ARQ Assisted Short-Packet Communications for NOMA Networks Over Nakagami-\(m\) Fading Channels

FENGLIAN CAO\(^1\), YI SONG\(^1,2\), AND YUCUI YANG\(^1,2\)

\(^1\)School of Physics and Electronic Electrical Engineering, Huaiyin Normal University, Huai’an 223001, China
\(^2\)Jiangsu Province Key Construction Laboratory of Modern Measurement Technology and Intelligent System, Huai’an 223300, China

Corresponding author: Fenglian Cao (cfl@hytc.edu.cn)

ABSTRACT Non-orthogonal multiple access (NOMA) is expected as a promising technology for satisfying massive connectivity requirements in Internet of Things (IoT). Besides, short-packet communications are a notable characteristic in IoT for low delay transmissions or massive machine-type communications (MTC). The co-channel interference from NOMA and packet error caused by finite packet will reduce system performance. In view of this, this article adopts automatic repeat request (ARQ) to assist short-packet communications for NOMA networks where a near user (NU) and a far user (FU) are paired to share the same non-orthogonal communication resources. The closed-form expressions for average packet error rate (APER) and effective throughput (ET) of both NU and FU are derived over Nakagami-\(m\) channels. For gaining an insight into system performance, the performance of the scheme without ARQ is also studied. Results show that ARQ improves APER for both NU and FU, and an optimal power allocation factor or packet length exists for achieving higher ET.

INDEX TERMS Automatic repeat request, short-packet communications, non-orthogonal multiple access.

I. INTRODUCTION

Internet of Things (IoT) provides an envision that everything in this world is connected to a unified network. Allowing ubiquitous devices to be connected into networks is one of the major challenges for the development of IoT because the massive devices will impose a heavy burden on the limited radio resources [1]. Non-orthogonal multiple access (NOMA) has been applied to IoT because it can further exploit the power domain for improving network connectivity and spectrum efficiency [2]. On the other hand, the deployment of a large number of machine-type communication (MTC) devices is the main driving force for realizing ubiquitous connectivity [3]. Sensors and other devices involved in MTC usually demand short packet for transmitting several bits to the controller. Besides, in the emerging IoT applications, such as smart transportation system, the control messages with short packet should be transmitted with low delay [4]. Thus, short-packet communications are a typical characteristic of IoT. How to guarantee the short packet to be transmitted reliably is another challenge for IoT.

Power domain NOMA allows multiple users to share the same non-orthogonal communication resources by allocating different powers to users [5]–[8]. The users’ messages with diversified power levels can be distinguished by using successive interference cancellation (SIC). Thus, NOMA is expected to be a promising technique for improving connectivity in IoT [9]. Xiang et al. [10] considered the diversified communication requirements in IoT and adopted NOMA to serve the delay-sensitive user and security-required user on the same non-orthogonal communication resource, simultaneously; the results show that the proposed NOMA scheme outperforms the benchmark OMA scheme. A cellular massive IoT was investigated in [11] where the throughput and energy efficiency are investigated in the NOMA scenarios. Song et al. [12], [13] employed a cognitive radio inspired NOMA scheme to millimeter wave networks for further improving network connectivity. In cooperative IoT networks, NOMA also shows the potential to improve network connectivity [14].

Recently, NOMA also has been presented in short-packet communications for improving spectrum efficiency and reducing transmission delay [15]–[20]. Sun et al. [15] considered a two-user downlink NOMA network with short packet...
constraints, where the transmit rates and power allocation factors are optimized; the results show that NOMA can achieve higher effective throughput (ET) than orthogonal multiple access (OMA) with the same packet length or lower delay with the same ET. Similarly, [16] also demonstrated that NOMA achieves lower delay than OMA when the power allocation factors and block length are optimized. In uplink communications, Choi [18] employed an opportunistic NOMA mode for short-packet communications and the upper bound of the session error probability is derived; results show that the proposed NOMA mode achieves lower error probability than OMA. Cooperative NOMA assisted short-packet communications are investigated in [20], where the cell-edge user cooperating with the central user communicates with the source. The average packet error rate (APER) of the cell-edge user is derived and the results show that the cooperative NOMA scheme achieves lower APER than the traditional NOMA scheme and cooperative OMA scheme.

The key to NOMA is using SIC to distinguish signals. When the SIC process is failed to eliminate the interference, the signal detection performance will be deteriorated seriously [21], [22]. Xiang et al. [23], [24] adopted automatic repeat request (ARQ) to alleviate the SIC error on the detection performance of the NOMA users. Besides, when block length is sufficiently long, the packet error can be made arbitrarily small if the transmit rate is lower than channel capacity. However, due to the constraint of packet length, the packet error cannot be neglected in short-packet communications [25]. Thus, reliable transmission becomes one of the major problems in short-packet communications. In addition to alleviate the effect of SIC error, ARQ can also reduce packet error rate by allowing retransmission.

In our previous work [26], ARQ is adopted to short-packet communications in cooperative networks for improving reliable transmission performance. In this article, we employ ARQ to improve the performance of short-packet communications in NOMA networks. Due to the retransmission bring by ARQ, it will bring extra transmission delay. Thus, only one retransmission is allowed. A similar ARQ scheme is proposed in [22]. However, different from [22] where relay system is considered and infinite packet length is assumed, we focus on recruiting ARQ to improve transmission reliability and efficiency of short-packet communications for NOMA networks. The contributions of this article are briefly summarized as follows.

- We design an ARQ scheme to assist short-packet communications for NOMA networks where a near user (NU) and a far user (FU) are paired to share the same non-orthogonal communication resources. For reducing transmission delay, only one retransmission is allowed for both NU and FU.
- For providing a theory analysis, the closed-form expressions for the APER and ET are derived for both NU and FU over Nakagami-$m$ fading channels. Besides, the performance of the traditional scheme without ARQ is also analyzed.

Results show that ARQ reduces APER for both NU and FU. Besides, simulation results demonstrate that ET is improved in high transmit power region when the channels become stable and an optimal power allocation factor or packet length exists for achieving better ET.

The rest of this article is organized as follows. The network model and the transmission scheme are presented in Section II. The closed-form expressions for the APER and ET of the NU and FU are derived at Section III. Then, in Section IV, simulations are conducted to verify the analysis results and gain an insight into system performance. Finally, concluding remarks are given in Section V.

II. NETWORK MODEL AND TRANSMISSION SCHEME

As illustrated in Fig. 1, we consider a NOMA network where a NU near to the transmitter and an FU far away from the transmitter are paired to occupy the same non-orthogonal communication resources. Considering the short-packet communication characteristics in IoT, the transmitter conveys the short-packet to the users, where the packet size is usually lower than $10^3$ bytes [25]. According to Shannon’s coding theorem, the decoding error is negligible when the code rate is below the channel capacity and the packet-length tends to infinite. However, in short-packet communications, due to the finite packet-length, the decoding error is inevitable. Besides, the SIC is also challenged by the short-packet communications. ARQ provides an alternative for improving transmission reliability. When NU or FU cannot decode its intended messages correctly, it will feed back a negative acknowledgment (NACK) for requesting retransmission. Otherwise, it will feed back an acknowledgment (ACK). The transmitter will stop transmission when it receives ACKs from both NU and FU or the number of retransmission times reaches the maximum allowed numbers [23], [24]. For reducing transmission delay, only one retransmission is allowed. For the ease of notation, the previous transmission is denoted as the first transmission and the retransmission is denoted as the second transmission. We adopt the generalized block Nakagami-$m$ fading to model the channels in the proposed network where the channels between the transmitter and NU and FU are denoted as $h_{N}^{j}$ ($j \in \{1, 2\}$) and $h_{F}^{j}$ ($j \in \{1, 2\}$) with fading parameters $m_{N}$ and $m_{F}$ and means $\Omega_{N}$ and $\Omega_{F}$, respectively. All the nodes in the network are equipped with a single antenna and working in the half-duplex model. The additive white Gaussian noise (AWGN) powers at the users
are assumed as $\sigma^2$. The transmitter requires to transmit $B_N$ and $B_F$ information bits to the NU and FU, respectively.

In short-packet communication systems, the devices generally possess constrained resources. Due to the high complexity of maximal ratio combining (MRC) scheme, selection combining (SC) scheme is adopted in this article to receive multiple transmitted signals. The transmission scheme is detailed as follows.

1) THE FIRST TRANSMISSION SLOT

In the first transmission slot, the transmitter adopts superposition coding to send signal $x_N$ and $x_F$ to the NU and FU, respectively. The received signals at the users are expressed as

$$y_i^1 = h_i^1 \left( \sqrt{P_{\alpha_F} x_F} + \sqrt{P_{\alpha_N} x_N} \right) + n_i,$$

(1)

where $i \in \{F, N\}$, $\alpha_N$ and $\alpha_F$ are the power allocation factors for NU and FU, respectively, $P$ is the transmit power of the transmitter and $n_i$ is the AWGN. According to the NOMA principle [27], more power should be allocated to the user with lower channel gains, i.e., $\alpha_F > \alpha_N$.

The NU should eliminate the interference from $x_F$ first before detecting its intended message. Utilizing (1), the received signal to interference plus noise ratio (SINR) at the NU for decoding $x_F$ is given by

$$\gamma_{N \rightarrow F}^1 = \frac{P_{\alpha_F} |h_F^1|^2}{P_{\alpha_N} |h_N^1|^2 + \sigma^2}.$$  

(2)

As the packet is short, the packet cannot always be detected correctly, and the packet error rate (PER) is denoted as [10]

$$\varepsilon_{N \rightarrow F}^1 = Q \left( \ln 2 \frac{N_F}{V (\gamma_{N \rightarrow F}^1)} \left( \log_2 \left( 1 + \gamma_{N \rightarrow F}^1 \right) - R_F \right) \right),$$  

(3)

where $N_F$ is the packet length of $x_F$, $R_F = B_F / N_F$ is the code rate of $x_F$. $Q (x) = \int_1^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2} dt$, and $V (x) = 1 - (1 + x)^{-2}$.

It is worthy of noting that (3) represents a performance bound. The PER can be also obtained by using practical coding and modulation schemes [28]–[30].

When the NU eliminates the interference from $x_F$ completely, the signal to noise ratio (SNR) for detecting $x_N$ is expressed as

$$\gamma_N^1 = \frac{P_{\alpha_N} |h_N^1|^2}{\sigma^2}.$$  

(4)

Similar to (3), the PER of $x_N$ is given by

$$\varepsilon_N^1 = Q \left( \ln 2 \frac{N_N}{V (\gamma_N^1)} \left( \log_2 \left( 1 + \gamma_N^1 \right) - R_N \right) \right),$$  

(5)

where $N_N$ is the packet length of $x_N$, $R_N = B_N / N_N$ is the code rate of the $x_N$.

When the NU cannot eliminate the interference from $x_F$, the interference will impose a severe effect on the detection of $x_N$. Considering the worst case, the interference will result in packet error directly. Thus, for the NU, the PER is given by

$$\Psi_{N,1} = \varepsilon_{N \rightarrow F}^1 + \left( 1 - \varepsilon_{N \rightarrow F}^1 \right) \varepsilon_N^1.$$  

(6)

In a real communication system, $\varepsilon_{N \rightarrow F}^1$ and $\varepsilon_N^1$ are generally small. Therefore, the PER for the NU can be rewritten as

$$\Psi_{N,1} \approx \min \left\{ 1, \varepsilon_{N \rightarrow F}^1 + \varepsilon_N^1 \right\}.$$  

(7)

When the FU detects its desired message, the signal intended to the NU is viewed as interference. The SINR for $x_F$ is given by

$$\gamma_F^1 = \frac{P_{\alpha_F} |h_F^1|^2}{P_{\alpha_N} |h_N^1|^2 + \sigma^2}.$$  

(8)

Similarly, the PER for FU is expressed as

$$\varepsilon_F^1 = Q \left( \ln 2 \frac{N_F}{V (\gamma_F^1)} \left( \log_2 \left( 1 + \gamma_F^1 \right) - R_F \right) \right).$$  

(9)

As the SIC is not required for the FU, the total PER in the first transmission slot is given by

$$\Psi_{F,1} = \varepsilon_F^1.$$  

(10)

2) THE SECOND TRANSMISSION SLOT

The first transmission cannot always guarantee that the receivers detect their message without error. When packet error occurs, the receivers feed back a NACK for requesting retransmission, otherwise, they feed back an ACK for stopping transmission. According to the receiving condition of the receivers, there are four cases given in Table 1. It is worthy of noting that for improving transmission reliability, all the communication resources will be allocated to one user if another user obtains it messages correctly in the first transmission slot.

In case 1 and case 3, packet error occurs at the NU and the retransmission is requested.

For case 1, in the second transmission slot, the received signals at the NU are expressed as

$$y_N^{1,2} = h_N^2 \left( \sqrt{P_{\alpha_F} x_F} + \sqrt{P_{\alpha_N} x_N} \right) + n_N.$$  

(11)

Similar to the first transmission slot, NU should eliminate the interference from $x_F$ firstly and the SINR for detecting $x_F$
and the PER is expressed as
\[ \epsilon_{\text{PER}}^{1,2} = Q \left( \ln 2 \frac{N_F}{\left( \frac{1}{y_{N,F}^{1,2}} \right)} \left( \log_2 \left( 1 + y_{N,F}^{1,2} \right) - R_F \right) \right). \]  

Similarly, the packet error rate is given by
\[ \epsilon_{\text{PER}}^{1,2} = Q \left( \ln 2 \frac{N_F}{\left( \frac{1}{y_{N,F}^{1,2}} \right)} \left( \log_2 \left( 1 + y_{N,F}^{1,2} \right) - R_F \right) \right). \]  

When \( x_F \) is decoded without error, the interference from \( x_F \) can be eliminated by SIC. Thus, the SNR for \( x_N \) is given by
\[ \gamma_{N}^{1,2} = \frac{P_{\text{tx}} |h_N^2|^2}{\sigma^2}, \]  

and the packet error rate is expressed as
\[ \epsilon_{\text{PER}}^{1,2} = Q \left( \ln 2 \frac{N_F}{\left( \frac{1}{y_{N,F}^{1,2}} \right)} \left( \log_2 \left( 1 + y_{N,F}^{1,2} \right) - R_F \right) \right). \]  

For case 3, the received signals at the NU are expressed as
\[ y_{N}^{3,2} = h_N^2 \sqrt{P_{\text{tx}} x_N} + n_N. \]  

Because the retransmitted signals only contain the messages for the NU, the SNR for \( x_N \) is given by
\[ \gamma_{N}^{3,2} = \frac{P_{\text{tx}} |h_N^2|^2}{\sigma^2}, \]  

and the PER is expressed as
\[ \epsilon_{\text{PER}}^{3,2} = Q \left( \ln 2 \frac{N_F}{\left( \frac{1}{y_{N,F}^{3,2}} \right)} \left( \log_2 \left( 1 + y_{N,F}^{3,2} \right) - R_N \right) \right). \]  

According to case 1 and case 2, the PER in the second transmission slot is given by
\[ \Psi_{N,2} \approx \Psi_{F,1} \min \left\{ 1, \epsilon_{N,F}^{1,2} + \epsilon_{N}^{1,2} \right\} + (1 - \Psi_{F,1}) \epsilon_{N}^{3,2}. \]  

A. AVERAGE PACKET ERROR RATE OF THE NU

When the NU decodes its intended messages with error, it will request for one retransmission. After retransmission, the NU still cannot obtain its intended messages correctly, the packet error is inevitable. Thus, the APER of the NU is given by
\[ \text{APER}_{\text{NU}} = E \left[ \Psi_{N,1} \Psi_{N,2} \right]. \]  

Utilizing (23), the SNR for demodulating \( x_F \) is given by
\[ \gamma_{F}^{2,2} = \frac{P_{\text{tx}} |h_F^2|^2}{\sigma^2}. \]  

and the PER is written as
\[ \epsilon_{F}^{2,2} = Q \left( \ln 2 \frac{N_F}{\left( \frac{1}{y_{F,F}^{2,2}} \right)} \left( \log_2 \left( 1 + y_{F,F}^{2,2} \right) - R_F \right) \right). \]  

According to case 1 and case 2, the PER in the second transmission slot is given by
\[ \Psi_{F,2} = \Psi_{F,1} \epsilon_{F}^{2,2} + (1 - \Psi_{F,1}) \epsilon_{F}^{3,2}. \]  

III. PERFORMANCE ANALYSIS OF THE PROPOSED SCHEME

In this section, we adopt APER and ET to evaluate the reliability and efficiency of the proposed system, respectively. The closed-form expressions for APER and ET of the NU and FU are derived.
Lemma 2: The closed-form expression for \( E[e_N] \) is expressed as

\[
E[e_N] = \left( \frac{1}{2} + g_N \sqrt{N} h_N \right) F_{|h|^2} \left( \frac{q_N}{\rho \alpha_N} \right) + \left( \frac{1}{2} - g_N \sqrt{N} h_N \right) F_{|h|^2} \left( \frac{p_N}{\rho \alpha_N} \right) - \frac{q_N \sqrt{N}}{\rho \alpha_N} \left( \frac{m_N}{\Omega_N} \right)^{-1} \times g_N \left( \gamma \left( m_N + 1, \frac{q_N m_N}{\rho \alpha_N} \Omega_N \right) - \gamma \left( m_N + 1, \frac{p_N m_N}{\rho \alpha_N} \Omega_N \right) \right).
\]  

(31)

where \( h_N = 2^h - 1, g_N = (2 \pi (2^{2h} - 1))^{-\frac{1}{2}}, p_N = h_N + \frac{1}{2} g_N \left( N - \frac{1}{2} N^\frac{1}{2} \right), q_N = h_N + \frac{1}{2} g_N \left( N - \frac{1}{2} N^\frac{1}{2} \right), \) and \( \gamma (\alpha, x) = \int_0^x e^{-t} t^{\alpha-1} dt. \)

Proof: See Appendix B.

Comparing (3) and (13), we can observe that \( E[e_N] \) and \( E[e_N^2] \) have the same expression. Thus, we can deduce that \( E[e_N^2] = E[e_N] \). Besides, (5) and (15) also have the same expression and \( E[e_N^2] = E[e_N] \). Furthermore, we can also find that (18) has a similar expression as (5). The closed-form expression for \( E[e_N^2] \) can be derived by replacing \( \alpha_N \) by 1 in (31). The closed-form expression for \( E[e_N^2] \) is expressed as

\[
E[e_N^2] = \left( \frac{1}{2} + g_N \sqrt{N} h_N \right) F_{|h|^2} \left( \frac{q_N}{\rho} \right) + \left( \frac{1}{2} - g_N \sqrt{N} h_N \right) F_{|h|^2} \left( \frac{p_N}{\rho} \right) - \frac{q_N \sqrt{N}}{\rho} \left( \frac{m_N}{\Omega_N} \right)^{-1} \times g_N \left( \gamma \left( m_N + 1, \frac{q_N m_N}{\rho} \Omega_N \right) - \gamma \left( m_N + 1, \frac{p_N m_N}{\rho} \Omega_N \right) \right).
\]  

(32)

Observing (9) and (3), we can find that \( E[e_F] \) has a similar expression as \( E[e_N] \). The closed-form expression for \( E[e_F] \) can be derived by replacing \( m_N \) by \( m_F \), \( \Omega_N \) by \( \Omega_F \), \( \mu_N \) by \( \mu_F \), and \( F_{|h|^2}(x) \) by \( F_{|h|^2}(x) \) in (30). The closed-form expression for \( E[e_F] \) is expressed as (33), as shown at the bottom of the next page, where \( \mu_F = \frac{m_F}{\rho \alpha_N \Omega_F} \) and \( F_{|h|^2}(x) = \left( 1 - \sum_{r=0}^{m_F-1} \left( \frac{m_F}{\mu_F} \right)_r \frac{1}{r!} e^{-\frac{m_F}{\mu_F}} \right) \). Substituting \( E[e_{N-F}], E[e_N], E[e_{N-F}], E[e_{N-F}], E[e_{N-F}^2], \) and \( E[e_{N-F}^2] \) into (29), the closed-form expression for the APER of the NU is obtained.

From (30) and (31), we can find that APER of the NU is a decreasing function of transmit power or packet length, which demonstrates that improving transmit power or increasing packet length will improve transmission reliability.

Besides, in traditional networks without ARQ, the APER of NU is expressed as

\[
A_{\text{non-ARQ}} = \min \left\{ 1, E[e_N], E[e_N^2] + E[e_N] \right\}.
\]  

(34)

Compared (34) with (29), we can find that \( P_{\text{non-ARQ}}^\text{NU} > P_{\text{ARQ}}^\text{NU} \). It demonstrates that ARQ is beneficial to reliable transmissions.

**B. AVERAGE PACKET ERROR RATE OF THE FU**

Similar to the NU, if the FU cannot obtain its intended messages correctly it will request retransmission. When the FU still cannot detect its desired messages in the second transmission slot, packet error occurs. The average packet error rate of the FU is expressed as

\[
A_{\text{per}}^\text{FU} = E\left[ \Psi_{\text{F},1} \Psi_{\text{F},2} \right].
\]  

(35)

Utilizing (10) and (26), (35) can be rewritten as

\[
A_{\text{per}}^\text{FU} = E\left[ e_F \left( \Psi_{\text{F},1} e_F^{1.2} + (1 - \Psi_{\text{F},1}) e_F^{2.2} \right) \right].
\]  

(36)

Due to the independence between the two transmission slots, (36) can be further expressed as

\[
A_{\text{per}}^\text{FU} = E\left[ e_F \right] E\left[ e_F^{1.2} \right] \min \left\{ 1, E\left[ e_{N-F} \right] + E\left[ e_N \right] \right\} + E\left[ e_F \right] E\left[ e_F^{2.2} \right] \left( 1 - \min \left\{ 1, E\left[ e_{N-F} \right] + E\left[ e_N \right] \right\} \right).
\]  

(37)

We can also find that \( E[e_F^{1.2}] \) has the same expression as \( E[e_F] \). Thus, \( E[e_F^{1.2}] = E[e_F] \). Besides, observing (26) and (18), we can also find that \( E[e_F^{2.2}] \) has the same expression to \( E[e_F^2] \). After replacing \( g_N \) by \( g_F \), \( N_N \) by \( N_F \), \( h_N \) by \( h_F \), \( q_N \) by \( q_F \), \( p_N \) by \( p_F \), \( m_N \) by \( m_F \), \( \Omega_N \) by \( \Omega_F \), and
\[ E \left[ \frac{2.2}{F} \right] \]

\[ = \left( \frac{1}{2} + g_F \sqrt{N_F h_F} \right) F_{\left| \beta F \right|^2} \left( \frac{\gamma F}{\rho} \right) \]

\[ + \left( \frac{1}{2} - g_F \sqrt{N_F h_F} \right) F_{\left| \beta F \right|^2} \left( \frac{p_F}{\rho} - g_F \frac{\sqrt{N_F}}{\Gamma(m_F)} \frac{m_F}{\Omega_F} \right)^{-1} \]

\[ \times \left( \gamma \left( m_F + 1, \frac{q_F m_F}{\rho \Omega_F} \right) - \gamma \left( m_F + 1, \frac{p_F m_F}{\rho \Omega_F} \right) \right), \quad (38) \]

Substituting \( E \left[ \frac{\epsilon F}{\beta} \right], E \left[ \frac{\epsilon 1.2 F}{\beta} \right], \) and \( E \left[ \frac{2.2 F}{\beta} \right] \) into (37), the closed-form expression for the APER of the FU is obtained.

Similar to the NU, the APER of the FU also can be decreased by improving transmit power or increasing packet length. It shows that high transmit power and long packet help to improve transmission reliability.

In the tradition system without ARQ, the APER is given by

\[ A_{\text{non-ARQ}}^{\text{FU}} = E \left[ \frac{\epsilon 1 F}{\beta} \right]. \quad (39) \]

Comparing (39) with (37), we can also find that ARQ improves transmission reliability.

C. EFFECTIVE THROUGHPUT OF THE NU

The above analysis shows that ARQ improves transmission reliability. However, the transmission efficiency will be decreased when the same message is transmitted multiple times. ET is adopted as the metric for evaluating the transmission efficiency, which is formulated as

\[ \eta_N = R_N \left( 1 - E \left[ \Psi_{N,1} \right] \right) \left( 1 - E \left[ \Psi_{F,1} \right] \right) \]

\[ + \frac{1}{2} R_N \left( 1 - E \left[ \Psi_{N,1} \right] \right) E \left[ \Psi_{F,1} \right] \]

\[ + \frac{1}{2} R_N E \left[ \Psi_{N,1} \right] \left( 1 - E \left[ \Psi_{N,2} \right] \right). \quad (40) \]

Utilizing (7), (10) and (19), (40) can be further expressed as

\[ \eta_N = R_N \left( 1 - \min \left\{ 1, E \left[ \epsilon_{N,F} \right] + E \left[ \epsilon_{1,F} \right] \right\} \right) \left( 1 - E \left[ \epsilon_{1,F} \right] \right) \]

\[ + \frac{1}{2} R_N \left( 1 - \min \left\{ 1, E \left[ \epsilon_{N,F} \right] + E \left[ \epsilon_{1,F} \right] \right\} \right) E \left[ \epsilon_{1,F} \right] \]

\[ + \frac{1}{2} R_N \min \left\{ 1, E \left[ \epsilon_{N,F} \right] + E \left[ \epsilon_{1,F} \right] \right\}. \]

D. EFFECTIVE THROUGHPUT OF THE FU

Similarly, the ET of the FU is expressed as

\[ E \left[ \epsilon_{1,F} \right] = \left( \frac{1}{2} - g_F h_F \sqrt{N_F} \right) \left( U \left( p_F - \frac{\alpha_F}{\alpha_N} \right) + U \left( \frac{\alpha_F}{\alpha_N} - p_F \right) \right) F_{\left| \beta F \right|^2} \left( \varphi \left( p_F \right) \right) \]

\[ + \left( \frac{1}{2} + g_F h_F \sqrt{N_F} \right) \left( U \left( q_F - \frac{\alpha_F}{\alpha_N} \right) + U \left( \frac{\alpha_F}{\alpha_N} - q_F \right) \right) F_{\left| \beta F \right|^2} \left( \varphi \left( q_F \right) \right) \]

\[ \times \sum_{s=0}^{m_F} \left( m_F \right)_s \left( m_F \right)^{m_F} \rho \alpha_F g_F h_F \sqrt{N_F} \left( -1 \right)^{m_F-s} \frac{\mu_N^s}{\left( \mu_N \alpha_N \right)^{m_F+s+1}} \frac{\Gamma \left( s, \mu_N u \left( p_N \right) \right)}{\Gamma \left( s, \mu_F u \left( q_N \right) \right)} \]. \quad (33)
From (44) and (45), we can find that improving transmit power or packet length will improve the transmission efficiency of the FU. Besides, when $\eta_{N,1}$ is small, the proposed scheme with ARQ achieves better performance than the traditional scheme.

### IV. NUMERICAL RESULTS

In this section, we provide numerical results to verify our analysis results and investigate the effects of the system parameters on the reliability and efficiency performance. Part of the simulation parameters is given in Table 1. In the simulations, the units for ET and packet length are bit per channel use and channel use, respectively.

Fig. 2 plots the APER versus $P$ with $m = 2$ and $\alpha_N = 0.2$, in the proposed ARQ assisted scheme and traditional scheme without ARQ. Firstly, we can find that the analysis points match well with the simulation curves, which verifies the correctness of the derivations. Secondly, the APERs of both NU and FU drop dramatically with the increase of $P$ in both the ARQ assisted scheme and traditional scheme without ARQ. It shows that high transmit power at the relay benefits to the reliable transmission of NOMA in short-packet communications. We can also find that with the increase of packet length, the APER also decreases, which demonstrates that short-packet deteriorates the reliable transmission. However, Fig. 2 shows that the proposed ARQ assisted scheme achieves better reliability performance.

Fig. 3 depicts APER versus $\alpha_F/\alpha_N$ with $m = 1$ and $P = 10$dBm, in the proposed ARQ assisted scheme and traditional scheme without ARQ. We can obtain from Fig. 3 that the APER of the NU first decreases and then increases with the increase of the power allocation factor ratio $\alpha_F/\alpha_N$. However, the APER of the FU keeps dropping when increasing $\alpha_F/\alpha_N$. It can be explained that increase $\alpha_F/\alpha_N$ will improve the power allocated to FU, which will improve the reliability of the FU and the SIC performance of the NU. Thus, when $\alpha_F/\alpha_N$ is small, increasing $\alpha_F/\alpha_N$ will decrease APER of the NU. Increasing $\alpha_F/\alpha_N$ also decreases the power allocated to the NU which will increase the APER when $\alpha_F/\alpha_N$ is bigger enough. Ultimately, Fig. 3 also shows that the proposed ARQ assisted scheme achieves better reliability performance than the traditional scheme without ARQ for different $\alpha_F/\alpha_N$.

Fig. 4 illustrates the ET versus $P$ with $N = 150$ and $\alpha_N = 0.2$, in the schemes with and without ARQ. We can find that...
the ET keeps increasing with the improvement of the transmit power, which shows that high transmit power also improves transmission efficiency. Due to the fixed packet length and the number of information bits, improving transmit power cannot increase ET infinitely. Besides, in the low transmit power region, the ET of the FU is higher than that of the NU because the NU is allocated less power and the performance of the NU is restricted in the low transmit power region. However, in the high transmit power region, due to the better channel condition of the NU, it achieves higher ET than that of the FU. We can also find that increasing the fading parameter $m$ will improve ET in the high transmit power region. On the contrary, in the low transmit power region, the fluctuation of channels will improve transmission efficiency. Ultimately, the scheme with ARQ shows the advantage of improving the ET of the NU. It is because that the FU does not need to employ SIC and easy to obtain its message correctly in the first transmission slot, while the retransmission is required to the NU.

Fig. 5 plots ET versus $\alpha_F \backslash \alpha_N$ with $N = 150$ and $P = 10$dBm, in the proposed ARQ assisted scheme. Firstly, we can observe from Fig. 5 that the ET first increases and then decreases with the increase of $\alpha_F \backslash \alpha_N$ for both NU and FU. Increasing $\alpha_F \backslash \alpha_N$ firstly improves the SIC performance at the NU which will increase ET. However, keeping increasing $\alpha_F \backslash \alpha_N$ will reduce the power allocated to the NU and increasing the number of transmission times, resulting in the decreasing of ET at the NU. Although the APER keeps decreasing when increasing $\alpha_F \backslash \alpha_N$, the number of transmission times also increases which decreases the ET of FU, which reduces the ET of the FU. It also demonstrates that an optimal power allocation factor exists for achieving the highest ET of both NU and FU. We can also find that the fading parameters can improve the ET for both the NU and FU.

In Fig. 6, we plots the ET versus $N$ in the SC and MRC schemes for $P_T = 8$dBm and $P_R = 4$dBm. First of all, we can find that the ET first increases and then decreases with the increase of $N$ for both NU and FU. From Fig. 2 we can obtain that increasing packet length $N$ will improve transmission reliability. However, for the fixed information bit required to transmit, increasing $N$ will reduce the transmit rate, which will decrease ET. We can also find that increasing transmit power from 5dBm to 15dBm will improve ET for different $N$. It demonstrates that high transmit power is beneficial to improve transmission efficiency.

V. CONCLUSION

This article adopts ARQ to improve the short-packet communication performance for NOMA networks, where the NU and FU are paired to share the same non-orthogonal communication resources. The closed-form expressions for the APER and ET are derived over Nakagami-$m$ fading channels. Results show that ARQ improves the transmission reliability for both the NU and FU. Besides, an optimal power allocation factor exists for the NU for achieving the lowest APER and the optimization method will be left as future work. The results also demonstrate that increasing fading parameters the ET will be improved in the high transmit power region. Ultimately, the packet length can be optimized for obtaining the highest ET of both NU and FU.

APPENDIX A

PROOF OF LEMA 1

Using (3), $E \left[ \epsilon_{N \rightarrow F} \right]$ can be approximated as [15]

$$E \left[ \epsilon_{N \rightarrow F} \right] = \begin{cases} 0, & x \geq q_F \\ \frac{1}{2} - g_F \sqrt{N_F} (x - h_F), & p_F < x < q_F \\ 1, & x \leq p_F. \end{cases} \quad (46)$$

Using (2) and (46), the average block error rate is expressed as

$$E \left[ \epsilon_{N \rightarrow F} \right] = \int_0^\infty \epsilon_{N \rightarrow F} \left( \frac{\rho \alpha_F}{\rho \alpha_N} + 1 \right) \frac{f_{|h_N|^2}}{h_N} (x) dx, \quad (48)$$
where \( f_{|h_N|^2}(x) = \left( \frac{m_N}{\Omega_N} \right)^{m_N} \frac{x^{m_N-1} e^{-\frac{m_N x}{\Omega_N}}}{\Gamma(m_N)} \) is the probability density function (PDF) of \(|h_N|^2\).

According to (46), the closed-form expression derivation about (48) is divided into three conditions.

**Condition 1:** When \( \frac{m_N}{\Omega_N} \leq p_F \), \( E\left[\varepsilon_{N,F}^1\right] = 1 \), and (48) is rewritten as

\[
E\left[\varepsilon_{N,F}^1\right] = \int_0^{\infty} f_{|h_N|^2}(x) dx. \tag{49}
\]

In this condition, we can obtain \( E\left[\varepsilon_{N,F}^1\right] = 1 \).

**Condition 2:** When \( \frac{m_N}{\Omega_N} > q_N \), (46) is given by (47), as shown at the bottom of the page, where \( F_{|h_N|^2}(x) \) is the cumulative distribution function (CDF) of \(|h_N|^2\). Utilizing [31, eq.(3.381.3)], \( \Xi_2 \) is derived as

\[
\Xi_2 = \sum_{s=0}^{m_N} \left( \frac{m_N}{s} \right) \left( \frac{m_N}{\Omega_N} \right)^s \frac{\rho \alpha x}{\rho \alpha N x + 1} \frac{\sqrt{N}}{\mu N} \left( \Gamma(s, \mu N u(q_F)) - \Gamma(s + 1, \mu N u(q_F)) \right). \tag{51}
\]

**Condition 3:** When \( p_F < \frac{m_N}{\Omega_N} < q_F \), (46) is rewritten as

\[
\varepsilon_{N,F}^1 = \begin{cases} 
1, & x \leq \varphi(p_F) \\
\frac{1}{2} - g_N \sqrt{N} (g_F(x) - h_F), & \varphi(p_F) < x \leq \varphi(q_F) \\
0, & x > \varphi(q_F)
\end{cases} \tag{52}
\]

where \( g(x) = \frac{\rho \alpha x}{\rho \alpha N x + 1} \).

Substituting (52) into (48) and after some mathematical manipulations, the average block error rate is given by

\[
E\left[\varepsilon_{N,F}^1\right] = \int_0^{\infty} f_{|h_N|^2}(x) dx
\]

for \( \varphi(p_N) \)

\[
+ \int_{\varphi(p_N)}^{\infty} \varepsilon_{N,F}^1 f_{|h_N|^2}(x) dx. \tag{53}
\]


Following the derivation in **Condition 2**, the closed-form expression for (53) is similarly derived as

\[
E\left[\varepsilon_{N,F}^1\right] = \frac{1}{2} + g_F h_F \sqrt{N} f_{|h_N|^2}(\varphi(q_F)) + \frac{1}{2} - g_F h_F \sqrt{N} F_{|h_N|^2}(\varphi(p_F))
\]

\[
\Xi_1 = \int_{\varphi(p_N)}^{\infty} \left( \frac{\rho \alpha x}{\rho \alpha N x + 1} \right)^{m_N} \frac{e^{-\frac{m_N x}{\Omega_N}}}{\Gamma(m_N)} dx. \tag{47}
\]

Combining the above three conditions, the closed-form expression of \( E\left[\varepsilon_{N,F}^1\right] \) is obtained in (30).

**APPENDIX B**

**PROOF OF LEMMA 2**

Utilizing (5), \( \varepsilon_R^1 \) can be approximated as [15]

\[
\varepsilon_R^1 = \begin{cases} 
1, & y_N^1 \leq p_N \\
\frac{1}{2} - g_N \sqrt{N} (y_N^1 - h_N), & p_N < y_N^1 < q_N \\
0, & y_N^1 \geq q_N
\end{cases} \tag{54}
\]

Utilizing (4) and (55), the average block error rate can be expressed as

\[
E\left[\varepsilon_R^1\right] = \int_0^{\infty} \varepsilon_R^1 f_{|h_N|^2}(x) dx, \tag{56}
\]

where \( f_{|h_N|^2}(x) = \left( \frac{m_N}{\Omega_N} \right)^s \frac{\rho \alpha x}{\rho \alpha N x + 1} e^{-\frac{m_N x}{\Omega_N}} \) is the probability density function (PDF) of \(|h_N|^2\).

According to (55), \( E\left[\varepsilon_R^1\right] \) can be further expressed as

\[
E\left[\varepsilon_R^1\right] = \int_0^{\frac{m_N}{\Omega_N}} \frac{1}{2} - g_N \sqrt{N} (\rho \alpha x - h_N) f_{|h_N|^2}(x) dx
\]

\[
+ \int_0^{\frac{m_N}{\Omega_N}} \varepsilon_R^1 f_{|h_N|^2}(x) dx. \tag{57}
\]

Substituting the PDF of \(|h_N|^2\) into (57), which can be expressed as

\[
E\left[\varepsilon_R^1\right] = \left( \frac{1}{2} + g_N \sqrt{N} h_N \right) F_{|h_N|^2} \left( \frac{q_N}{\rho \alpha N} \right)
\]

\[
\Xi_2 = \int_{\varphi(p_N)}^{\infty} \left( \frac{\rho \alpha x}{\rho \alpha N x + 1} \right)^{m_N} \frac{e^{-\frac{m_N x}{\Omega_N}}}{\Gamma(m_N)} dx. \tag{47}
\]
Utilizing [31, eq.(3.351.1)], $E\left[ r_1^2 \right]$ is derived as (31).

REFERENCES

[1] Y. Qu, S. Yu, W. Zhou, S. Peng, G. Wang, and K. Xiao, “Privacy of things: Emerging challenges and opportunities in wireless Internet of Things,” IEEE Wireless Commun., vol. 25, no. 6, pp. 91–97, Dec. 2018.

[2] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, “Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations,” IEEE Commun. Mag., vol. 55, no. 9, pp. 55–61, Sep. 2017.

[3] A. Mukherjee, “Physical-layer security in the Internet of Things: Sensing and communication confidentiality under resource constraints,” Proc. IEEE, vol. 103, no. 10, pp. 1747–1761, Oct. 2015.

[4] G. Durisi, T. Koch, and P. Popovski, “Toward massive, ultra-reliable, and low-latency wireless communication with short packets,” Proc. IEEE, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[5] Z. Xiang, W. Yang, G. Pan, Y. Cai, and Y. Song, “Physical layer security in cognitive radio inspired NOMA network,” IEEE J. Sel. Topics Signal Process., vol. 13, no. 3, pp. 700–714, Jun. 2019.

[6] W. Wu, F. Zhou, R. Q. Hu, and B. Wang, “Energy-efficient resource allocation for secure NOMA-enabled mobile edge computing networks,” IEEE Trans. Commun., vol. 68, no. 1, pp. 493–505, Jan. 2020.

[7] W. Wu, X. Yin, P. Deng, T. Guo, and B. Wang, “Transceiver design for downlink SWIPT NOMA systems with cooperative full-duplex relaying,” IEEE Access, vol. 7, pp. 33464–33472, 2019.

[8] F. Zhou, Y. Wu, Y.-C. Liang, Z. Li, Y. Wang, and K.-K. Wong, “State of the art, taxonomy, and open issues on cognitive radio networks with NOMA,” IEEE Wireless Commun., vol. 25, no. 2, pp. 100–108, Apr. 2018.

[9] Z. Ding, L. Dai, and H. V. Poor, “MIMO-NOMA design for small packet transmission in the Internet of Things,” IEEE Access, vol. 4, pp. 1393–1405, 2016.

[10] Z. Xiang, W. Yang, Y. Cai, J. Xiong, Z. Ding, and Y. Song, “Secure transmission in a NOMA assisted IoT network with diversified communication requirements,” IEEE Internet Things J., early access, May 19, 2020, doi: 10.1109/JIOT.2020.2995609.

[11] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, “On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT,” IEEE J. Sel. Areas Commun., vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[12] Y. Song, W. Yang, X. Yang, Z. Xiang, and B. Wang, “Physical layer security in cognitive millimeter wave networks,” IEEE Access, vol. 7, pp. 109162–109180, 2019.

[13] Y. Song, W. Yang, Z. Xiang, N. Sha, H. Wang, and Y. Yang, “An analysis on secure millimeter wave NOMA communications in cognitive radio networks,” IEEE Access, vol. 8, pp. 78965–78978, 2020.

[14] M. Alkhawatreh, Y. Gong, G. Chen, S. Lambotharan, and J. A. Chambers, “Buffer-aided relay selection for cooperative NOMA in the Internet of Things,” IEEE Internet Things J., vol. 6, no. 3, pp. 5722–5731, Jun. 2019.

[15] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, “Short-packet downlink transmission with non-orthogonal multiple access,” IEEE Wireless Commun., vol. 17, no. 7, pp. 4550–4564, Jul. 2018.

[16] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, “On the performance of non-orthogonal multiple access in short-packet communications,” IEEE Commun. Lett., vol. 22, no. 3, pp. 590–593, Mar. 2018.

[17] J. Zheng, Q. Zhang, and J. Qin, “Average block error rate of downlink NOMA short-packet communication systems in Nakagami-m fading channels,” IEEE Commun. Lett., vol. 23, no. 10, pp. 1712–1716, Oct. 2019.

[18] J. Choi, “Opportunistic NOMA for uplink short-message delivery with a delay constraint,” IEEE Trans. Wireless Commun., vol. 19, no. 6, pp. 3727–3737, Jun. 2020.

[19] Z. Wang, T. Lv, Z. Lin, J. Zeng, and P. T. Mathiopoulos, “Outage performance of URLLC NOMA systems with wireless power transfer,” IEEE Wireless Commun. Lett., vol. 9, no. 3, pp. 380–384, Mar. 2020.

[20] X. Lai, Q. Zhang, and J. Qin, “Cooperative NOMA short-packet communications in flat Rayleigh fading channels,” IEEE Trans. Veh. Technol., vol. 68, no. 6, pp. 6182–6186, Jun. 2019.

[21] Z. Xiang, W. Yang, Y. Cai, Z. Ding, Y. Song, and Y. Zou, “NOMA-assisted secure short-packet communications in IoT,” IEEE Wireless Commun., vol. 27, no. 4, pp. 8–15, Aug. 2020.

[22] Z. Yu, C. Zhai, and J. Liu, “Non-orthogonal multiple access relaying with truncated ARQ,” IET Commun., vol. 11, no. 4, pp. 514–521, Mar. 2017.

[23] Z. Xiang, W. Yang, G. Pan, Y. Cai, Y. Song, and Y. Zou, “Secure transmission in HARQ-assisted non-orthogonal multiple access networks,” IEEE Trans. Inf. Forensics Security, vol. 15, pp. 2171–2182, 2020.

[24] Z. Xiang, W. Yang, Y. Cai, Z. Ding, and Y. Song, “Secure transmission design in HARQ assisted cognitive NOMA networks,” IEEE Trans. Inf. Forensics Security, vol. 15, pp. 2528–2541, 2020.

[25] H.-M. Wang, Q. Yang, Z. Ding, and H. V. Poor, “Secure short-packet communications for mission-critical IoT applications,” IEEE Trans. Wireless Commun., vol. 18, no. 5, pp. 2565–2578, May 2019.

[26] Y. Yang, Y. Song, and F. Cao, “HARQ assisted short-packet communications for cooperative networks over Nakagami-m fading channels,” IEEE Access, to be published.

[27] 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.

[28] H. Mukhtar, A. Al-Dweik, M. Al-Mualla, and A. Shami, “Low complexity power optimization algorithm for multimedia transmission over wireless networks,” IEEE J. Sel. Topics Signal Process., vol. 9, no. 1, pp. 113–124, Feb. 2015.

[29] T. Assaf, A. Al-Dweik, M. E. Moursi, and H. Zeineidln, “Exact BER performance analysis for downlink NOMA systems over Nakagami-m fading channels,” IEEE Access, vol. 7, pp. 134539–134555, 2019.

[30] T. Assaf, A. Al-Dweik, M. ElMoursi, H. Zeineidln, and M. Al-Jarah, “Exact bit error-rate analysis of two-user NOMA using QAM with arbitrary modulation orders,” 2020, techRxiv:12310169.v1. [Online]. Available: https://doi.org/10.36227/techrxiv.12310169.v1

[31] I. S. Gradstein and I. M. Ryzhik, Table of Integral, Series and Products, 7th ed. New York, NY, USA: Academic, 2007.

FENGLIAN CAO received the M.S. degree from the School of Electronic Science and Engineering, Nanjing University, Nanjing, in 2006. She is currently working with the School of Physics and Electronic Electrical Engineering, Huaqian Normal University. Her current research interests include cooperative communications, non-orthogonal multiple access, and physical-layer security.

YI SONG received the M.S. degree from the Nanjing University of Aeronautics and Astronautics, in 2011, and the Ph.D. degree from the Institution of Communications Engineering, Army Engineering University of PLA, in 2020. He is currently an Associate Professor with the School of Physics and Electronic Electrical Engineering, Huaqian Normal University, and the Jiangsu Province Key Construction Laboratory of Modern Measurement Technology and Intelligent System. His research interests include cooperative communications, non-orthogonal multiple access, and cognitive radio.

YUCUI YANG received the M.S. degree from the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, in 2006. She is currently with the School of Physics and Electronic Electrical Engineering, Huaqian Normal University, and the Jiangsu Province Key Construction Laboratory of Modern Measurement Technology and Intelligent System. Her current research interests include cooperative communications, pattern recognition, and image retrieval.

VOLUME 8, 2020

158272