_{|\mathcal{A}|=i} &\left(\prod_{m \in \mathcal{A}} \theta_m [1 - \omega_m(\mathcal{A})] \right) \\ &\times \left[\prod_{m \in \{1,...,M\} \backslash \mathcal{A}} (1 - \theta_m) \right]. \end{aligned} \tag{49}$$

On the other hand, an outage is declared for RUs when one of the following outage events happens:

- Outage event $\mathfrak{O}^{\text{RU}}_1 \triangleq \{|\mathcal{S}_R| < K, \mathcal{R}_p = \emptyset\}$, which means that not all RUs are successful after the first phase and no successful AU can be a potential relay.
- Outage event $\mathfrak{O}^{\text{RU}}_2 \triangleq \{|\mathcal{S}_R| < K, \mathcal{R}_p \neq \emptyset, \cup_{k \in \{1,...,K\} \backslash \mathcal{S}_R} \mathfrak{o}^{\text{RU}}_{2,k}\}$ with $\mathfrak{o}^{\text{RU}}_{2,k} \triangleq \{\gamma_{A_{m^{\dagger}} \to R_k, H} < \tilde{\tau}_H\}$, meaning that some RUs remain unsuccessful after receiving the forwarded signal from $A_{m^{\dagger}}$.

The following lemma provides exact expressions for probabilities $\Pr(\mathfrak{O}^{\text{RU}}_1)$ and $\Pr(\mathfrak{O}^{\text{RU}}_2)$.

Lemma 3: In OC-NOMA multicast, the outage event $\mathfrak{O}^{\text{RU}}_1$ happens with the probability being

$$\Pr(\mathfrak{O}^{\text{RU}}_1) = \left(1 - \prod_{k=1}^{K} \zeta_k \right)$$

[5] Thus, overall duration of all $(M+K+3)$ minislots is much less than the duration of a transmission block, and their impacts on the system performance are ignored.

$$\times\sum_{i=0}^{M}\sum_{|\mathcal{A}|=i}\left(\prod_{m\in\mathcal{A}}\theta_m[1-\omega_m(\mathcal{A})]\right)$$
$$\times\left[\prod_{m\in\{1,\ldots,M\}\setminus\mathcal{A}}(1-\theta_m)\right], \quad (50)$$

while the outage event $\mathfrak{O}_2^{\mathrm{RU}}$ happens with the probability being

$$\Pr(\mathfrak{O}_2^{\mathrm{RU}})=\sum_{i=1}^{M}\sum_{|\mathcal{A}|=i}\left(\prod_{m\in\mathcal{A}}\theta_m\right)\left[\prod_{m\in\{1,\ldots,M\}\setminus\mathcal{A}}(1-\theta_m)\right]$$
$$\times\sum_{j=0}^{K-1}\sum_{|\mathcal{B}|=j}\left(\prod_{k\in\mathcal{B}}\zeta_k\right)\left[\prod_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}(1-\zeta_k)\right]$$
$$\times\sum_{t=1}^{i}\sum_{|\mathcal{C}|=t}\left(\prod_{m\in\mathcal{A}\setminus\mathcal{C}}[1-\omega_m(\mathcal{A})]\right)$$
$$\times\left(\prod_{m\in\mathcal{C}}\omega_m(\mathcal{A})[1-\xi_m(\mathcal{B})]\right), \quad (51)$$

where $\xi_m(\mathcal{B}) \triangleq e^{-\frac{\varphi_1}{\mu\rho}\sum_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}1/\Omega^h_{m,k}}$.

Proof: Please refer to Appendix C. ■

As the outage events $\mathfrak{O}_1^{\mathrm{RU}}$ and $\mathfrak{O}_2^{\mathrm{RU}}$ are mutually exclusive, the overall outage probability of RUs can be expressed as $P_{\mathrm{out}}^{\mathrm{RU}} = \Pr(\mathfrak{O}_1^{\mathrm{RU}}) + \Pr(\mathfrak{O}_2^{\mathrm{RU}})$. Combining this fact with Lemma 3, a closed-form expression of the overall outage probability of RUs is derived.

D. Diversity Order Analysis

***Theorem** 5:* In the OC-NOMA multicast, AUs achieve a diversity order of M, while RUs achieve a diversity order of $(M+1)$.

Proof: Using the results in (49) and then following similar steps to those in the proof of Theorem 3, the high-SNR asymptotic outage probability for AUs can be obtained as

$$P_{\mathrm{out}}^{\mathrm{AU}} \overset{\rho\to\infty}{\simeq} \rho^{-M}\Upsilon, \quad (52)$$

with

$$\Upsilon\triangleq\sum_{i=0}^{M-1}\sum_{|\mathcal{A}|=i}\frac{[\max(\varphi_1,\varphi_2)]^M}{\mu^i}$$
$$\times\left[\prod_{m\in\mathcal{A}}\left(\sum_{m'\in\{1,\ldots,M\}\setminus\mathcal{A}}\frac{1}{\Omega^g_{m,m'}}\right)\right]$$
$$\times\left(\prod_{m\in\{1,\ldots,M\}\setminus\mathcal{A}}\frac{1}{\Omega^f_{S,m}}\right). \quad (53)$$

It is observed that $P_{\mathrm{out}}^{\mathrm{AU}} \propto \rho^{-M}$ for $\rho\to\infty$, showing that a diversity order of M is achieved by AUs.

Following the same rationale, the high-SNR asymptotic outage probability for RUs can be expressed based on (50) and (51) as

$$P_{\mathrm{out}}^{\mathrm{RU}} \overset{\rho\to\infty}{\simeq} \rho^{-(M+1)}\Theta, \quad (54)$$

with

$$\Theta\triangleq\left(\sum_{k=1}^{K}\frac{\varphi_1}{\Omega^f_{S,k}}\right)\sum_{i=0}^{M}\sum_{|\mathcal{A}|=i}\left(\prod_{m\in\{1,\ldots,M\}\setminus\mathcal{A}}\frac{\max(\varphi_1,\varphi_2)}{\Omega^f_{S,m}}\right)$$
$$\times\left(\prod_{m\in\mathcal{A}}\sum_{m'\in\{1,\ldots,M\}\setminus\mathcal{A}}\frac{\max(\varphi_1,\varphi_2)}{\mu\Omega^g_{m,m'}}\right)$$
$$+\sum_{i=1}^{M}\sum_{|\mathcal{A}|=i}\sum_{t=1}^{i}\sum_{|\mathcal{C}|=t}\left[\sum_{k=1}^{K}\frac{\varphi_1}{\Omega^f_{S,k}}\left(\prod_{m\in\mathcal{C}}\frac{\varphi_1}{\mu\Omega^h_{m,k}}\right)\right]$$
$$\times\left(\prod_{m\in\mathcal{A}\setminus\mathcal{C}}\sum_{m'\in\{1,\ldots,M\}\setminus\mathcal{A}}\frac{\max(\varphi_1,\varphi_2)}{\mu\Omega^g_{m,m'}}\right)$$
$$\times\left(\prod_{m\in\{1,\ldots,M\}\setminus\mathcal{A}}\frac{\max(\varphi_1,\varphi_2)}{\Omega^f_{S,m}}\right). \quad (55)$$

It can be seen that, $P_{\mathrm{out}}^{\mathrm{RU}} \propto \rho^{-(M+1)}$ for $\rho\to\infty$, indicating the RUs achieve a diversity order of $M+1$. ■

***Remark** 3:* When the number of AUs is larger than 2, the OC-NOMA multicast outperforms the DC-NOMA multicast in terms of diversity order. Furthermore, it is noteworthy that, in the OC-NOMA multicast, the RUs achieve one order higher diversity than AUs. The reason is as follows. When message forwarding to AUs is required, there exists at least one unsuccessful AU, which means that at most $(M-1)$ AUs can be eligible for serving as potential relays. On the other hand, when RUs require message forwarding, all AUs may be able to serve as potential relays, thus leading to a higher diversity order.

E. Minimal Energy Consumption for Guaranteed Reliability

Similar to the ECPB of DC-NOMA multicast strategy, the ECPB of OC-NOMA multicast strategy can be expressed as [20]

$$E^{\mathrm{OC}} = \frac{P_ST_0}{2}+\frac{P_AT_0}{2}=\frac{(1+\mu)\rho\sigma^2T_0}{2}, \quad (56)$$

where the second equality uses the facts $\rho = P_S/\sigma^2$ and $P_A = \mu P_S$.

***Theorem** 6:* Consider that the outage probabilities are constrained by $P_{\mathrm{out}}^{\mathrm{AU}} \le p_o$ and $P_{\mathrm{out}}^{\mathrm{RU}} \le \beta p_o$ $(\beta > 0)$. When multicast services are highly reliability-sensitive, i.e., $p_o \to 0$, the minimal ECPB that ensures $P_{\mathrm{out}}^{\mathrm{AU}} \le p_o$ and $P_{\mathrm{out}}^{\mathrm{RU}} \le \beta p_o$ can be asymptotically expressed as

$$E_{\min}^{\mathrm{OC}} \overset{p_o\to 0}{\simeq} \frac{(1+\mu)\sigma^2T_0}{2{p_o}^{1/M}}\max\left(\Upsilon^{\frac{1}{M}},\ \left(\frac{\Theta {p_o}^{1/M}}{\beta}\right)^{\frac{1}{M+1}}\right). \quad (57)$$

Proof: When $p_o \to 0$, the transmit SNR should be sufficiently high to achieve $P_{\mathrm{out}}^{\mathrm{AU}} \le p_o$ and $P_{\mathrm{out}}^{\mathrm{RU}} \le \beta p_o$ $(\beta > 0)$. From the high-SNR asymptotic outage probability of AUs given in (52), it is known that $P_{\mathrm{out}}^{\mathrm{AU}} \le p_o$ holds for $\rho \ge \left(\frac{\Upsilon}{p_o}\right)^{\frac{1}{M}}$. Hence, the minimal transmit SNR to achieve $P_{\mathrm{out}}^{\mathrm{AU}} \le p_o$ can be asymptotically expressed as $\rho_{\min}^{\mathrm{AU}} \overset{p_o\to 0}{\simeq} \left(\frac{\Upsilon}{p_o}\right)^{\frac{1}{M}}$. Likewise, using (54) with the same rationale, the minimal transmit SNR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2874239, IEEE Transactions on Communications

that ensures $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ can be asymptotically expressed as $\rho_{\min}^{\text{RU}} \overset{p_o \to 0}{\simeq} \left(\frac{\Theta}{\beta p_o}\right)^{\frac{1}{M+1}}$. Based on the results above, the minimal SNR that ensures both $P_{\text{out}}^{\text{AU}} \leq p_o$ and $P_{\text{out}}^{\text{RU}} \leq \beta p_o$ hold for $p_o \to 0$ can be expressed as

$$\begin{aligned}\rho_{\min} &= \max\left(\rho_{\min}^{\text{AU}}, \rho_{\min}^{\text{RU}}\right) \\ &\overset{p_o \to 0}{\simeq} \frac{1}{p_o^{1/M}} \max\left(\Upsilon^{\frac{1}{M}}, \left(\frac{\Theta p_o^{1/M}}{\beta}\right)^{\frac{1}{M+1}}\right). \quad (58)\end{aligned}$$

Applying (58) into (56), the minimal ECPB of OC-NOMA multicast strategy is given in (57). This completes the proof. ■

***Corollary** 2:* The asymptotic energy saving gain achieved by the OC-NOMA multicast over the direct NOMA multicast is given by

$$\begin{aligned}\frac{E_{\min}^{\text{direct}}}{E_{\min}^{\text{OC}}} \overset{p_o \to 0}{\simeq} &\frac{2}{(1+\mu) p_o^{\frac{M-1}{M}}} \\ &\times \frac{\Psi}{\max\left(\Upsilon^{\frac{1}{M}}, \left(\frac{\Theta p_o^{1/M}}{\beta}\right)^{\frac{1}{M+1}}\right)}. \quad (59)\end{aligned}$$

Proof: Combining Theorems 2 and 6, this corollary is proved. ■

***Remark** 4:* From Corollary 2 we know that, compared with the direct NOMA multicast, the energy saving gain achieved by the OC-NOMA multicast scales with a factor $p_o^{-\frac{M-1}{M}}$, which is much higher than that of DC-NOMA multicast for $M > 2$. This fact demonstrates that, when more than two AUs are served and the layered multicast is highly reliability-sensitive ($p_o \to 0$), the energy saving gain of OC-NOMA multicast increases faster than that of DC-NOMA multicast.

F. Further Discussions

Similar to the DC-NOMA multicast strategy, the RUs may further decode the LP message from the forwarded signal in the OC-NOMA multicast strategy. For RU R_k, if message x_H has been decoded and removed by SIC, the SNR for decoding message x_L is $\gamma_{A_{m^\dagger} \to R_k, L} = \mu\rho\alpha_L |h_{m^\dagger,k}|^2$. In the OC-NOMA multicast strategy, all RUs will successfully decode both HP and LP messages only if the following four events happen simultaneously: 1) some AUs are successful after the first phase, i.e., $\mathcal{S}_A \neq \emptyset$, 2) the potential relay set is non-empty, i.e., $\mathcal{R}_p \neq \emptyset$, 3) all successful RUs correctly decode the LP message during the second phase, i.e., $\cap_{k \in \mathcal{S}_R}\{\gamma_{A_{m^\dagger} \to R_k, L} \geq \tilde{\tau}_L\}$, and 4) all unsuccessful RUs correctly decode both HP and LP messages during the second phase, i.e., $\cap_{k \in \{1,...,K\} \setminus \mathcal{S}_R}\{\gamma_{A_{m^\dagger} \to R_k, H} \geq \tilde{\tau}_H, \gamma_{A_{m^\dagger} \to R_k, L} \geq \tilde{\tau}_L\}$. Otherwise, an advanced outage is declared for RUs. Based on the above facts, the advanced outage probability of RUs can be expressed as (60), shown at the top of the next page. In (60), the term $\chi_{\mathcal{B},\mathcal{C}}$ is further expressed as (61), also shown at the top of the next page.

If $\mathcal{B} \neq \{1,...,K\}$, we define $X_m = \min_{k \in \{1,...,K\} \setminus \mathcal{B}} |h_{m,k}|^2$ for $m \in \mathcal{C}$. The cumulative distribution function (CDF) and probability density function (PDF) of X_m can be expressed as $F_{X_m}(x) = 1 - e^{-x\Lambda_{m,\mathcal{B}}^h}$ and $f_{X_m}(x) = \Lambda_{m,\mathcal{B}}^h e^{-x\Lambda_{m,\mathcal{B}}^h}$, where $\Lambda_{m,\mathcal{B}}^h \triangleq \sum_{k \in \{1,...,K\} \setminus \mathcal{B}} 1/\Omega_{m,k}^h$. Then, using the selection criterion (45) with letting $\mathcal{R}_p = \mathcal{C}$ and $\mathcal{S}_R = \mathcal{B}$, $\Pr(m^\dagger = m)$ can be derived as

$$\begin{aligned}\Pr(m^\dagger = m) &= \Pr\left(\bigcap_{m' \in \mathcal{C} \setminus \{m\}} X_{m'} < X_m\right) \\ &= \int_0^\infty \prod_{m' \in \mathcal{C} \setminus \{m\}} F_{X_{m'}}(x) f_{X_m}(x) \mathrm{d}x \\ &= \int_0^\infty \prod_{m' \in \mathcal{C} \setminus \{m\}} \left(1 - e^{-x\Lambda_{m',\mathcal{B}}^h}\right) \Lambda_{m,\mathcal{B}}^h e^{-x\Lambda_{m,\mathcal{B}}^h} \mathrm{d}x \\ &= \int_0^\infty \left(\sum_{u=0}^{t-1} \sum_{|\mathcal{D}|=u, \mathcal{D} \subseteq \mathcal{C} \setminus \{m\}} (-1)^u \right. \\ &\qquad \left. \times e^{-x \sum_{m' \in \mathcal{D}} \Lambda_{m',\mathcal{B}}^h}\right) \Lambda_{m,\mathcal{B}}^h e^{-x\Lambda_{m,\mathcal{B}}^h} \mathrm{d}x \\ &= \sum_{u=0}^{t-1} \sum_{|\mathcal{D}|=u, \mathcal{D} \subseteq \mathcal{C} \setminus \{m\}} (-1)^u \\ &\qquad \times \frac{\Lambda_{m,\mathcal{B}}^h}{\Lambda_{m,\mathcal{B}}^h + \sum_{m' \in \mathcal{D}} \Lambda_{m',\mathcal{B}}^h}. \quad (62)\end{aligned}$$

If $\mathcal{B} = \{1,...,K\}$, recall that the selection criterion in (45) will reduce to the random selection. This fact indicates that $\Pr(m^\dagger = m) = 1/|\mathcal{C}| = 1/t$ holds for $m \in \mathcal{C}$.

Applying the above derived probability expressions of $\Pr(m^\dagger = m)$ into (61), $\chi_{\mathcal{B},\mathcal{C}}$ is derived in closed form. Finally, substituting the closed-form expression of $\chi_{\mathcal{B},\mathcal{C}}$ and the results in (20), (22) and (47) into (60), a closed-form expression of the advanced outage probability of RUs is obtained.

VI. Simulation Results

This section provides numerical results to verify our theoretical analysis and evaluate the performance of our proposed strategies. The BS is located at coordinate $(0, 0)$. For each simulated combination of (M, K), we randomly generate locations of M AUs within a circle centered at $(50, 0)$ and with radius 30, and K RUs within another circle centered at $(100, 0)$ and with radius 30. Once the locations of AUs and RUs are generated, they keep unchanged over all numerical trials. The small-scale Rayleigh fading varies independently over each numerical trial. Denoting the distance between nodes i and j as d_{ij} and the path-loss reference distance as d_0, the path-loss attenuation between node i and j is modelled as $(d_{ij}/d_0)^{-\kappa}$ for $d_{ij} \geq d_0$ [36, sec. 2.6], where κ is the path-loss exponent. Moreover, if $d_{ij} < d_0$, we assume that there is no path-loss attenuation between nodes i and j. Following typical parameters for urban cellular networks [36], we set the path-loss reference distance as $d_0 = 20$ and path-loss exponent as $\kappa = 3$. Since the BS is expected to own higher transmit power than successful AUs, we set $\mu = P_A/P_S = 0.1$ in the simulation. Other parameters in the simulations are $\alpha_H = 0.8$, $\alpha_L = 0.2$, $r_H = 0.7$ bps/Hz, $r_L = 1$ bps/Hz, and $T_0 = 1 \times 10^{-3}$.

$$\begin{aligned}
P_{\text{adv,out}}^{\text{RU}} =& 1-\Pr\left(\mathcal{S}_A\neq\emptyset,\mathcal{R}_p\neq\emptyset,\bigcap_{k\in\mathcal{S}_R}\gamma_{A_{m^\dagger}\to R_k,L}\geq\tilde{\tau}_L,\bigcap_{k\in\{1,\ldots,K\}\backslash\mathcal{S}_R}\left\{\gamma_{A_{m^\dagger}\to R_k,H}\geq\tilde{\tau}_H,\gamma_{A_{m^\dagger}\to R_k,L}\geq\tilde{\tau}_L\right\}\right)\\
=&1-\sum_{i=1}^{M}\sum_{|\mathcal{A}|=i}\sum_{t=1}^{i}\sum_{|\mathcal{C}|=t}\Pr\left(\mathcal{S}_A=\mathcal{A}\right)\Pr\left(\mathcal{R}_p=\mathcal{C}\middle|\mathcal{S}_A=\mathcal{A}\right)\sum_{j=0}^{K}\sum_{|\mathcal{B}|=j}\Pr\left(\mathcal{S}_R=\mathcal{B}\right)\\
&\times\underbrace{\Pr\left(\bigcap_{k\in\mathcal{S}_R}|h_{m^\dagger,k}|^2\geq\frac{\varphi_2}{\mu\rho},\bigcap_{k\in\{1,\ldots,K\}\backslash\mathcal{S}_R}|h_{m^\dagger,k}|^2\geq\frac{\max(\varphi_1,\varphi_2)}{\mu\rho}\middle|\mathcal{S}_R=\mathcal{B},\mathcal{R}_p=\mathcal{C}\right)}_{\triangleq\chi_{\mathcal{B},\mathcal{C}}}.
\end{aligned}\tag{60}$$

$$\begin{aligned}
\chi_{\mathcal{B},\mathcal{C}}=&\Pr\left(\min_{k\in\mathcal{B}}|h_{m^\dagger,k}|^2\geq\frac{\varphi_2}{\mu\rho}\middle|\mathcal{R}_p=\mathcal{C}\right)\Pr\left(\min_{k\in\{1,\ldots,K\}\backslash\mathcal{B}}|h_{m^\dagger,k}|^2\geq\frac{\max(\varphi_1,\varphi_2)}{\mu\rho}\middle|\mathcal{R}_p=\mathcal{C}\right)\\
=&\left[\sum_{m\in\mathcal{C}}\Pr(m^\dagger=m)\Pr\left(\min_{k\in\mathcal{B}}|h_{m,k}|^2\geq\frac{\varphi_2}{\mu\rho}\right)\right]\Pr\left(\max_{m\in\mathcal{C}}\left[\min_{k\in\{1,\ldots,K\}\backslash\mathcal{B}}|h_{m,k}|^2\right]\geq\frac{\max(\varphi_1,\varphi_2)}{\mu\rho}\right)\\
=&\left[\sum_{m\in\mathcal{C}}\Pr\left(m^\dagger=m\right)e^{-\frac{\varphi_2}{\mu\rho}\sum_{k\in\mathcal{B}}1/\Omega_{m,k}^h}\right]\left[1-\prod_{m\in\mathcal{C}}\left(1-e^{-\frac{\max(\varphi_1,\varphi_2)}{\mu\rho}\sum_{k\in\{1,\ldots,K\}\backslash\mathcal{B}}1/\Omega_{m,k}^h}\right)\right].
\end{aligned}\tag{61}$$

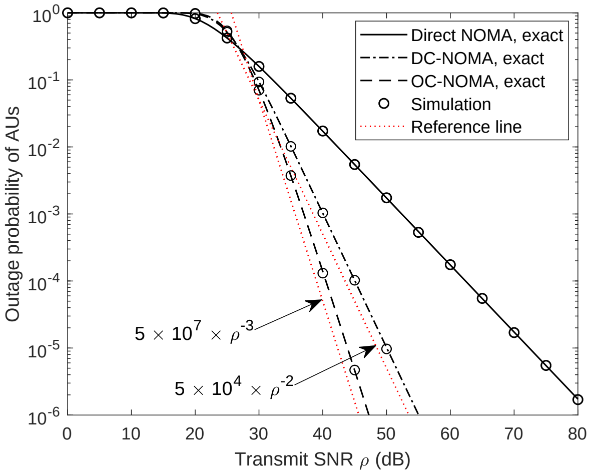


Fig. 2. Theoretical and simulated outage performance of AUs in the direct NOMA multicast strategy, the DC-NOMA multicast strategy, and the OC-NOMA multicast strategy, where $M=3$ and $K=4$.

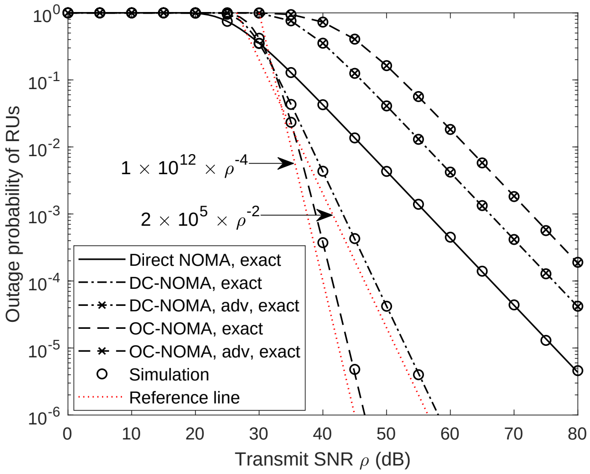


Fig. 3. Theoretical and simulated outage performance of RUs in the direct NOMA multicast strategy, the DC-NOMA multicast strategy, and the OC-NOMA multicast strategy, where $M=3$ and $K=4$.

A. *Verification of Theoretical Results*

In this subsection, we verify the derived theoretical results of proposed DC-NOMA multicast and OC-NOMA multicast strategies in terms of outage probability and energy consumption. Note that to obtain theoretical results, distance information of the transmission links is needed.

Figs. 2 and 3 demonstrate the theoretical and simulated outage performance of AUs and RUs in the direct NOMA multicast strategy, the DC-NOMA multicast strategy, and the OC-NOMA multicast strategy, where $M=3$ and $K=4$. The red dotted lines are reference lines, and "adv" in the legend means advanced outage probability. As shown in both figures, the simulated values perfectly coincide with the exact values obtained from the derived closed-form expressions. When transmit SNR ρ ranges from 0 dB to 20 dB, the outage probabilities of AUs and RUs approach one for each strategy in both figures, indicating that outage events of AUs and RUs always happen. Recall that ρ is the ratio of transmit power to noise power. Thus, when ρ ranges from 0 dB to 20 dB (low SNR regime), the transmit power is not enough to combat the path-loss and/or fading, thereby almost no AU/RU becoming successful. When transmit SNR ρ ranges from 20 dB to 30 dB, the direct NOMA multicast strategy slightly outperforms the proposed cooperation strategies in terms of outage probabilities of both AUs and RUs, since the proposed cooperation strategies suffer from $1/2$ spectral efficiency loss and the diversity orders cannot take effect in medium SNR

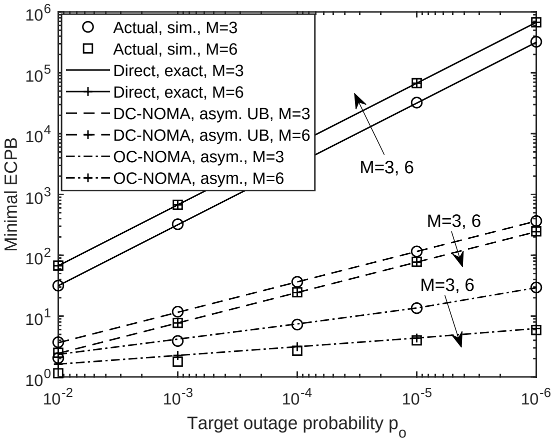


Fig. 4. The comparison of minimal ECPB among the direct NOMA multicast strategy, the DC-NOMA multicast strategy, and the OC-NOMA multicast strategy, where $M = 3$ and 6, $K = 3$, and $\beta = 1$.

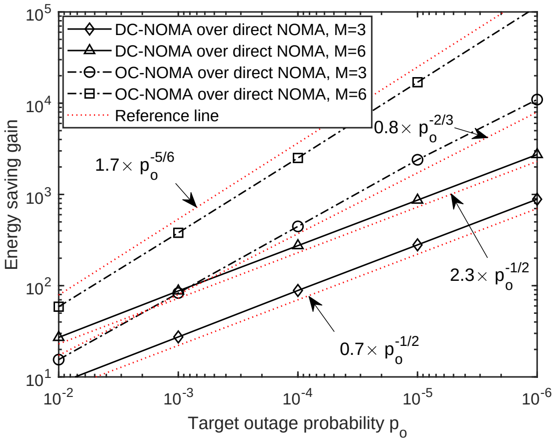


Fig. 5. Energy saving gain of the DC-NOMA multicast strategy and the OC-NOMA multicast strategy over the direct NOMA multicast strategy, where $M = 3$ and 6, $K = 3$, and $\beta = 1$.

regime. When transmit SNR ρ increases to higher than 30 dB, the two proposed cooperation strategies achieve much better outage performance than the direct NOMA multicast strategy due to the diversity orders. Moreover, by comparing with reference lines $5 \times 10^4 \times \rho^{-2}$ and $2 \times 10^5 \times \rho^{-2}$ in Figs. 2 and 3, respectively, it is seen that, when DC-NOMA multicast is adopted, the outage probabilities of AUs and RUs decay with ρ at rate ρ^{-2} in high-SNR regime. This observation demonstrates that a diversity order of two is provided by DC-NOMA multicast strategy for both AUs and RUs, as shown in Theorem 3. Further, when OC-NOMA multicast strategy is employed, we observe that, in high-SNR regime, the outage probabilities of AUs and RUs decrease at the same rate as those of the reference lines $5 \times 10^7 \times \rho^{-3}$ and $1 \times 10^{12} \times \rho^{-4}$, respectively. Since the number of AUs is $M = 3$, it is concluded that the OC-NOMA multicast strategy achieves a diversity order of M at AUs and a diversity order of $(M + 1)$ at RUs, as demonstrated in Theorem 5.

Fig. 3 also shows the advanced outage probability of RUs. It can be seen that, in terms of advanced outage probability, the RUs have a unit diversity order in both proposed strategies. The reasons are as follows. 1) In both proposed strategies, RUs do not try to decode LP message from the BS's signal due to the weak direct link. 2) In DC-NOMA multicast, the received signals at each RU during the second phase may be a destructive combination of the signals from the relays, and thus, a unit diversity order is achieved. 3) In OC-NOMA multicast, the relay selection criterion (45) does not consider channel gains to those RUs that have successfully decoded HP message from the BS's signal but need to decode LP message from the selected relay's signal. Thus, relay selection only provides a unit diversity when all RUs try to decode LP message from the selected relay's signal. Fig. 3 shows that DC-NOMA multicast has a lower advanced outage probability than the OC-NOMA multicast. This is because both strategies have unit diversity order in terms of advanced outage probability, and thus, more relays participating in cooperation (as DC-NOMA multicast does) benefits reception at RUs.

Fig. 4 shows the derived minimal ECPB of direct NOMA multicast strategy, asymptotic upper bound (marked as "asym. UB") of minimal ECPB of the DC-NOMA multicast strategy, and the asymptotic minimal ECPB of the OC-NOMA multicast strategy, where $M = 3$ and 6, $K = 3$, and $\beta = 1$. As seen from this figure, the actual minimal ECPB of direct NOMA multicast complies with the exact minimal ECPB derived in (9), while the actual minimal ECPB of DC-NOMA multicast and OC-NOMA multicast strategies are tightly upper bounded by the asymptotic upper-bound (37) and asymptotic expression (57), respectively. It can also be seen that the minimal ECPB of direct NOMA multicast strategy increases with the number of AUs. On the other hand, the minimal ECPB of DC-NOMA multicast and OC-NOMA multicast strategies decrease with the number of AUs. This is because, in the proposed cooperation strategies, having more AUs can provide more available relays, thus achieving the target outage probability at a lower transmit SNR.

Fig. 5 demonstrates the energy saving gain of the DC-NOMA multicast strategy and the OC-NOMA multicast strategy over the direct NOMA multicast strategy, where $M = 3$ and 6, $K = 3$, and $\beta = 1$. As observed from this figure, the energy saving gains achieved by both proposed cooperation strategies increase as the target outage probability p_o tends to zero, implying that more energy saving can be attained as the system becomes more reliability-sensitive. In particular, by comparing with reference lines $0.7 \times p_o^{-\frac{1}{2}}$ and $2.3 \times p_o^{-\frac{1}{2}}$, it is seen that the energy saving gain of DC-NOMA multicast strategy increases at rate $p_o^{-\frac{1}{2}}$, as demonstrated in Remark 2. It can also be seen that, for $M = 3$ and $M = 6$, the energy saving gain of OC-NOMA multicast strategy increases at the same rate as those of reference lines $0.8 \times p_o^{-\frac{2}{3}}$ and $1.7 \times p_o^{-\frac{5}{6}}$, respectively. This observation coincides with Remark 4 that the energy saving gain of OC-NOMA multicast scales with p_o at rate $p_o^{-\frac{M-1}{M}}$.

B. Performance Comparison

In this subsection, we compare our proposed cooperation strategies with the following four cooperation strategies. Note that, the first phase of each following strategy is the same as those in the proposed cooperation strategies. Thus, here we only describe the second phase of each strategy.

- **Strategy-1**: This strategy employs the best average reception selection (BARS) scheme in [37], in which the multicast user with best average channel gain from the BS is selected. If the selected multicast user is successful after the first phase, it will serve as a relay in the second phase. In this paper, as the AUs are assumed to serve as relays, the BARS should be performed as $m^\dagger = \arg\max_{m\in\{1,\ldots,M\}} \Omega^f_{S,m}$, namely, the AU that owns the best average channel gain from the BS is selected.
- **Strategy-2**: This strategy randomly picks up at most $Q(\le M)$ successful AUs to serve as relays. If the number of successful AUs is more than Q, a number Q of AUs are randomly selected from all successful AUs to simultaneously forward both HP and LP messages during the second phase. Otherwise, all successful AUs simultaneously forward both messages during the second phase. This strategy represents a wide range of cooperation strategies that do not require CSI. For example, the work in [38] randomly selects a successful receiver to serve as relay, which is Strategy-2 with $Q = 1$. The work in [39] randomly selects up to a fixed number of successful receivers to serve as relays, which is Strategy-2 with $Q > 1$.
- **Strategy-3**: This strategy employs the conventional proactive relay selection (PRS) scheme proposed in [40]. Specifically, the best AU $A_{m^\dagger}$ is proactively selected to maximize the worst reception quality of the two phases, i.e.,

$$m^\dagger = \arg\max_{m=1,\ldots,M} \min\Bigg(|f_{S,m}|^2, \min_{k=1,\ldots,K}|h_{m,k}|^2, \min_{m'=1,\ldots,M,m'\neq m}|g_{m,m'}|^2\Bigg). \quad (63)$$

  If the selected AU is successful after the first phase, it will forward both HP and LP messages during the second phase
- **Strategy-4**: This strategy employs the best user selection (BUS) scheme for fixed power allocation [24], termed as *F-BUS scheme*, which selects the successful AU that maximizes the worst normalized reception quality of unsuccessful AUs and unsuccessful RUs, i.e.,

$$m^\dagger = \arg\max_{m\in\mathcal{S}_A} \min\Bigg(\bigg(\min_{m'\in\{1,\ldots,M\}\setminus\mathcal{S}_A}|g_{m,m'}|^2\bigg) \times \min\bigg(\frac{1}{\varphi_1},\frac{1}{\varphi_2}\bigg), \bigg(\min_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}|h_{m,k}|^2\bigg)\cdot\frac{1}{\varphi_1}\Bigg). \quad (64)$$

  During the second phase, the selected AU will forward both HP and LP messages.

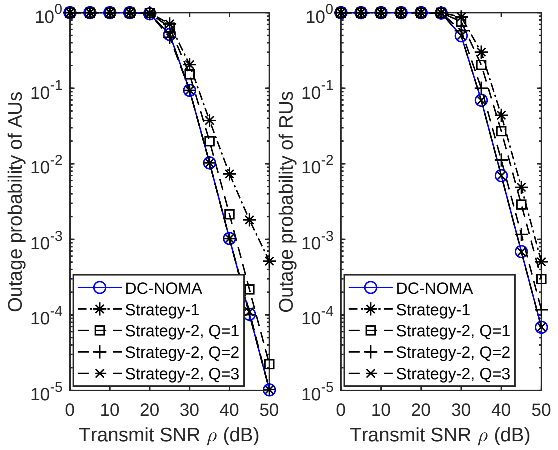


Fig. 6. Outage performance of the proposed DC-NOMA multicast strategy, Strategy-1, and Strategy-2, where $M = 3$ and $K = 6$.

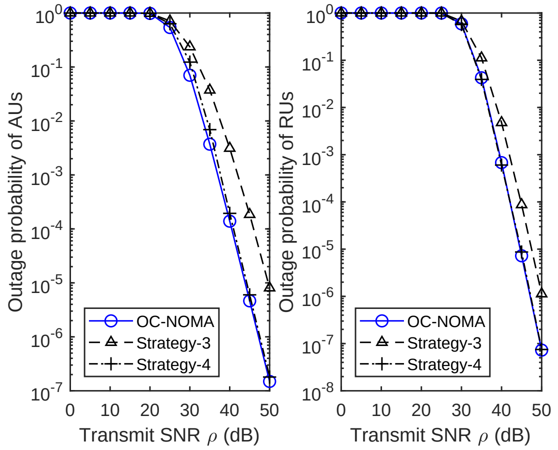


Fig. 7. Outage performance of the proposed OC-NOMA multicast strategy, Strategy-3, and Strategy-4, where $M = 3$ and $K = 6$.

For fairness of comparison, the proposed DC-NOMA multicast is compared with Strategy-1 and Strategy-2, as they do not need instantaneous CSI, while the proposed OC-NOMA multicast is compared with Strategy-3 and Strategy-4, as they all select one of successful AUs based on instantaneous CSI.

Figs. 6 compares the outage performance among the proposed DC-NOMA multicast strategy, Strategy-1, and Strategy-2, while Fig. 7 compares the outage performance among the proposed OC-NOMA multicast strategy, Strategy-3, and Strategy-4. It is seen in both figures that, when $\rho < 25$ dB (low SNR regime), the outage probabilities of AUs and RUs approach one in all strategies, showing that outage events of AUs and RUs always happen in low SNR regime. In Fig. 6, when transmit SNR $\rho \geq 25$ dB, the proposed DC-NOMA multicast strategy always outperforms Strategy-1. This is because the BARS scheme used in Strategy-1 is carried out based on average channel quality, thus always selecting the same AU to forward information. When $\rho \geq 25$ dB, the proposed DC-NOMA multicast strategy outperforms Strategy-2 with $Q = 1$, in terms of outage probabilities of AUs and RUs.

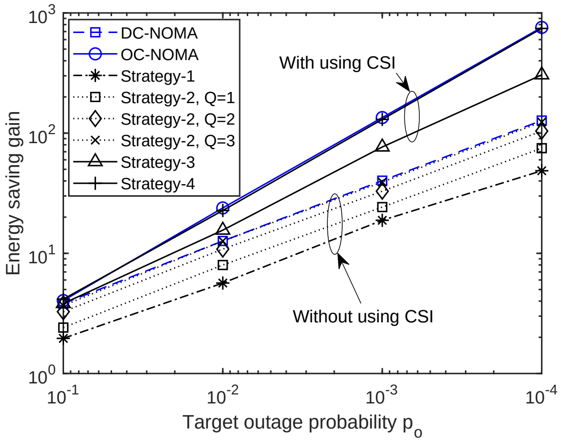


Fig. 8. Energy saving gains of the proposed strategies and Strategy 1∼4 over the direct NOMA multicast strategy, where $M = 3$, $K = 6$ and $\beta = 1$.

Further, as the value of Q increases, the outage performance of Strategy-2 is shown to approach that of the proposed DC-NOMA multicast strategy. Recall that Strategy-2 represents a wide range of cooperation strategies without using CSI. Thus, the proposed DC-NOMA multicast strategy provides the best achievable outage performance for the cooperation strategies without exploiting CSI. Moreover, in Fig. 7, the proposed OC-NOMA multicast strategy always outperforms Strategy-3 for $\rho \geq 25$ dB. Further, compared with Strategy-4, the proposed OC-NOMA multicast achieves a bit lower outage probability for AUs and almost the same outage probability for RUs. These observations can be explained as follows. Recall that the OC-NOMA multicast strategy employs a two-step selection, which ensures the reliability of AUs in the first step, and then, improves the reception quality of RUs in the second step. On the other hand, by observing the selection criterion given in (64), it is known that the F-BUS scheme employed in Strategy-4 targets improving the reception quality of AUs and RUs simultaneously. Therefore, it is expected that a lower outage probability of AUs is achieved by the proposed OC-NOMA multicast strategy.

Fig. 8 compares the energy saving gains of the proposed DC-NOMA multicast strategy, the proposed OC-NOMA multicast strategy, and Strategy 1∼4 over the direct NOMA multicast strategy, where $M = 3$, $K = 6$, and $\beta = 1$. Among the strategies without using instantaneous CSI, the proposed DC-NOMA multicast strategy achieves a higher energy saving gain than Strategy-1 and Strategy-2 with $Q = 1, 2$. Moreover, Strategy-2 with $Q = 3$ achieves the same energy saving gain with the proposed DC-NOMA multicast strategy. This is because, when the number of AUs is $M = 3$, Strategy-2 with $Q = 3$ always recruits all successful AUs as relays, which is same as the proposed DC-NOMA multicast. On the other hand, for the comparison among the strategies with exploiting instantaneous CSI, the OC-NOMA multicast strategy achieves a slightly higher energy saving gain than Strategy-4, while both strategies have higher energy saving gains than that of Strategy-3.

VII. Conclusion

In this paper, we have proposed two cooperative NOMA strategies for layered multicast: the DC-NOMA multicast and OC-NOMA multicast. For each proposed strategy, we have theoretically derived the exact outage probability and achieved diversity order. It has been shown that, the DC-NOMA multicast strategy achieves a diversity order of two, while the OC-NOMA multicast strategy achieves a diversity order not less than the number of AUs. On the other hand, considering that the layered multicast is outage-constrained, we have theoretically evaluated the energy consumption of proposed cooperation strategies and further demonstrated their energy saving gains over the direct NOMA multicast strategy. Finally, numerical results have been provided to validate our analysis and show the superior performance of the proposed cooperation strategies.

Appendix A
Proof of Lemma 1

First, the probability for $\mathbf{O}_1^{\mathrm{AU}}$ happening can be expressed as

$$\Pr\left(\mathbf{O}_1^{\mathrm{AU}}\right) = \Pr\left(\mathcal{S}_A = \emptyset\right) = \prod_{m=1}^{M} \left(1 - \theta_m\right), \qquad \text{(A.1)}$$

where the second equality comes from (20) with letting $\mathcal{A} = \emptyset$. Thus, (23) is obtained.

On the other hand, with the help of Total Probability Theorem [41, sec. 3.3.8], $\Pr(\mathbf{O}_2^{\mathrm{AU}})$ can be expressed as

$$\Pr(\mathbf{O}_2^{\mathrm{AU}}) = \sum_{i=1}^{M-1} \sum_{|\mathcal{A}|=i} \Pr\left(\mathcal{S}_A = \mathcal{A}\right) \times \Pr\left(\bigcup_{m \in \{1,\ldots,M\} \setminus \mathcal{S}_A} \mathbf{o}_{2,m}^{\mathrm{AU}} \middle| \mathcal{S}_A = \mathcal{A} \right). \qquad \text{(A.2)}$$

As the exact expressions for $\Pr\left(\mathcal{S}_A = \mathcal{A}\right)$ has been derived in (20), we only need to further derive the conditional probability $\Pr\left(\bigcup_{m \in \{1,\ldots,M\} \setminus \mathcal{S}_A} \mathbf{o}_{2,m}^{\mathrm{AU}} \middle| \mathcal{S}_A = \mathcal{A} \right)$. Using the SINR/SNR expressions in (16) and (17), $\mathbf{o}_{2,m}^{\mathrm{AU}}$ can be rewritten as

$$\begin{aligned} \mathbf{o}_{2,m}^{\mathrm{AU}} &= \{X_{\mathcal{S}_A,m} < \varphi_1/\mu\rho\} \cup \{X_{\mathcal{S}_A,m} < \varphi_2/\mu\rho\} \\ &= \{X_{\mathcal{S}_A,m} < \max(\varphi_1,\varphi_2)/\mu\rho\}. \end{aligned}$$

Since $g_{l,m} \sim \mathcal{CN}(0, \Omega_{l,m}^g)$, we have $\sum_{l \in \mathcal{S}_A} g_{l,m} \sim \mathcal{CN}(0, \sum_{l \in \mathcal{S}_A} \Omega_{l,m}^g)$, indicating that $X_{\mathcal{S}_A,m} = |\sum_{l \in \mathcal{S}_A} g_{l,m}|^2$ follows exponential distribution with a parameter being $1/\sum_{l \in \mathcal{S}_A} \Omega_{l,m}^g$ [36, sec. 3.2.2]. Thus, the conditional probability $\Pr\left(\cup_{m \in \{1,\ldots,M\} \setminus \mathcal{S}_A} \mathbf{o}_{2,m}^{\mathrm{AU}} \middle| \mathcal{S}_A = \mathcal{A}\right)$ can be derived as

$$\begin{aligned} &\Pr\left(\bigcup_{m \in \{1,\ldots,M\} \setminus \mathcal{S}_A} \mathbf{o}_{2,m}^{\mathrm{AU}} \middle| \mathcal{S}_A = \mathcal{A} \right) \\ &= \Pr\left(\bigcup_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} X_{\mathcal{A},m} < \frac{\max(\varphi_1,\varphi_2)}{\mu\rho} \right) \\ &= 1 - \prod_{m \in \{1,\ldots,M\} \setminus \mathcal{A}} \Pr\left(X_{\mathcal{A},m} \geq \frac{\max(\varphi_1,\varphi_2)}{\mu\rho} \right) \end{aligned}$$

$$=1-\underbrace{e^{-\frac{\max(\varphi_1,\varphi_2)}{\mu\rho}\sum_{m\in\{1,\ldots,M\}\setminus\mathcal{A}}\Lambda^g_{\mathcal{A},m}}}_{=\varpi(\mathcal{A})}, \tag{A.3}$$

where $\Lambda^g_{\mathcal{A},m}=1/\sum_{l\in\mathcal{A}}\Omega^g_{l,m}$. Combining (20) and (A.3) with the help of (A.2), a closed-form expression of $\Pr(\mathbf{O}_2^{\text{AU}})$ is obtained as shown in (24). This completes the proof.

Appendix B
Proof of Lemma 2

The probability of outage event $\mathbf{O}_1^{\text{RU}}$ happening can be expressed as

$$\begin{aligned}P(\mathbf{O}_1^{\text{RU}})=&\Pr\left(|\mathcal{S}_R|<K,\mathcal{S}_A=\emptyset\right)\\=&\Pr\left(|\mathcal{S}_R|<K\right)\Pr\left(\mathcal{S}_A=\emptyset\right),\end{aligned} \tag{B.1}$$

in which $\Pr(\mathcal{S}_A=\emptyset)$ has been derived in (A.1), and $\Pr\left(|\mathcal{S}_R|<K\right)$ can be further expressed as

$$\begin{aligned}\Pr\left(|\mathcal{S}_R|<K\right)=&1-\Pr\left(\mathcal{S}_R=\{1,...,K\}\right)\\=&1-\prod_{k=1}^{K}\zeta_k,\end{aligned} \tag{B.2}$$

where the second equality in (B.2) uses the results in (22) with letting $\mathcal{B}=\{1,...,K\}$. Combining (A.1) and (B.2) with the help of (B.1), (26) is obtained.

On the other hand, the probability $\Pr(\mathbf{O}_2^{\text{RU}})$ can be expressed as

$$\begin{aligned}\Pr(\mathbf{O}_2^{\text{RU}})=&\sum_{i=1}^{M}\sum_{|\mathcal{A}|=i}\sum_{j=0}^{K-1}\sum_{|\mathcal{B}|=j}\Pr\left(\mathcal{S}_A=\mathcal{A}\right)\Pr\left(\mathcal{S}_R=\mathcal{B}\right)\\&\times\Pr\left(\bigcup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathbf{o}_{2,k}^{\text{RU}}\middle|\mathcal{S}_A=\mathcal{A},\mathcal{S}_R=\mathcal{B}\right),\end{aligned} \tag{B.3}$$

in which $\Pr\left(\mathcal{S}_A=\mathcal{A}\right)$ and $\Pr\left(\mathcal{S}_R=\mathcal{B}\right)$ have been derived in (20) and (22), respectively. Thus, we only need to further derive the conditional probability $\Pr\big(\cup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathbf{o}_{2,k}^{\text{RU}}\big|\mathcal{S}_A=\mathcal{A},\mathcal{S}_R=\mathcal{B}\big)$. Using the SINR expression in (18), we have $\mathbf{o}_{2,k}^{\text{RU}}=\{Y_{\mathcal{S}_A,k}<\varphi_1/\mu\rho\}$. As $h_{l,k}\sim\mathcal{CN}(0,\Omega^h_{l,k})$, it is known that $\sum_{l\in\mathcal{R}}h_{l,k}\sim\mathcal{CN}(0,\sum_{l\in\mathcal{S}_A}\Omega^h_{l,k})$ [36, sec. (3.2.2)]. Thus, we have $Y_{\mathcal{S}_A,k}=|\sum_{l\in\mathcal{S}_A}h_{l,k}|^2$ follows exponential distribution with parameter being $1/\sum_{l\in\mathcal{S}_A}\Omega^h_{l,k}$. Then, following the derivations in (A.3), the conditional probability $\Pr\big(\cup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathbf{o}_{2,k}^{\text{RU}}\big|\mathcal{S}_A=\mathcal{A},\mathcal{S}_R=\mathcal{B}\big)$ can be derived as

$$\begin{aligned}&\Pr\left(\bigcup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathbf{o}_{2,k}^{\text{RU}}\middle|\mathcal{S}_A=\mathcal{A},\mathcal{S}_R=\mathcal{B}\right)\\=&\Pr\left(\bigcup_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}Y_{\mathcal{A},k}<\frac{\varphi_1}{\mu\rho}\right)\\=&1-\prod_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}\Pr\left(Y_{\mathcal{C},k}\geq\frac{\varphi_1}{\mu\rho}\right)\\=&1-\underbrace{e^{-\frac{\varphi_1}{\mu\rho}\sum_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}\Lambda^h_{\mathcal{A},k}}}_{\eta(\mathcal{A},\mathcal{B})},\end{aligned} \tag{B.4}$$

where $\Lambda^h_{\mathcal{A},k}=1/\sum_{l\in\mathcal{A}}\Omega^h_{l,k}$. Finally, applying (20), (22) and (B.4) into (B.3), a closed-form expression of $\Pr(\mathbf{O}_2^{\text{RU}})$ is obtained as shown in (27). This completes the proof.

Appendix C
Proof of Lemma 3

First, the probability for $\mathfrak{O}_1^{\text{RU}}$ can be expressed as

$$\begin{aligned}\Pr(\mathfrak{O}_1^{\text{RU}})&=\Pr\left(|\mathcal{S}_R|<K\right)\Pr\left(\mathcal{R}_p=\emptyset\right)\\&=\left[1-\Pr\left(\mathcal{S}_R=\{1,...,K\}\right)\right]\Pr\left(\mathcal{R}_p=\emptyset\right),\end{aligned} \tag{C.1}$$

in which the probability $\Pr(\mathcal{S}_R=\{1,...,K\})$ can be derived by using (22) with letting $\mathcal{B}=\{1,...,K\}$ as

$$\Pr\left(\mathcal{S}_R=\{1,...,K\}\right)=\prod_{k=1}^{K}\zeta_k. \tag{C.2}$$

Further, the probability $\Pr(\mathcal{R}_p=\emptyset)$ can be expressed as

$$\begin{aligned}\Pr(\mathcal{R}_p=\emptyset)=&\sum_{i=0}^{M}\sum_{|\mathcal{A}|=i}\Pr\left(\mathcal{S}_A=\mathcal{A}\right)\Pr\left(\mathcal{R}_p=\emptyset|\mathcal{S}_A=\mathcal{A}\right)\\=&\sum_{i=0}^{M}\sum_{|\mathcal{A}|=i}\left(\prod_{m\in\mathcal{A}}\theta_m[1-\omega_m(\mathcal{A})]\right)\\&\times\left[\prod_{m\in\{1,\ldots,M\}\setminus\mathcal{A}}(1-\theta_m)\right],\end{aligned} \tag{C.3}$$

where the second equality uses the results in (20) and (47). Combining (C.2) and (C.3) with the help of (C.1), (50) is obtained.

On the other hand, based on the Total Probability Theorem [41, sec. 3.3.8], the probability for $\mathfrak{O}_2^{\text{RU}}$ happening can be expressed as

$$\begin{aligned}\Pr(\mathfrak{O}_2^{\text{RU}})=&\Pr\left(|\mathcal{S}_R|<K,\mathcal{R}_p\neq\emptyset,\bigcup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathfrak{o}_{2,k}^{\text{RU}}\right)\\=&\sum_{i=1}^{M}\sum_{|\mathcal{A}|=i}\Pr\left(\mathcal{S}_A=\mathcal{A}\right)\sum_{j=0}^{K-1}\sum_{|\mathcal{B}|=j}\Pr\left(\mathcal{S}_R=\mathcal{B}\right)\\&\times\sum_{t=1}^{i}\sum_{|\mathcal{C}|=t}\Pr\left(\mathcal{R}_p=\mathcal{C}|\mathcal{S}_A=\mathcal{A}\right)\\&\times\Pr\left(\bigcup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathfrak{o}_{2,k}^{\text{RU}}\middle|\mathcal{S}_R=\mathcal{B},\mathcal{R}_p=\mathcal{C}\right),\end{aligned} \tag{C.4}$$

in which the probabilities $\Pr(\mathcal{S}_A=\mathcal{A})$, $\Pr(\mathcal{S}_R=\mathcal{B})$ and $\Pr\left(\mathcal{R}_p=\mathcal{C}|\mathcal{S}_A=\mathcal{A}\right)$ have been derived in (20), (22) and (47), respectively. In the following, we focus on deriving the conditional probability $\Pr\left(\cup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathfrak{o}_{2,k}^{\text{RU}}\big|\,\mathcal{S}_R=\mathcal{B},\mathcal{R}_p=\mathcal{C}\right)$. As $\mathfrak{o}_{2,k}^{\text{RU}}=\{\gamma_{A_{m^\dagger}\to R_k,H}<\tilde{\tau}_H\}=\{|h_{m^\dagger,k}|^2<\varphi_1/\mu\rho\}$, the conditional probability can be derived as

$$\begin{aligned}&\Pr\left(\bigcup_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}\mathfrak{o}_{2,k}^{\text{RU}}\middle|\mathcal{S}_R=\mathcal{B},\mathcal{R}_p=\mathcal{C}\right)\\=&\Pr\left(\min_{k\in\{1,\ldots,K\}\setminus\mathcal{S}_R}|h_{m^\dagger,k}|^2<\frac{\varphi_1}{\mu\rho}\middle|\mathcal{S}_R=\mathcal{B},\mathcal{R}_p=\mathcal{C}\right)\end{aligned}$$

$$
\begin{aligned}
&=\Pr\left(\max_{m\in\mathcal{C}}\min_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}|h_{m,k}|^2<\frac{\varphi_1}{\mu\rho}\right)\\
&=\prod_{m\in\mathcal{C}}\left[1-\prod_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}\Pr\left(|h_{m,k}|^2\geq\frac{\varphi_1}{\mu\rho}\right)\right]\\
&=\prod_{m\in\mathcal{C}}\left(1-\underbrace{e^{-\frac{\varphi_1}{\mu\rho}\sum_{k\in\{1,\ldots,K\}\setminus\mathcal{B}}1/\Omega^h_{m,k}}}_{=\xi_m(\mathcal{B})}\right), \qquad \text{(C.5)}
\end{aligned}
$$

where the second equality comes from (45). Finally, substituting (C.5) into (C.4), and then combining the results with (20), (22) and (47), the expression in (51) is obtained. This completes the proof.

REFERENCES

[1] L. Yang, Q. Ni, L. Lv, J. Chen, X. Xue, H. Zhang, H. Jiang, and J. Shi, "Cooperative NOMA for wireless layered multicast," in *Proc. IEEE Globecom 2018*, accepted.

[2] *Study on Downlink Multiuser Superposition Transmission for LTE*, 3rd Generation Partnership Project, Shanghai, China, Mar. 2015.

[3] "5G radio access: Requirements, concept and technologies," NTT DOCOMO, Inc., Tokyo, Japan, 5G White Paper, Jul. 2014.

[4] *Proposed Solutions for New Radio Access, Mobile and Wireless Communications Enablers for the 2020 Information Society (METIS), Deliverable D.2.4*, METIS, Feb. 2015.

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

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

[7] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[8] Y. Xu, C. Shen, Z. Ding, X. Sun, S. Yan, G. Zhu, and Z. Zhong, "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4874–4886, Sep. 2017.

[9] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in HetNets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5825–5837, Sep. 2017.

[10] Q. Yang, H. M. Wang, D. W. K. Ng, and M. H. Lee, "NOMA in downlink SDMA with limited feedback: performance analysis and optimization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2281–2294, Oct. 2017.

[11] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.

[12] 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.

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

[14] L. Zhang, J. Liu, M. Xiao, G. Wu, Y.-C. Liang, and S. Li, "Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2398–2412, Oct. 2017.

[15] 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.

[16] 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.

[17] R. Jiao, L. Dai, R. MacKenzie, and M. Hao, "On the performance of NOMA-based cooperative relaying systems over Rician fading channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11409–11413, Dec. 2017.

[18] 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.

[19] 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.

[20] Y. Zhou, H. Liu, Z. Pan, L. Tian, J. Shi, and G. Yang, "Two-stage cooperative multicast transmission with optimized power consumption and guaranteed coverage," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 274–284, Feb. 2014.

[21] K. Xiao, L. Gong, and M. Kadoch, "Opportunistic multicast NOMA with security concerns in a 5G massive MIMO system," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 91–95, Mar. 2018.

[22] J. Montalban, P. Scopelliti, M. Fadda, E. Iradier, C. Desogus, P. Angueira, M. Murroni, and G. Araniti, "Multimedia multicast services in 5G networks: Subgrouping and non-orthogonal multiple access techniques," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 96–103, Mar. 2018.

[23] 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.

[24] L. Yang, J. Chen, Q. Ni, J. Shi, and X. Xue, "NOMA-enabled cooperative unicast-multicast: Design and outage analysis," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7870–7889, Dec. 2017.

[25] Q. Zhang, Q. Guo, Q. Ni, W. Zhu, and Y.-Q. Zhang, "Sender-adaptive and receiver-driven layered multicast for scalable video over the Internet," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 4, pp. 482–495, Apr. 2005.

[26] Z. Zhang, Z. Ma, M. Xiao, G. Liu, and P. Fan, "Modeling and analysis of non-orthogonal MBMS transmission in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2221–2237, Oct. 2017.

[27] Z. Zhang, Z. Ma, Y. Xiao, M. Xiao, G. K. Karagiannidis, and P. Fan, "Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1794–1808, Aug. 2017.

[28] Z. Zhang, Z. Ma, X. Lei, M. Xiao, C.-X. Wang, and P. Fan, "Power domain non-orthogonal transmission for cellular mobile broadcasting: Basic scheme, system design, and coverage performance," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 90–99, Apr. 2018.

[29] "Digital video broadcasting (DVB); Framing structure, channel coding and modulation for satellite services to handheld devices (SH) below 3 GHz," Sophia-Antipolis Cedex, France, ETSI EN 302 583 V1.1.1 (2008–03), 2008.

[30] 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.

[31] L. Fay, L. Michael, D. Gomez-Barquero, N. Ammar, and M. W. Caldwell, "An overview of the ATSC 3.0 physical layer specification," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 159–171, Mar. 2016.

[32] P. Patel and J. Holtzman, "Analysis of a simple successive interference cancellation scheme in a DS/CDMA system," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 5, pp. 796–807, June 1994.

[33] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 7th Ed. New York, NY, USA: Academic, 2007.

[34] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[35] 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, July 2016.

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

[37] Y. Chen, L. Wang, and B. Jiao, "Cooperative multicast non-orthogonal multiple access in cognitive radio," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–6.

[38] C.-H. Liu and J. G. Andrews, "Multicast outage probability and transmission capacity of multihop wireless networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4344–4358, July 2011.

[39] 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, July 2010.

[40] A. Bletsas, H. Shin, and M. Z. Win, "Cooperative communications with outage-optimal opportunistic relaying," *IEEE Trans. Wireless Commun.*, vol. 6, no. 9, pp. 3450–3460, Sep. 2007.

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

Long Yang (M'18) received the B.Sc. and Ph.D. degrees from Xidian University, Xi'an, China, in 2010 and 2015, respectively. Since December 2015, he has been a faculty member with Xidian University, Xi'an, China, where he is currently an Lecturer at the School of Telecommunications Engineering. Since November 2017, he has also been a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada. His current research interests include cooperative communications, non-orthogonal multiple access (NOMA), wireless multicast, and wireless physical-layer security.

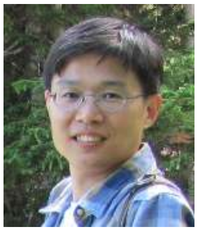
Qiang Ni (M'04-SM'08) received the B.Sc., M.Sc., and Ph.D. degrees from the Huazhong University of Science and Technology, China, all in engineering. He is currently a Professor and the Head of the Communication Systems Group, School of Computing and Communications, Lancaster University, Lancaster, U.K. His research interests include the area of future generation communications and networking, including green communications and networking, cognitive radio network systems, non-orthogonal multiple access (NOMA), heterogeneous networks, 5G and 6G, SDN, cloud networks, energy harvesting, wireless information and power transfer, IoTs, cyber physical systems, machine learning, big data analytics, and vehicular networks. He has authored or co-authored over 200 papers in these areas. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a contributor to the IEEE Wireless Standards.

Lu Lv received his Ph.D degree in Communication and Information Systems from Xidian University, China, in 2018. From March 2016 to September 2016, he was a visiting Ph.D student at the School of Computing and Communications, Lancaster University, UK. From November 2016 to November 2018, he was a visiting Ph.D student at the Department of Electrical and Computer Engineering, University of Alberta, Canada. He is currently a research fellow at the School of Telecommunications Engineering, Xidian University, China. His research interests include cooperative communications, non-orthogonal multiple access, and physical layer security.

Jian Chen (M'14) received the B.Sc. degree from Xi'an Jiaotong University, China, in 1989, the M.S. degree from Xi'an Institute of Optics and Precision Mechanics of Chinese Academy of Sciences in 1992, and the Ph.D. degree in telecommunications engineering in Xidian University, China, in 2005. He is a Professor at the school of Telecommunications Engineering, Xidian University, Xi'an, China. He was a Visiting Scholar at The University of Manchester from 2007 to 2008, and a Senior Visiting Scholar at University of Alberta from 2017 to 2018. His current research interests include cognitive radio, OFDM, wireless sensor networks, non-orthogonal multiple access.

Xuan Xue (M'18) received the B.Sc. 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, London, Ontario, Canada. Since January 2018, she has been a faculty member with Xidian University, Xi'an, China, where she is currently an Lecturer at the School of Telecommunications Engineering. Her research interests include massive MIMO, millimeter-wave communications, cooperative communications, and non-orthogonal multiple access (NOMA).

Hailin Zhang (M'98) received B.Sc. and M.S. degrees from Northwestern Polytechnic University, Xi'an, China, in 1985 and 1988 respectively, and the Ph.D. from Xidian University, Xi'an, China, in 1991. In 1991, he joined School of Telecommunications Engineering, Xidian University, where he is currently a Senior Professor and the Dean of this school. He is also currently the Director of the Key Laboratory in Wireless Communications Sponsored by China Ministry of Information Technology, a Key Member of the State Key Laboratory of Integrated Services Networks, one of the state government specially compensated scientists and engineers, a Field Leader in Telecommunications and Information Systems in Xidian University, an Associate Director for National 111 Project. Dr. Zhang's current research interests include key transmission technologies and standards on broadband wireless communications for 5G wireless access systems. He has published over 100 papers in journals and conferences.



Hai Jiang (SM'15) received the B.Sc. and M.Sc. degrees in electronics engineering from Peking University, Beijing, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, Ontario, Canada, in 2006. Since July 2007, he has been a faculty member with the University of Alberta, Edmonton, Alberta, Canada, where he is currently a Professor at the Department of Electrical and Computer Engineering. His research interests include radio resource management, cognitive radio networking, and cooperative communications.