Wireless Communications and Mobile Computing

Architectures, Challenges and Opportunities within 6G Emerging Technologies

Lead Guest Editor: Wei Duan

Guest Editors: Miaowen Wen and Shahid Mumtaz

Architectures, Challenges and Opportunities within 6G Emerging Technologies

Wireless Communications and Mobile Computing

Architectures, Challenges and Opportunities within 6G Emerging Technologies

Lead Guest Editor: Wei Duan
Guest Editors: Miaowen Wen and Shahid Mumtaz

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9803249, 1 page
https://doi.org/10.1155/2021/9803249

Corrigendum

Corrigendum to "Collaborative Computing and Resource Allocation for LEO Satellite-Assisted Internet of Things"

Tao Leng,[1,2] Xiaoyao Li,[1,2] Dongwei Hu,[3] Gaofeng Cui,[1,2,3] and Weidong Wang[1,2]

[1] *School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2] *Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3] *Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang, China*

Correspondence should be addressed to Gaofeng Cui; cuigaofeng@bupt.edu.cn

Received 6 December 2021; Accepted 6 December 2021; Published 23 December 2021

In the article titled "Collaborative Computing and Resource Allocation for LEO Satellite-Assisted Internet of Things" [1], a statement noting the equal contribution of authors Tao Leng and Xiaoyao Li was omitted in error.

Authors' Contributions

Tao Leng and Xiaoyao Li contributed equally to this work.

References

[1] T. Leng, X. Li, D. Hu, G. Cui, and W. Wang, "Collaborative computing and resource allocation for LEO satellite-assisted internet of things," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 4212548, 12 pages, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5519786, 9 pages
https://doi.org/10.1155/2021/5519786

Research Article

Spatial Property of Optical Wave Propagation through Anisotropic Atmospheric Turbulence

Bing Guan,[1] Haiyang Yu,[1] Wei Song,[2] and Jaeho Choi[3]

[1] *School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China*
[2] *Department of Electronic Information and Communication Engineering, Applied Technology College of Soochow University, Suzhou 215325, China*
[3] *Department of Electronic Engineering, CAIIT, Jeonbuk National University, Jeonju 54896, Republic of Korea*

Correspondence should be addressed to Wei Song; songw3015@suda.edu.cn

Received 31 January 2021; Revised 10 June 2021; Accepted 22 November 2021; Published 21 December 2021

Academic Editor: Xiangjie Kong

For the free-space optical (FSO) communication system, the spatial coherence of a laser beam is influenced obviously as it propagates through the atmosphere. This loss of spatial coherence limits the degree to which the laser beam is collimated or focused, resulting in a significant decrease in the power level of optical communication and radar systems. In this work, the analytic expressions of wave structure function for plane and spherical wave propagation through anisotropic non-Kolmogorov turbulence in a horizontal path are derived. Moreover, the new expressions for spatial coherence radius are obtained considering different scales of atmospheric turbulence. Using the newly obtained expressions for the spatial coherent radius, the effects of the inner scales and the outer scales of the turbulence, the power law exponent, and the anisotropic factor are analyzed. The analytical simulation results show that the wave structure functions are greatly influenced by the power law exponent α, the anisotropic factor ζ, the turbulence strength $\tilde{\sigma}_R^2$, and the turbulence scales. Moreover, the spatial coherence radiuses are also significantly affected by the anisotropic factor ζ and the turbulence strength $\tilde{\sigma}_R^2$, while they are gently influenced by the power law exponent α and the inner scales of the optical waves.

1. Introduction

In recent years, the traffic carried by the telecommunication network is growing significantly, especially the wireless network. The popularity of wireless data and mobile Internet is much faster than anyone imagined, and it enhances voice communication with much richer multimedia content. Mobile data traffic and mobile service spectrum have increased by many orders of magnitude from 2010 to 2020 [1]. This is an important topic of the sixth generation (6G) wireless communication. On 6G communication, optical wireless communication (OWC) technology has many advantages in frequency spectrum, security, and transmission rate, which can be used as a potential replacement and supplement of radio frequency-based wireless communication technology. OWC technology provides a basic combination of the various advantages necessary to deliver high-speed services to optical backbone networks. It provides an unlicensed spectrum, almost unlimited data rate, low-cost development, and convenient installation [2]. On the other hand, the ground-based point-to-point OWC system, also known as free-space optical (FSO) communication system, works at near-infrared frequency. Beam spreading caused by atmospheric turbulence occupies a very important position in FSO communication systems, because it determines the loss of power at the receiver plane [3]. For optical wave propagation, the classic Kolmogorov model has been widely used in theoretical researches due to its simple mathematical form [3–5]. Over the years, the Kolmogorov model is extended and several non-Kolmogorov turbulence models have been also proposed [6–13]. Toselli et al. [6] is one of them, and they analyze the angle of arrival fluctuations by using the generalized exponent factor α instead of the standard exponent value 11/3. The anisotropic factor is also used to describe anisotropy of the atmosphere turbulence [7], and the generalized non-Kolmogorov von Karman spectrum of the anisotropic atmospheric turbulence is available [8–10]. In addition, there are also numerous studies on beam wanders, loss of spatial coherence, temporal frequency spread, and the angle of arrival fluctuation [14–21], which are all related to the random fluctuation of optical waves propagating through random media.

Lately, more research attention is drawn to the theoretical survey of wave structure function (WSF) for the long-exposure modulation transfer function (MTF) and spatial coherence radius (SCR) [22–29]. Based on the Rytov approximation method, a researcher like Young proposes new expressions for the WSFs of optical waves, which fit the moderate to strong fluctuation regimes [22, 23]. Lu et al. have derived new expressions for the WSFs and the SCRs for plane waves and spherical waves propagating through homogeneous and isotropic oceanic turbulence [25]. Moreover, Cui et al. consider the turbulence scales and have derived the long-exposure MTFs for plane waves and spherical waves propagating through anisotropic non-Kolmogorov atmospheric turbulence [26, 27]; Kotiang and Choi and Guan et al. also have derived a new long-exposure MTF for Gaussian waves propagating through isotropic non-Kolmogorov atmospheric turbulence and anisotropic maritime turbulence [28, 29].

In this study, we derive new WSF and SCR expressions for the plane waves and the spherical waves which propagate in the anisotropic non-Kolmogorov atmosphere turbulence. Here, the generalized von Karman model is used by incorporating the anisotropic factor $\zeta^{2-\alpha}$. In the simulation analyses, using the newly derived WSF and SCR expressions, the effects of the inner scale and the outer scale of the eddy size are investigated. In addition to the influences of the power law exponent, the turbulence strength and the anisotropic factor, which are all affecting parameters of the WSFs and SCRs, are also carefully analyzed.

2. Anisotropic Non-Kolmogorov Spectrum with Inner and Outer Scales

Stribling et al. [30] developed a power spectrum in non-Kolmogorov turbulence as the power law for the spectrum of the index of refraction fluctuations is varied from 3 to 4. This power spectrum, which we call the conventional isotropic non-Kolmogorov spectrum:

$$\begin{aligned} \Phi_n(\kappa,\alpha) &= A(\alpha)\tilde{C}_n^2\kappa^{-\alpha}\ (\kappa>0, 3<\alpha<4), \\ A(\alpha) &= \frac{\Gamma(\alpha-1)}{4\pi^2}\cos\left(\frac{\alpha\pi}{2}\right), \end{aligned} \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function, κ is the spatial wave number, α is the power law exponent, and $\tilde{C}_n^2$ is the generalized structure parameter with $m^{3-\alpha}$ as its unit. The function $A(\alpha)$ maintains the consistency between the index structure function and its power spectrum.

Toselli in [7] proposed a new power spectrum by introducing an effective anisotropic parameter ζ. When there are non-Kolmogorov power law and anisotropy along the propagation direction, it is helpful to simulate optical turbulence. Also, the concept of atmospheric turbulence anisotropy at different scales is introduced:

$$\begin{aligned} \Phi_n(\kappa,\alpha,\zeta) = {} & A(\alpha)\tilde{C}_n^2\zeta^2\left[\zeta^2\kappa_{xy}^2+\kappa_z^2+\kappa_0^2\right]^{-\alpha/2} \\ & \cdot \exp\left(-\frac{\zeta^2\kappa_{xy}^2+\kappa_z^2}{\kappa_m^2}\right)(\kappa>0, 3<\alpha<4), \end{aligned} \tag{2}$$

where ζ is the anisotropic factor; $\kappa_0 = 2\pi/L_0$ and L_0 is the outer scale parameter; $\kappa_m = C(\alpha)/l_0$ and l_0 is the inner scale parameter; $\kappa = \sqrt{\zeta^2(\kappa_x^2+\kappa_y^2)+\kappa_z^2} = \sqrt{\zeta^2\kappa_{xy}^2+\kappa_z^2}$ and κ_x, κ_y, and κ_z are the components of κ in $x, y,$ and z direction; and $C(\alpha) = \left[\Gamma((5-\alpha)/2)A(\alpha)2/3\pi\right]^{1/(\alpha-5)}$.

In this case, Equation (2) can be defined as the one in [6], which is formed by multiplying the generalized von Karman model with the anisotropic factor $\zeta^{2-\alpha}$. The resulting expression is then the modified anisotropic non-Kolmogorov power spectrum, and it is defined as follows [31]:

$$\begin{aligned} \Phi_n(\kappa,\alpha,\zeta) = {} & A(\alpha)\tilde{C}_n^2\zeta^{2-\alpha}\left[\kappa^2+\tilde{\kappa}_0^2\right]^{-\alpha/2} \\ & \cdot \exp\left(-\frac{\kappa^2}{\tilde{\kappa}_m^2}\right)(\kappa>0, 3<\alpha<4), \end{aligned} \tag{3}$$

where $\tilde{\kappa}_0^2 = \kappa_0^2/\zeta^2$ and $\tilde{\kappa}_m^2 = \kappa_m^2/\zeta^2$.

3. The Expressions for Wave Structure Functions

In order to evaluate the performance of optical wave structure function and spatial coherence radius in atmospheric turbulence, we need to derive the wave structure function of optical waves first. In this paper, we first derive the wave structure function of plane wave and spherical wave and then use these equations to derive their spatial coherence radius.

3.1. The Plane Wave Structure Function. The WSF of optical waves propagating in isotropic non-Kolmogorov turbulence is proposed in [3], and it can be used to calculate the spatial coherence radius of the optical waves; the formulae are shown as follows:

$$D_{\mathrm{pl}}(\rho,\alpha) = 8\pi^2k^2L\int_0^\infty \kappa\Phi_n(\kappa,\alpha)\left[1-J_0(\kappa\rho)\right]d\kappa, \tag{4}$$

$$D_{\mathrm{sp}}(\rho,\alpha) = 8\pi^2k^2L\int_0^1\int_0^\infty \kappa\Phi_n(\kappa,\alpha)\left[1-J_0(\kappa\xi\rho)\right]d\kappa d\xi, \tag{5}$$

where $D_{\mathrm{pl}}(\rho,\alpha)$ is the plane wave structure function; $D_{\mathrm{sp}}(\rho,\alpha)$ is the spherical wave structure function; ρ is the scalar separation distance between two points in the 2D plane; $k = 2\pi/\lambda$ is the optical wave number; $\xi = 1-z/L$ and L is the path length; z is the propagation distance; and $J_0(\cdot)$ is the zero-order Bessel function of the first kind.

In this paper, Equation (3) is used as the expression for the anisotropic non-Kolmogorov power spectrum in our derivation of new wave structure function expressions for optical waves. Then, Equations (4) and (5) can be rewritten as follows:

$$D_{\mathrm{pl}}(\rho,\alpha,\zeta)=8\pi^2k^2L\int_0^\infty \kappa\Phi_n(\kappa,\alpha,\zeta)[1-J_0(\kappa\rho)]d\kappa, \tag{6}$$

$$D_{\mathrm{sp}}(\rho,\alpha,\zeta)=8\pi^2k^2L\int_0^1\int_0^\infty \kappa\Phi_n(\kappa,\alpha,\zeta)[1-J_0(\kappa\xi\rho)]d\kappa d\xi. \tag{7}$$

By substituting Equation (3) into Equation (6) and expanding $J_0(\cdot)$ as a series representation, one can obtain the new WSF expression for the plane waves as follows:

$$\begin{aligned}D_{\mathrm{pl}}(\rho,\alpha,\zeta)&=8\pi^2k^2LA(\alpha)\tilde{C}_n^2\zeta^{2-\alpha}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n!\cdot(1)_n}\cdot\left(\frac{\rho}{2}\right)^{2n}\\&\times\int_0^\infty\kappa^{2n+1}\left[\kappa^2+\tilde{\kappa}_0^2\right]^{-\alpha/2}\cdot\exp\left(-\frac{\kappa^2}{\tilde{\kappa}_m^2}\right)d\kappa.\end{aligned} \tag{8}$$

Here, the integration can be resolved by using the confluent hypergeometric function of the second kind defined as follows [32]:

$$U(a,c,z)=\frac{1}{\Gamma(a)}\int_0^\infty e^{-zt}t^{a-1}(1+t)^{c-a-1}dt,\quad a>0, \mathrm{Re}\ (z)>0, \tag{9}$$

$$U(a,c,z)\approx\frac{\Gamma(1-c)}{\Gamma(1+a-c)}+\frac{\Gamma(c-1)}{\Gamma(a)}z^{1-c},\quad |z|<<1. \tag{10}$$

Then, Equation (8) can be derived as follows:

$$\begin{aligned}D_{\mathrm{pl}}(\rho,\alpha,\zeta)&=4\pi^2k^2LA(\alpha)\tilde{C}_n^2\zeta^{2-\alpha}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n!\cdot(1)_n}\cdot\left(\frac{\rho}{2}\right)^{2n}\tilde{\kappa}_0^{2n+2-\alpha}\\&\times\left\{\frac{n!(-1)^n\Gamma(\alpha/2-1)}{(2-\alpha/2)_n\Gamma(\alpha/2)}+\left(1-\frac{\alpha}{2}\right)_n\Gamma\left(1-\frac{\alpha}{2}\right)\frac{\tilde{\kappa}_0^{\alpha-2n-2}}{\tilde{\kappa}_m^{\alpha-2n-2}}\right\}.\end{aligned} \tag{11}$$

Finally, the new expression of the wave structure function for the plane waves is defined as follows:

$$\begin{aligned}D_{\mathrm{pl}}(\rho,\alpha,\zeta)&=4\pi^2k^2LA(\alpha)\tilde{C}_n^2\zeta^{2-\alpha}\left\{\tilde{\kappa}_m^{2-\alpha}\Gamma\left(1-\frac{\alpha}{2}\right)[1-J(\alpha)]\right.\\&\left.+\frac{\rho^2\tilde{\kappa}_0^{4-\alpha}}{(\alpha-2)(\alpha-4)}\right\},\end{aligned} \tag{12}$$

$$J(\alpha)={}_1F_1\left(1-\frac{\alpha}{2};1;-\frac{\rho^2\tilde{\kappa}_m^2}{4}\right), \tag{13}$$

where ${}_1F_1(\cdot)$ is the confluent hypergeometric function of the first kind [32].

3.2. The Spherical Wave Structure Function. By substituting Equation (3) into Equation (7) and expanding $J_0(\cdot)$ as a series representation, one can obtain the new expression of the wave structure function for the spherical waves. It is expressed as follows:

$$\begin{aligned}D_{\mathrm{sp}}(\rho,\alpha,\zeta)&=8\pi^2k^2LA(\alpha)\tilde{C}_n^2\zeta^{2-\alpha}\sum_{n=0}^{\infty}\frac{(-1)^{n-1}}{n!(1)_n}\left(\frac{\rho}{2}\right)^{2n}\\&\times\int_0^1\xi^{2n}d\xi\int_0^\infty\kappa^{2n+1}\left[\kappa^2+\tilde{\kappa}_0^2\right]^{-\alpha/2}\\&\cdot\exp\left(-\frac{\kappa^2}{\tilde{\kappa}_m^2}\right)d\kappa d\xi.\end{aligned} \tag{14}$$

Using Equations (9) and (10), the final expression of the WSF for the spherical waves can be derived and it is rewritten as follows:

$$\begin{aligned}D_{\mathrm{sp}}(\rho,\alpha,\zeta)&=4\pi^2k^2LA(\alpha)\tilde{C}_n^2\zeta^{2-\alpha}\left\{\tilde{\kappa}_m^{2-\alpha}\Gamma\left(1-\frac{\alpha}{2}\right)[1-K(\alpha)]\right.\\&\left.+\frac{\rho^2\tilde{\kappa}_0^{4-\alpha}}{3(\alpha-2)(\alpha-4)}\right\},\end{aligned} \tag{15}$$

$$K(\alpha)={}_2F_2\left(\frac{1}{2},1-\frac{\alpha}{2};\frac{3}{2},1;-\frac{\rho^2\tilde{\kappa}_m^2}{4}\right), \tag{16}$$

where ${}_pF_q(\cdot)$ is the generalized hypergeometric function and p and q are nonnegative integers [32].

4. New Expressions for Spatial Coherence Radius

The main purpose of this section is to calculate the spatial coherence radius of the optical waves. It is used to determine the spatial coherence radius of the wave at the receiver pupil plane. As we know, the spatial coherence radius defines the effective receiver aperture size in a heterodyne detection system. The new expressions for SCRs of the plane waves and the spherical waves are derived in this section. The SCR derivations begin with the new WSFs derived in the previous section. Those WSFs need to be approximated and simplified to be used effectively in numerical computations when one performs computer simulations.

Consider the WSFs in Equations (12) and (15). Those equations involve the confluent hypergeometric functions ${}_1F_1(\cdot)$ and ${}_2F_2(\cdot)$. Those hypergeometric function can be approximately expressed as follows [32]:

$${}_1F_1(a;c;-z)\approx\begin{cases}1-\dfrac{az}{c}, & |z|<<1,\\[2ex] \dfrac{\Gamma(c)}{\Gamma(c-a)}z^{-a}, & \mathrm{Re}\ (z)>>1,\end{cases} \tag{17}$$

$$
{}_2F_2(a, b\,; c, d;-z) \approx \begin{cases} 1-\dfrac{abz}{cd}, & |z|<<1, \\ \dfrac{\Gamma(c)\Gamma(d)\Gamma(b-a)}{\Gamma(b)\Gamma(c-a)\Gamma(d-a)}z^{-a}+\dfrac{\Gamma(c)\Gamma(d)\Gamma(a-b)}{\Gamma(a)\Gamma(c-b)\Gamma(d-b)}z^{-b}, & \mathrm{Re}\,(z)>>1. \end{cases} \tag{18}
$$

Equations (17) and (18) can be substituted into Equations (12) and (15), respectively. The WSFs of the optical waves can be rewritten as follows:

$$
D_{\mathrm{pl}}(\rho, \alpha, \zeta) \approx \begin{cases} R(\alpha)\tilde{\sigma}_R^2(\alpha)\left[\tilde{\kappa}_m^{2-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)-\dfrac{2^{2-\alpha}\Gamma(1-\alpha/2)}{\Gamma(\alpha/2)}\rho^{\alpha-2}-\dfrac{\rho^2\tilde{\kappa}_0^{4-\alpha}}{(\alpha-2)(4-\alpha)}\right], & \rho>>l_0, \\ R(\alpha)\tilde{\sigma}_R^2(\alpha)\rho^2\left[\dfrac{2-\alpha}{8}\tilde{\kappa}_m^{4-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)-\dfrac{\tilde{\kappa}_0^{4-\alpha}}{(\alpha-2)(4-\alpha)}\right], & \rho<<l_0, \end{cases} \tag{19}
$$

$$
D_{\mathrm{sp}}(\rho, \alpha, \zeta) \approx \begin{cases} R(\alpha)\tilde{\sigma}_R^2(\alpha)\left[\tilde{\kappa}_m^{2-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)-\dfrac{2^{2-\alpha}\Gamma(1-\alpha/2)}{(\alpha-1)\Gamma(\alpha/2)}\rho^{\alpha-2}-\dfrac{\rho^2\tilde{\kappa}_0^{4-\alpha}}{3(\alpha-2)(4-\alpha)}\right], & \rho>>l_0, \\ R(\alpha)\tilde{\sigma}_R^2(\alpha)\rho^2\left[\dfrac{2-\alpha}{24}\tilde{\kappa}_m^{4-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)-\dfrac{\tilde{\kappa}_0^{4-\alpha}}{3(\alpha-2)(4-\alpha)}\right], & \rho<<l_0, \end{cases} \tag{20}
$$

where

$$
R(\alpha)=-0.5\alpha\left(\sin\frac{\pi\alpha}{4}\right)^{-1}\zeta^{2-\alpha}k^{\alpha/2-1}L^{1-\alpha/2}\left[\Gamma\left(1-\frac{\alpha}{2}\right)\right]^{-1}, \tag{21}
$$

and $\tilde{\sigma}_R^2(\alpha)$ is the non-Kolmogorov Rytov variance defined by the plane wave scintillation index in non-Kolmogorov turbulence [33] as follows:

$$
\tilde{\sigma}_R^2(\alpha)=-8\pi^2A(\alpha)\frac{1}{\alpha}\Gamma\left(1-\frac{\alpha}{2}\right)\tilde{C}_n^2k^{3-\alpha/2}L^{\alpha/2}\sin\frac{\pi\alpha}{4}. \tag{22}
$$

From the WSF of the optical waves, the SCR ρ_0 is defined by the $1/e$ point of the complex degree of coherence [3] and $D(\rho_0, L)=2$.

Based on the approximation expression of the WSF for the plane waves defined in Equation (19), the new SCR expression of plane waves for the case of $L_0=\infty$ is derived, and it is defined as follows:

$$
\rho_0 \equiv \rho_{\mathrm{pl}} \approx \begin{cases} \left\{\dfrac{2^{\alpha-2}\Gamma(\alpha/2)}{\Gamma(1-\alpha/2)}\left[\tilde{\kappa}_m^{2-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)-\dfrac{2}{R(\alpha)\tilde{\sigma}_R^2(\alpha)}\right]\right\}^{1/(\alpha-2)}, & l_0<<\rho_{\mathrm{pl}}<<L_0, \\ \left\{R(\alpha)\tilde{\sigma}_R^2(\alpha)\dfrac{2-\alpha}{16}\tilde{\kappa}_m^{4-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)\right\}^{-1/2}, & \rho_{\mathrm{pl}}<<l_0<<L_0. \end{cases} \tag{23}
$$

Similarly, based on the approximation expression of the WSF for the spherical waves defined in Equation (20), the new SCR expression of the spherical waves for the case of $L_0=\infty$ is derived, and it is defined as follows:

$$
\rho_0 \equiv \rho_{\mathrm{sp}} \approx \begin{cases} \left\{\dfrac{(\alpha-1)2^{\alpha-2}\Gamma(\alpha/2)}{\Gamma(1-\alpha/2)}\left[\tilde{\kappa}_m^{2-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)-\dfrac{2}{R(\alpha)\tilde{\sigma}_R^2(\alpha)}\right]\right\}^{1/(\alpha-2)}, & l_0<<\rho_{\mathrm{sp}}<<L_0, \\ \left\{R(\alpha)\tilde{\sigma}_R^2(\alpha)\dfrac{2-\alpha}{48}\tilde{\kappa}_m^{4-\alpha}\Gamma\left(1-\dfrac{\alpha}{2}\right)\right\}^{-1/2}, & \rho_{\mathrm{sp}}<<l_0<<L_0, \end{cases} \tag{24}
$$

while $\alpha=11/3$ and $\zeta=1$, which makes a special case of the well-known isotropic Kolmogorov turbulence, Equations (23) and (24) can be reduced to the expressions as follows:

$$
\rho_0 \equiv \rho_{\mathrm{pl}} \approx \begin{cases} \left(1.46\tilde{C}_n^2k^2L\right)^{-3/5}, & l_0<<\rho_{\mathrm{pl}}<<L_0, \\ \left(1.64\tilde{C}_n^2k^2Ll_0^{-1/3}\right)^{-1/2}, & \rho_{\mathrm{pl}}<<l_0<<L_0, \end{cases} \tag{25}
$$

$$
\rho_0 \equiv \rho_{\mathrm{sp}} \approx \begin{cases} \left(0.55\tilde{C}_n^2k^2L\right)^{-3/5}, & l_0<<\rho_{\mathrm{sp}}<<L_0, \\ \left(0.55\tilde{C}_n^2k^2Ll_0^{-1/3}\right)^{-1/2}, & \rho_{\mathrm{sp}}<<l_0<<L_0, \end{cases} \tag{26}
$$

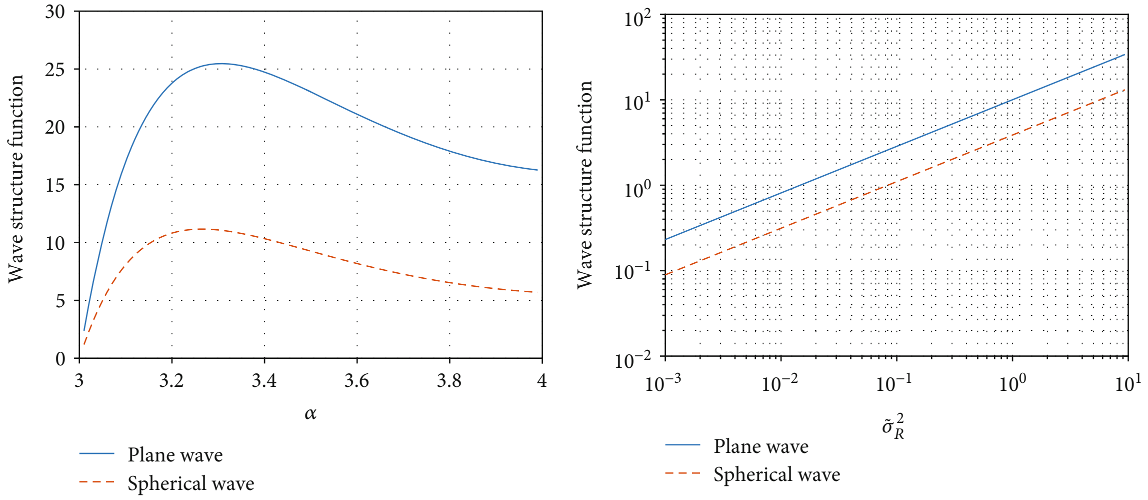
and the result of Equations (25) and (26) is consistent with the research in [3].

5. Evaluations Using Numerical Analysis

Based on the above derived new analytic expressions, we analyze the wave structure function and spatial coherence radius for the plane and spherical wave propagation in anisotropic non-Kolmogorov turbulence. There are two sets of new expressions derived, and they are set for evaluation with respect to various characterizing parameters. Those are WSFs defined in Equations (12) and (15) and SCRs defined in Equations (23) and (24), respectively. We have made some general assumptions in the numerical simulations: the optical waves propagate with the generalized structure parameter $\tilde{C}_n^2=1.4\times10^{-14}m^{3-\alpha}$; the wavelength $\lambda=1.65\times1$ 0^{-6} m; the scalar separation distance is $\rho=3$ cm, the inner scale of the eddy size is 1 mm, and the outer scale of the eddy size is 10 m; the optical path lengths vary from 100 m to 8 km; the power law exponent α varies from 3 to 4; and the case of $l_0< <\rho<<L_0$ is used for the SCR simulations.

5.1. Evaluations on WSFs. The first set of simulations are performed using the new expressions of wave structure function defined in Equations (12) and (15). The focus of the evaluation is to analyze the behaviors of the WSFs in terms of various characterization parameters. Those include the power law exponent α, the turbulence strength $\tilde{\sigma}_R^2$, and the anisotropic factor ζ.

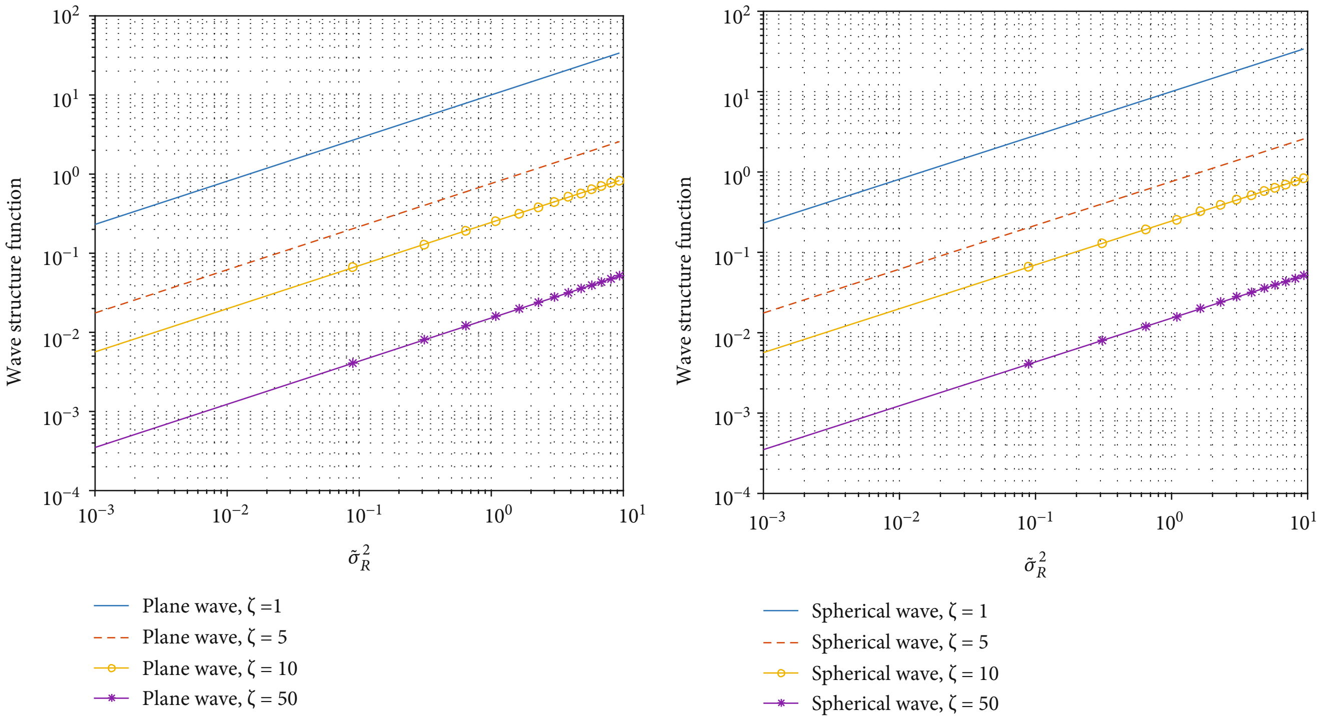
Figure 1(a) shows the behavior of the WSF with respect to the increasing power law exponent α, when the anisotropic factor $\zeta=1$, which actually makes the turbulence isotropic. The WSFs increase when α varies from 3 to 3.3 and then decrease gently afterwards. One can see the smooth bumps that the WSFs get to their maximum when $\alpha\approx3.3$. On the other hand, Figure 1(b) shows the behaviors of the WSF with respect to the turbulence strength $\tilde{\sigma}_R^2$. The WSFs monotonically increase as the turbulence strength increases. One can observe that the values of WSF for the plane waves are always bigger than those of spherical waves in both figures.



(a)

(b)

Figure 1: WSF as a function of increasing (a) power exponent α and (b) turbulence strength.



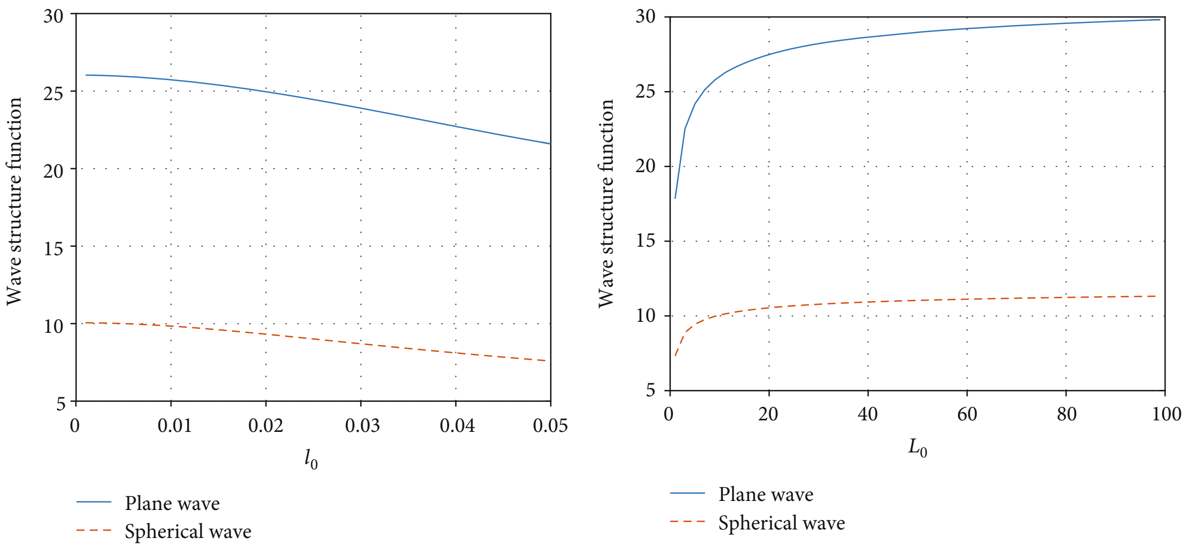
(a)

(b)

Figure 2: WSF as a function of increasing turbulence strength for a varying anisotropic factor: (a) plane wave and (b) spherical wave.

Figure 2 shows the behavior of the WSF with respect to the increasing turbulence strength $\tilde{\sigma}_R^2$ and the increasing anisotropic factor ζ. Figure 2(a) is for the plane waves, and Figure 2(b) is for the spherical waves. In both cases, the WSFs increase as the turbulence strength gets stronger. Also, it is also important to note that the WSFs are significantly influenced by the anisotropic factor ζ that the strength of the WSFs gets as much as 400 times weaker as the anisotropy increases from 1 to 50.

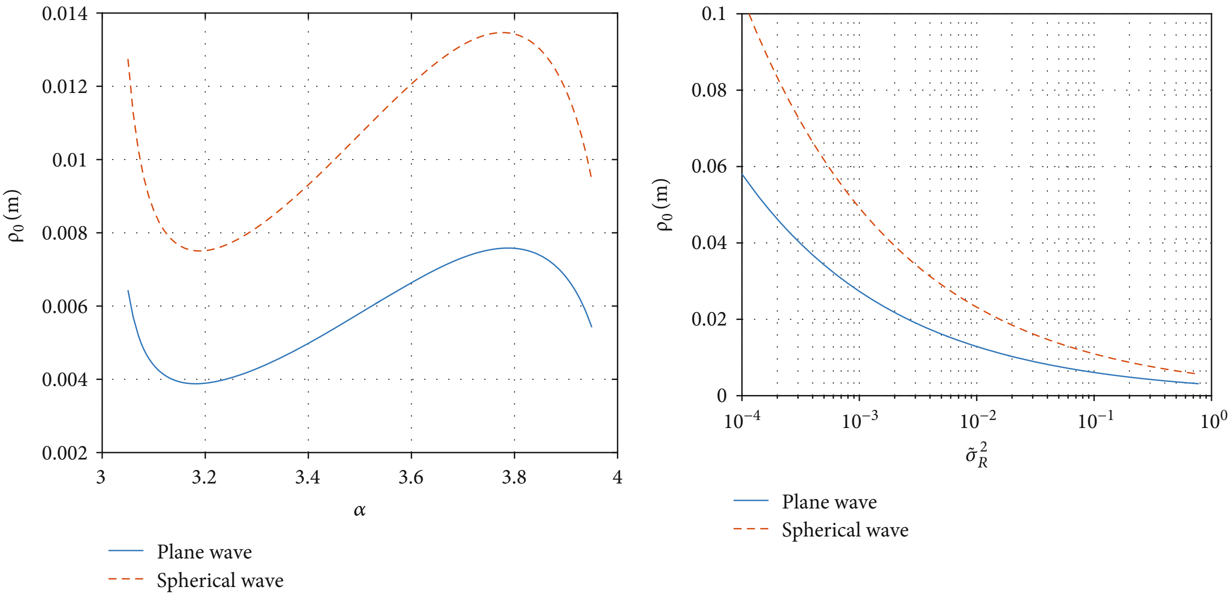
Figure 3 shows the behavior of the WSF with respect to the scales of the eddy sizes. In these simulations, we have set that the anisotropic factor $\zeta = 1$ and the power law exponent $\alpha = 11/3$. In Figure 3(a), one can observe that the WSFs for both optical waves slowly decrease as the inner scale of the eddy size l_0 increases. On the other hand, in Figure 3(b), the WSFs increase asymptotically as the outer scale of the eddy size L_0 increases. Note that the WSFs increase sheer over the outer scale of the eddy size up to 20 meters.



(a)

(b)

FIGURE 3: WSF as a function of increasing the value of (a) inner scale and (b) outer scale.



(a)

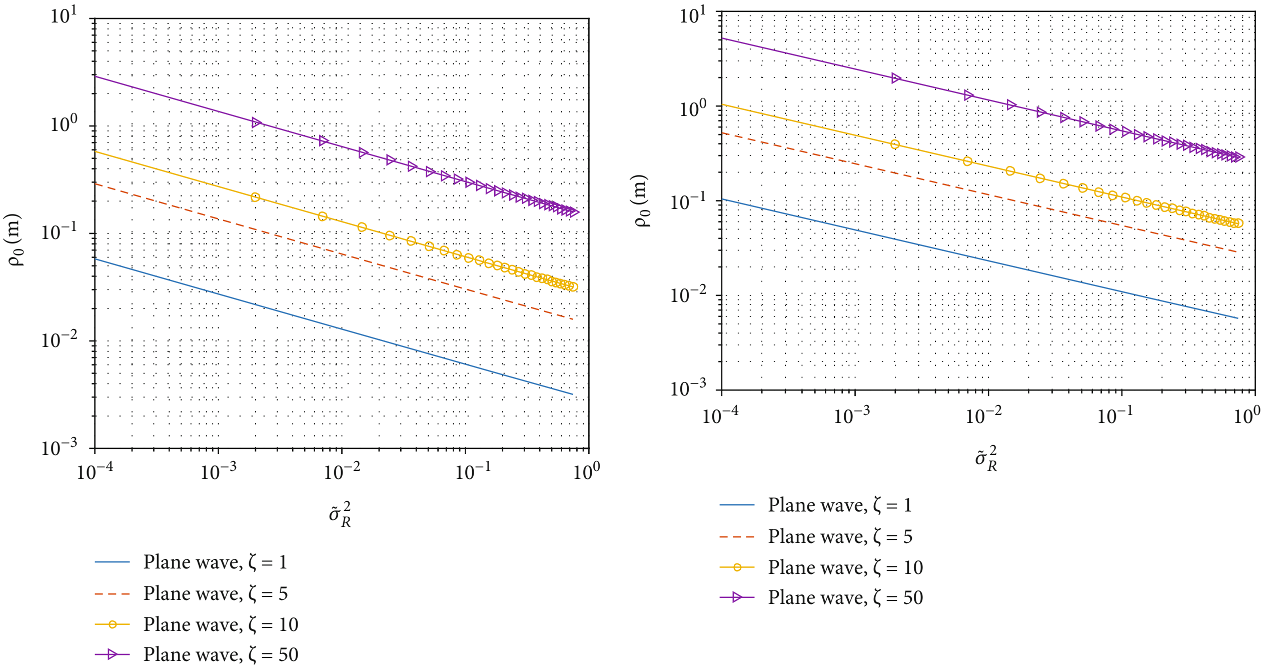
(b)

FIGURE 4: Spatial coherence radius as a function of increasing (a) power exponent α and (b) turbulence strength.

6. Evaluation on SCRs

The second set of simulations is performed using the new expressions of spatial coherence radiuses defined in Equations (23) and (24). Similar to the simulations using the WSFs, the focus of the evaluation is also to analyze the behaviors of SCRs in terms of various characterization parameters. Those include the power law exponent α, the turbulence strength $\tilde{\sigma}_R^2$, and the anisotropic factor ζ.

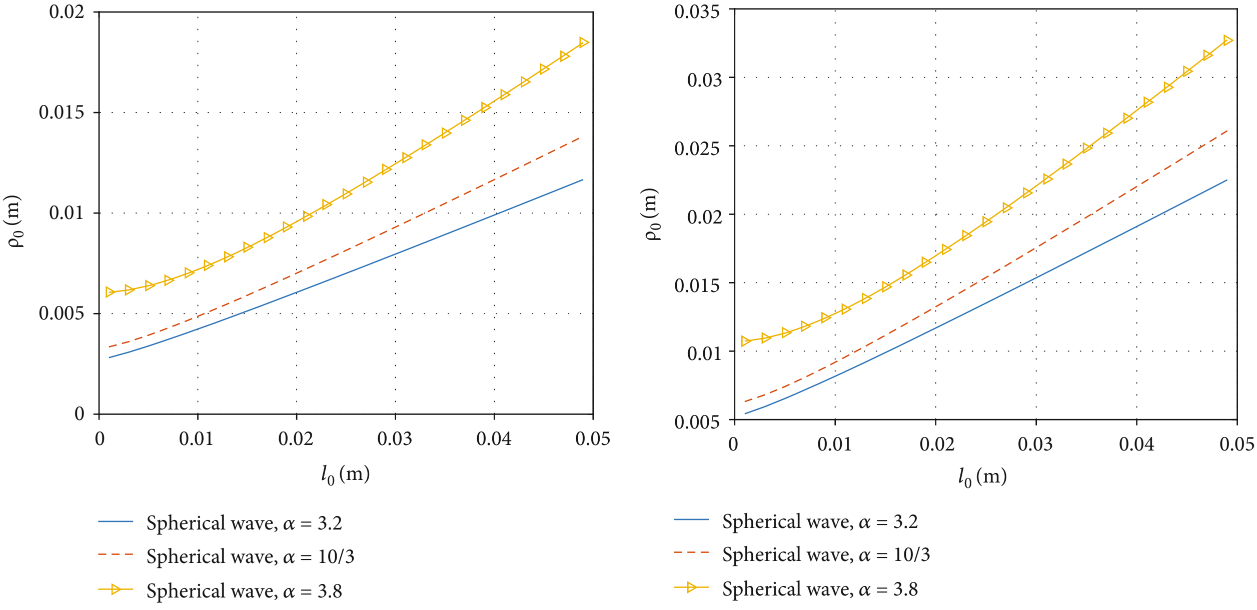
Figure 4(a) shows the behavior of the SCR with respect to the increasing power law exponent α when the anisotropic factor $\zeta = 1$, which actually makes the turbulence isotropic. Power law exponent α is related to altitude of the propagation path of optical waves [34]. It is clear that the height will influence the atmosphere condition and the curves change when the atmosphere changes. The SCRs decrease sheer when α varies from 3 to 3.2 and increase gently when α varies from 3.3 to 3.8 and finally decrease sheer afterwards as α goes to 4. On the other hand, Figure 4(b) shows the behavior of the SCR with respect to the turbulence strength $\tilde{\sigma}_R^2$. The SCRs also monotonically decrease as the turbulence strength increases. It is notable that the strengths of SCRs for two optical waves become almost the same when the turbulence strength gets moderate to strong, i.e., $\tilde{\sigma}_R^2 > >1$. The physical



(a)

(b)

Figure 5: Spatial coherence radius as a function of increasing turbulence strength for a varying anisotropic factor: (a) plane wave and (b) spherical wave.



(a)

(b)

Figure 6: Spatial coherence radius as a function of increasing inner scale for a varying power exponent: (a) plane wave and (b) spherical wave.

reason for this phenomenon is that the strong turbulence will weaken the optical wave signals; thus, the SCR will reduce when the turbulence strength increases. Moreover, as in the simulations using the SCRs, the magnitudes of the plane waves are always smaller than those of the spherical waves.

Figure 5 shows the behavior of the SCR with respect to the increasing turbulence strength $\tilde{\sigma}_R^2$ and the increasing anisotropic factor ζ. Figure 5(a) is for the plane waves, and Figure 5(b) is for the spherical waves. On the contrary to the WSF cases, the SCRs decrease as the turbulence strength gets stronger. Also, it is important to note that the SCRs are also significantly influenced by the anisotropic factor ζ that the strength of the SCRs gets as much as 40 times stronger as the anisotropy increases from 1 to 50. The physical reason for this can be found in the anisotropic property of eddies; the eddies work as lenses with a larger radius of curvature in anisotropic turbulence than in isotropic turbulence, and the larger curvature lenses can focus an optical wave better than in isotropic turbulence [35].

Finally, Figure 6 shows the behavior of the SCR with respect to the inner scale of the eddy size l_0. In these simulations, we have set that the anisotropic factor $\zeta = 1$ and the power law exponent α equals to 3.2, 10/3, and 3.8, respectively. The simulation results show that the SRCs for both of the optical waves are significantly influenced by the inner scale of the eddy size; the curves increase monotonically as l_0 increases. Moreover, the plane waves are more affected by the power exponent factor α at the fixed inner scale of the eddy size; its magnitude increases as much as 50 percent as varies from 3.2 to 3.8. As the SCR gets smaller, it means the receiver needs more work to equalize the channel distortion.

7. Conclusion

In this work, we have presented new sets of expressions for the wave structure functions and also for the spatial coherence radiuses of the free-space optical waves such as the plane waves and the spherical waves propagating in a horizontal path of a free space, which is disturbed by anisotropic turbulence. Those newly derived analytic expressions of WSFs and SCRs are evaluated, and their behaviors are observed by varying five major characterizing parameters, which are the power law exponent α, the turbulence strength $\tilde{\sigma}_R^2$, the anisotropic factor ζ, and the inner scale and the outer scale of the eddy size, l_0 and L_0, respectively.

Those five parameters individually or in their combinations have extensive impacts on the magnitudes of the WSFs and SCRs. The behaviors of the WSFs and the SCRs come out differently with respect to the power law exponent. Also, with respect to the increasing turbulence strength, the WSFs and SCRs show an inverse relation that the WSFs increase while the SCRs decrease. Similarly, the anisotropic factor affects the WSFs and the SCRs inversely. In other words, the WSFs increase as the anisotropy increases while the SCRs decrease on the contrary. Finally, the scales of the eddy size gently affect the WSFs and the SCRs that both of them monotonically increase as the scales of the eddy sizes increase regardless of the inner scale or the outer scale. Particularly, for the SCRs, the plane waves are more significantly affected by the turbulence power than the spherical waves. Moreover, the bigger power law exponents shrink the size of SCRs more than those of the smaller ones; the plane waves are also more affected than the spherical waves in this case as well.

We have also found that the wave structure function can be used to analyze the temporal frequency spreads of optical waves and MTF of an imaging system. In the optical communication system, in order to recognize the target effectively, it is necessary to evaluate the micro-Doppler shift caused by the background noise. This kind of noise also includes the wave caused by optical wave propagation in turbulent atmosphere. These random changes cause frequency spread in the spectrum of laser signal, which is manifested as additional Doppler frequency shift. In addition, the significant frequency spread can eliminate the micro-Doppler shift caused by the target. Therefore, the temporal frequency spread of optical waves can be used in lidar system, optical detection, ranging system, and other target detection and recognition fields. In conjunction with the current work presented in this paper, we are also looking at the effects of short-exposure MTFs on imaging systems of optical waves propagating in a free space with anisotropic maritime turbulence; the propagation paths can be slant and horizontal. Our results have important theoretical and practical significance for optical communication and imaging and sensing systems involving turbulent atmospheric channels on 6G communication. Also, our research has some limitations; the work is mainly based on the power-law spectrum derived from the mathematical formula and no outdoor experiments. In future work, we will try to compare the results with the measured data from the outdoor experiment, because the outdoor results can better represent the actual atmospheric conditions.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the Science and Technology Planning Project of Guangdong Province of China (Grant No. 2020A1414010285), the Natural Science Foundation of Jiangsu Province of China (Grant No. BK20201201), the Brain Korea 21 Program for Leading Universities & Students, the Professorial and Doctoral Scientific Research Foundation of Applied Technology College of Soochow University, Suzhou (Grant No. RC0006), and the Professorial and Doctoral Scientific Research Foundation of Huizhou University (Grant Nos. 2020JB014 and 2019JB019).

References

[1] C. Uysal, B. Ghassemlooy, and E. G. Udvary, *Optical wireless communications- an emerging technology*, Springer Publishing Company, 2016.

[2] Z. Ghassemlooy, W. Popoola, and S. Rajbhandari, *Optical Wireless Communications : System and Channel Modelling with MATLAB*, CRC Press, Inc., 2012.

[3] L. C. Andrews and R. L. Phillips, *Laser Beam Propagation through Random Media*, SPIE Press, Bellingham, WA, USA, 2005.

[4] V. I. Tatarskii, *The Effects of the Turbulent Atmosphere on Wave Propagation*, Israel Program for Scientific Translations, Jerusalem, Israel, 1971.

[5] A. Ishimaru, *Wave Propagation and Scattering in Random Media*, John Wiley & Sons, 1999.

[6] I. Toselli, L. C. Andrews, R. L. Phillips, and V. Ferrero, "Angle of arrival uctuations for free space laser beam propagation through non Kolmogorov turbulence," in *Atmospheric Propagation IV*, no. article 65510E, 2007International Society for Optics and Photonics, 2007.

[7] I. Toselli, "Introducing the concept of anisotropy at different scales for modeling optical turbulence," *Josa A*, vol. 31, no. 8, p. 1868, 2014.

[8] L. Cui, B. Xue, and F. Zhou, "Generalized anisotropic turbulence spectra and applications in the optical waves propagation through anisotropic turbulence," *Optics Express*, vol. 23, no. 23, 2015.

[9] Y. Baykal, "Intensity fluctuations of asymmetrical optical beams in anisotropic turbulence," *Applied Optics*, vol. 55, no. 27, p. 7462, 2016.

[10] Y. Baykal, Y. Luo, and X. Ji, "Scintillations of higher order laser beams in anisotropic atmospheric turbu-lence," *Applied Optics*, vol. 55, no. 33, p. 9422, 2016.

[11] J. Ma, Y.-L. Fu, S.-Y. Yu, X. Xie, and L. Tan, "Further analysis of scintillation index for a laser beam propagating through moderate-to-strong non-kolmogorov turbulence based on generalized effective atmospheric spectral model," *Chinese Physics B*, vol. 27, no. 3, article 034201, 2018.

[12] J. M. Cheng, L. Guo, J. Li, X. Yan, R. Sun, and Y. You, "Effects of asymmetry atmospheric eddies on spreading and wander of bessel-gaussian beams in anisotropic turbulence," *IEEE Photonics Journal*, vol. 10, no. 3, pp. 1–10, 2018.

[13] L. Tang, H. Wang, X. Zhang, and S. Zhu, "Propagation properties of partially coherent Lommel beams in non-Kolmogorov turbulence," *Optics Communications*, vol. 427, pp. 79–84, 2018.

[14] L. Andrews, W. Miller, and J. Ricklin, "Spatial coherence of a gaussian-beam wave in weak and strong optical turbulence," *The Journal of the Optical Society of America A*, vol. 11, no. 5, pp. 1653–1660, 1994.

[15] W. Wen, Y. Jin, M. Hu et al., "Beam wander of coherent and partially coherent airy beam arrays in a turbulent atmosphere," *Optics Communications*, vol. 415, pp. 48–55, 2018.

[16] Y. Jin, M. Hu, M. Luo et al., "Beam wander of a partially coherent airy beam in oceanic turbulence," *Journal of the Optical Society of America. A*, vol. 35, no. 8, pp. 1457–1464, 2018.

[17] G. Wu, W. Dai, H. Tang, and H. Guo, "Beam wander of random electromagnetic Gaussian-Shell model vortex beams propagating through a Kolmogorov turbulence," *Optics Communications*, vol. 336, pp. 55–58, 2015.

[18] J. Strohbehn and S. Clifford, "Polarization and angle-of-arrival fluctuations for a plane wave propagated through a turbulent medium," *IEEE Transactions on Antennas and Propagation*, vol. 15, no. 3, pp. 416–421, 1967.

[19] X. Ke and Z. Tan, "Effect of angle-of-arrival fluctuation on heterodyne detection in slant atmospheric turbulence," *Applied Optics*, vol. 57, no. 5, pp. 1083–1090, 2018.

[20] J. Borgnino, F. Martin, and A. Ziad, "Effect of a finite spatial-coherence outer scale on the covariances of angle- of-arrival fluctuations," *Optics Communications*, vol. 91, no. 3-4, pp. 267–279, 1992.

[21] B. Guan and J. Choi, "Temporal frequency spread of optical waves propagating in anisotropic maritime atmospheric turbulence," *Applied Optics*, vol. 58, no. 11, pp. 2913–2919, 2019.

[22] C. Young, A. J. Masino, F. E. Thomas, and C. J. Subich, "The wave structure function in weak to strong fluctuations: an analytic model based on heuristic theory," *Waves in Random Media*, vol. 14, no. 1, pp. 75–96, 2004.

[23] R. L. Lucke and C. Y. Young, "Theoretical wave structure function when the effect of the outer scale is significant," *Applied Optics*, vol. 46, no. 4, p. 559, 2007.

[24] X. Ji, X. Li, and G. Ji, "Propagation of second-order moments of general truncated beams in atmospheric turbulence," *New Journal of Physics*, vol. 13, no. 10, article 103006, 2011.

[25] L. Lu, X. Ji, and Y. Baykal, "Wave structure function and spatial coherence radius of plane and spherical waves propagating through oceanic turbulence," *Optics Express*, vol. 22, no. 22, pp. 27112–27122, 2014.

[26] L. Cui, B. Xue, X. Cao, and F. Zhou, "Atmospheric turbulence MTF for optical waves′ propagation through anisotropic non-Kolmogorov atmospheric turbulence," *Optics & Laser Technology*, vol. 63, pp. 70–75, 2014.

[27] L. Cui and B. Xue, "Influence of anisotropic turbulence on the long-range imaging system by the MTF model," *Infrared Physics & Technology*, vol. 72, pp. 229–238, 2015.

[28] S. Kotiang and J. Choi, "Wave structure function and long-exposure MTF for laser beam propagation through non-Kolmogorov turbulence," *Optics & Laser Technology*, vol. 74, pp. 87–92, 2015.

[29] B. Guan, H. Yu, W. Song, and J. Choi, "Wave structure function and long-exposure MTF for Gaussian-beam waves propagating in anisotropic maritime atmospheric turbulence," *Applied Sciences*, vol. 10, no. 16, p. 5484, 2020.

[30] B. E. Stribling, B. M. Welsh, and M. C. Roggemann, "Optical propagation in non-Kolmogorov atmospheric turbulence," in *Atmospheric Propagation and Remote Sensing IV*, vol. 2471, pp. 181–197, International Society for Optics and Photonics, 1995.

[31] S. Kotiang and J. Choi, "Temporal frequency spread of optical wave propagation through anisotropic non-Kolmogorov turbulence," *Journal of Optics*, vol. 17, no. 12, article 125606, 2015.

[32] L. C. Andrews, *Special Functions of Mathematics for Engineers*, McGraw-Hill, 2nd ed edition, 1997.

[33] I. Toselli, L. C. Andrews, R. L. Phillips, and V. Ferrero, "Scintillation index of optical plane wave propagating through non-Kolmogorov moderate-strong turbulence," in *Optics in Atmospheric Propagation and Adaptive Systems X*, vol. 6747, International Society for Optics and Photonics, 2007.

[34] A. Zilberman, E. Golbraikh, and N. S. Kopeika, "Propagation of electromagnetic waves in Kolmogorov and non-Kolmogorov atmospheric turbulence: three-layer altitude model," *Applied Optics*, vol. 47, no. 34, pp. 6385–6391, 2008.

[35] I. Toselli, B. Agrawal, and S. Restaino, "Light propagation through anisotropic turbulence," *JOSA A*, vol. 28, no. 3, p. 483, 2011.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5569006, 9 pages
https://doi.org/10.1155/2021/5569006

Research Article

Reconfigurable Intelligent Surface-Based Space-Time Block Transmission on 6G

Wei Song[1] **and Bing Guan**[2]

[1]*Department of Electronic Information and Communication Engineering, Applied Technology College of Soochow University, Suzhou 215325, China*
[2]*School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China*

Correspondence should be addressed to Bing Guan; binggwan@hotmail.com

Received 31 January 2021; Revised 17 March 2021; Accepted 25 November 2021; Published 16 December 2021

Academic Editor: Miaowen Wen

Reconfigurable intelligent surface (RIS) is considered to be a new technology with great potential and is being studied extensively and deeply. And the application extension of STBC in the RIS-aided scheme provides a new train of thought for the research of channel coding. In this paper, we propose we extend the scheme of using the RIS to adjust the phase and reconfigure the reflected signal and propose the design of the RIS-aided QO-STBC scheme and the RIS-aided QO-STBC scheme with interference cancellation. Particularly in the RIS-aided QO-STBC scheme with interference cancellation, the design can achieve the transmission of the full rate and full diversity using an auxiliary reflection group to eliminate the influence of interference term. Also, the advantages and disadvantages of the schemes are analyzed in the paper, and the decoding algorithms with different complexity used in the proposed schemes are described. The simulation results show that the performance of the RIS-aided QO-STBC scheme with interference cancellation is better than that of the RIS-aided QO-STBC scheme and the RIS-aided Alamouti scheme by about 5 dB and 7 dB at 10^{-3} BER because of diversity gain and coding gain.

1. Introduction

Over the last 30 years, wireless communication technologies have achieved revolutionary development. The application of multiple-input and multiple-output (MIMO) [1] systems fundamentally solved the problem of insufficient wireless channel capacity. On this basis, the research of OFDM (orthogonal frequency division multiplex) [2] promotes more rapid growth of communication.

Form 2019, the fifth-generation (5G) mobile communication are worldwide tested and applied. The performance of 5G communication with a higher rate and low latency can provide communication support for Internet-of-Things (IoT), intelligent manufacturing (IM), massive machine-type communications (MTC), etc. 5G base stations (BSs) with massive MIMO antennas supply the enhanced mobile broadband (eMBB) communication service. And massive MIMO achieves huge diversity gain, path gain, and spatial reusability [3] due to large-scale antenna arrays. The International Telecommunication (ITU) predicated that the whole mobile data flows will reach 60 zettabytes (ZB) [4] per year. As everybody knows, 5G has a superiority unmatched compared with traditional cellular communication. The European Telecommunications Standards Institute (ETSI) published the target peak rate is 10 Gb/s and 20 Gb/s in the uplink and downlink, respectively. 5G communication brings the possibility to realize the application of new technology in various fields. Meanwhile, the ITU also estimated the 5G will reach its limit in 2030, because some applications need higher transmission rate and lower time-delay and reliability, and 5G communication can not satisfy its requirements, obviously. It means that the new challenge will be coming on 6G technology. The focus of the research of 6G is being made not only to solve the insufficiency of performance of 5G but also to meet the needs of the continuous development and application of wireless communication, especially in ultrabroadband (uMUB), ultrahigh-speed-with-low-latency communications (uHSLLC), and ultrahigh data density (uHDD) [5].

Although the research of 6G wireless communication is in an early stage, the direction of revolutionary-innovative research can not be determined completely. However, it can be predicted that the research of 6G will continue many 5G technologies to make our living environment more intelligent. In the future, the main problem of 6G communication will focus on improving the intelligent service level between devices and further enhancing the intelligence level of the whole society under the condition of small samples. Internet of Intelligent Things (IoIT) and Artificial Intelligence (AI) will be widely used at all levels. At the same time, the great improvement of big data and computing power also requires more reliable and faster communication support, especially in wireless communication. The researchers have already studied the new technology of 6G in order to achieve the promised goal, which connects every smart device to the Internet. Also, the 6G will offer precisely and higher Quality of Service (QoS) such as holographic communication, augmented reality/virtual reality. It is worth mentioning that 6G will focus on Quality of Experience (QoE) to provide rich experiences. Meanwhile, the green technology solutions for energy saving and environmental protection of future mobile communication have become the key research direction of sustainable wireless communication development.

In the last two years, a new technology with great potential for significant energy consumption reductions is being studied extensively and deeply called reconfigurable intelligent surface (RIS) [6–8]. The RIS is a new concept with massive radiating and sending elements, in an ideal line-of-sight environment; the entire surface is used as a receiving antenna array. Under the condition that the surface area is large enough, the received signal after matched filtering operation can be approximated as a SINC-like intersymbol interference channel and go beyond contemporary large-scale MIMO technology. In [9], the authors proposed a dual-hop RIS-aided (RIS-DH) scheme and RIS-aided transmit (RIS-T) scheme, and neither of these two schemes can provide multiplexing gain; only diversity gain can be obtained. The channel of the RIS-T scheme looks like a transmission diversity structure, and the RIS-DH scheme is more like a keyhole MIMO channel; meanwhile, the outage probability, bit error rate, and average channel capacity are given. As a promising 6G auxiliary transmission scheme, RIS can provide higher array gain and achieve more accurate channel estimation by using lower hardware cost and energy loss compared with large-scale MIMO technology [10]. Recent results showed that nonorthogonal multiple access (NOMA) using RIS-aided [11] can make some uncontrollable factors of affected received signal quality of wireless communication become controllable, which provides an effective transmission guarantee for wireless communication. In addition, using reconfigurable passive elements, the RIS-aided scheme can allow to build a programmable wireless environment [12]. An RIS-based space shift keying (SSK) [13] scheme also can adjust the phase of the reflected signal effectively to improve energy efficiency and high transmission reliability. In [14], a new transmission protocol for wideband RIS-aided single-input multiple-output (SIMO) orthogonal frequency division multiplexing (OFDM) communication systems was proposed. And the passive beamforming of the RIS was fine-tuned to improve the achievable rate for data transmission. In conclusion, the RIS-aided transmission strategies can greatly raise the effectiveness of performance of wireless communication. Recently, from the application level, some researchers have studied the degree of the isolation of the antenna [15], the accurate characterization of probability density function (PDF) [16], the two-dimensional physical structure of RIS [17], the communication between two nodes [18], the capability of extending cellular coverage [19–25], etc., which further verifies the feasibility and applicability of RIS-aided communication. Also, the security performance of the RIS-aided wireless communication system is widely concerned [26, 27].

With the in-depth study of RIS technology, many original technologies of wireless communication have the new application scenarios and the direction of research. It is known that MIMO technology fundamentally solves the problem of channel capacity limitation of single input single output (SISO). And the space time-block codes (STBC) can provide an effective capacity gain and spatial diversity gain in MIMO systems [28, 29]. Next, the relay nodes were considered in a multiply-hop MIMO system [30–34] where several relay stations transmit jointly to the same destination node. However, the most difficult problems to solve with relay nodes are the cost and intelligent control of relay devices. In these aspects, the RIS technology opens up new ideas for researchers and makes the application of cooperative communication more feasible. In [35, 36], the authors present two different approaches to achieve classic Alamouti space-time coding wireless transmission through an RIS-aided transmitter, respectively; two proposed approaches convincingly validate the feasibility of RIS-based Alamouti space-time coding. The key difference is that two proposed schemes use different number of antennas on the source node.

Next, let us analyze the differences between the two schemes in detail. The first proposed scheme [35] uses one transmission antenna on the source node and N reflection elements on the RIS side, and RIS element was divided into two parts where each part applies phase modulation to reconfigure the singles and forward it. Thus, the RIS provides diversity gain. Compared to the first design, the second proposed scheme [36] uses two transmission antennas on the source side to realize the diversity gain. Anyway, if the time loss from source to RIS is ignored, both two plans make good use of RIS to assist the implementation of the characteristic of Alamouti code with full rate and full diversity. This application extension of STBC in the RIS-aided scheme provides a new train of thought.

As everyone knows, the Alamouti code is the only orthogonal complex coding matrix with full rate and full diversity and diversity gain is two. In this paper, we propose new extended quasiorthogonal STBC (QO-STBC) applications to improve the diversity gain of the scheme. To be specific, the new scheme applies one group of RIS to assist QO-STBC coding transmission, and another group of RIS is used to eliminate the nonorthogonal interference caused by the coding matrix simultaneously.

1.1. Prior Work. A perfect STC (space time code) should be an orthogonal coding with full rate and linear decoding complexity. The classical Alamouti [28] with these features can be expressed as

$$A_{ij} = \begin{bmatrix} x_i & x_j \\ -x_j^* & x_i^* \end{bmatrix}, \tag{1}$$

where $j = i + 1$ for any positive integer i. It is proved mathematically in [37] that complex O-STBCs with the full rate did not exist for more than two transmission antennas. Consequently, Laneman et al. presented a QO-STBC scheme [30] with the full rate. But the QO-STBC scheme reduces the diversity gain and uses the more complex algorithm with pairwise decoding. However, the QO-STBC has better performance than the orthogonal codes with a rate less than 1 at low signal-to-noise (SNR). Jafarkhani's code used Alamouti code as block element of the coding matrix for four transmission antennas denoted as

$$S_J = \begin{bmatrix} A_{12} & A_{34} \\ -A_{34}^* & A_{12}^* \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ -x_2^* & x_1{}^* & -x_4{}^* & x_3^* \\ -x_3^* & -x_4{}^* & x_1{}^* & x_2^* \\ x_4 & -x_3 & -x_2 & x_1 \end{bmatrix}. \tag{2}$$

We can get the character matrix Q_J [38] as

$$Q_J = S_J^H S_J = \begin{bmatrix} \alpha & 0 & 0 & \beta_J \\ 0 & \alpha & -\beta_J & 0 \\ 0 & -\beta_J & \alpha & 0 \\ \beta_J & 0 & 0 & \alpha \end{bmatrix}, \tag{3}$$

where $\alpha = \sum_{i=1}^{4} |X_i|^2$ and $\beta_J = (X_1 X_4^* + X_1^* X_4) - (X_2 X_3^* + X_2^* X_3)$. From Equation (3), it is obvious that the orthogonality of Jafarkhani's code was destroyed by nonmain diagonal elements. On the condition of the full rate, the diversity gain was reduced.

1.2. Organization and Notation. The rest of this paper is organized as follows. In Section 2, we present the common scheme model and the assumption of wireless communication environment. In Section 3, we introduce the Alamouti RIS-aided scheme and extend it to the QO-STBC RIS-aided scheme and analyze advantages and disadvantages of performance of each scheme. According to the analysis results of the previous section, a new model with interference cancellation is proposed and its performance is analyzed in detail in Section 4. The simulation results are provided in Section 5. Finally, the conclusions of the research are summarized in Section 6.

In this paper, A^*, A^T, and A^H denote the complex conjugate, the transpose, and Hermitian transpose of matrix A, respectively. $\mathbb{C}$ denotes the complex-valued set. The Diag[.] denotes a diagonal matrix, and the $N \times N$ identity matrix is denoted by I_N. $\mathscr{CN} \sim (\mu, \sigma^2)$ denotes the distribution of circularly symmetric complex random vector with mean μ and variance σ^2, and $\sim$ represents "distributed as."

2. System Model

The proposed RIS-aided wireless transmission scheme has one source node S and one destination node D, as illustrated in Figure 1, and the RIS consisting of an N reflecting elements is deployed to assist transmission from the node S equipped with N_t transmitting antennas to the node D with N_r receiving antennas. In the proposed scheme, we assume and the wireless communication system is performed over a complex, additive white Gaussian noise (AWGN) coherent fading channel, in which the RIS is close to the node S, and the line of sight (LoS) channel does not exist from the node S to the node D because of some obstructions. So the node S can not communicate with the node D directly, and the signals must select the reflection channels to transmit.

At the source node S, information sequences are modulated into $L = 2L_0$ length complex symbol vector $X = \{x_1, x_2, \cdots, x_L\}$, where every two neighboring symbols become one group, and L_0 denotes the number of groups as $G_i = \{x_i, x_{i+1}\}$ for $i = 1, \cdots, L_0$. Each time interval, a complex symbol is successively transmitted from the node S to the RIS nodes, and the RIS is connected to the node S through control link, which is in charge of all controlling information from the node S. The controlling information can assist the RIS elements to reconfigure the phase of reflection signals [35].

The $S \longrightarrow$ RIS channels are assumed to be a frequency selective channel with independent propagation paths. Besides, we assume that the channel state information is quasistatic, so we can write the channel impulse response as

$$h_{S,R_i}^j(t) = \alpha_{S,R_i}^j \delta(t - \tau_{i,j}), \tag{4}$$

where α_{S,R_i}^j is the channel coefficient from the j^{th} transmission antenna of the source node S to the i^{th} element of the RIS, and $\tau_{i,j}$ denotes the corresponding path delay. The h_{S,R_i}^j denotes the channel fading from the j^{th} transmission antenna of the source node S to the i^{th} element of the RIS, and $h_{S,R_i}^j \in \mathbb{C}^{N_t \times N}$, where $i \in N$ and $j \in N_t$.

Also, the channel fading coefficient from the i^{th} element of the RIS to the k^{th} receiver antenna of the node D denotes as $h_{R_i,D}^k$, and $h_{R_i,D}^k \in \mathbb{C}^{N \times N_r}$ and $k \in N_r$. Both of h_{S,R_i}^j and $h_{R_i,D}^k$ are distributed as complex Rayleigh random distribution. So $h_{S,R_i}^j \sim \mathscr{CN}$ (μ_1, σ_1^2) and $h_{R_i,D}^k \sim \mathscr{CN}$ (μ_2, σ_2^2), respectively.

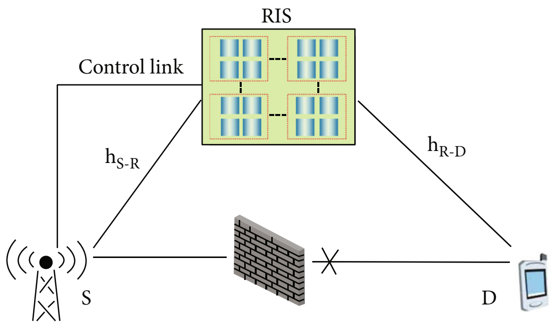


Figure 1: RIS-aided wireless transmission scheme with one source and one destination and N RIS elements.

Accordingly, the received signal at the node D is given by

$$r^k = \sqrt{E_0}\left(\sum_{i=1}^{N}\sum_{j=1}^{N_t} h^j_{S,R_i} h^k_{R_i,D}\right)x + n, \tag{5}$$

where x denotes the transmitted symbol and n is the complex AWGN ~ $\mathscr{CN}$ $(0, N_0)$. The E_0 denotes average energy.

Without losing generality, we simply select only one transmission antenna at the node S and one receiver antenna at the node D and analyze the performance of broadcast symbols through the RIS auxiliary scheme. We get

$$r = \sqrt{E_0}\left(\sum_{i=1}^{N} h_{S,R_i} h_{R_i,D}\right)x + n = \sqrt{E_0}\mathbf{H}x + n, \tag{6}$$

where $\mathbf{H} = [h_1, h_2, \cdots, h_i]^T$ denotes the channel matrix of the $S \longrightarrow$ RIS $\longrightarrow D$. The SNR gets

$$\mathbf{SNR} = \frac{E_0 P |\mathbf{H}|^2}{N_0}. \tag{7}$$

In our proposed scheme model, the following suppositions have been given:

(i) The source node S broadcasts the symbol to each element of RIS. And each element of RIS reconfigures the symbol according to control information from the node C and reflect to the node D. And only the destination node D knows all the channel state information (CSI) h_{S,R_i} and $h_{R_i,D}$

(ii) The channel state information h_{S,R_i} and $h_{R_i,D}$ are the complex Rayleigh random variable and independent and identically distributed (i.i.d.), and all of them keep constant during one period time. And $h_{S,R_i} \sim \mathscr{CN}$ (μ_1, σ_1^2) and $h_{R_i,D} \sim \mathscr{CN}$ (μ_2, σ_2^2). The noise n is the complex AWGN ~ $\mathscr{CN}$ $(0, N_0)$

(iii) All transmission antennas have been subjected to half-duplex transmission mode. At the node S, the binary signal sequences are modulated into complex MPSK symbols and the maximum likelihood (ML) decoding method is used at the node D. And the RIS is close to the node S

Suppose that the codeword belongs to the codebook X, the ML decoder should find the minimum distance following the expression to detect the symbols:

$$< x, \hat{x} > = \operatorname{argmin}_{x_i} |r - \mathbf{H}x|^2, \tag{8}$$

where $x_i \in X$. The ML method with linear complexity can achieve better performance for the proposed scheme.

3. RIS-Aided QO-STBC Model

In the previous section, we already introduce the RIS-aided Alamouti scheme [35], in which the RIS elements were divided into two parts to adjust the reflection phase. In two time slots, the RIS realizes the transmission of orthogonal coding matrix by reflecting different types of signals. The complex signal can denote as

$$x_i = \sqrt{p}e^{j\theta_i} = \sqrt{p}[\cos\theta_i + j\sin\theta_i], \tag{9}$$

where p denotes average power, and $\sqrt{j} = -1$. From the MPSK constellation map, we know the modulated signals have the same power, and the two parts can use $-(\theta_{i+1} + \pi$ $)$ and $-\theta_i$ to adjust the phase to realize the RIS-aided Alamouti's scheme, as follows:

$$\begin{gathered}\sqrt{p}[\cos(-\theta_i) + j\sin(-\theta_i)] = \sqrt{p}[\cos\theta_i - j\sin\theta_i] = x_i^*,\\ \sqrt{p}[\cos(-\theta_{i+1} - \pi) + j\sin(-\theta_{i+1} - \pi)] = -x_{i+1}^*.\end{gathered} \tag{10}$$

Thus, the 2×2 OSTBC with linear complexity was achieved in MPSK modulation mode. We consider that the adjacent reflection elements will have the similar channel state information, so, the RIS with N-elements can be divided into $N/2$ groups; each group contains a pair of non-adjacent reflection elements, as shown in Figure 2. The expression is

$$\begin{aligned} r_1 &= x_1 h_1 + x_2 h_2 + n_1,\\ r_2 &= -x_2^* h_1 + x_1^* h_2 + n_2,\end{aligned} \tag{11}$$

where $h_1 = \sum_{i=1}^{N/2}(h_{S,R_{2i-1}} h_{R_{2i-1},D})$ and $h_2 = \sum_{i=1}^{N/2}(h_{S,R_{2i}} h_{R_{2i},D})$. And n_1 and n_2 are AWGN by different channels. It is thus clear that the equivalent diversity gain of the scheme comes from the reflection elements of the RIS. In the same mode [35], the proposed scheme costs more time slots because of using multiantennas at the source node S.

Although the first scheme has some limitations in modulation mode, its performance advantage can not be ignored. So we can extend this method to QO-STBC, and it is called the RIS-aided QO-STBC scheme. In the proposed RIS-aided QO-STBC model, we also divided the RIS with N-elements into $N/4$ groups. Each group has four reflection elements,

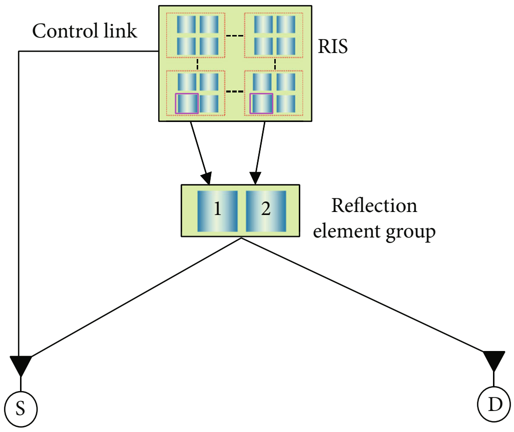


Figure 2: Alamouti RIS-aided reflection element.

in which the four reflection elements use $-\theta_1$, $-\theta_2 - \pi$, $-\theta_3$, and $-\theta_4 - \pi$ to modulate the phase of reflection signals at element 1, element 2, element 3, and element 4, respectively, as shown in Figure 3. And the results of MPSK modulation of reflection elements are shown in Table 1.

Although the design of the RIS-aided QO-STBC scheme is realized in this way, it can be seen from Equations (2) and (3) that the performance of the proposed RIS-aided scheme is not very superior. The reason is that Jafarkhani's code is a quasiorthogonal code with full rate and not full rank; the nonmain diagonal elements of character matrix are the interference term and the key factor of diversity loss. In fact, Jafarkhani's code can be considered to have orthogonal performance, if the BPSK modulation is used. The rank of the code is only 2 using the others. In terms of decoding complexity, the pairwise decoding algorithm is used because Jafarkhani's code is not orthogonal. Anyway, due to the scarcity of complex orthogonal full rate coding matrix, the design of the RIS-aided QO-STBC scheme opens up a new research direction for the application of STBC technology on 6G communication.

4. RIS-Aided QO-STBC Model with Interference Cancellation

In this section, we try to design a simply RIS-aided scheme based on the RIS-aided QO-STBC model to eliminate the interference term in the character matrix Equation (3) as much as possible. Let us think about changing the transmission matrix as $\widehat{S}_J = S_J \Lambda$, where $\Lambda = \text{Diag}[1, 1, -1, -1]$. The new character matrix is

$$\widehat{Q}_J = \widehat{S}_J^H \widehat{S}_J = \Lambda S_J^H S_J \Lambda = \begin{bmatrix} \alpha & 0 & 0 & -\beta_J \\ 0 & \alpha & \beta_J & 0 \\ 0 & \beta_J & \alpha & 0 \\ -\beta_J & 0 & 0 & \alpha \end{bmatrix}. \tag{12}$$

Combined the $\widehat{Q}_J$ and Q_J, we get $\widehat{Q}_J + Q_J = 2\alpha I_4$. Mathematically, this operation completely cancels out the interference term of the character matrix. Next, we design the wireless communication scheme according to this idea. At the RIS, a pair of adjacent reflection elements are selected, one of which is used for normal signal transmission, and the other is provided with interference elimination function besides ensuring transmission. In this way, two reflection groups A and B are built in RIS as illustrated in Figure 4. Meanwhile, in order to ensure the independence of the two reflection groups, the node D needs to be equipped with two adjacent receiving antennas in the assumed scheme, and one-to-one transmission is carried out through the beamforming. Because the distance of adjacent A_1 in the reflection group A and B_1 in the reflection group B and adjacent receiving antenna RE_1 and RE_2 in the node D is very closer, we assume that the corresponding transmission channels have the same CSI. During four time slots, the respective received signals can be written as

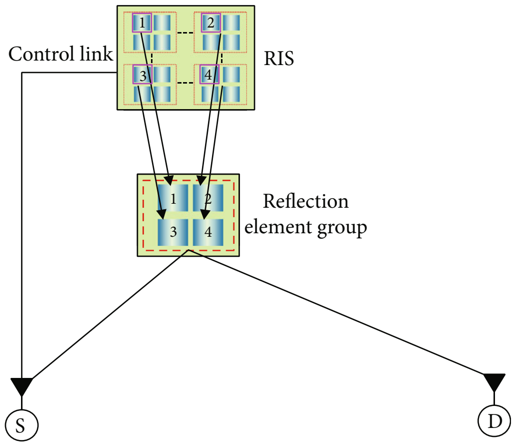


Figure 3: QO-STBC RIS-aided reflection element.

Table 1: Results of MPSK modulation of reflection elements.

Time	Element 1	Element 2	Element 3	Element 4
Slot 1	x_1	x_2	x_3	x_4
Slot 2	$-x_2^*$	x_1^*	$-x_4^*$	x_3^*
Slot 3	$-x_3^*$	$-x_4^*$	x_1^*	x_2^*
Slot 4	x_4	$-x_3$	$-x_2$	x_1

$$\begin{aligned} \mathbf{r}_{RE1} &= S_J H + \mathbf{n}_1, \\ \mathbf{r}_{RE2} &= \widehat{S}_J H + \mathbf{n}_2, \end{aligned} \tag{13}$$

where H is the CSI vector, and $H = [h_1, h_2, h_3, h_4]^T$, in which $h_i = \sum_{j=0}^{N/8-1} (h_{S,R_{A_{4j+i}}} h_{R_{A_{4j+i}},D_{RE1}})$ or $h_i = \sum_{i=0}^{N/8-1} (h_{S,R_{B_{4j+i}}} h_{R_{B_{4j+i}},D_{RE2}})$. $n_1 = [n_1, n_2, n_3, n_4]^T$ and $n_2 = [n_5, n_6, n_7, n_8]^T$. The matched reflection symbols at the reflection group A and B are reconfigured as Tables 2 and 3, respectively.

Before decoding, we first combine the received signals in (15) and (16) from the two receiving antennas; the

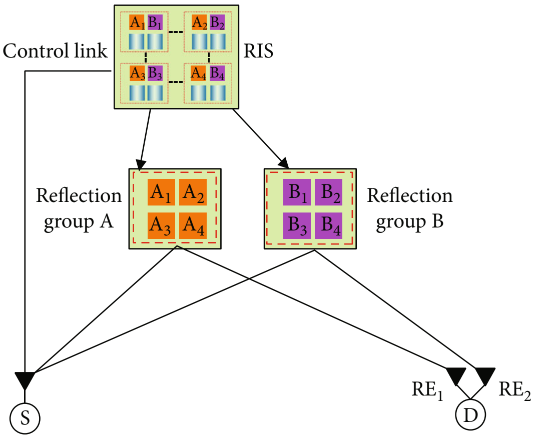


Figure 4: QO-STBC RIS-aided reflection element with interference cancellation.

Table 2: Results of MPSK modulation of reflection element group *A*.

Time	Elem-A_1	Elem-A_2	Elem-A_3	Elem-A_4
Slot 1	x_1	x_2	x_3	x_4
Slot 2	$-x_2^*$	x_1^*	$-x_4^*$	x_3^*
Slot 3	$-x_3^*$	$-x_4^*$	x_1^*	x_2^*
Slot 4	x_4	$-x_3$	$-x_2$	x_1

Table 3: Results of MPSK modulation of reflection element group *B*.

Time	Elem-B_1	Elem-B_2	Elem-B_3	Elem-B_4
Slot 1	x_1	x_2	$-x_3$	$-x_4$
Slot 2	$-x_2^*$	x_1^*	x_4^*	$-x_3^*$
Slot 3	$-x_3^*$	$-x_4^*$	$-x_1^*$	$-x_2^*$
Slot 4	x_4	$-x_3$	x_2	$-x_1$

equivalent transmission matrix has

$$S = \begin{bmatrix} X_1 & -X_2{}^* & -X_3{}^* & X_4 & X_1 & -X_2{}^* & -X_3{}^* & X_4 \\ X_2 & X_1{}^* & -X_4{}^* & -X_3 & X_2 & X_1{}^* & -X_4{}^* & -X_3 \\ X_3 & -X_4{}^* & X_1{}^* & -X_2 & -X_3 & X_4{}^* & -X_1{}^* & X_2 \\ X_4 & X_3{}^* & X_2{}^* & X_1 & -X_4 & -X_3{}^* & -X_2{}^* & -X_1 \end{bmatrix}^T. \tag{14}$$

Also, the equivalent transmission matrix can be written as

$$S = \begin{bmatrix} S_J \\ \widehat{S}_J \end{bmatrix}, \tag{15}$$

and the equivalent character matrix has

$$S^H S = \begin{bmatrix} S_J^H \widehat{S}_J^H \end{bmatrix} \begin{bmatrix} S_J \\ \widehat{S}_J \end{bmatrix} = 2\alpha I_4. \tag{16}$$

Obviously, the equivalent transmission matrix S is a new type O-STBC transmission matrix with full rate and full diversity. The difference is that the proposed RIS-aided QO-STBC model with interference cancellation transfers the loss of time slots to an increase in the number of reflection elements by reasonable transformation. So, the use of the linear decoding algorithm [39] can greatly reduce the decoding complexity of the receiver. The decision statistics $\tilde{x}_i$ of the transmitted signal x_i can be constructed as

$$\tilde{x}_i = \sum_{t\in\eta(i)} \mathrm{sgn}_t(i) \cdot \tilde{y}_t \cdot \tilde{h}^*_{\epsilon_t(i)}, \tag{17}$$

where $\eta(i)$ is the set of rows of the transmission matrix including x_i,

$$\begin{aligned} \tilde{y}_t(i) &= \begin{cases} y_t, & x_i \text{ belongs to the } t^{\text{th}} \text{ row of } S; \\ (y_t)^*, & x_i^* \text{ belongs to the } t^{\text{th}} \text{ row of } S, \end{cases} \\ \tilde{h}^*_{\epsilon_t(i)} &= \begin{cases} h^*_{\epsilon_t(i)}, & x_i \text{ belongs to the } t^{\text{th}} \text{ row of } S; \\ h_{\epsilon_t(i)}, & x_i^* \text{ belongs to the } t^{\text{th}} \text{ row of } S. \end{cases} \end{aligned} \tag{18}$$

To minimize each individual decision metric, we get

$$|\tilde{x}_i - x_i|^2 + \left(2\sum_{j=1}^{4} |h_j|^2 - 1\right)|x_i|^2. \tag{19}$$

5. Simulation Results and Discussions

In this section, we present computer simulation results to evaluate the performance of the RIS-aided Alamouti scheme, the RIS-aided QO-STBC scheme, and the RIS-aided QO-STBC scheme with interference cancellation in terms of bit error ratio (BER) and analysis of the results. In the simulations, we considered the assumption of frequency-selective Rayleigh flat fading channels by employing the ML-based decoding algorithm. Considering the large number of reflection elements in RIS, the number of reflection function groups divided in different ways is consistent. Without losing generality, only one reflection function group of the RIS was selected to participate in the computer simulation.

In this paper, all of the proposed schemes are based on the reflection elements of the RIS to reconstruct the reflection signals by phase modulation. The time loss from the source S to the RIS can be ignored when calculating the transmission rate. So, the transmission rate of all proposed schemes is approximately equal to one. Figure 5 shows the simulation of the proposed schemes, in which we compared the performance of the RIS-aided Alamouti scheme, the RIS-aided QO-STBC scheme, and the RIS-aided QO-STBC

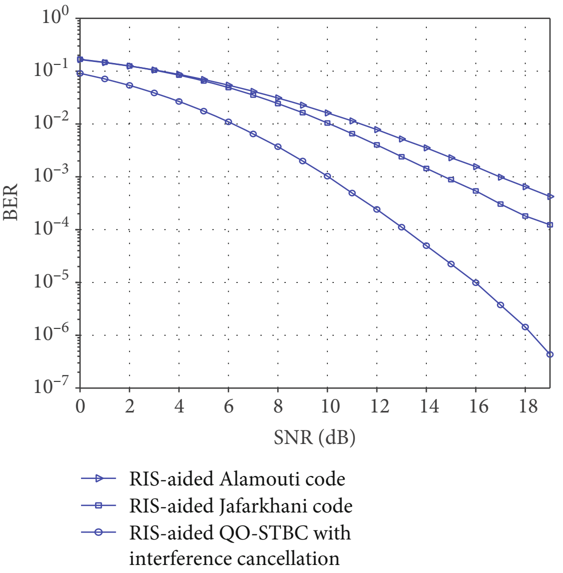


Figure 5: Performance of the proposed RIS-aided scheme with frequency-selective Rayleigh flat fading channel (transmission rate 2 bits/s/Hz).

scheme with interference cancellation for the transmission of two bits/s/Hz (QPSK).

From the simulation results, in Figure 5, we can see that the RIS-aided Alamouti scheme and the RIS-aided QO-STBC scheme achieved the same full rate characteristics as the traditional STBC transmission matrix. Due to the characteristic of nonfull rank of QO-STBC, the diversity gain can not reach the maximum order of matrix. However, compared with the RIS-aided Alamouti scheme with diversity 2, the diversity gain of the RIS-aided QO-STBC scheme is improved. The proposed RIS-aided QO-STBC scheme outperforms the RIS-aided Alamouti scheme about 2 dB at 10^{-3} BER. And the RIS-aided QO-STBC scheme with interference cancellation achieves the full rate and the full diversity and improves the coding gain due to the participation of reflection group *B*. The result of simulation shows that the performance of the RIS-aided QO-STBC scheme with interference cancellation is better than that of the RIS-aided QO-STBC scheme and the RIS-aided Alamouti scheme by about 5 dB and 7 dB at 10^{-3} BER, respectively. The analysis fits well in with simulation results.

6. Conclusions

In this paper, we extend the scheme of using the RIS to adjust the phase and reconfigure the reflected signal and propose the design of the RIS-aided QO-STBC scheme. In particular, aiming at the problem that the diversity of QO-STBC is reduced due to the nonorthogonality, a scheme called the RIS-aided QO-STBC scheme with interference cancellation using auxiliary reflection group to eliminate the influence of interference term is designed. And the advantages and disadvantages of the scheme are analyzed in detail; also, the decoding algorithms with different complexity used in the proposed schemes are described.

Through the reasonable assumption of the design of the RIS-aided QO-STBC scheme with interference cancellation, the wireless communication of full rate and full diversity is realized. The difference is that the proposed RIS-aided QO-STBC model with interference cancellation transfers the loss of time slots to an increase in the number of reflection elements by transformation.

Simulation results show that the performance of the proposed schemes has remarkable improvement due to the increase of diversity gain and coding gain. Also, the design of the RIS-aided QO-STBC scheme opens up a new research direction for the application of STBC technology on 6G communication. And the future research is to improve the received diversity gain.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request. The authors declare that there are no conflicts of interest.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The authors would like to extend their sincere gratitude to Dr. Huaisheng Wang, for his instructive advice and useful suggestions and help on this paper. This work was supported in part by the Natural Science Foundation of Jiangsu Province, China (Grant No. BK20201201), the Professorial and Doctoral Scientific Research Foundation of Applied Technology College of Soochow University, Suzhou (Grant No. RC0006), and the Professorial and Doctoral Scientific Research Foundation of Huizhou University (Grant No. 2020JB014).

References

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[2] S. Nanda, R. Walton, J. Ketchum, M. Wallace, and S. Howard, "A highperformance MIMO OFDM wireless LAN," *IEEE Communications Magazine*, vol. 43, no. 2, pp. 101–109, 2005.

[3] C. Han, J. M. Jornet, and I. F. Akyildiz, "Ultra-massive MIMO channel modeling for graphene-enabled terahertz-band communications," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, Porto, Portugal, 2018.

[4] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 118–125, 2020.

[5] B. Zong, C. Fan, X. Wang, X. Duan, B. Wang, and J. Wang, "6G technologies: key drivers, core requirements, system

architectures, and enabling technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 18–27, 2019.

[6] S. Hu, F. Rusek, and O. Edfors, "Beyond massive MIMO: the potential of data transmission with large intelligent surfaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2746–2758, 2018.

[7] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," 2018, http://arxiv.org/abs/1809.01423.

[8] C. Liaskos, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "Using any surface to realize a new paradigm for wireless communications," *Communications of the ACM*, vol. 61, no. 11, pp. 30–33, 2018.

[9] L. Yang, F. Meng, Q. Wu, D. B. da Costa, and M.-S. Alouini, "Accurate closed-form approximations to channel distributions of RIS-aided wireless systems," *IEEE Wireless Communications Letters*, vol. 9, no. 11, pp. 1985–1989, 2020.

[10] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2621–2636, 2020.

[11] A. Hemanth, K. Umamaheswari, A. C. Pogaku, D.-T. Do, and B. M. Lee, "Outage performance analysis of reconfigurable intelligent surfaces-aided NOMA under presence of hardware impairment," *IEEE Access*, vol. 8, pp. 212156–212165, 2020.

[12] H. Guo, Y. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3064–3076, 2020.

[13] A. E. Canbilen, E. Basar, and S. S. Ikki, "Reconfigurable intelligent surface-assisted space shift keying," *IEEE Wireless Communications Letters*, vol. 9, no. 9, pp. 1495–1499, 2020.

[14] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, F. Chen, and S. Mumtaz, "Adaptive transmission for reconfigurable intelligent surface-assisted OFDM wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2653–2665, 2020.

[15] T. Manafi, M. A. al-Tarifi, and D. S. Filipovic, "Isolation improvement techniques for wideband millimeter-wave repeaters," *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 2, pp. 355–358, 2018.

[16] L. Kong, Y. Ai, S. Chatzinotas, and B. Ottersten, "Effective rate evaluation of RIS-assisted communications using the sums of cascaded α-μ random variates," *IEEE Access*, vol. 9, pp. 5832–5844, 2021.

[17] L. Dai, B. Wang, M. Wang et al., "Reconfigurable intelligent surface-based wireless communications: antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45913–45923, 2020.

[18] S. Atapattu, R. Fan, P. Dharmawansa, G. Wang, J. Evans, and T. A. Tsiftsis, "Reconfigurable intelligent surface assisted two-way communications: performance analysis and optimization," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6552–6567, 2020.

[19] I. C. B. Garcia, A. Sibille, and M. Kamoun, "Reconfigurable intelligent surfaces: bridging the gap between scattering and reflection," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2538–2547, 2020.

[20] B. Li, Z. Zhang, Z. Hu, and Y. Chen, "Joint array diagnosis and channel estimation for RIS-aided mmWave MIMO system," *IEEE Access*, vol. 8, pp. 193992–194006, 2020.

[21] W. Yan, X. Yuan, Z.-Q. He, and X. Kuai, "Passive beamforming and information transfer design for reconfigurable intelligent surfaces aided multiuser MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1793–1808, 2020.

[22] L. Yang, Y. Yang, M. O. Hasna, and M.-S. Alouini, "Coverage, probability of SNR gain, and DOR analysis of RIS-aided communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 8, pp. 1268–1272, 2020.

[23] E. Basar, "Reconfigurable intelligent surface-based index modulation: a new beyond MIMO paradigm for 6G," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 3187–3196, 2020.

[24] M. Nemati, J. Park, and J. Choi, "RIS-assisted coverage enhancement in millimeter-wave cellular networks," *IEEE Access*, vol. 8, pp. 188171–188185, 2020.

[25] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, "Reconfigurable intelligent surface (RIS) assisted wireless coverage extension: RIS orientation and location optimization," *IEEE Communications Letters*, vol. 25, no. 1, pp. 269–273, 2021.

[26] L. Yang, J. Yang, W. Xie, M. O. Hasna, T. Tsiftsis, and M. D. Renzo, "Secrecy performance analysis of RIS-aided wireless communication systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12296–12300, 2020.

[27] A. Almohamad, A. M. Tahir, A. al-Kababji et al., "Smart and secure wireless communications via reflecting intelligent surfaces: a short survey," *Communications Society*, vol. 1, pp. 1442–1456, 2020.

[28] A. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, 1998.

[29] H. Jafarkhani, *Space-Time Coding: Theory and Practice*, Cambridge University Press, Cambridge, 2005.

[30] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[31] J. N. Laneman and G. W. Wornell, "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2425, 2003.

[32] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity part I: system description," *IEEE Transactions on communications*, vol. 51, no. 11, pp. 1927–1938, 2003.

[33] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity part II: implementation aspects and performance analysis," *IEEE Transactions on communications*, vol. 51, no. 11, pp. 1939–1948, 2003.

[34] Y. Jing and B. Hassibi, "Distributed space-time coding in wireless relay networks," *IEEE Wireless Communications*, vol. 5, no. 12, pp. 3524–3536, 2006.

[35] A. Khaleel and E. Basar, "Reconfigurable intelligent surface-empowered MIMO systems," *IEEE Systems Journal*, vol. 15, no. 3, pp. 4358–4366, 2021.

[36] W. Tang, J. Y. Dai, M. Z. Chen et al., "Realization of reconfigurable intelligent surface-based Alamouti space-time transmission," in *2020 International Conference on Wireless*

Communications and Signal Processing (WCSP), pp. 904–909, Nanjing, 2020.

[37] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, 1999.

[38] Jia Hou, Moon Ho Lee, and Ju Yong Park, "Matrices analysis of quasi-orthogonal space-time block codes," *IEEE Communications Letters*, vol. 7, no. 8, pp. 385–387, 2003.

[39] J. G. Proakis, *Digital Communication*, McGraw-Hill, New York, NY, USA, 5th ed. edition, 2008.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5550825, 7 pages
https://doi.org/10.1155/2021/5550825

Research Article

Tradeoff-HARQ Scheme for Full-Duplex SWIPT DF Relay

Xiaoye Shi, Haiting Zhu, Fei Ding, Zhaowei Zhang, and Nan Bao

School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Correspondence should be addressed to Xiaoye Shi; shixy187@njupt.edu.cn

Received 28 January 2021; Accepted 15 October 2021; Published 25 October 2021

Academic Editor: Miaowen Wen

The SWIPT (simultaneous wireless information and power transfer) DF (decoding and forwarding) relay system could achieve the purpose of both increasing revenue and reducing expenditure. By analysing the system model and transmission characteristics of full-duplex relay, this paper optimizes the retransmission slot structure to enhance the system performance. Firstly, the state transition model is established based on the analysis of the retransmission slot structure. Secondly, the state probability of each state and the transition probability between states are calculated to obtain the total data passing rate, energy transmission efficiency, and total transmission time. Thirdly, in order to compare the performance of various HARQ (hybrid automatic repeat request) schemes more effectively, JNTP (joint normalized throughput of information transmission and energy transmission) is constructed. Monte Carlo simulations finally confirm that the proposed tradeoff-HARQ scheme outperforms the regular-HARQ scheme in terms of JNTP: the performance of the tradeoff-HARQ scheme is 0.03883 higher than that of the regular-HARQ scheme when the total power limit is 20 dB and 0.00651 higher than that of the regular-HARQ scheme when the total power limit is 30 dB.

1. Introduction

The SWIPT (simultaneous wireless information and power transfer) relay system could promote both increasing revenue and reducing expenditure of energy [1]. It is an important technology of green communication [2–4]. Xiaomi Corporation recently officially released its self-developed space isolation SWIPT technology, realizing the practical application of indoor SWIPT technology. The SWIPT relay system is regarded as the ultimate solution for mobile devices by collecting the energy in radiofrequency signals to charge the equipment and prolong the service life of the equipment [5–7]. Meanwhile, the relay can be easily deployed in the near range of equipment, which can improve the efficiency of information transmission and energy transmission at the same time [8]. In reference [9], the two-hop cooperative transmission of SWIPT relay was studied. Multiple antennas at the relay are divided into two disjoint groups for information decoding and energy acquisition, respectively. Reference [10] studied the feasibility of using relay-assisted large-capacity MIMO (multiple input multiple output) to improve the performance of wireless SWIPT. The key idea is to use the redundant degrees of freedom provided by a large number of base station antenna arrays to transmit power and information to the direct/relay users at the same time.

In order to ensure the reliable transmission of high-speed data, the system includes effective error detection and retransmission technology. The HARQ (hybrid automatic repeat request) strategy consists of two parts: (1) error detection, the receiving port will receive the data error detection, if the data is wrong, start retransmission and (2) retransmission, using the reverse link to request the source side to send again [11, 12]. In reference [13], the influence of low-power transmission strategy and incomplete channel state information on outage probability of HARQ-assisted nonorthogonal multiple access systems was obtained by using integral domain partition method and extended to the scenarios with any number of users. Reference [14] revealed the retransmission scheme of hybrid automatic repeat request for uplink transmission in large-scale cellular networks and deeply understands the influence of network parameters (such as power control parameters) on the uplink coverage performance. A new HARQ scheme for the internet of things relay was proposed: if the acknowledgement signal is negative, the relay will select an appropriate high-order modulation constellation to transmit the source signal in one transmission slot or several time slots; if there is no error, the source will continue to transmit [15]. In reference [16], a generalized two-dimensional discrete time Markov chain model was established, and the state transition probability is analysed. By calculating the steady-state distribution of Markov chain, the closed expressions of throughput and energy efficiency of future vehicular mobile networks are derived.

Although the research on HARQ technology of SWIPT system has broad research potential, the current research is not sufficient. In reference [17], HARQ was introduced into the SWIPT direct link, and an optimal strategy aiming at the minimum expected retransmission times is proposed. The strategy only uses the received RF signal to obtain energy or accumulate mutual information. Based on the deliberate thoughts on the slot structure of the full-duplex SWIPT DF (decoding and forwarding) relay, this paper proposes a tradeoff-HARQ scheme and analyses its performance considering information transmission and energy acquisition at the same time. The main contributions in this paper are summarized as follows: (a) This paper analyses the retransmission slot structure according to the system model of the full-duplex SWIPT DF relay. (b) The state transition model is established, the transition probability between states is calculated, and then, the state transition matrix is constructed. (c) According to the initial state and state transition matrix, the state probability is calculated to obtain the total data passing rate, energy collection efficiency, and transmission time. (d) In order to compare the performance of various HARQ schemes more effectively, joint normalized throughput of information transmission and energy transmission is constructed after the parameters are normalized.

2. System Model of the Full-Duplex SWIPT DF Relay

As shown in Figure 1, the full-duplex DF relay system consists of base station B, relay R, and destination user D. Both the base station and the user are in a single-antenna half-duplex mode. The relay is equipped with two antennas in full-duplex mode; that is, it can receive the signal sent by the base station and relay the signal to the destination user at the same time. The channel fading of BR and RD is represented by h_{SR} and h_{RD}, respectively, which is a complex Gaussian distribution with a mean value of zero and variance of σ_{SR}^2 and σ_{RD}^2, respectively. The equivalent noise $n_i(i \in \{R, D\})$ at the receiver i is a complex Gaussian distribution with a mean value of 0 and a variance of N_0. Some negative factors, such as the error of the nonlinearity of the amplifier and self-interference estimation and reconstruction, may lead to the self-interference SI_R caused by full duplex. Due to the recent development of self-interference cancellation technology, the full-duplex transceiver can simultaneously transmit and receive signals in the same frequency band, and the performance gap caused by self-interference is far less than other interference signals [18, 19]. h_{ii} represent the channel fading of the self-interference channel, which is the Gaussian distribution of mean zero and variance σ_{RR}^2.

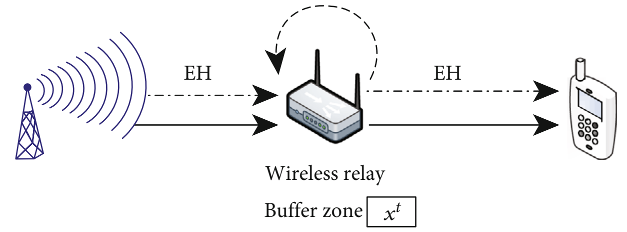


Figure 1: The system model graph of the full-duplex SWIPT DF relay.

A buffer is set at relay R to store the received data x^t temporarily. Without losing generality, it is assumed that the feedback signal can be received without error and delay and the number of retransmissions of HARQ is limited (it is assumed that the number is $M-1$). The transmission signal of S is the complex Gaussian distribution with a mean value of 0 and variance of 1, and the error correcting code can reach the channel capacity. The receiver of each link performs CRC (cyclic redundancy check) verification, and the verification result is fed back to the transmitter through the feedback link:

(a) The identifier θ_1^t reflects whether the relay has successfully received the feedback signal of the tth frame data, $\theta_1^t = 0$ indicates the state that the relay has successfully received the tth frame data, and $\theta_1^t = 1$ indicates the unsuccessful state

(b) The identifier θ_2^t reflects whether the t-1 frame data in the R-D link is successfully received, $\theta_2^t = 0$ indicates the state that the relay has successfully received the previous frame data, and $\theta_2^t = 1$ indicates the unsuccessful state

3. Slot Structure and State Transition Model

To take into account both information transmission and energy acquisition, each transmission frame (assuming that the frame length is FL, and the time required to transmit a frame is FT, which remains unchanged in the system) is divided into two time slots: information transmission slot and energy transmission slot. The occupied time of information transmission slot is αFT, and the energy transmission slot is $(1\text{-}\alpha)$FT $(0 \leq \alpha \leq 1$ is the proportion of information transmission slot).

According to the state of the two links, the system is divided into four states (τ is used to mark the actual number of transmitted data frames; $\tau \leq t$ is used to mark the actual number of transmitted data frames):

(a) Double normal data frames, denoted as $G_{0,0}$, i.e., $\theta_1^t = 0$ and $\theta_2^t = 0$. Both links BR and RD continues to transmit new data frames. The expression of the received signals of link *RD* and link *BR* is as follows:

$$\begin{aligned} y_{RD}^{t} &= h_{RD}^{t}\sqrt{E_R}x^{\tau\text{-}1} + n_{RD}^{t}, \\ y_{BR}^{t} &= h_{BR}^{t}\sqrt{E_B}x^{\tau} + h_{RR}^{t}\sqrt{E_R}x^{r-1} + n_{BR}^{t} \end{aligned} \tag{1}$$

(b) Double retransmit frame, denoted as $G_{1,1}$, i.e. $\theta_1^t = 1$ and $\theta_2^t = 1$. Links *BR* and *RD* goes on retransmit the data of the previous frame. The expression of the received signals of link *RD* and link *BR* is as follows:

$$\begin{aligned} y_{RD}^{t} &= h_{RD}^{t}\sqrt{E_R}x^{\tau\text{-}1} + n_{RD}^{t}, \\ y_{BR}^{t} &= h_{BR}^{t}\sqrt{E_B}x^{\tau} + h_{RR}^{t}\sqrt{E_R}x^{\tau\text{-}1} + n_{BR}^{t} \end{aligned} \tag{2}$$

(c) Relay energy collection-relay retransmits frame, denoted as $G_{0,1}$, i.e., $\theta_1^t = 0$ and $\theta_2^t = 1$. In this case, the receiving antenna of the relay is used to collect energy, and the transmitting antenna is used to retransmit the data of the previous frame. The expression of the received signal of the link *RD* is as follows:

$$y_{RD}^{t} = h_{RD}^{t}\sqrt{E_R}x^{\tau\text{-}1} + n_{RD}^{t} \tag{3}$$

(d) Destination user energy collection-base station retransmits frame, which is recorded as $G_{1,0}$, that is, $\theta_1^t = 1$ and $\theta_2^t = 0$. In this case, the user's receiving antenna is used to collect energy, and the base station retransmits the data of the previous frame. The expression of the received signal of link *BR* is as follows:

$$y_{BR}^{t} = h_{BR}^{t}\sqrt{E_B}x^{\tau} + h_{RR}^{t}\sqrt{E_R}x^{\tau\text{-}1} + n_{BR}^{t} \tag{4}$$

Without losing generality, it is assumed that all channels obey quasistatic Rayleigh fading, that is, the transmission quality of the channel does not change in the retransmission frame. According to the received signal expression of each frame, the average SNR of links *BR* and *RD* are illustrated in equations (5) and (6), respectively.

$$\bar{\mu}_{BR} = \frac{E\left[E_B|h_{BR}|^2\right]}{\left(E\left[E_R|h_{RR}|^2\right] + N_0\right)} = \frac{E_B\sigma_{BR}^2}{E_R\sigma_{RR}^2 + N_0}, \tag{5}$$

$$\bar{\mu}_{RD} = \frac{E\left[E_R|h_{RD}|^2\right]}{N_0} = \frac{E_R\sigma_{RD}^2}{N_0}. \tag{6}$$

The data rate threshold required by the system is recorded as T. According to the definition of channel capacity and outage probability, the outage probability $P_{RD}(R)$ on link *RD* is $1 - \exp\left(-(2^T - 1)/\bar{\mu}_{RD}\right)$. Therefore, after transmitting the same data m times in the *RD* channel and combining the received signals with the maximum ratio, the outage probability $P_{RD}^{(m)}(T)$ is approximately equal to $(2^T - 1)^m/m! * (\bar{\mu}_{RD})^m$.

Figure 2 shows the system state transition diagram. According to whether R and D receive the data packet successfully or not, it can be divided into five types of states:

(a) $F_{0,0}$ indicates the normal transmission state and links *BR* and *RD* transmit new data

(b) $F_{m,0}$ represents the relay retransmission state; that is, after the relay transmits m times, the relay successfully receives the τ frame data, but D still does not receive the τ-1 frame data successfully

(c) $F_{0,m}$ indicates the retransmission status of the base station; that is, D successfully receives the τ-1 frame data after m times of transmission, but the relay still fails to receive the τ frame data

(d) $F_{m,m}$ indicates the retransmission status of the base station and relay; that is, the R transmits m times, and B transmits m times, but D fails to receive the τ-1 frame data and R fails to receive the τ frame data

(e) $S_{m,m}$ indicates R and D successfully receive the data after m times of transmission

The state transition probability is calculated according to the independent event formula and conditional probability formula:

$$\begin{aligned} P\left(F_{m\text{-}1,m\text{-}1} \longrightarrow F_{m,m}\right) &= \frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)}\frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)}, \\ P\left(F_{m\text{-}1,m\text{-}1} \longrightarrow F_{m,0}\right) &= \frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)}\frac{\left(1\text{-}P_{RD}^{(m)}(T)\right)}{P_{RD}^{(m\text{-}1)}(T)}, \\ P\left(F_{m\text{-}1,m\text{-}1} \longrightarrow F_{0,m}\right) &= \frac{\left(1\text{-}P_{BR}^{(m)}(T)\right)}{P_{BR}^{(m\text{-}1)}(T)}\frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)}, \\ P\left(F_{m\text{-}1,m\text{-}1} \longrightarrow S_{m,m}\right) &= \frac{\left(1\text{-}P_{BR}^{(m)}(T)\right)}{P_{BR}^{(m\text{-}1)}(T)}\frac{\left(1\text{-}P_{RD}^{(m)}(T)\right)}{P_{RD}^{(m\text{-}1)}(T)}. \end{aligned} \tag{7}$$

Therefore, the state transfer equation is as follows:

$$\begin{bmatrix} P(S_{m,m}) \\ P(F_{m,0}) \\ P(F_{0,m}) \\ P(F_{m,m}) \end{bmatrix} = H_{m\text{-}1} * \begin{bmatrix} P(S_{m\text{-}1,m\text{-}1}) \\ P(F_{m\text{-}1,0}) \\ P(F_{0,m\text{-}1}) \\ P(F_{m\text{-}1,m\text{-}1}) \end{bmatrix}, \tag{8}$$

where the state transition matrix is

$$H_{m\text{-}1} = \begin{bmatrix} 0 & \left(1\text{-}\frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)}\right) & \left(1\text{-}\frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)}\right) & \left(1\text{-}\frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)}\right) * \left(1\text{-}\frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)}\right) \\ 0 & \frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)} & 0 & \frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)} * \left(1\text{-}\frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)}\right) \\ 0 & 0 & \frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)} & \left(1\text{-}\frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)}\right) * \frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)} \\ 0 & 0 & 0 & \frac{P_{BR}^{(m)}(T)}{P_{BR}^{(m\text{-}1)}(T)} * \frac{P_{RD}^{(m)}(T)}{P_{RD}^{(m\text{-}1)}(T)} \end{bmatrix}. \tag{9}$$

$F_{0,0}$ is the initial state, and its state probability is $P(F_{0,0}) = 1$.

The probability of each state is calculated by the state transition equation:

$$\begin{aligned} P(F_{m,m}) &= P_{BR}^{(m)}(T) P_{RD}^{(m)}(T), \\ P(F_{m,0}) &= P_{BR}^{(m)}(T)\left(1\text{-}P_{RD}^{(m)}(T)\right), \\ P(F_{0,m}) &= \left(1\text{-}P_{BR}^{(m)}(T)\right) P_{RD}^{(m)}(T), \end{aligned}$$

$$\begin{aligned} P(S_{m,m}) = {} & \left(P_{BR}^{(m\text{-}1)}(T)\text{-}P_{BR}^{(m)}(T)\right)\left(1\text{-}P_{RD}^{(m\text{-}1)}(T)\right) \\ & + \left(P_{RD}^{(m\text{-}1)}(T)\text{-}P_{RD}^{(m)}(T)\right)\left(1\text{-}P_{BR}^{(m\text{-}1)}(T)\right) \\ & + \left(P_{BR}^{(m\text{-}1)}(T)\text{-}P_{BR}^{(m)}(T)\right)\left(P_{RD}^{(m\text{-}1)}(T)\text{-}P_{RD}^{(m)}(T)\right). \end{aligned} \tag{10}$$

4. Joint Normalized Throughput of Information Transmission and Energy Transmission

No matter whether the link *BR* or *RD* fails to decode the data after reaching the limit of transmission times, the transmission data will be discarded. Thus, after M-1 retransmissions (i.e., M transmissions), the states that the system still fails including $F_{M,0}$, $F_{0,M}$, and $F_{M,M}$. The outage probability is the sum of the above state probabilities. Substituting the probabilities of each state, the outage probability of the system is obtained as follows:

$$\begin{aligned} P_{\text{out}} &= P(F_{M,0}) + P(F_{0,M}) + P(F_{M,M}) \\ &= \left(1\text{-}P_{BR}^{(M)}(T)\right) P_{RD}^{(M)}(T) + P_{BR}^{(M)}(T)\left(1\text{-}P_{RD}^{(M)}(T)\right) + P_{BR}^{(M)}(T) P_{RD}^{(M)}(T) \\ &= P_{RD}^{(M)}(T) + P_{BR}^{(M)}(T)\text{-}P_{BR}^{(M)}(T) P_{RD}^{M}(T). \end{aligned} \tag{11}$$

The states of the system successfully receiving the data from the base station include all $S_{m,m}$, that is, $S_{1,1}, S_{2,2}, \cdots, S_{m,m}, \cdots, S_{M,M}$. Then, the total data pass rate of the system is

$$\begin{aligned} P_{pass} &= \sum_{m=1}^{M} S_{m,m} \\ &= \sum_{m=1}^{M} \begin{pmatrix} \left(P_{BR}^{(m\text{-}1)}(T)\text{-}P_{BR}^{(m)}(T)\right)\left(1\text{-}P_{RD}^{(m\text{-}1)}(T)\right) \\ +\left(P_{RD}^{(m\text{-}1)}(T)\text{-}P_{RD}^{(m)}(T)\right)\left(1\text{-}P_{BR}^{(m\text{-}1)}(T)\right) \\ +\left(P_{BR}^{(m\text{-}1)}(T)\text{-}P_{BR}^{(m)}(T)\right)\left(P_{RD}^{(m\text{-}1)}(T)\text{-}P_{RD}^{(m)}(T)\right) \end{pmatrix} \\ &= \left(1\text{-}P_{BR}^{(M)}(T)\text{-}P_{RD}^{(M)}(T) + P_{BR}^{(M)}(T) P_{RD}^{(M)}(T)\right). \end{aligned} \tag{12}$$

The transmission times of state $F_{m,0}$, $F_{0,m}$, $F_{m,m}$, and $S_{m,m}$ are all m times. When the system stops retransmission, the possible states are $F_{M,0}$, $F_{0,M}$, $F_{M,M}$, and all $S_{m,m}$ states. Therefore, the total transmission time of each frame is

$$\begin{aligned} ST &= \mathrm{FT} * \left(M * P(F_{M,0}) + M * P(F_{0,M}) + M * P(F_{M,M}) + \sum_{m=1}^{M} (m * P(S_{m,m}))\right) \\ &= \mathrm{FT} * \begin{pmatrix} M * \left(P_{RD}^{(M)}(T) + P_{BR}^{(M)}(T)\text{-}P_{BR}^{(M)}(T) P_{RD}^{M}(T)\right) \\ + \sum_{m=1}^{M} \left(m * \begin{pmatrix} P_{BR}^{(m\text{-}1)}(T)\text{-}P_{BR}^{(m)}(T) + P_{RD}^{(m\text{-}1)}(T)\text{-}P_{RD}^{(m)}(T) \\ \text{-}P_{RD}^{(m\text{-}1)}(T) P_{BR}^{(m\text{-}1)}(T) + P_{BR}^{(m)}(T) P_{RD}^{(m)}(T) \end{pmatrix}\right) \end{pmatrix}. \end{aligned} \tag{13}$$

When the system is in this state $F_{m,0} (m < M)$, the relay will be in the state of energy collection in the next frame. In addition, the first a part of each frame of the system is also used for energy collection, and the energy collected at this time is $\mathrm{FT} * \eta\sigma_{BR}^2 E_R$. Therefore, the total power of relay energy collection is

$$EH_R = (\alpha)\mathrm{FT} * \eta\sigma_{BR}^2 E_B \sum_{m=1}^{M-1} P(F_{m,0}) + (1-\alpha)\mathrm{FT} * \eta\sigma_{BR}^2 E_B, \tag{14}$$

where η represents the received energy conversion efficiency.

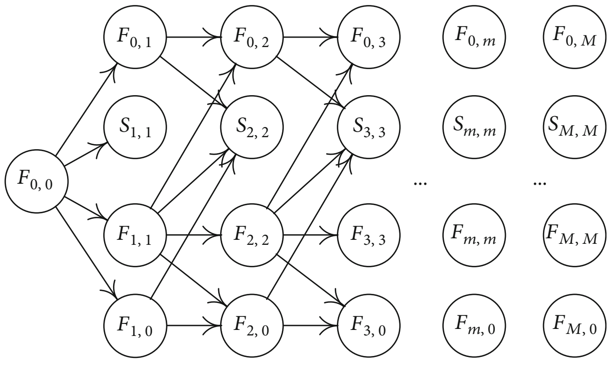


FIGURE 2: The state transition model diagram.

Similarly, the total power of D energy collection in each frame is

$$EH_D = (\alpha)\mathrm{FT} * \eta\sigma_{RD}^2 E_R \sum_{m=1}^{M\text{-}1} P(F_{0,m}) + (1-\alpha)\mathrm{FT} * \eta\sigma_{RD}^2 E_R. \tag{15}$$

In order to better study the relationship between energy transmission and information transmission, the energy transmission power is normalized:

$$\begin{aligned} EH_R^{\mathrm{norm}} &= \frac{EH_R}{EH_R^{\max}} = (1\text{-}\alpha) + \alpha \sum_{m=1}^{M\text{-}1} P(F_{m,0}), \\ EH_D^{\mathrm{norm}} &= \frac{EH_D}{EH_D^{\max}} = (1\text{-}\alpha) + \alpha \sum_{m=1}^{M\text{-}1} P(F_{0,m}), \end{aligned} \tag{16}$$

where $EH_R^{\max} = \mathrm{FT} * \eta\sigma_{BR}^2 E_B$ and $EH_D^{\max} = \mathrm{FT} * \eta\sigma_{RD}^2 E_R$.

Without losing generality, the JNTP (joint normalized throughput of information transmission and energy transmission) is defined as

$$TP = \frac{\left(EH_R^{\mathrm{norm}} + EH_{\mathrm{D}}^{\mathrm{norm}} + \alpha P_{\mathrm{pass}}\right)}{ST}. \tag{17}$$

In the multi-SWIPT relay scenario, relay selection is a very important problem, which can effectively improve the information transmission rate and energy transmission rate of the system [20]. The channel fading of radio wave is related to the transmission distance of information. We assume that the distance between B and R is d_{BR}, the distance between R and D is d_{RD}, and the path loss exponent γ_{L}. According to the large-scale fading multislope model, the relationship between the variance of channel fading and the terminal distance is as follows: $\sigma_{BR}^2 \sim (d_{BR})^{-\gamma_{\mathrm{L}}}$ and $\sigma_{RD}^2 \sim (d_{RD})^{-\gamma_{\mathrm{L}}}$. Therefore, the problem of relay selection can be solved by finding the best relay location (d_{BR}^*, d_{RD}^*):

$$d_{BR}^*, d_{RD}^* = \arg\max TP(d_{BR}, d_{RD}). \tag{18}$$

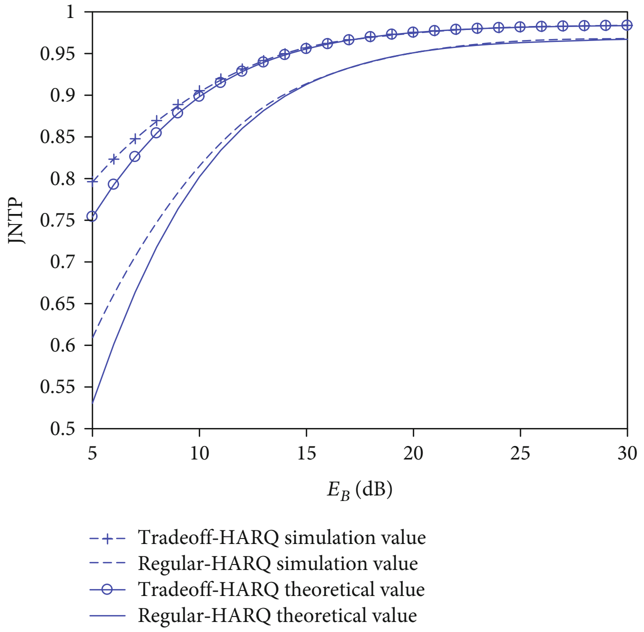


FIGURE 3: The JNTP of the proposed tradeoff-HARQ scheme and the regular-HARQ scheme against E_B.

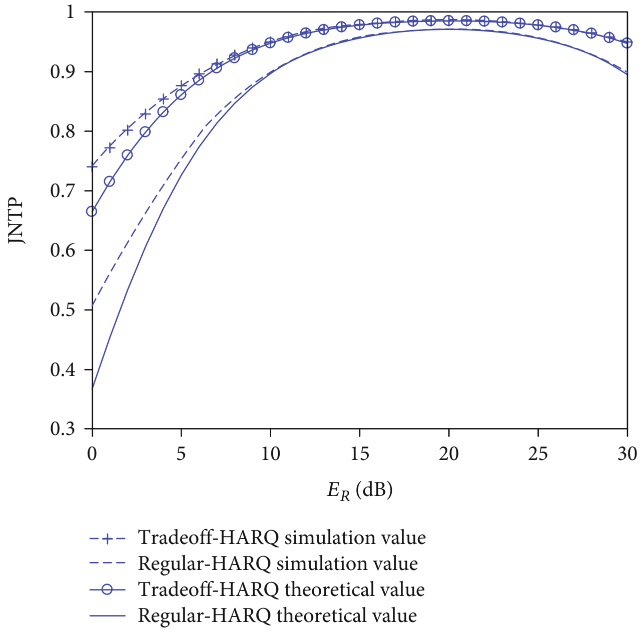


FIGURE 4: The JNTP with E_R under different schemes.

5. Monte Carlo Simulations

In this section, Monte Carlo simulation is used to verify the theoretical value. The simulation software is MATLAB, and the general simulation parameters are as follows: the number of transmission is limited to $M = 3$ (in other words, the number of retransmission is limited to 2), and data rate threshold $T = 1$, $\sigma_{BR}^2 = 1$, $\sigma_{RD}^2 = 1$, and $\sigma_{RR}^2 = 0.01$.

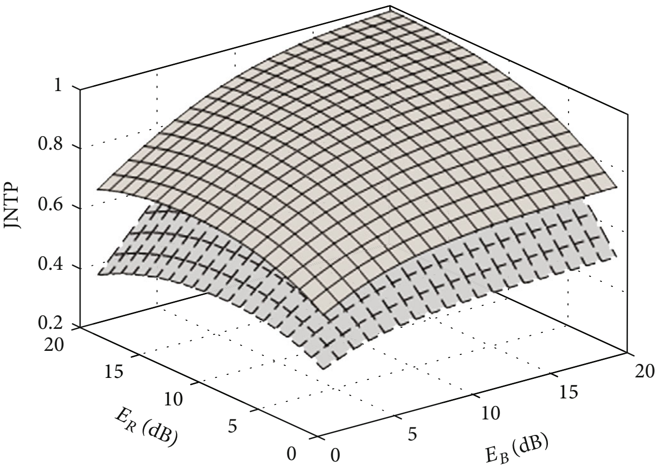


Figure 5: The JNTP curved surface of different schemes.

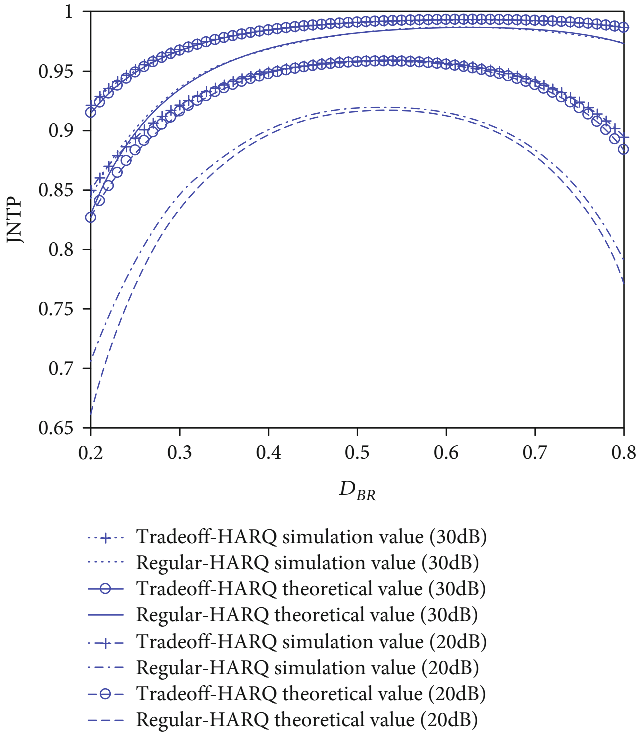


Figure 6: The trend chart of JNTP and distance parameters under different total power limits.

The JNTP of the proposed tradeoff-HARQ scheme and the regular-HARQ scheme against E_B is given in Figure 3. The simulation condition is $E_R = 15$ dB, $\alpha = 0.5$. The "+" and "O" in this figure are the simulation curve and theoretical curve of the JNTP under the tradeoff-HARQ scheme. The unmarked curve is the regular-HARQ scheme. Comparing the theoretical value with the simulation value, it can be seen that there is a small amount of error at low SNR (less than 10 dB), but the error can be ignored at other SNR, so it can be considered that the theoretical value can approximate the simulation value of the JNTP. This figure shows that the JNTP performance of the tradeoff-HARQ scheme is better than that of the regular-HARQ scheme. At low SNR, such as $E_B = 5$ dB, the JNTP of the tradeoff-HARQ scheme is 0.22 higher than that of the regular-HARQ scheme. At high SNR, such as $E_B > 20$ dB, the tradeoff-HARQ scheme is still better than the regular-HARQ scheme, and its advantage is reduced to 0.02.

Figure 4 describes the joint normalized throughput under different schemes with E_R. The simulation condition is $E_B = 15$ dB, $\alpha = 0.5$, and the mark is the same as Figure 3. As shown in this figure, when E_R is close to 20 dB, the JNTP of the proposed tradeoff-HARQ scheme reaches the maximum. The self-interference signal of the relay increase with the relay transmission power, and the received signal to interference noise ratio of the relay decrease with the increase of E_R.

For better insights, the JNTP curved surface of different schemes is illustrated in Figure 5. The dotted line is the regular-HARQ scheme. As shown in this figure, the proposed tradeoff-HARQ scheme is superior to the regular-HARQ scheme in most cases of E_B and E_R. The advantage is obvious in low SNR.

Figure 6 elaborates the change trend of JNTP and distance parameters of different schemes under different total power limits. The simulation condition is that the total power is limited to 20 dB and 30 dB, and the curve mark is shown in this figure. In order to investigate the relationship between relay location and performance, we assume that the distance between B and D is 1, and the distance between B and R is normalized to the distance parameter D_{BR}. For convenience of description, the simulation in Figure 6 assumes that the R is located on the connection between B and D. When the total power is limited to 20 dB and the distance parameter is 0.54, the JNTP of the tradeoff-HARQ scheme and the regular-HARQ reaches the maximum of 0.95937 and 0.92054, and the performance of the tradeoff-HARQ scheme is 0.03883 higher than that of the regular-HARQ. When the total power is limited to 30 dB, the maximum values are 0.99345 and 0.98694, respectively (when the distance parameter is about 0.63). The performance of the tradeoff-HARQ scheme is 0.00651 higher than that of the regular-HARQ scheme.

6. Conclusions

In order to develop the potential of SWIPT DF relay, this paper proposes the tradeoff-HARQ scheme of full-duplex SWIPT DF relay. First of all, the system model and retransmission slot structure are optimized by using the characteristics of full-duplex relay. Then, the state transition model is established by analysing the slot structure. Furthermore, the state probability and transition probability of each state are calculated to obtain the total passing rate, energy collection efficiency, and the total transmission time. Ultimately, in order to compare the performance of various HARQ schemes more effectively, the joint normalized throughput is constructed. Simulation results show that the joint normalized throughput of the proposed tradeoff-HARQ scheme is better than that of the regular-HARQ scheme. When the total power is limited to 20 dB, the performance of the tradeoff-HARQ scheme is 0.03883 higher than that of the regular-HARQ scheme.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (61701251, 62071244, 62071249, 61801236, 61806100, and 61872423), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (20KJB630003), and the Natural Science Foundation of Jiangsu Province (Grant No. BK20160903).

References

[1] I. Krikidis, S. Timotheou, S. Nikolaou, G. Zheng, D. W. Ng, and R. Schober, “Simultaneous wireless information and power transfer in modern communication systems,” *IEEE Communications Magazine*, vol. 52, no. 11, pp. 104–110, 2014.

[2] Z. Xie, “Secured green communication scheme for interference alignment based networks,” *Journal of Communications and Networks*, vol. 22, no. 1, pp. 23–36, 2020.

[3] A. Abrol and R. K. Jha, “Power optimization in 5G networks: a step towards green communication,” *IEEE Access*, vol. 4, pp. 1355–1374, 2016.

[4] Z. Chen, H. Xu, L. X. Cai, and Y. Cheng, “Beamforming design for max-min fair SWIPT in green cloud-RAN with wireless fronthaul,” in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Abu Dhabi, United Arab Emirates, December 2018.

[5] I. Krikidis, “Relay selection in wireless powered cooperative networks with energy storage,” *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 12, pp. 2596–2610, 2015.

[6] X. Shi, J. Sun, D. Li, F. Ding, and Z. Zhang, “Rate-energy tradeoff for wireless simultaneous information and power transfer in full-duplex and half-duplex systems,” *CMC-Computer Materials and Continua*, vol. 65, no. 2, pp. 1373–1384, 2020.

[7] N. Ashraf, S. A. Sheikh, S. A. Khan, I. Shayea, and M. Jalal, “Simultaneous wireless information and power transfer with cooperative relaying for next-generation wireless networks: a review,” *IEEE Access*, vol. 9, pp. 71482–71504, 2021.

[8] G. Pan, “On secrecy performance of MISO SWIPT systems with TAS and imperfect CSI,” *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3831–3843, 2016.

[9] Y. Liu, “Joint resource allocation in SWIPT-based multiantenna decode-and-forward relay networks,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9192–9200, 2017.

[10] D. Kudathanthirige, R. Shrestha, and G. A. A. Baduge, “Wireless information and power transfer in relay-assisted downlink massive MIMO,” *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 789–805, 2019.

[11] C. Tseng and S. Wu, “Selective and opportunistic AF relaying for cooperative ARQ: an MLSD perspective,” *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 124–139, 2019.

[12] S. Luo and K. C. Teh, “Amplify-and-forward based two-way relay ARQ system with relay combination,” *IEEE Communications Letters*, vol. 19, no. 2, pp. 299–302, 2015.

[13] Z. Shi, C. Zhang, Y. Fu, H. Wang, G. Yang, and S. Ma, “Achievable diversity order of HARQ-aided downlink NOMA systems,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 471–487, 2020.

[14] X. Lu, E. Hossain, H. Jiang, and G. Li, “On coverage probability with type-II HARQ in large-scale uplink cellular networks,” *IEEE Wireless Communications Letters*, vol. 9, no. 1, pp. 3–7, 2020.

[15] M. Otaru, M. Ajiya, A. Adinoyi, M. Aljlayl, and H. Yanikomeroglu, “An ARQ-based cooperative relaying scheme for 5G IoT slice,” in *IEEE Global Conference on Internet of Things (GCIoT)*, pp. 1–5, Dubai, United Arab Emirates, 2019.

[16] S. Li, F. Wang, J. Gaber, and X. Chang, “Throughput and energy efficiency of cooperative ARQ strategies for VANETs based on hybrid vehicle communication mode,” *IEEE Access*, vol. 8, pp. 114287–114304, 2020.

[17] M. S. H. Abad, O. Ercetin, T. ElBatt, and M. Nafie, “SWIPT using hybrid ARQ over time varying channels,” *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 1087–1100, 2018.

[18] J. Liu, Q. Rong, and M. Wang, “Adaptive self-interference cancellation in broadband full-duplex MIMO relays,” in *Asia-Pacific International Symposium on Electromagnetic Compatibility*, pp. 570–572, Shenzhen, China, May 2016.

[19] R. P. Sirigina and A. S. Madhukumar, “Interference cancellation through interference forwarding in relay-assisted systems,” *IEEE Systems Journal*, vol. 12, no. 3, pp. 2373–2384, 2018.

[20] S. Gautam, E. Lagunas, S. K. Sharma, S. Chatzinotas, and B. Ottersten, “Relay selection strategies for SWIPT-enabled cooperative wireless systems,” in *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications*, pp. 1–7, Montreal, QC, Canada, October 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4099918, 11 pages
https://doi.org/10.1155/2021/4099918

Research Article

Design of Nonbinary LDPC Cycle Codes with Large Girth from Circulants and Finite Fields

Hengzhou Xu,[1,2] **Huaan Li**,[1] **Jixun Gao**,[3] **Guixiang Zhang**,[1] **Hai Zhu**,[1] **and Xiao-Dong Zhang**[2]

[1] *School of Network Engineering, Zhoukou Normal University, Zhoukou 466000, China*
[2] *School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*
[3] *School of Computer, Henan University of Engineering, Zhengzhou 451191, China*

Correspondence should be addressed to Hengzhou Xu; hzxu@zknu.edu.cn

Received 21 May 2021; Accepted 8 September 2021; Published 23 September 2021

Academic Editor: Miaowen Wen

In this paper, we study a class of nonbinary LDPC (NBLDPC) codes whose parity-check matrices have column weight 2, called NBLDPC cycle codes. We propose a design framework of $(2,\rho)$-regular binary quasi-cyclic (QC) LDPC codes and then construct NBLDPC cycle codes of large girth based on circulants and finite fields by randomly choosing the nonzero field elements in their parity-check matrices. For enlarging the girth values, our approach is twofold. First, we give an exhaustive search of circulants with column/row weight ρ and design a masking matrix with good cycle distribution based on the edge-node relation in undirected graphs. Second, according to the designed masking matrix, we construct the exponent matrix based on finite fields. The iterative decoding performances of the constructed codes on the additive white Gaussian noise (AWGN) channel are finally provided.

1. Introduction

Nonbinary low-density parity-check (NBLDPC) codes based on modulo arithmetics were first discovered by Gallager in 1960s [1] and redefined over finite fields GF (q) by Davey and MacKay in 1998 [2]. Similar to binary LDPC codes, NBLDPC codes also have the ability of approaching capacity when decoded with the iterative algorithms. Moreover, NBLDPC codes have much better performance than binary LDPC codes for the short and moderate code lengths. As much more low-complexity decoding algorithms were proposed [3–8], NBLDPC codes provide a promising coding scheme for 6G communications [9].

As shown in [10], NBLDPC codes over larger finite fields will have much better performance for a constant code length. However, when the finite field size is sufficiently large, the performance improvement is little. Moreover, when the finite field size is equal or greater than 64, the column weights of the parity-check matrices of good NBLDPC codes tend to 2. Since NBLDPC cycle codes perform well over various channels [11–13], it is worth studying NBLDPC codes over large finite fields whose parity-check matrices have column weight 2, referred to as NBLDPC cycle codes. As an important cycle codes, $(2,\rho)$-regular NBLDPC codes also perform well under iterative decoding; lots of methods for constructing such codes were proposed [14–17]. Among these works on the construction of NBLDPC codes, the codes can be mainly classified into two categories: the first one is constructed by means of computer search under the algorithms satisfying certain rules, and the other one is constructed based on combinatorial designs, graph theory, matrix theory, and finite fields [18]. Simulation results show that they all have good performance. For a given code rate and length, it is of great interest to study which one of them has the best error performance.

Cycle structure plays an important role in binary/nonbinary LDPC codes. Research results show that NBLDPC codes with large girth have good iterative performance [19]. In general, NBLDPC codes with large girth have large

Hamming minimum distance, and it can be ensured that NBLDPC codes have good performance in the waterfall and error-floor region. Hence, it is interesting to construct LDPC cycle codes with large girth.

In this paper, we focus on the construction of $(2, \rho)$-regular quasi-cyclic LDPC (QC-LDPC) codes with large girth. We first proposed the construction framework of $(2, \rho)$-regular QC-LDPC codes based on the edge-node relation in undirected graphs and transfer the construction of $(2, \rho)$-regular QC-LDPC codes into two main parts, i.e., circulants and exponent matrices. In the first part, we find circulants with good cycle distribution by performing an exhaustive search. In order to prune the search space of circulants, isomorphism theory of circulants is proposed. For the second part, we directly employ finite fields to construct exponent matrices of QC-LDPC codes. Here, the employed finite fields are divided two types, i.e., prime fields and finite fields of characteristic 2. Finally, numerical results to show the good performance of our proposed codes are provided.

The rest of this paper is organized as follows. Section 2 introduces the definitions of LDPC codes and their associated Tanner graphs. Section 3 presents the design framework of $(2, \rho)$-regular QC-LDPC codes. Design of NBLDPC cycle codes with large girth is proposed in Section 4, and numerical results are also provided in this section. Finally, Section 5 concludes this paper.

2. Preliminaries

2.1. LDPC Codes. A binary (γ, ρ)-regular LDPC code is generated by the null space of an $m \times n$ sparse parity-check matrix $\mathbf{H}$ over GF (2), and the matrix $\mathbf{H}$ has the following properties: (1) each column has γ 1's; (2) each row has ρ 1's; (3) $\gamma \ll m$ and $\rho \ll n$. If the sparse matrix $\mathbf{H}$ is over GF (q) for q being a prime power, then LDPC codes generated by such $\mathbf{H}$ are called nonbinary codes or q-ary codes. Binary LDPC codes are referred to as quasi-cyclic (QC) [20], if their parity-check matrices $\mathbf{H}$ have the following form

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}(p_{1,1}) & \mathbf{I}(p_{1,2}) & \mathbf{I}(p_{1,3}) & \cdots & \mathbf{I}(p_{1,\rho}) \\ \mathbf{I}(p_{2,1}) & \mathbf{I}(p_{2,2}) & \mathbf{I}(p_{2,3}) & \cdots & \mathbf{I}(p_{2,\rho}) \\ \mathbf{I}(p_{3,1}) & \mathbf{I}(p_{3,2}) & \mathbf{I}(p_{3,3}) & \cdots & \mathbf{I}(p_{3,\rho}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}(p_{\gamma,1}) & \mathbf{I}(p_{\gamma,2}) & \mathbf{I}(p_{\gamma,3}) & \cdots & \mathbf{I}(p_{\gamma,\rho}) \end{bmatrix}_{\gamma Q \times \rho Q}, \tag{1}$$

where for $1 \le i \le \gamma, 1 \le j \le \rho, -1 \le p_{i,j} \le Q-1$, $\mathbf{I}(p_{i,j})$ is a $Q \times Q$ circulant permutation matrix (CPM) formed by cyclically shifting each row of a $Q \times Q$ identity matrix $\mathbf{I}$ to the right (or left) by $p_{i,j} (\mathrm{mod}\ Q)$ positions, and $\mathbf{I}(-1)$ is a zero matrix of size $Q \times Q$. Obviously, $\mathbf{I}(0)$ is an identity matrix of size $Q \times Q$. For example, if $Q = 4$, then

$$\begin{aligned} \mathbf{I}(0) &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \\ \mathbf{I}(1) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \\ \mathbf{I}(2) &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \\ \mathbf{I}(3) &= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \end{aligned} \tag{2}$$

We can easily see that the positions of 1's of $\mathbf{H}$ in (1) are uniquely determined by the following matrix $\mathbf{P}$, called exponent matrix (or permutation shift matrix),

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots & p_{1,\rho} \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots & p_{2,\rho} \\ p_{3,1} & p_{3,2} & p_{3,3} & \cdots & p_{3,\rho} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{\gamma,1} & p_{\gamma,2} & p_{\gamma,3} & \cdots & p_{\gamma,\rho} \end{bmatrix}. \tag{3}$$

It is not hard to see that the correspondence between $\mathbf{P}$ and $\mathbf{H}$ is one-to-one. It is noticeable that the parameter Q is called *expansion factor* (or *lifting degree*) [21]. By replacing 1's in a CPM $\mathbf{I}(p_{i,j})$ of $\mathbf{H}$ in (1) with the same nonzero field element in finite field GF (q), the resulting code is nonbinary QC-LDPC codes [22].

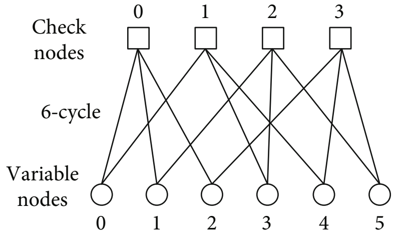
2.2. Tanner Graph. Apart from the matrix representation, an LDPC code can be also described in a simple and intuitive way, i.e., a graphical model called Tanner graph [23]. In fact, the Tanner graph of an LDPC code with the parity-check matrix $\mathbf{H} = [h_{s,t}]$ is a bipartite graph in which the two classes of nodes are variable nodes (representing the code-bit nodes) and check nodes (representing the constraint nodes),

$$H_b = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

(a)

$$H_{nb} = \begin{bmatrix} \alpha^0 & \alpha^0 & \alpha^1 & 0 & 0 & 0 \\ \alpha^1 & 0 & 0 & \alpha^2 & \alpha^1 & 0 \\ 0 & \alpha^2 & 0 & \alpha^0 & 0 & \alpha^0 \\ 0 & 0 & \alpha^0 & 0 & \alpha^2 & \alpha^1 \end{bmatrix}$$

(b)



(c)

Figure 1: Tanner graph of $\mathbf{H}_b$ (or $\mathbf{H}_{nb}$): (a) $\mathbf{H}_b$ over GF (2); (b) $\mathbf{H}_{nb}$ over GF (4); (c) Tanner graph.

respectively. An edge in a Tanner graph connects the check node s to the variable node t if and only if the row -s and column -t element $h_{s,t}$ in $\mathbf{H}$ is nonzero. A cycle in a Tanner graph is a sequence of the connected check nodes and variable nodes which start and end at the same node in the graph and contain no other nodes more than once. The cycle length is simply the number of the contained edges (or nodes), and the length of the shortest cycle is referred to as girth of the Tanner graph (or an LDPC code). As an example, Figure 1 shows the Tanner graph of $\mathbf{H}_b$ (or $\mathbf{H}_{nb}$) and an associate cycle of length 6 (6-cycle for short).

It is well-known that the iterative decoding algorithm converges to the optimal solution provided that the Tanner graph of an LDPC code is free of cycles [24]. In other words, short cycles, especially, the cycles of length 4, affect the decoding performance when decoded with the iterative algorithms based on belief propagation. In fact, there exist many cycles in an LDPC code with finite length. Hence, in order to avoid short cycles or obtain LDPC codes with large girth, many construction methods and techniques are proposed [25–33].

3. Design Framework of $(2, \rho)$-Regular Binary QC-LDPC Codes

3.1. Edge-Node Relation in Undirected Graphs. Let $G = (V, E)$ be an undirected graph, where V is a set of nodes and E is some subset of the pairs (called edges) $\{\{a, b\}: a, b \in V, a \neq b\}$. A cycle of $G = (V, E)$ has distinct nodes (or edges), and an edge in a cycle has two distinct nodes. If we treat the nodes and edges of $G = (V, E)$ as the check nodes and variable nodes, respectively, then a bipartite graph G_B can be obtained. Consider a cycle of length k (denoted by k -cycle for short) in $G = (V, E)$. This k-cycle will be turned into a $2k$-cycle in the above bipartite graph G_B. In other words, the girth of G_B is double that of $G = (V, E)$. Based on this process, we can construct bipartite graphs (or Tanner graphs) with large girth from an undirected graph. In order to make it clearly, we give an example.

Consider the following 4×4 matrix

$$B_{4\times4} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}. \tag{4}$$

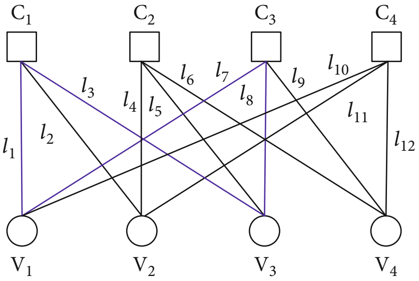


Figure 2: Tanner graph of $B_{4\times4}$.

It is easy to plot the Tanner graph of $B_{4\times4}$, given in Figure 2. By treating the nodes and edges of the Tanner graph in Figure 2 as the check nodes and variable nodes, respectively, we can construct a new bipartite graph, given in Figure 3. We can see from Figures 2 and 3 that a 4-cycle in the Tanner graph of $B_{4\times4}$ becomes a 8-cycle in the bipartite graph.

3.2. Construction Framework of $(2, \rho)$-Regular Binary QC-LDPC Codes. In this subsection, we will present the framework for constructing $(2, \rho)$-regular binary QC-LDPC codes by using the edge-node relation in an undirected graph in Section 3.1. In order to design $(2, \rho)$-regular codes, the node degree of $G = (V, E)$ should be ρ. Furthermore, to guarantee $(2, \rho)$-regular codes are quasi-cyclic, the incidence matrix of $G = (V, E)$ should possess quasi-cyclic structure. In conclusion, the incidence matrix of $G = (V, E)$ is (ρ, ρ)-regular and quasi-cyclic. Hence, in order to obtain $(2, \rho)$-regular binary QC-LDPC codes with large girth, we need to design a (ρ, ρ)-regular quasi-cyclic matrix with large girth. For convenience, this (ρ, ρ)-regular quasi-cyclic matrix is called base matrix. Next, we will give the construction framework.

First, we design a (ρ, ρ)-regular base matrix $\mathbf{B}$ of size $L \times L$. By employing the edge-node relation in Section 3.1, we can transfer the Tanner graph of $\mathbf{B}$ into a new bipartite graph, and the incidence matrix $\mathbf{B}_M$ of such a bipartite graph is obtained. It is obvious that $\mathbf{B}_M$ is a $(2, \rho)$-regular quasi-cyclic matrix of size $2L \times \rho L$. Second, we construct an exponent matrix $\mathbf{P}$ of size $2L \times \rho L$, and the corresponding expansion factor is Q. Third, we use $\mathbf{B}_M$ to mask the exponent matrix $\mathbf{P}$, and a $2L \times \rho L$ array $\mathbf{H}_M$ of $Q \times Q$ CPMs is constructed. The null space of $\mathbf{H}_M$ gives a $(2, \rho)$-regular binary QC-LDPC code of length ρLQ.

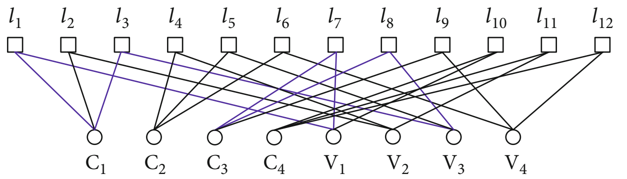


Figure 3: A new bipartite graph constructed from the Tanner graph of $B_{4\times4}$.

4. Design of Nonbinary LDPC Cycle Codes with Large Girth

In order to construct $(2, \rho)$-regular binary QC-LDPC codes with large girth, we only design a base matrix and a corresponding exponent matrix based on the construction framework in Section 3.2. By replacing the nonzero element in the parity-check matrices of binary QC-LDPC cycle codes with the nonzero field elements, nonbinary LDPC cycle codes can be obtained. In this paper, we do not optimize the nonzero field elements and adopt the optimized row elements in [34]. Next, we will provide the construction of the base matrices and exponent matrices.

4.1. Exhaustive Search of Circulants Based on Isomorphism Theory. In this paper, we employ the circulant as the base matrix. It can be seen from the construction framework in Section 3.2 that the size of the base matrix is not too large since the code lengths of NBLDPC codes are short or moderate. In the following, we will give the design of the circulants.

A circulant is a square matrix whose i-th row is generated by cyclically shifting the first row to the right (or left) by $(i-1)$ positions. Hence, the first row of a circulant is referred to as the generator of the circulant. For a circulant of size $L \times L$, each row (or column) is a rightward (or downward) cyclic-shift of its above (or left) row (or column), and the first row (or column) is the rightward (or downward) cyclic-shift of the last row (or column). Therefore, the rows and columns of a circulant have the same weight. It is clear that the row (or column) weight is associated with the row weight of the generator.

Consider a circulant C of size $L \times L$, and its generator is $G = (g_1, g_2, \cdots, g_L)$ where $g_i \in \{0, 1\}$ for $1 \le i \le L$. Let ρ be the number of the nonzero components of G. Hence, C is (ρ, ρ)-regular. We select the nonzero components from $g_1, g_2, \cdots, g_L$ and record their subscripts in a set S, called location set in this paper. Then, the location set S has ρ elements. Without loss of generality, the location set S is denoted by

$$S = \{s_1, s_2, \cdots, s_\rho\}, \quad (5)$$

where $1 \le s_i < s_j \le L$ for $1 \le i < j \le \rho$. It is obvious that the generator G and the location set S have a one-to-one correspondence. Based on the isomorphism theory of LDPC codes (or their parity-check matrices) in [16, 35, 36], we can directly give the isomorphism theory of the circulants as follows.

Theorem 1. *Let $S_1 = \{s_{1,1}, s_{1,2}, \cdots, s_{1,\rho}\}$ and $S_2 = \{s_{2,1}, s_{2,2}, \cdots, s_{2,\rho}\}$ be two location sets of the circulants C_1 and C_2 of size $L \times L$, respectively. Then, C_1 is isomorphic to C_2, denoted by $C_1 \cong C_2$, if S_2 is derived from S_1 with either of the following two methods.*

(1) For $r \in \{0, 1, \cdots, L-1\}$, the elements of S_2 are derived from these of S_1 by adding a constant r to the elements of S_1 modulo L, i.e., $s_{2,i} = s_{1,i} + r (\mathrm{mod}\ L)$ for $1 \le i \le \rho$

(2) Suppose that r and L are coprime. The elements of S_2 are derived from these of S_1 using $s_{2,i} = r \cdot s_{1,i} (\mathrm{mod}\ L)$ for $1 \le i \le \rho$

Note that in the calculation process, if the element in S_1 and S_2 equals 0, it actually equals L. Moreover, in the case (2) of Theorem 1, the number of r can be determined by a well-known function, called Euler's phi function, i.e.,

$$\phi(L) = L \prod_{r|L} \left(1 - \frac{1}{r}\right). \quad (6)$$

If $C_1 \cong C_2$, we say S_1 is isomorphic to S_2, denoted by $S_1 \cong S_2$.

In general, the size of the employed circulants in this paper is not large. Hence, we can make an exhaustive search of the circulants by using the computer. The search space of the location sets of the $L \times L$ circulants with row/column weight ρ is

$$\binom{L}{\rho} = \frac{L!}{(L-\rho)! \cdot \rho!}. \quad (7)$$

Based on the case (1) in Theorem 1, we can see that all the location sets of the $L \times L$ circulants have the following isomorphic form: $S = \{s_1 (= 1), s_2, \cdots, s_\rho\}$, where $2 \le s_i < s_j \le L$ for $2 \le i < j \le \rho$. Hence, the search space of such location sets S is

$$\binom{L-1}{\rho-1} = \frac{(L-1)!}{(L-\rho)! \cdot (\rho-1)!}. \quad (8)$$

That is, an exhaustive search of such location sets (or circulants) is feasible. Here, we do not provide the specific exhaustive search algorithm. Combined with the cycle-counting algorithms [37–39], the optimal $L \times L$ circulants with row/column weight ρ can be found. In this paper, the optimal ones are such circulants whose Tanner graphs have fewer short cycles and larger girths. In order to facilitate the readers, some optimal circulants are presented in Table 1.

4.2. Review of Finite Fields Based QC-LDPC Codes and Their Exponent Matrices. In this subsection, we will review the

Table 1: Some optimal $L \times L$ circulants with row/column weight ρ and location set S.

L	ρ	Location set S	L	ρ	Location set S	L	ρ	Location set S	L	ρ	Location set S
4	4	{0, 1, 2, 3}	5	5	{0, 1, 2, 3, 4}	6	6	{0, 1, 2, 3, 4, 5}	7	7	{0, 1, 2, 3, 4, 5, 6}
5	4	{0, 1, 2, 3}	6	5	{0, 1, 2, 3, 4}	7	6	{0, 1, 2, 3, 4, 5}	8	7	{0, 1, 2, 3, 4, 5, 6}
6	4	{0, 1, 2, 3}	7	5	{0, 1, 2, 3, 4}	8	6	{0, 1, 2, 3, 4, 5}	9	7	{0, 1, 2, 3, 4, 5, 6}
7	4	{0, 1, 2, 4}	8	5	{0, 1, 2, 3, 5}	9	6	{0, 1, 2, 3, 4, 6}	10	7	{0, 1, 2, 3, 4, 5, 7}
8	4	{0, 1, 2, 4}	9	5	{0, 1, 2, 3, 5}	10	6	{0, 1, 2, 3, 5, 6}	11	7	{0, 1, 2, 3, 4, 6, 7}
9	4	{0, 1, 2, 4}	10	5	{0, 1, 2, 3, 6}	11	6	{0, 1, 2, 4, 5, 7}	12	7	{0, 1, 2, 3, 4, 7, 9}
10	4	{0, 1, 2, 5}	11	5	{0, 1, 2, 4, 7}	12	6	{0, 1, 2, 3, 5, 8}	13	7	{0, 1, 2, 3, 4, 6, 9}
11	4	{0, 1, 2, 5}	12	5	{0, 1, 2, 4, 7}	13	6	{0, 1, 2, 3, 5, 9}	14	7	{0, 1, 2, 3, 5, 6, 9}
12	4	{0, 1, 3, 7}	13	5	{0, 1, 2, 4, 7}	14	6	{0, 1, 2, 3, 5, 9}	15	7	{0, 1, 2, 4, 5, 8, 10}
13	4	{0, 1, 3, 9}	14	5	{0, 1, 2, 4, 7}	15	6	{0, 1, 2, 3, 6, 10}	16	7	{0, 1, 2, 3, 5, 8, 12}
14	4	{0, 1, 4, 6}	15	5	{0, 1, 2, 4, 7}	16	6	{0, 1, 2, 3, 6, 10}	17	7	{0, 1, 2, 3, 5, 8, 12}
15	4	{0, 1, 3, 7}	16	5	{0, 1, 2, 5, 8}	17	6	{0, 1, 2, 3, 6, 10}	18	7	{0, 1, 2, 3, 5, 8, 12}
16	4	{0, 1, 3, 7}	17	5	{0, 1, 2, 4, 12}	18	6	{0, 1, 2, 4, 8, 13}	19	7	{0, 1, 2, 3, 5, 9, 14}
17	4	{0, 1, 3, 7}	18	5	{0, 1, 2, 5, 11}	19	6	{0, 1, 2, 4, 7, 11}	20	7	{0, 1, 2, 3, 6, 10, 15}
18	4	{0, 1, 3, 7}	19	5	{0, 1, 2, 6, 9}	20	6	{0, 1, 2, 4, 7, 12}	21	7	{0, 1, 2, 4, 8, 11, 16}
19	4	{0, 1, 3, 8}	20	5	{0, 1, 2, 5, 14}	21	6	{0, 1, 2, 4, 7, 12}	22	7	{0, 1, 2, 4, 6, 14, 17}
20	4	{0, 1, 3, 14}	21	5	{0, 1, 4, 14, 16}	22	6	{0, 1, 2, 4, 8, 13}	23	7	{0, 1, 2, 3, 8, 13, 17}
21	4	{0, 1, 3, 9}	22	5	{0, 1, 3, 7, 12}	23	6	{0, 1, 2, 4, 7, 15}	24	7	{0, 1, 2, 4, 7, 15, 19}
22	4	{0, 1, 3, 9}	23	5	{0, 1, 3, 8, 14}	24	6	{0, 1, 2, 4, 12, 19}	25	7	{0, 1, 2, 4, 7, 12, 16}
23	4	{0, 1, 3, 10}	24	5	{0, 1, 3, 11, 20}	25	6	{0, 1, 2, 4, 9, 15}	26	7	{0, 1, 2, 4, 7, 13, 18}
24	4	{0, 1, 3, 10}	25	5	{0, 1, 3, 15, 21}	26	6	{0, 1, 2, 5, 9, 15}	27	7	{0, 1, 2, 4, 7, 12, 21}
25	4	{0, 1, 3, 10}	26	5	{0, 1, 3, 7, 12}	27	6	{0, 1, 2, 5, 13, 22}	28	7	{0, 1, 2, 4, 7, 17, 21}
26	4	{0, 1, 3, 11}	27	5	{0, 1, 3, 7, 18}	28	6	{0, 1, 4, 15, 20, 22}	29	7	{0, 1, 2, 4, 7, 12, 16}
27	4	{0, 1, 4, 10}	28	5	{0, 1, 3, 13, 24}	29	6	{0, 1, 2, 4, 18, 23}	30	7	{0, 1, 2, 4, 7, 15, 25}
28	4	{0, 1, 3, 12}	29	5	{0, 1, 3, 7, 19}	30	6	{0, 1, 2, 5, 14, 24}	30	5	{0, 1, 3, 12, 25}

finite field based method for constructing QC-LDPC codes, and provide two classes of exponent matrices of these QC-LDPC codes [18].

4.2.1. A General Construction of QC-LDPC Codes Based on Finite Fields of Characteristic 2. Let GF (q) be a finite field with $q = 2^t$ with $t \geq 2$, and let α be a primitive element of GF (q). For each nonzero element α^i with $0 \leq i \leq q-2$, we define the location vector $v(\alpha^i)$ as a $(q-1)$-tuple over GF (2),

$$v(\alpha^i) = (v_1, v_2, \cdots, v_{q-1}), \tag{9}$$

whose components correspond to the nonzero elements $\alpha_0, \alpha_1, \cdots, \alpha_{q-2}$ of GF (q), where the i-th component v_i is set to 1 and all the other $(q-2)$ components are 0. Hence, based on the nonzero element α^i of GF (q), we can uniquely form a $(q-1) \times (q-1)$ square matrix $M(\alpha^i)$ whose j-th row is obtained by cyclically shifting every component of the $(q-1)$-ary location vector $v(\alpha^i)(j-1)$ places to the right (or left) for $0 \leq i \leq q-2, 1 \leq j \leq q-1$. The resulting square matrix $M(\alpha^i)$ is a CPM, and it is also referred to as the $(q-1)$-fold matrix dispersion (or expansion) of the nonzero field element α^i over GF (2) [18].

Consider a $\gamma \times \rho$ matrix $\mathbf{P}$ over GF (q),

$$P = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_\gamma \end{bmatrix} = \begin{bmatrix} p_{1,2} & p_{1,3} & \cdots & p_{1,\rho} \\ p_{2,2} & p_{2,3} & \cdots & p_{2,\rho} \\ p_{3,2} & p_{3,3} & \cdots & p_{3,\rho} \\ \vdots & \vdots & \ddots & \vdots \\ p_{\gamma,2} & p_{\gamma,3} & \cdots & p_{\gamma,\rho} \end{bmatrix}, \tag{10}$$

whose rows satisfy the following two constraints: (1) for $1 \leq i < \gamma, 0 \leq k, l \leq q-2$ and $k \neq l$, $\alpha^k p_i$ and $\alpha^l p_i$ have at most one position where both of them have the same element of GF (q) (i.e., they differ in at least $(\rho - 1)$ positions); (2) for $1 \leq i, j < \gamma, i \neq j$, and $0 \leq k, l \leq q-2, \alpha^k p_i$ and $\alpha^l p_j$ differ in at least $(\rho - 1)$ positions. These constraints are called α-multiplied row constraints in [18]. By replacing each of its entry $p_{i,j}$ with the binary matrix $M(p_{i,j})$, we obtain a $\gamma \times \rho$ array $H_{\gamma,\rho}$ of $(q-1) \times (q-1)$ CPMs. Then, the null space of $H_{\gamma,\rho}$ gives a binary (γ, ρ)-regular QC-LDPC code. Moreover, the α-multiplied row constraints ensure the Tanner graph of this code free of cycle of length 4. Hence, the constructed QC-LDPC codes have girth at least 6.

Table 2: The nonzero field elements in the parity-check matrix of the proposed 64-ary (496,248) LDPC cycle code in Example 1.

Row index	Nonzero field elements				Row index	Nonzero field elements				Row index	Nonzero field elements			
1	37	9	22	0	2	7	0	18	44	3	0	37	9	19
4	0	37	19	9	5	7	18	44	0	6	7	18	0	44
7	7	18	44	0	8	0	18	44	7	9	19	9	0	37
10	0	18	44	7	11	0	9	37	22	12	18	44	7	0
13	9	19	37	0	14	18	0	7	44	15	22	37	0	9
16	0	44	7	18	17	9	19	0	37	18	37	22	0	9
19	44	18	7	0	20	9	37	22	0	21	19	37	9	0
22	0	22	37	9	23	44	7	0	18	24	0	7	44	18
25	9	37	0	22	26	19	9	37	0	27	22	37	0	9
28	9	22	37	0	29	9	37	22	0	30	37	9	22	0
31	37	9	22	0	32	22	9	37	0	33	7	18	44	0
34	37	0	9	22	35	0	37	22	9	36	7	44	0	18
37	0	44	7	18	38	9	0	37	19	39	37	9	22	0
40	0	9	19	37	41	44	7	18	0	42	37	19	9	0
43	9	37	0	19	44	37	9	22	0	45	0	9	19	37
46	19	9	37	0	47	0	19	37	9	48	18	0	44	7
49	0	19	9	37	50	22	37	9	0	51	37	0	9	22
52	18	7	0	44	53	0	9	37	22	54	44	7	0	18
55	7	0	44	18	56	9	0	37	22	57	9	19	37	0
58	22	9	37	0	59	9	19	37	0	60	0	44	18	7
61	18	7	0	44	62	9	37	0	19	63	22	0	9	37
64	7	0	18	44	65	18	44	0	7	66	19	37	9	0
67	9	37	22	0	68	18	7	0	44	69	0	9	22	37
70	37	22	0	9	71	18	7	44	0	72	7	0	18	44
73	0	22	37	9	74	7	0	44	18	75	9	0	19	37
76	37	0	9	19	77	9	37	0	22	78	0	37	9	22
79	37	9	0	22	70	9	0	37	19	81	7	44	18	0
82	0	44	18	7	83	9	37	22	0	84	44	0	7	18
85	9	22	37	0	86	9	19	37	0	87	22	9	0	37
88	0	18	7	44	89	9	37	22	0	90	18	0	7	44
91	37	22	9	0	92	0	22	37	9	93	37	0	9	22
94	9	22	37	0	95	44	0	7	18	96	0	9	19	37
97	7	44	18	0	98	9	19	37	0	99	7	44	18	0
100	37	22	0	9	101	0	22	9	37	102	22	0	37	9
103	44	0	7	18	104	22	9	0	37	105	22	0	37	9
106	44	7	18	0	107	37	19	0	9	108	0	9	37	22
109	9	37	19	0	110	44	18	0	7	111	37	22	9	0
112	7	0	18	44	113	0	18	44	7	114	37	22	0	9
115	0	9	37	19	116	7	18	0	44	117	9	0	22	37
118	0	37	19	9	119	0	9	22	37	120	19	37	9	0
121	37	0	22	9	122	37	0	9	22	123	0	7	44	18
124	19	37	0	9	125	7	44	18	0	126	7	44	0	18
127	9	37	22	0	128	44	18	0	7	129	7	18	0	44
130	0	37	22	9	131	7	0	44	18	132	0	37	9	22
133	0	44	18	7	134	44	18	7	0	135	9	37	0	22
136	9	37	19	0	137	0	9	37	22	138	18	44	7	0
142	0	18	7	44	143	44	18	0	7	144	9	37	0	19
145	0	18	7	44	146	18	0	7	44	147	19	9	37	0

Table 2: Continued.

Row index	Nonzero field elements				Row index	Nonzero field elements				Row index	Nonzero field elements			
148	44	18	0	7	149	37	0	19	9	150	19	0	9	37
151	0	37	22	9	152	7	18	44	0	153	44	0	7	18
154	0	37	9	22	155	18	0	7	44	156	44	7	0	18
157	0	37	9	19	158	22	37	9	0	159	7	18	44	0
160	0	37	9	19	161	0	18	7	44	162	0	37	9	22
163	7	18	44	0	164	37	22	9	0	165	0	9	37	22
166	0	37	19	9	167	18	44	7	0	168	9	0	22	37
169	0	37	9	19	170	44	0	18	7	171	9	0	19	37
172	9	0	37	19	173	9	37	19	0	174	44	7	18	0
175	19	0	37	9	176	0	22	9	37	177	9	0	37	19
178	0	18	7	44	179	0	18	44	7	180	37	0	9	22
181	9	37	0	22	182	0	37	22	9	183	37	0	22	9
184	0	9	22	37	185	9	19	37	0	186	44	7	18	0
187	19	9	0	37	188	37	0	9	22	189	18	44	7	0
190	22	37	9	0	191	37	19	0	9	192	37	9	19	0
193	0	37	19	9	194	22	0	37	9	195	7	0	18	44
196	37	19	0	9	197	9	37	22	0	198	22	9	0	37
199	44	18	0	7	200	22	37	9	0	201	7	18	0	44
202	0	19	9	37	203	0	44	18	7	204	9	0	19	37
205	7	18	0	44	206	37	19	0	9	207	7	44	18	0
208	44	0	18	7	209	19	37	0	9	210	44	18	0	7
211	19	37	0	9	212	18	0	7	44	213	37	9	22	0
214	0	9	37	22	215	22	37	0	9	216	19	37	9	0
217	7	44	18	0	218	44	0	7	18	219	44	0	7	18
220	37	0	19	9	221	22	9	0	37	222	44	18	7	0
223	19	9	37	0	224	18	0	44	7	225	37	22	9	0
226	22	9	37	0	227	19	37	9	0	228	18	44	7	0
229	9	19	37	0	230	18	44	7	0	231	18	7	0	44
232	44	7	18	0	233	44	18	0	7	234	37	22	9	0
235	22	9	37	0	236	22	9	0	37	237	44	0	18	7
238	0	37	22	9	239	37	22	0	9	240	37	19	0	9
241	0	18	7	44	242	19	37	0	9	243	22	9	0	37
244	44	7	18	0	245	9	37	19	0	246	37	22	0	9
247	44	7	18	0	248	19	9	0	37					

According to the definition in Section 2, we can see that the above $\gamma \times \rho$ matrix $\mathbf{P}$ is the exponent matrix, and the associate expansion factor is $(q-1)$. A framework for constructing such matrix $\mathbf{P}$ based on two arbitrary subsets of a finite field was proposed in [40]. Let η be a nonzero element in GF (q) and α be a primitive element. For $1 \leq \gamma, \rho \leq q$, let $T_1 = \{\alpha^{i_1}, \alpha^{i_2}, \cdots, \alpha^{i_\gamma}\}$ and $T_2 = \{\alpha^{j_1}, \alpha^{j_2}, \cdots, \alpha^{j_\rho}\}$ be two arbitrary subsets of elements in GF (q) with i_k and j_l in the set $\{-\infty, 0, 1, 2, \cdots, q-2\}$, $i_1 < i_2 < \cdots < i_\gamma$, and $j_1 < j_2 < \cdots < j_\rho$. The $\gamma \times \rho$ matrix $\mathbf{P}$ can be formed by

$$P(\eta) = \left[\eta \alpha^{i_k} + \alpha^{j_l}\right]_{1 \leq k \leq \gamma, 1 \leq l \leq \rho}. \tag{11}$$

Under this framework, some well-known constructions of QC-LDPC codes based on finite fields and combinatorial designs are special cases [14, 41, 42]. For example, when $T_1 = T_2 = \mathrm{GF}(q)$, $P(\eta)$ is a Latin square over GF (q) [43].

4.2.2. Construction of QC-LDPC Codes Based on Prime Fields. Let p be a prime. Consider a $p \times p$ matrix

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,p} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ p_{p,1} & p_{p,2} & \cdots & p_{p,p} \end{bmatrix}, \tag{12}$$

where $p_{i,j} = (i-1) \cdot (j-1) (\mathrm{mod}\, p)$ for $1 \leq i \leq p$ and $1 \leq j \leq p$. We select a $\gamma \times \rho$ submatrix from P and replace its elements $p_{s,t}$ with the CPMs $\mathbf{I}(p_{s,t})$ for $1 \leq s \leq \gamma$ and $1 \leq t \leq \rho$. The

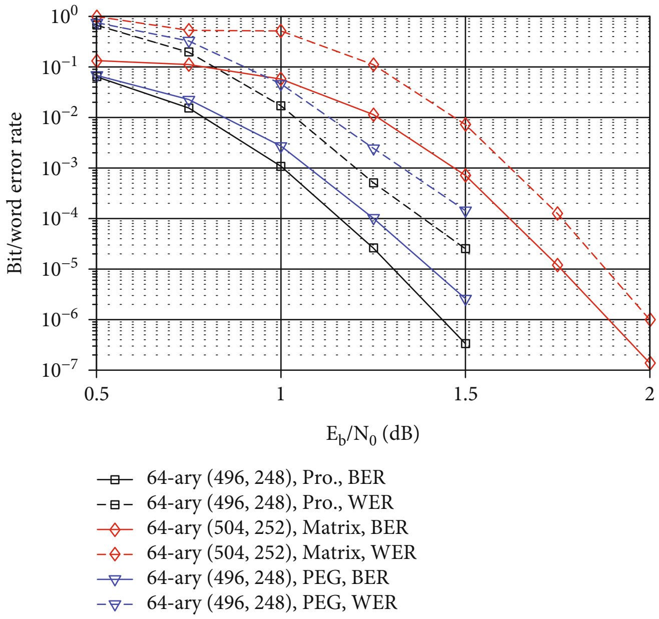


Figure 4: The error performances of the proposed (496,248) LDPC cycle code over GF (64), the comparable (504,252) irregular QC-LDPC code over GF (64) constructed from finite field GF (64) in [15], and (2,4)-regular (496,248) LDPC code over GF (64) constructed based on the progressive edge-growth (PEG) algorithm in [46]. The transmissions on the BPSK modulated AWGN channel are assumed.

following $\gamma \times \rho$ array $H(P_{\gamma\times\rho})$ of $p \times p$ CPMs over GF (2) is obtained.

$$H(P_{\gamma\times\rho}) = [I(p_{s,t})]_{1\le s\le\gamma,1\le t\le\rho}. \quad (13)$$

Actually, the null space of $H(P_{\gamma\times\rho})$ gives a (γ, ρ)-regular QC-LDPC code of length $p\rho$ with girth at least 6. Notice that the exponent matrix of this code is $P_{\gamma\times\rho}$ and the expansion factor is p. How to select good γ rows and ρ columns of CPMs from P in (12) can be found in [44, 45].

4.3. Nonbinary LDPC Cycle Codes and Numerical Results.

Combined with Sections 3.2, 4.1, and 4.2, we can easily construct a binary QC-LDPC cycle code. Based on the replacement of the nonzero elements in finite fields, nonbinary LDPC cycle codes can be constructed. In order to show the advantages of our proposed construction methods, we next provide some examples.

Table 3: The nonzero field elements in the corresponding CPMs of the parity-check matrices $H_{M,256}(P_{8\times 16,2})$ in Example 2.

Row index	Nonzero field elements			
1	0	182	8	172
2	173	0	182	8
3	40	167	0	127
4	40	127	0	169
5	40	169	0	128
6	172	8	182	0
7	8	172	182	0
8	40	169	128	0

Example 1. Consider the all-one matrix (or circulant) of size 4×4. It is clear that there is only one such circulant for $\rho = 4$ and $L = 4$. The following 8×16 matrix $B_{8\times 16}$ can be obtained based on the construction framework in Section 3.2.

$$7ptB_{8\times 16} = \begin{bmatrix} 1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \\ 1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0 \\ 0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0 \\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0 \\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1 \end{bmatrix}. \quad (14)$$

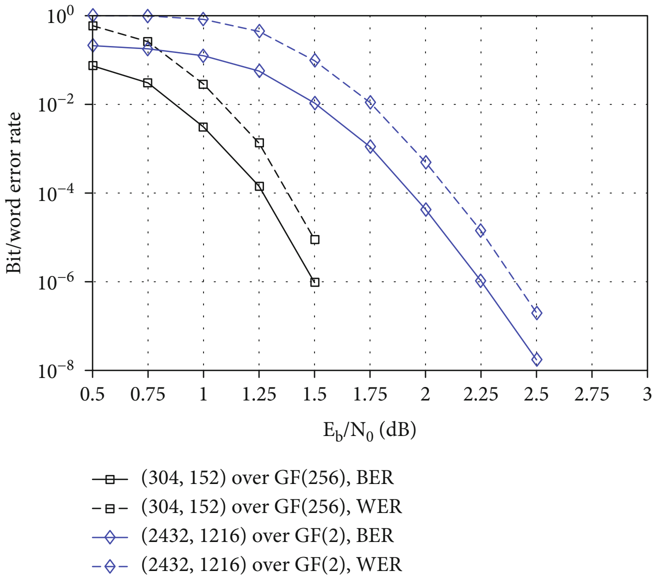


Figure 5: The error performances of the constructed (304,152) QC-LDPC cycle code over GF (256) and the comparable (2432,1216) LDPC code over GF (2) constructed based on the progressive edge-growth (PEG) algorithm in [46]. The transmissions on the BPSK modulated AWGN channel are assumed.

Based on the prime field GF (31), we can construct an 8×16 array $H(P_{8\times16,1})$ of 31×31 CPMs in the form of (13) by choosing the 8 rows and 16 columns from the exponent matrix P in equation (12). The indices of the chosen 8 rows and 16 columns from P are {3, 4, 5, 10, 15, 17, 24, 28} and {1, 4, 5, 6, 7, 8, 9, 10, 11, 15, 16, 19, 26, 28, 29, 30}, respectively. The row and column selection method is based on the proposed method in [45]. By employing $B_{8\times16}$ to mask $H(P_{8\times16,1})$, we can obtain an array $H_{\mathrm{M}}(P_{8\times16,1})$ of 31×31 CPMs. By replacing the 1's of each row in $H_{\mathrm{M}}(P_{8\times16,1})$ with nonzero elements of GF (64) in the corresponding rows of Table 2, a $(2,4)$-regular matrix $H_{\mathrm{M},64}(P_{8\times16,1})$ over GF (64) is obtained. Note that the numbers in Table 2 are the power numbers of α, where α is a primitive element of GF (64) created by using the primitive polynomial $p(x) = 1 + x + x^6$. The null space of $H_{\mathrm{M},64}(P_{8\times16})$ gives a $(2,4)$-regular (496,248) LDPC cycle code over GF (64). The girth of this code is 16, and the number of 16-cycles is 775. For comparison, we reconstruct the irregular (504,252) QC-LDPC code over GF (64) based on finite field GF (64) and the masking matrix $B_{\mathrm{mask}}(4,8)$ in [15]. The average column and row weights in the parity-check matrix of (504,252) QC-LDPC code over GF (64) are 2.5 and 5, respectively. Moreover, we also design a comparable (2,4)-regular (496,248) LDPC code over GF (64) based on the progressive edge-growth (PEG) algorithm in [46], called PEG-LDPC code. Note that the nonzero elements of the parity-check matrix of the (2,4)-regular (496,248) PEG-LDPC code over GF (64) are randomly chosen. When decoded with the fast-Fourier-transform (FFT) based Q-ary sum-product algorithm (QSPA) with 50 iterations, the bit/word error rates (BERs/WERs) of these three nonbinary LDPC codes are shown in Figure 4. In the simulations, the transmissions on the BPSK modulated AWGN channel are assumed. We can see that at the BER of 2×10^{-6}, the proposed nonbinary (496,248) LDPC cycle code outperforms the (504,252) QC-LDPC code over GF (64) and the (2,4)-regular (496,248) PEG-LDPC code over GF (64) about 0.45 dB and 0.1 dB, respectively.

In the above example, we can see that the constructed nonbinary LDPC cycle code is not quasi-cyclic because of the randomly chosen nonzero field elements in its parity-check matrix. Hence, we next show the performances of the proposed nonbinary QC-LDPC cycle codes.

Example 2. Consider the prime field GF (19), we can easily construct an 8×16 matrix $P_{8\times16,2}$ over GF (19) based on the exponent matrix P in equation (12) and the construction framework in Section 3.2. The indices of the chosen 8 rows and 16 columns from P are {1, 2, 4, 5, 7, 8, 14, 15} and {1, 2, 3, 4, 5, 6, 7, 9, 10, 13, 14, 15, 16, 17, 18, 19}, respectively. The row and column selection method is also based on the proposed method in [45]. By dispersing each entry of $P_{8\times16,2}$ into the corresponding CPMs of size 19×19, an 8×16 array $H(P_{8\times16,2})$ of 19×19 CPMs is obtained. By employing $B_{8\times16}$ in equation (14) to mask the array $H(P_{8\times16,2})$, we can construct an array $H_M(P_{8\times16,2})$ of 19×19 CPMs. By replacing the 1's of each CPM in $H_M(P_{8\times16,2})$ with the same nonzero element of GF (256) in the corresponding row of Table 3, a (2,4)-regular matrix $H_{M,256}(P_{8\times16,2})$ over GF (256) is obtained. It is noticeable that the numbers in Table 3 are the power numbers of α, where α is a primitive element of GF (256) created by using the primitive polynomial $p(x) = 1 + x^2 + x^3 + x^4 + x^8$. The null space of $H_{M,256}(P_{8\times16,2})$ gives a (2,4)-regular (304,152) LDPC cycle code over GF (256). The girth of this code is 16, and the number of 16-cycles is 969. Since the nonzero field elements of a

CPM in $H_{M,256}(P_{8\times16,2})$ are the same, and the resulting (304,152) LDPC cycle code over GF (256) is quasi-cyclic.

For comparison, we construct the (3,6)-regular (2432, 1216) LDPC code over GF (2) based on the progressive edge-growth (PEG) algorithm in [46]. The error performances of these two LDPC codes are shown in Figure 5. In the simulations, the employed decoding algorithms of the constructed (304,152) QC-LDPC cycle code over GF (256) and the (2432,1216) LDPC code over GF (2) are the QSPA (50 iterations) and the sum-product algorithm (SPA) with 50 iterations, respectively. The transmissions on the BPSK modulated AWGN channel are assumed. We can see that the constructed (304,152) QC-LDPC cycle code over GF (256) can outperform the (2432,1216) LDPC code over GF (2) about 0.75 dB at the BER of 10^{-6}.

5. Conclusion

This paper proposed a design framework of binary QC-LDPC cycle codes and constructed nonbinary LDPC (NBLDPC) cycle codes based on circulants and finite fields. The presented construction method consists of three parts. First, the masking matrices are designed based on circulants and the point-line relation in graph theory. Second, the exponent matrices of binary QC-LDPC cycle codes are constructed from finite fields and the designed masking matrices. Third, by replacing 1's in the parity-check matrices of binary QC-LDPC cycle codes with the nonzero field elements, NBLDPC cycle codes are obtained. Numerical results show that the constructed NBLDPC cycle codes have good iterative decoding performance.

Data Availability

The data used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61801527 and 11971311, the TianYuan Special Funds of the National Natural Science Foundation of China under Grants 12026230 and 12026231, and the Development Project of Henan Provincial Department of Science and Technology under Grants 212102310544 and 212102310551.

References

[1] R. Gallager, "Low-density parity-check codes," *IRE Transactions on Information Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[2] M. Davey and D. MacKay, "Low-density parity check codes over GF(q)," *IEEE Communications Letters*, vol. 2, no. 6, pp. 165–167, 1998.

[3] S. Wang, Q. Huang, and Z. Wang, "Symbol flipping decoding algorithms based on prediction for non-binary LDPC codes," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 1913–1924, 2017.

[4] Q. Huang, L. Song, and Z. Wang, "Set message-passing decoding algorithms for regular non-binary LDPC codes," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5110–5122, 2017.

[5] B. Dai, R. Liu, C. Gao, and Z. Mei, "Symbol flipping algorithm with self-adjustment strategy for LDPC codes over GF(q)," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7189–7193, 2019.

[6] Z. Liu, R. Liu, and L. Zhao, "GPU-based non-binary LDPC decoder with weighted bit-reliability based algorithm," *China Communications*, vol. 17, no. 5, pp. 78–88, 2020.

[7] S. Song, J. Tian, J. Lin, and Z. Wang, "An improved reliability-based decoding algorithm for NB-LDPC codes," *IEEE Communications Letters*, vol. 25, no. 4, pp. 1153–1157, 2021.

[8] V. B. Wijekoon, E. Viterbo, and Y. Hong, "Decoding of NB-LDPC codes over subfields," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 716–727, 2021.

[9] S. Chen, Y. C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, "Vision, requirements, and technology trend of 6G: how to tackle the challenges of system coverage, capacity, user datarate and movement speed," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 218–228, 2020.

[10] X. Y. Hu and E. Eleftheriou, "Binary representation of cycle Tanner-graph GF(2^b) codes," in *2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577)*, pp. 528–532, Paris, France, June 2004.

[11] H. Song, J. Liu, and B. V. K. VijayaKumar, "Large girth cycle codes for partial response channels," *IEEE Transactions on Magnetics*, vol. 40, no. 4, pp. 3084–3086, 2004.

[12] Ronghui Peng and Rong-Rong Chen, "Application of nonbinary LDPC cycle codes to MIMO channels," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2020–2026, 2008.

[13] X. Liu, F. Xiong, Z. Zhou, Y. Yin, and L. Zhang, "Construction of QC LDPC cycle codes over GF(q) based on cycle entropy and applications on patterned media storage," *IEEE Transactions on Magnetics*, vol. 51, no. 11, pp. 1–5, 2015.

[14] Shumei Song, Bo Zhou, Shu Lin, and K. Abdel-Ghaffar, "A unified approach to the construction of binary and nonbinary quasi-cyclic LDPC codes based on finite fields," *IEEE Transactions on Communications*, vol. 57, no. 1, pp. 84–93, 2009.

[15] J. Li, K. Liu, S. Lin, and K. Abdel-Ghaffar, "A matrix-theoretic approach to the construction of non-binary quasi-cyclic LDPC codes," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1057–1068, 2015.

[16] H. Xu, C. Chen, M. Zhu, B. Bai, and B. Zhang, "Nonbinary LDPC cycle codes: efficient search, design, and code optimization," *Science China Information Sciences*, vol. 61, no. 8, pp. 089303:1–089303:3, 2018.

[17] H. Xu, H. Li, M. Xu, D. Feng, and H. Zhu, "Two classes of QC-LDPC cycle codes approaching Gallager lower bound," *Science China Information Sciences*, vol. 62, no. 10, pp. 209305:1–209305:3, 2019.

[18] W. E. Ryan and S. Lin, *Channel Codes: Classical and Modern*, Cambridge Univ. Press, New York, USA, 2009.

[19] C. Chen, B. Bai, G. Shi, X. Wang, and X. Jiao, "Nonbinary LDPC codes on cages: structural property and code

optimization," *IEEE Transactions on Communications*, vol. 63, no. 2, pp. 364–375, 2015.

[20] M. P. C. Fossorier, "Quasi-cyclic low-density parity-check codes from circulant permutation matrices," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1788–1793, 2004.

[21] J. Li, K. Liu, S. Lin, K. Abdel-Ghaffar, and R. E. Ryan, "An unnoticed strong connection between algebraic-based and protograph-based LDPC codes, part I: binary case and interpretation," in *2015 Information Theory and Applications Workshop (ITA)*, pp. 36–50, San Diego, CA, USA, February 2015.

[22] S. Zhao and X. Ma, "Construction of high-performance array-based non-binary LDPC codes with moderate rates," *IEEE Communications Letters*, vol. 20, no. 1, pp. 13–16, 2016.

[23] R. Tanner, "A recursive approach to low complexity codes," *IEEE Transactions on Information Theory*, vol. 27, no. 5, pp. 533–547, 1981.

[24] R. J. McEliece, D. J. C. MacKay, and Jung-Fu Cheng, "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, pp. 140–152, 1998.

[25] X. Jiang, X. G. Xia, and M. H. Lee, "Efficient progressive edge-growth algorithm based on Chinese remainder theorem," *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 442–451, 2014.

[26] Q. Diao, J. Li, S. Lin, and I. Blake, "New classes of partial geometries and their associated LDPC codes," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 2947–2965, 2016.

[27] X. Jiang, H. Hai, H. M. Wang, and M. H. Lee, "Constructing large girth QC protograph LDPC codes based on PSD-PEG algorithm," *IEEE Access*, vol. 5, pp. 13489–13500, 2017.

[28] X. Qin, C. Yang, Z. Zheng, and Z. Wang, "Optimization of QC-LDPC codes by edge exchange method based on ACE," *IEEE Photonics Technology Letters*, vol. 31, no. 17, pp. 1401–1404, 2019.

[29] H. Xu, H. Li, B. Bai, M. Zhu, and B. Zhang, "Tanner (J, L) quasi-cyclic LDPC codes: girth analysis and derived codes," *IEEE Access*, vol. 7, pp. 944–957, 2019.

[30] G. Wu, Y. Lv, and J. He, "Design of high-rate LDPC codes based on matroid theory," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2146–2149, 2019.

[31] X. Tao, Y. Xin, B. Wang, and L. Chang, "Layered construction of quasi-cyclic LDPC codes," *IEEE Communications Letters*, vol. 24, no. 5, pp. 946–950, 2020.

[32] M. Majdzade and M. Gholami, "On the class of high-rate QC-LDPC codes with girth 8 from sequences satisfied in GCD condition," *IEEE Communications Letters*, vol. 24, no. 7, pp. 1391–1394, 2020.

[33] A. Dehghan and A. H. Banihashemi, "On finding bipartite graphs with a small number of short cycles and large girth," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6024–6036, 2020.

[34] C. Poulliat, M. Fossorier, and D. Declercq, "Design of regular $(2, d_c)$-LDPC codes over GF(q) using their binary images," *IEEE Transactions on Communications*, vol. 56, no. 10, pp. 1626–1635, 2008.

[35] A. Tasdighi, A. H. Banihashemi, and M. R. Sadeghi, "Efficient search of girth-optimal QC-LDPC codes," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1552–1564, 2016.

[36] H. Xu, B. Bai, M. Zhu, B. Zhang, and Y. Zhang, "Construction of short-block nonbinary LDPC codes based on cyclic codes," *China Communications*, vol. 14, no. 8, pp. 1–9, 2017.

[37] M. Karimi and A. H. Banihashemi, "Counting short cycles of quasi cyclic protograph LDPC codes," *IEEE Communications Letters*, vol. 16, no. 3, pp. 400–403, 2012.

[38] M. Karimi and A. H. Banihashemi, "Message-passing algorithms for counting short cycles in a graph," *IEEE Transactions on Communications*, vol. 61, no. 2, pp. 485–495, 2013.

[39] A. Dehghan and A. H. Banihashemi, "On computing the number of short cycles in bipartite graphs using the spectrum of the directed edge matrix," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6037–6047, 2020.

[40] J. Li, K. Liu, S. Lin, and K. Abdel-Ghaffar, "Algebraic quasi-cyclic LDPC codes: construction, low error-floor, large girth and a reduced-complexity decoding scheme," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2626–2637, 2014.

[41] J. Kang, Q. Huang, L. Zhang, B. Zhou, and S. Lin, "Quasi-cyclic LDPC codes: an algebraic construction," *IEEE Transactions on Communications*, vol. 58, no. 5, pp. 1383–1396, 2010.

[42] L. Zhang, S. Lin, K. Abdel-Ghaffar, Z. Ding, and B. Zhou, "Quasi-cyclic LDPC codes on cyclic subgroups of finite fields," *IEEE Transactions on Communications*, vol. 59, no. 9, pp. 2330–2336, 2011.

[43] L. Zhang, Q. Huang, S. Lin, K. Abdel-Ghaffar, and I. F. Blake, "Quasi-cyclic LDPC codes: an algebraic construction, rank analysis, and codes on Latin squares," *IEEE Transactions on Communications*, vol. 58, no. 11, pp. 3126–3139, 2010.

[44] H. Xu, D. Feng, R. Luo, and B. Bai, "Construction of quasi-cyclic LDPC codes via masking with successive cycle elimination," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2370–2373, 2016.

[45] H. Zhu, B. Zhang, M. Xu, H. Li, and H. Xu, "Array based quasi-cyclic LDPC codes and their tight lower bounds on the lifting degree," *Physical Communication*, vol. 36, p. 100765, 2019.

[46] Xiao-Yu Hu, E. Eleftheriou, and D. M. Arnold, "Regular and irregular progressive edge-growth Tanner graphs," *IEEE Transactions on Information Theory*, vol. 51, no. 1, pp. 386–398, 2005.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4212548, 12 pages
https://doi.org/10.1155/2021/4212548

Research Article

Collaborative Computing and Resource Allocation for LEO Satellite-Assisted Internet of Things

Tao Leng,[1,2] Xiaoyao Li,[1,2] Dongwei Hu,[3] Gaofeng Cui,[1,2,3] and Weidong Wang[1,2]

[1] *School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2] *Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3] *Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang, China*

Correspondence should be addressed to Gaofeng Cui; cuigaofeng@bupt.edu.cn

Received 13 May 2021; Revised 5 August 2021; Accepted 27 August 2021; Published 15 September 2021

Academic Editor: Miaowen Wen

Satellite-assisted internet of things (S-IoT), especially the S-IoT based on low earth orbit (LEO) satellite, plays an important role in future wireless systems. However, the limited on-board communication and computing resource and high mobility of LEO satellites make it hard to provide satisfied service for IoT users. To maximize the task completion rate under latency constraints, collaborative computing and resource allocation among LEO networks are jointly investigated in this paper, and the joint task offloading, scheduling, and resource allocation is formulated as a dynamic mixed-integer problem. To tack the complex problem, we decouple it into two subproblems with low complexity. First, the max-min fairness is adopted to minimize the maximum latency via optimal resource allocation with fixed task assignment. Then, the joint task offloading and scheduling is formulated as a Markov decision process with optimal communication and computing resource allocation, and deep reinforcement learning is utilized to obtain long-term benefits. Simulation results show that the proposed scheme has superior performance compared with other referred schemes.

1. Introduction

Internet of things (IoT) plays an important role in future intelligent society, and many techniques have been evaluated and implemented to provide better service for the data transmission in IoT network. Although the fifth generation (5G) wireless systems can support massive machine type communication (mMTC), it mainly focuses on the terrestrial network-based IoT. For depopulated areas where lack telecommunication infrastructures, satellite communication has been adopted as an important component for 5G beyond or the sixth generation (6G) wireless systems [1, 2]. Moreover, edge computing refers to the techniques that shift the computing units to the access nodes near the user equipment, or the data is processed at the user equipment locally [3]. With edge computing, the access delay can be reduced, and the radio resource can also be utilized efficiently. Benefit from the development of satellite on-board processing techniques [4], edge computing-enhanced satellite networks have also become a hot topic for integrated satellite and terrestrial networks [2, 5, 6]. Thus, edge computing-enhanced satellite-assisted IoT (S-IoT) receives lots of attention in both industrial and academic areas [7, 8].

For edge computing-enhanced S-IoT networks, the IoT devices and satellites are both resource-limited. Therefore, the joint computing and communication resource allocation for tasks generated by users (IoT devices) is important for improving the performance of the systems [9]. Moreover, the characteristics of the satellite networks will also affect the mechanisms used for the resource management [10]. Generally, the existing satellite communication systems can be divided into three categories, which are geosynchronous earth orbit (GEO), medium earth orbit (MEO), and low earth orbit (LEO). Among the three categories of satellite systems, LEO satellites with the lowest propagation delay are emerging as an important component for future integrated satellite and terrestrial networks [11–13]. In this paper, we also focus on the edge computing-enhanced LEO satellite networks, and the advantages of LEO satellite-assisted IoT over the other satellite systems can be summarized as follows:

(1) The propagation delay introduced by the LEO satellite is low. For example, the one-trip propagation delay is about 5 ms for LEO satellites located at the height of 1500 km. While it is about 120 ms for GEO satellites

(2) Though the satellite on-board processing capability is usually limited, multiple satellites in the LEO networks, especially the mega-constellation LEO networks, can form a virtual resource pool, which can be utilized to improve the performance of the edge computing enhanced S-IoT

(3) Multiple LEO satellites can generate overlapped coverage areas, and the communication and computing resource can be allocated flexibly

From the above analysis, the LEO S-IoT can benefit from the low propagation delay and collaborative processing among multiple satellites. However, the edge computing-enhanced LEO S-IoT still faces several problems. First, the varying topology of LEO networks makes it difficult to manage the limited communication and computing resource dynamically [14]. Second, the resource management should jointly consider multiple types of links, such as satellite-to-ground and satellite-to-satellite. Last but not the least, the communication and computing resource should be jointly allocated.

In this paper, edge computing-enhanced LEO S-IoT is considered, and the task generated by the users needs to be handled locally or via the LEO networks collaboratively. Different from the existing studies, collaborative computing among multiple satellites is utilized to reduce the on-board processing latency. Moreover, the satellite-to-ground and satellite-to-satellite links are considered jointly for communication and computing resource allocation. The main contributions are listed as below.

(i) A framework for collaborative computing among multiple LEO satellites with varying topology is provided, and the effects of satellite-to-ground and satellite-to-satellite links on the processing latency are jointly considered

(ii) The collaborative computing and resource allocation for user tasks are formulated as a joint task offloading, scheduling, and multidimensional resource allocation problem to maximize the completion rate of tasks, and the complex problem is divided into two subproblems with low complexity

(iii) Deep reinforcement learning (DRL) and max-min fairness optimization are adopted to achieve long-term benefits in terms of task completion rate, and simulation results verify the performance of the proposed algorithms

The remainder of this paper is organized as follows. Section 2 summarizes the related works. The system model and problem formulation are described in Section 3. In Section 4, resource allocation based on max-min fairness and task scheduling and offloading with DRL is analyzed, respectively. Section 5 evaluates the proposed algorithms, and Section 6 concludes this paper.

2. Related Works

Joint task offloading, scheduling, and resource allocation plays an important role in edge computing enhanced S-IoT networks. Papa et al. evaluate the reconfigurable software-defined network with LEO constellation and propose an optimal controller placement and satellite-to-controller assignment method which can minimize the average flow setup time [15]. Liu et al. propose a task-orient network architecture for edge computing enhanced space-air-ground-aqua integrated networks [9]. Xie et al. analyze the joint caching, communication, and computing resource management for space information networks [2]. Although the joint task offloading, scheduling, and resource allocation for edge computing enhanced S-IoT is highlighted in the existing works [2, 9, 15], the resource management methods are not given in detail.

Cheng et al. propose a computing offloading method for IoT applications in space-air-ground integrated network with fixed data rate [16]. Cao et al. propose an edge-cloud architecture based on software-defined networking and network function virtualization for the space-air-ground integrated network [17]. Wang et al. introduce the hardware and software structure for the edge computing-enhanced S-IoT [18]. A fine-grained resource management scheme is introduced by Wang et al. for edge computing-enhanced satellite networks [19]. Yan et al. propose a 5G satellite edge computing framework based on microservice architecture with the embedded hardware platform [5]. LiWang et al. investigate the computing offloading methods with delay and cost constraints for satellite-ground internet of vehicles [20]. Jiao et al. analyze a joint network stability and resource allocation optimization problem for high-throughput satellite-based IoT [21]. An orbital edge computing architecture is introduced by Denby and Lucia, and the power and software optimization for the orbital edge are also analyzed [22]. Cui et al. propose a joint offloading and resource allocation for GEO satellite-assisted vehicle-to-vehicle communication [23]. A collaborative computing and resource allocation method among multiple user pairs is given by Zhang et al. for GEO S-IoT [24]. Song et al. propose a mobile edge computing framework for terrestrial-satellite IoT, and an energy-efficient computing offloading and resource allocation method is used to minimize the weighted sum energy consumption [25]. A learning-based queue-aware task offloading and resource allocation algorithm is analyzed by Liao et al. for space-air-ground-integrated power IoT [26]. Tang et al. present a hybrid cloud and edge computing LEO satellite network and investigate the computation offloading decisions to minimize the sum energy consumption of ground users [27].

Though some exiting works listed above investigate the task offloading and resource allocation for edge computing-enhanced S-IoT, none of these works consider the collaborative computing among multiple LEO satellites. Moreover, the joint optimization of satellite-to-ground and satellite-to-satellite links is not investigated either.

3. System Model and Problem Formulation

In this paper, we consider a typical scenario, as shown in Figure 1, consisting of multiple terrestrial users and a LEO satellite constellation. The LEO satellite constellation is deployed with intersatellite links (ISLs) for cooperative processing among satellites, and each satellite can exchange information with four adjacent satellites through ISLs. With the network topology described above, the tasks generated by the users arrive randomly as a time series, and the tasks can be processed locally or offloaded to satellites for processing. Although the computing resource of a single satellite is scarce due to the characteristics of the on-board devices, computing units on multiple satellites can form a collaborative computing pool, and the satellites which are overloaded can forward the tasks that need to be handled to the other satellites with light load. Thus, the task offloaded to the satellites can be handled by its serving satellite or other satellites available via ISLs. Moreover, the computing resource available for the IoT devices is also limited, and tasks that need to be processed locally should be handled one by one. While for satellites, tasks from multiple users can be handled in parallel, and the resource is shared among tasks belong to multiple users.

Moreover, every satellite needs to maintain the satellite-to-ground transmission queue and the on-board processing queue. For satellite-to-ground transmission, the tasks offloaded to satellites will be scheduled slot by slot, and the tasks cannot be partitioned. When the task arrives at the satellite used for data processing, it will enter the processing queue and wait for data processing. To guarantee the efficiency and reliability of transmission and data processing, the communication/computing resource allocated to user tasks will be occupied until the end of the transmission/processing. After the data processing, the results will be delivered to the users. During the task offloading processes, the resource occupancy and mobility of the LEO satellites will both affect the performance of the system. For tasks handled locally, the latency is mainly composed of waiting latency in queue and processing latency. While for tasks handled by satellites, several factors, which are transmission latency, propagation delay, and processing latency, need to be considered. Thus, the system models adopted and problem formulation are listed in the following subsections.

3.1. Satellite Orbit Model. In earth-centered inertial (ECI) coordinate system, the position of the satellite in space can be described by orbital elements, namely, eccentricity e, semimajor axis a, inclination i, right ascension of the ascending node (RAAN) Ω, argument of periapsis ω, and initial true anomaly φ. In this paper, we consider the satellite orbit to be a circular orbit with $e = 0$ and $\omega = 0$, so the ECI coordinate of satellite at time t can be expressed as

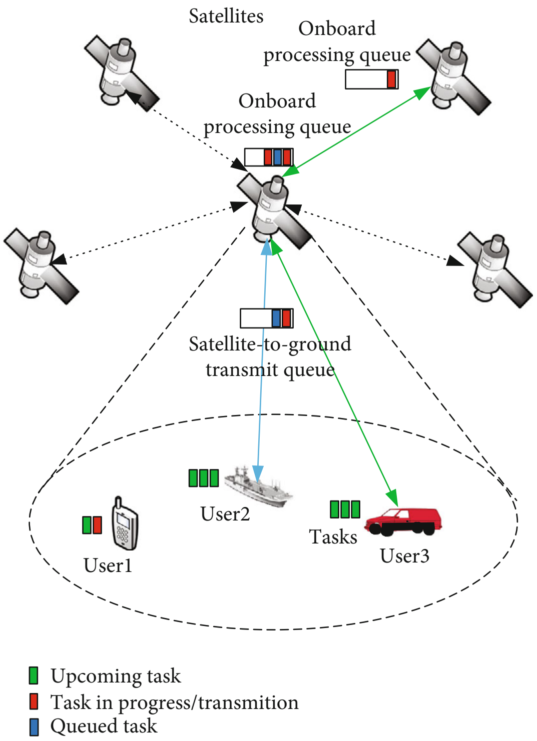


Figure 1: Collaborative computing among multiple LEO satellites.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_{\mathrm{ECI}} = (R+h)\begin{bmatrix} \cos\Omega\,\cos\left(\varphi'(t)\right) - \sin\Omega\,\sin\left(\varphi'(t)\right)\cos i \\ \sin\Omega\,\cos\left(\varphi'(t)\right) + \cos\Omega\,\sin\left(\varphi'(t)\right)\cos i \\ \sin\Omega\,\sin\left(\varphi'(t)\right)\sin i \end{bmatrix}, \tag{1}$$

where R is the earth radius, h is the height of satellite, and $a = R + h/2$. $\varphi'(t) = \varphi + \omega_s t$, where $\omega_s = \sqrt{GM_e/(R+h)^3}$ denotes the angular velocity of satellite, which is related to the altitude of satellite, gravitational constant G, and mass of the earth M_e. $t = \rho(l-1)$ denotes the running time of the satellite at the beginning of time slot l, in which ρ is the length of one time slot.

To obtain the coordinates of satellite n, the RAAN Ω_n and initial true anomaly φ_n also needs to be calculated. Since the Walker constellation is symmetrical and all satellites adopt circular orbits with the same height and the same inclination, the orbital plane is evenly distributed along the equator, and the satellites are evenly distributed in the orbital plane. The phase relationship of the satellites in different orbital planes can be expressed as

$$\begin{cases} \Omega_n = \dfrac{2\pi}{P}(P_n - 1), \\ \varphi_n = \dfrac{2\pi}{S}(N_n - 1) + \dfrac{2\pi}{N}F(P_n - 1), \end{cases} \tag{2}$$

where N denotes the number of satellites, P denotes the number of orbits, S denotes the number of satellites in each orbit, thus, $N = P \times S$. P_n is the serial number of the orbit where the satellite n is located, and $P_n = \lfloor n/S \rfloor + 1$. N_n is the serial number of satellite n in its orbit, and $N_n = n - (P_n - 1)S$. F denotes the phase factor of orbit. With (1) and (2), we can obtain the ECI coordinate of any satellite in the constellation at any time slot.

3.2. Coverage Model. Generally, the users' coordinates are expressed with longitude, latitude, and altitude (LLA) in the geographic coordinate system. To derive the situation of coverage of satellite at time slot l, we first need to convert the ECI coordinate of the satellite to earth centered earth fixed (ECEF) coordinate with the following formula

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_{\text{ECEF}} = \begin{bmatrix} \cos \eta_g & -\sin \eta_g & 0 \\ \sin \eta_g & \cos \eta_g & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{\text{ECI}}, \tag{3}$$

where $\eta_g = \eta_{g_0} + \omega_e t$, in which η_{g_0} denotes the Greenwich hour angle at the beginning of the first time slot, and ω_e denotes the angular velocity of the earth's rotation. Then, we need to convert the ECEF coordinate of the satellite to LLA coordinate (lon, lat, alt) according to

$$\begin{cases} lon = \arctan \dfrac{y}{x}, \\ lat = \arctan \dfrac{z + Je^2 \sin lat}{\sqrt{x^2 + y^2}}, \\ alt = \dfrac{z}{\sin lat} - J\left(1 - e^2\right), \end{cases} \tag{4}$$

where $J = a/1 - e^2 \sin^2 lat$. Since $e = 0$, $J = a$ can be achieved. With (3) and (4), we can obtain the LLA coordinate of users and satellites, and the elevation angle σ of user can be expressed as

$$\sigma = \arctan \frac{\cos \Delta lon \cos lat_u \cos lat_s + \sin lat_u \sin lat_s - R/R + h}{\sqrt{1 - (\cos \Delta lon \cos lat_u \cos lat_s + \sin lat_u \sin lat_s)^2}}, \tag{5}$$

where $\Delta lon = lon_u - lon_s$, lon_u and lat_u denote user's longitude and latitude, respectively, and lon_s and lat_s denote satellite's longitude and latitude, respectively. We consider user m is covered by satellite n when σ_n^m is larger than the minimum elevation angle $\sigma_{\min}$. At the beginning, user m will select the satellite n_a with the smallest elevation angle for association, and when σ_n^m is less than $\sigma_{\min}$ as the associated satellite moves, the user will select the satellite n_h according to the elevation angle at the current time slot for handover.

3.3. Channel Model. In the scenario shown in Figure 1, two kinds of links should be considered, which are satellite-to-ground (StG) and satellite-to-satellite (StS) links. For StG links, line-of-sight (LOS) channel is assumed to be always existing between satellites and users on ground. Since we focus on the resource management for the satellite systems, only the path loss affected by the distance between satellites and users is considered in this paper, and the free space path loss (FSPL) model is adopted. Moreover, the channel quality of the users on ground, which are indicated by signal-to-noise ratio (SNR), will be quantified and transmitted to the satellites via the control channels. For StS links, point-to-point optical links are assumed to be implemented, and the channel capacity of the StS links is assumed to be large enough for the data transmission between satellites.

3.4. Task Arrival Model. Generally, LEO satellite-assisted IoT is suitable for multiple kinds of services, such as object identification and tracking and assets monitoring. In this paper, the tasks of each user are assumed to arrive continuously, and the task arrival follows Poisson distribution. The probability of κ tasks arriving in l time slots can be expressed as

$$P(T(l) = \kappa) = \frac{(\lambda l)^\kappa e^{-\lambda l}}{\kappa!}, \tag{6}$$

where λ denotes the rate of task arrival. And the time interval of task arrival follows the exponential distribution with parameter λ. Moreover, multiple tasks belong to one single user can only be handled with a first-in-first-out policy, and a single task cannot be scheduled until the processing results of its previous task are sent back to the user. More complicated scenarios, where multiple tasks of a single user are scheduled simultaneously, will be considered in future work.

3.5. Latency Model. Since the task can be handled locally or offloaded to satellites, the latency τ_k^m of task k generated by user m will be analyzed with three possible cases.

If task k is handled locally by user m, the latency τ_k^m can be expressed as

$$\tau_k^m = \tau_{k,m}^{\text{wait}} + \tau_{k,m}^{\text{process}}, \tag{7}$$

where $\tau_{k,m}^{\text{wait}}$ denotes the waiting latency in queue and waiting latency due to the local resources being occupied by the task being processed. $\tau_{k,m}^{\text{process}} = T_k^m / X_m$ denotes the processing latency, T_k^m is the number of bits in task k of user m, and X_m is the local computing resource of user m.

If task k is handled by satellite n_a associated with user m, the latency τ_k^m can be expressed as

$$\tau_k^m = \tau_{k,\mathrm{m}}^{\text{off}} + \tau_{k,m}^{\text{process}} + \tau_{k,m}^{\text{return}}, \tag{8}$$

where $\tau_{k,m}^{\text{off}}$ is the latency of offloading task k from user m to its associated satellite n_a, and it consists of waiting latency $\tau_{k,m}^{w\_\text{trans}}$ for transmission, transmission latency $\tau_{k,m}^{\text{trans}}$, and propagation latency $\tau_{k,m}^{\text{prop}}$. Moreover, $\tau_{k,m}^{\text{trans}} = T_k^m / C_{n_a}^m$, where $C_{n_a}^m$ denotes the data rate allocated by satellite n_a to task k of user m, and $C_{n_a}^m$ is affected by the available communication resource, such as power, bandwidth. $\tau_{k,m}^{\text{prop}} = d_{m,n_a}/c$, where c is the speed of light, and d_{m,n_a} is the distance from user m to satellite n_a. $\tau_{k,m}^{\text{process}}$ composed of waiting latency in processing queue $\tau_{k,m}^{w\_\text{pro}}$ and processing latency $\tau_{k,m}^{\text{pro}}$, and $\tau_{k,m}^{\text{pro}} = T_k^m / X_{n_a}^m$, where $X_{n_a}^m$ denotes the computing resource allocated to task k of user m by satellite n_a. $\tau_{k,m}^{\text{return}}$ is the latency caused by sending the results back to user. Note that the transmission latency of return link is omitted, because the number of bits of computing results is usually very small. If the result can be returned to user within the coverage time of satellite n_a, $\tau_{k,m}^{\text{return}} = d_{n_a,m}/c$, otherwise, $\tau_{k,m}^{\text{return}} = d_{n_a,n_h} + d_{n_h,m}/c$, and d_{n_a,n_h} denotes the routing distance from satellite n_a to satellite n_h, and $d_{n_h,m}$ denotes the routing distance from satellite n_h to user m.

If task k is offloaded to satellite n_p through satellite n_a for processing, the latency τ_k^m can be expressed as

$$\tau_k^m = \tau_{k,m}^{\text{off}} + \tau_{k,m}^{\text{ISL}} + \tau_{k,m}^{\text{process}} + \tau_{k,m}^{\text{return}}, \tag{9}$$

where $\tau_{k,m}^{\text{off}}$ is the latency of offloading task k from user m to its associated satellite n_a. $\tau_{k,m}^{ISL}$ is the propagation latency of task k routing from satellite n_a to satellite n_p through ISLs. Suppose that optical links are adopted for intersatellite communication, and the ISLs data rate is high enough, the transmission latency of ISLs can be omitted. Therefore, $\tau_{k,m}^{\text{ISL}} = d_{n_a,n_p}/c$, in which d_{n_a,n_p} denotes the routing distance from satellite n_a to satellite n_p, and the routing strategy based on minimum distance is adopted. $\tau_{k,m}^{\text{process}}$ is composed of waiting latency $\tau_{k,m}^{w\_\text{pro}}$ in processing queue and processing latency $\tau_{k,m}^{\text{pro}}$ on satellite n_p. $\tau_{k,m}^{\text{pro}} = T_k^m / X_{n_p}^m$, where $X_{n_p}^m$ denotes the computing resource allocated to task k of user m by satellite n_p. $\tau_{k,m}^{\text{return}}$ is the propagation latency of returning the result back to user. First, the result will be routed from satellite n_p to satellite n_h, which is the serving satellite of user m. And then, the result will be sent by satellite n_h to user m. Thus, $\tau_{k,m}^{\text{return}} = d_{n_p,n_h} + d_{n_h,m}/c$.

3.6. Problem Formulation. In this paper, we intend to achieve the collaborative computing among multiple satellites through ISLs and maximize the completion rate of user tasks under latency constraints. Thus, the problem can be formulated as

$$\begin{aligned} \max_{\bar{\Omega}_l, \bar{\mathbf{\Psi}}_l, \bar{\mathrm{O}}_{\mathrm{l}}, \mathrm{C}_l, \mathrm{X}_l} \quad & \frac{\sum_l \sum_m \sum_k v_k^m}{\sum_l \sum_m \sum_k w_k^m} \\ s.t. \quad & \sum_{m \in \bar{\mathbf{\Phi}}_{n,l}} C_{n,l}^m \le \bar{C}_{n,l}, \\ & \sum_{m \in \bar{\mathbf{\Theta}}_{n,l}} X_{n,l}^m \le \bar{X}_{n,l}, \end{aligned} \tag{10}$$

where $w_k^m = 1$ when new task k arrives at user m, and $v_k^m = 1$ when $\tau_k^m \le \tau_{\max}$, otherwise, $v_k^m = 0$. $\tau_{\max}$ is the maximum latency constraint for the task k of user m. $\bar{\Omega}_l = \{\bar{\Omega}_{1,l}, \bar{\Omega}_{2,l}, \cdots, \bar{\Omega}_{M,l}\}$ and $\bar{\Omega}_{m,l} \in \{0,1\}$, $\bar{\Omega}_{m,l} = 1$ denotes that the task of user m will be handled locally at time slot l. $\bar{\mathbf{\Psi}}_l = \{\bar{\Psi}_{1,l}, \bar{\Psi}_{2,l}, \cdots, \bar{\Psi}_{M,l}\}$ and $\bar{\Psi}_{m,l} \in \{1, 2, \cdots, N\}$, $\bar{\Psi}_{m,l} = n$ denotes that the task k of user m will be offloaded to satellite n at time slot l for processing. $\bar{\mathrm{O}}_{\mathrm{l}} = \{\bar{\mathrm{O}}_{1,\mathrm{l}}, \bar{\mathrm{O}}_{2,\mathrm{l}}, \cdots, \bar{\mathrm{O}}_{\mathrm{M,l}}\}$ and $\bar{\mathrm{O}}_{\mathrm{m,l}} \in \{0,1\}$, $\bar{\mathrm{O}}_{\mathrm{m,l}} = 1$ denotes that the task k of user m will be scheduled to be transmitted or processed at time slot l. $\mathrm{C}_l = \{C_{1,l}, C_{2,l}, \cdots, C_{N,l}\}$ and $C_{n,l} = \{C_{n,l}^1, C_{n,l}^2, \cdots, C_{n,l}^M\}$ denote the communication resource allocated by satellite n to task k of user m at time slot l. $\mathrm{X}_l = \{X_{1,l}, X_{2,l}, \cdots, X_{N,l}\}$ and $X_{n,l} = \{X_{n,l}^1, X_{n,l}^2, \cdots, X_{n,l}^M\}$ denote the computing resource allocated by satellite n to task k of user m at time slot l. $\bar{C}_{n,l}$ and $\bar{X}_{n,l}$ are the available communication resource and computing resource of satellite n at time slot l, respectively. $\bar{\mathbf{\Phi}}_{n,l}$ and $\bar{\mathbf{\Theta}}_{n,l}$ denote the tasks scheduled in transmission queue and processing queue of satellite n at time slot l, respectively.

From (10), we can see that the completion rate of tasks is affected by the offloading decision, scheduling decision, and resource allocation at each time slot. In addition, the offloading decision, scheduling decision, and resource allocation at the current time slot l will affect the states of time slot $l+1$. For instance, if the task cannot complete the transmission from user to satellite at time slot l, it cannot be processed at time slot $l+1$, and the communication resource occupied cannot be released to other tasks. Thus, the problem formulated in (10) can be seen as a dynamic programming problem based on joint offloading decision, scheduling decision, and resource allocation, which is difficult to be solved with traditional methods. To tackle the problem, we will decompose the complex problem into two subproblems.

4. Task Completion Rate Optimization Based on Deep Reinforcement Learning

According to Section 3, task offloading and scheduling decision indicators are discrete, while the resource allocation variables are continuous. Therefore, the problem formulated in (10) is a dynamic mixed-integer problem, which is nonconvex and difficult to find the optimal solutions. To solve this problem, we decompose the problem into two subproblems to reduce its complexity. The first subproblem is communication and computing resource allocation with fixed offloading decision and scheduling decision, which will be solved based on max-min fairness. The second subproblem is the joint offloading decision and scheduling decision, which will be solved with a DRL-based algorithm. The two subproblems are analyzed in the following subsection A and subsection B, respectively.

4.1. Resource Allocation Based on Max-Min Fairness with Fixed Task Assignment. With the fixed offloading decision $\bar{\Omega}_l$, $\bar{\mathbf{\Psi}}_l$, and scheduling decision $\bar{\mathrm{O}}_{\mathrm{l}}$ at time slot l, the communication and computing resource allocation of satellite n can be formulated as two separated max-min fairness problems, which are listed as

$$\begin{aligned} \mathscr{P}1:\min_{C_{n,l}^{m}} \quad & \max_{\forall m} \tau_{n,m}^{\text{trans}}, \\ s.t. \quad & \sum_{m\in\bar{\mathbf{\Phi}}_{n,l}} C_{n,l}^{m} \le \bar{C}_{n,l}, \end{aligned} \tag{11}$$

$$\begin{aligned} \mathscr{P}2:\min_{X_{n,l}^{m}} \quad & \max_{\forall m} \tau_{n,m}^{\text{pro}}, \\ s.t. \quad & \sum_{m\in\bar{\mathbf{\Theta}}_{n,l}} X_{n,l}^{m} \le \bar{X}_{n,l}, \end{aligned} \tag{12}$$

where $\tau_{n,m}^{\text{trans}}$ is the transmission latency for a given task of user m associated with satellite n, and it can be calculated after the task being scheduled from transmission queue and assigned with communication resource. $\tau_{n,m}^{\text{pro}}$ is the processing latency a given task of user m processed at satellite n, and it can be obtained after task being scheduled from processing queue and assigned with computing resource. Here, minimize the maximum latency is adopted as an optimization objective, and it is helpful to guarantee the latency constraints of every task.

Take (11) as an example, introduce auxiliary variable χ, the problem can be rewritten as

$$\begin{aligned} \min_{C_{n,l}^{m},\chi} \quad & \chi, \\ s.t. \quad & \chi \ge \frac{T_m}{C_{n,l}^{m}}, \forall m, \\ & \sum_{m\in\bar{\mathbf{\Phi}}_{n,l}} C_{n,l}^{m} \le \bar{C}_{n,l}. \end{aligned} \tag{13}$$

Obviously, the objective function and the second constraint of (13) are both convex. Thus, we only need to prove that the first constraint is convex to prove that this problem is convex.

Let $F = T_m/\chi C_{n,l}^{m}$, the constraint can be rewritten as $F \le 1, \forall m$. Find the second partial derivative of F, and the Hessian matrix of F can be expressed as

$$H_F = \begin{bmatrix} \dfrac{2T_m}{C_{n,l}^{m}\chi^3} & \dfrac{T_m}{\left(C_{n,l}^{m}\right)^2\chi^2} \\ \dfrac{T_m}{\left(C_{n,l}^{m}\right)^2\chi^2} & \dfrac{2T_m}{\left(C_{n,l}^{m}\right)^3\chi} \end{bmatrix}. \tag{14}$$

With (14), we can easily get that all the principal minor of H_F are nonnegative, and H_F is a positive semidefinite matrix. Thus, the first constraint of (13) is convex, and problem (13) is a convex problem, which can be solved by the dual ascent method.

Construct Lagrangian functions $\mathscr{L}$ as

$$\mathscr{L}\left(C_{n,l}^{m}, \mu_m, \nu\right) = \chi + \sum_{M} \mu_m \left(\frac{T_m}{C_{n,l}^{m}} - \chi\right) + \nu \left(\sum_{m\in\bar{\mathbf{\Phi}}_{n,l}} C_{n,l}^{m} - \bar{C}_{n,l}\right), \tag{15}$$

where $\mu_m \ge 0$ and $\nu \ge 0$ are Lagrangian multipliers. Then, the dual function of $\mathscr{L}$ will be

$$\mathscr{D}(\mu_m, \nu) = \mathscr{L}\left(C_{n,l}^{m^*}(\mu_m, \nu), \mu_m, \nu\right), \tag{16}$$

where $C_{n,l}^{m^*} = \arg\min_{C_{n,l}^{m}} \mathscr{L}(C_{n,l}^{m}, \mu_m, \nu)$. Since the problem defined in (13) is convex, the maximum value of the $\mathscr{D}$ is equivalent to the minimum value of the problem defined in (13). Thus, we can find the optimal solution of problem defined in (13), which is also the solution of problem defined in (11), via the method proposed in Algorithm 1.

In this paper, μ_m and ν are seen to be converged when the value difference is less than 0.001 for 100 consecutive iterations. By continuously iterating the independent variable and Lagrangian multipliers alternately, we can find the optimal solution $C_{n,l}^{m}$ for communication resource allocation. Similarly, $X_{n,l}^{m}$ can be obtained.

4.2. Joint Task Offloading, Scheduling, and Resource Allocation. With Algorithm 1, we can obtain the resource allocation solution for each time slot. However, the joint offloading decision and scheduling decision is still a nonconvex problem with integer programming problem, which cannot be tackled directly with traditional methods based on optimization theory. To address the problem with affordable complexity, we model it as an Markov decision process (MDP) problem and propose a DRL-based method to achieve long-term rewards in terms of task completion rate.

The MDP corresponding to the problem defined in (10) can be expressed as

(1) *State*. The states are defined for every time slot, because the scheduling and resource allocation for tasks are managed slot by slot. The state at time slot l can be defined as $h_l = \{P_U(l), P_S(l), T_l, \tilde{\Xi}_l, \tilde{X}_{U,l}, \tilde{\mathrm{C}}_l, \tilde{\mathrm{X}}_l, \mathrm{Q}_{\text{trans},l}, \mathrm{Q}_{\text{pro},l}\}$. $P_U(l)$ and $P_S(l)$ denote the location of users and satellites, respectively. $T_l = \{T_{1,l}, T_{2,l}, \cdots, T_{M,l}\}$ denotes the bits of tasks currently waiting to be scheduled. $\tilde{\Xi}_l = \{\tilde{\Xi}_{1,l}, \tilde{\Xi}_{2,l}, \cdots, \tilde{\Xi}_{M,l}\}$, and $\tilde{\Xi}_{m,l} \in \{1, 2, \cdots, N\}$ denotes the satellite associated with user m at time slot l. $\tilde{X}_{U,l} = \{\tilde{X}_{1,l}, \tilde{X}_{2,l}, \cdots, \tilde{X}_{M,l}\}$ and $\tilde{X}_{m,l} \in \{0, 1\}$, $\tilde{X}_{m,l} = 1$ denotes that the local computing resource is occupied at time slot l. $\tilde{C}_l = \{\tilde{C}_{1,l}, \tilde{C}_{2,l}, \cdots, \tilde{C}_{N,l}\}$ denotes the communication resource of satellites occupied by users at time slot l. Similarly, $\tilde{X}_l = \{\tilde{X}_{1,l}, \tilde{\mathrm{X}}_{2,l}, \cdots, \tilde{X}_{N,l}\}$ denotes the computing resource of satellites occupied by users at time slot l. $\mathrm{Q}_{\text{trans},l} = \{Q_{\text{trans},l}^{1}, Q_{\text{trans},l}^{2}, \cdots, Q_{\text{trans},l}^{N}\}$ and $\mathrm{Q}_{\text{pro},l}$

```
Input:
  Lagrangian multipliers: $\mu_m$, $\nu$
Output:
  Resource allocation: $C_{n,l}^m$
Initialize $\mu_m$, $\nu$.
Repeat
    Find $C_{n,l}^{m^*} \longleftarrow \arg \min_{C_{n,l}^m} \mathscr{L}(C_{n,l}^m, \mu_m, \nu)$.
    Update $\chi \longleftarrow \max_{\forall m}(T_m/C_{n,l}^{m^*})$.
    Compute $d\mathscr{D}/d\mu_m = (d\mathscr{L}/d\lambda_m)(C_{n,l}^{m^*}, \lambda_m, \nu)$,
     $d\mathscr{D}/d\nu = (d\mathscr{L}/d\nu)(C_{n,l}^{m^*}, \lambda_m, \nu)$.
Update $\lambda_m \longleftarrow \lambda_m + \alpha(d\mathscr{D}/d\mu_m)$,
    $\nu \longleftarrow \nu + \alpha(d\mathscr{D}/d\nu)$.
Until $\mu_m$, $\nu$ converge.
```

Algorithm 1: Resource allocation based on maximum latency minimization (MLMRA).

$= \{Q_{\text{pro},l}^1, Q_{\text{pro},l}^2, \cdots, Q_{\text{pro},l}^N\}$ denote the total bits of tasks that wait for transmission and processing in the queue of satellites, respectively

(2) *Action*. For each time slot l, the action consists of offloading decision, scheduling decision, and resource allocation of user's current tasks. Since we can obtain the resource allocation with Algorithm 1, we only need to define the action space for offloading decision and scheduling decision with lower dimensions. Thus, the action at time slot l can be defined as $a_l = \{\bar{A}_{1,l}, \bar{A}_{2,l}, \cdots, \bar{A}_{M,l}\}$. $\bar{A}_{m,l} = \{A_{\text{off}}, A_{\text{exe}}\}$, in which $A_{\text{off}} \in \{0, 1, \cdots, Z\}$ denotes the satellite that will handle the current task of user m at time slot l, and $A_{\text{exe}} \in \{0, 1\}$, $A_{\text{exe}} = 1$ denotes that the current task of user m will be scheduled from the queue at time slot l. Otherwise, the task will be kept on waiting in the queue for scheduling. With a specific action a_l, the offloading decision and scheduling decision of all current tasks at time slot l can be obtained correspondingly

(3) *Transition Probability*. For MDP, the transition probability from one state to another is needed for any action a_l. However, it is difficult to get the accurate probability for all of states h_l and actions a_l, because the states space and action space are too large. In this paper, a method based on model-free DRL is considered

(4) *Reward*. To maximize the completion rate of tasks, the reward $R(h_l, a_l)$ at time slot l with state h_l and action a_l is defined as

$$R(h_l, a_l) = \sum_{k \in \bar{O}_l} \left(R_p - \tau_k\right) + dR_d, \tag{17}$$

where τ_k denotes the latency defined in (8) or (9) for task k at time slot l. R_p is a constant value that makes the $R_p - \tau_k$ positive. R_d is a positive completion reward, and d denotes the number of tasks which are completed within the latency constraints in time slot l.

Given the action policy π, value function $V(h \mid \pi)$, which can be used to evaluate the long-term performance of the policy π, is defined as

$$V(h \mid \pi) = \mathbb{E}\left[\sum_l \gamma^l R(h_l, a_l) \mid h_0 = h, \pi\right], \tag{18}$$

where γ denotes the discount factor, and the value function can be seen as an expectation of completion rate defined in (10) with $\gamma = 1$. Thus, the optimal policy π^* can be expressed as

$$\pi^*(h) = \arg\max_a \left[R(h, a) + \sum_{\bar{h}} \gamma V\left(\bar{h} \mid \pi^*\right)\right]. \tag{19}$$

where state $\bar{h}$ can be obtained with action a and state h. In this paper, deep Q-network (DQN) [28, 29], which is composed of target network and main network, is adopted to obtain the target Q-value $Q^*(h, a)$. Moreover, the approximated Q-function $Q(h, a\,; \theta)$ will approach $Q^*(h, a)$ via training process by minimizing the loss function, which can be defined as $L(\theta) = E[(Q^*(h, a) - Q(h, a\,; \theta))^2]$. And θ is the weight of network. The detailed description and analysis for the processes of DQN can be found in [23]. The proposed joint task offloading, scheduling, and resource allocation (JTOSRA) approach for collaborative computing among LEO satellites is shown in Algorithm 2, where G denotes maximum of training step, and ζ denotes the experience replay buffer. ε-greedy policy is utilized to balance the exploration and utilization of models [29], and ε will decay from 1.0 to 0.001 through 20000 steps.

Generally, it is hard to obtain the accurate computational complexity of a DRL-based algorithm. In Algorithm 2, the computational complexity of the DQN network mainly depends on the number of users and the network structure

```
Input:
  IoT terminal information: $P_U(l)$, $\tilde{X}_{U,l}$, $T_l$
  Satellite information: $P_S(l)$, $\tilde{\Xi}_l$, $\tilde{C}_l$, $\tilde{X}_l$
  Queue information: $Q_{trans,l}$, $Q_{pro,l}$
Output:
  Offloading and scheduling decisions: $\bar{\Omega}_l$, $\bar{\Psi}_l$, $\bar{O}_l$
  Resource allocation: $C_l$, $X_l$
Initialize network with $\gamma$, $\varepsilon$ and $\zeta$.
Initialize state $h_l$
While $l < G$
    Select an action $a_l$ according to $\varepsilon$-greedy policy.
    Allocate resource according to Algorithm 1.
    Calculate reward $R(h_l, a_l)$.
    Update next state $h_{l+1}$.
    Save $(h_l, a_l, R(h_l, a_l), h_{l+1})$, and update $\zeta$.
    Update $\theta$.
    $l$ ++.
End while
```

Algorithm 2: Joint task offloading, scheduling, and resource allocation based on DQN.

of neural network utilized by DQN. All layers of the neural network used in this algorithm are fully connected layers, and the number of input parameters of the network is determined by the state space, which can be represented as n_0. Assuming that the number of neural network layers is J, and the number of neurons in the j-th layer is expressed by n_j, the computational complexity for the DQN network can be expressed as $O(G \times (\sum_{j=0}^{J} n_j n_{j+1}))$ [30].

5. Simulation Results

5.1. Simulation Configurations. Simulation parameters are listed in Table 1. In the simulation, we focus on the users located in a specific area covered by the LEO satellites. The simulation time starts at 0 : 00 on October 1, 2020, and the Greenwich hour angle θ_{g_0} at this moment is 10.2. The communication capacity of satellite is set to 10 Gbps. The CPU cycle needed for processing is set to 1000 cycle/bit [31]. And we set the satellite computing capacity and local computing capacity to 10 GC/s [32] and 1.5 GC/s [33], respectively.

5.2. Convergence of JTOSRA. Figure 2 shows the convergence of the loss function. It can be seen that the loss function defined in $L(\theta)$ will converge when the training steps increase. As shown in Figure 3, the completion rate of tasks will also increase during the training processes. Figures 2 and 3 demonstrate that the JTOSRA based on DQN is applicable for the problem formulated in Section 3. Though a large amount of training steps is needed to achieve convergence, the training processes will only be implemented in the initial phase of the LEO network. Once the convergence is achieved, joint task offloading and scheduling decisions can be made step by step with low complexity. Moreover, the training processes can be implemented offline via pretraining processes to decrease the complexity further.

Table 1: Simulation parameters.

Parameter	Symbol	Value
Scene parameters		
Task size	T_m	8×10^4-1.2×10^5 bit
Rate of task arrival	λ	0.05
Latency constraint of task	$\tau_{\max}$	200 ms
Length of time slot	ρ	10 ms
Number of satellite orbits	P	18
Number of satellites per orbit	S	40
Height of LEO satellite	H	1200 km
Inclination of satellite orbit	i	87.9°
Phase factor of orbit	F	1
Minimum elevation angle of user	$\sigma_{\min}$	20°
Greenwich hour angel	θ_{g_0}	10.2°
CPU cycle needed for processing	ξ	1000 cycles/bit
Satellite communication capacity	C	10 Gbps
Satellite computing capacity	X_S	10 GC/s
Local computing capacity	X_U	1.5 GC/s
Algorithm parameters		
Step size of dual ascent	α	0.001
Maximum training episode	G	150000
Size of replay buffer	ζ	20000
Observation size	O_b	5000
Discount factor	γ	0.95
Positive reward	R_p	5
Completion reward	R_d	15
Learning rate	ι	0.001

5.3. Performance Analysis. To analyze the performance of MLMRA with fixed task assignment, two reference schemes are adopted and compared in terms of completion rate. The referred resource allocation scheme is listed as

(i) *Average Resource Allocation (ARA)*. The resource will be allocated evenly to tasks

(ii) *Resource Allocation Based on Latency Minimization (LMRA)*. The resource will be allocated to tasks by minimizing the sum latency of tasks

Figure 4 shows the influence of the number of tasks on the completion rate with different resource allocation method. It can be seen that the completion rate of the task will decrease along with the increase of the number of tasks. Moreover, the MLMRA performs better than the LMRA, and the ARA algorithm performs the worst. This shows that the MLMRA can allocate resources more equitably and minimize the maximum latency of the task. This is because that the LMRA focuses on reducing the sum latency of tasks, while the MLMRA will allocate more resources to tasks that are difficult to be completed within the limited latency. In general, the proposed MLMRA algorithm can effectively improve the completion rate of the task with fairness among tasks.

To evaluate the performance of the JTOSRA, the following two algorithms are introduced for task assignment:

(i) *Random*. Tasks will be offloaded and scheduled randomly, and resources will be allocated according to LMRA and MLMRA. And in Figures 5–7, the methods are labeled as random-LMRA and random-MLMRA, respectively

(ii) *Simulated Annealing (SA)*. Tasks will be offloaded and scheduled through the SA algorithm, and resources will be allocated by LMRA and MLMRA. In Figures 5–7, the methods are labeled as SA-LMRA and SA-MLMRA, respectively

In addition, in order to ensure the validity of simulation results, each point in the figures is obtained by taking the average value over multiple tests, and each test lasts for 200000 slots. Figure 5 shows the completion rate performance of the algorithms as the number of users increases. With the increase of the number of users, the number of tasks waiting for scheduling will increase, and the shortage of resources will lead to the decline of the task completion rate. Obviously, the proposed JTOSRA algorithm performs better than the SA algorithm. And for the JTOSRA, the task completion rate decreases slower than that of the SA algorithm as the number of users increases. This is because the SA algorithm tends to fall into local optimal solutions, resulting in poor algorithm performance. On the other hand, traditional algorithms such as the SA algorithm can only optimize the decision for a specific time slot, but cannot continuously optimize the offloading and scheduling decisions for multiple time slots, and the algorithm needs to iterate at each step, which brings high time cost. However, the proposed JTOSRA algorithm can continue to accumulate experience in the decision-making process to optimize the

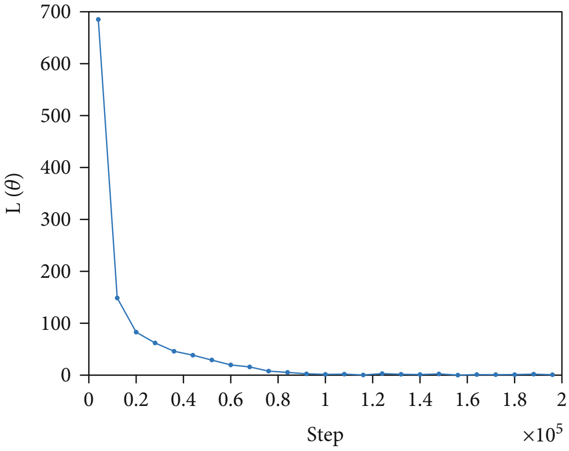


Figure 2: Convergence process of loss.

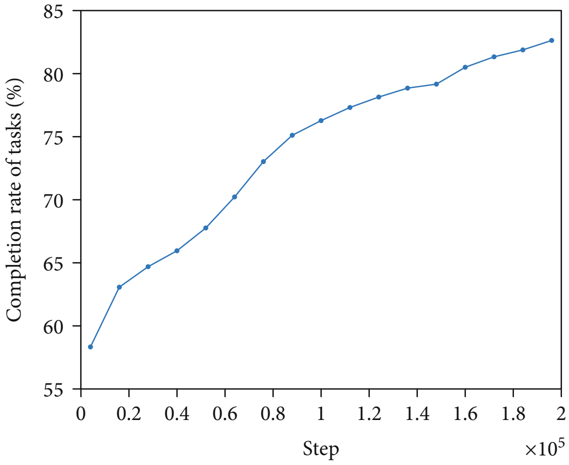


Figure 3: Convergence process of completion rate.

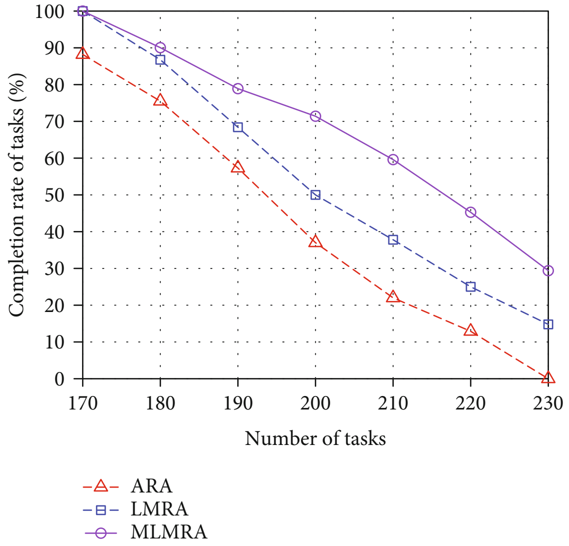


Figure 4: Completion rate of tasks vs. number of tasks.

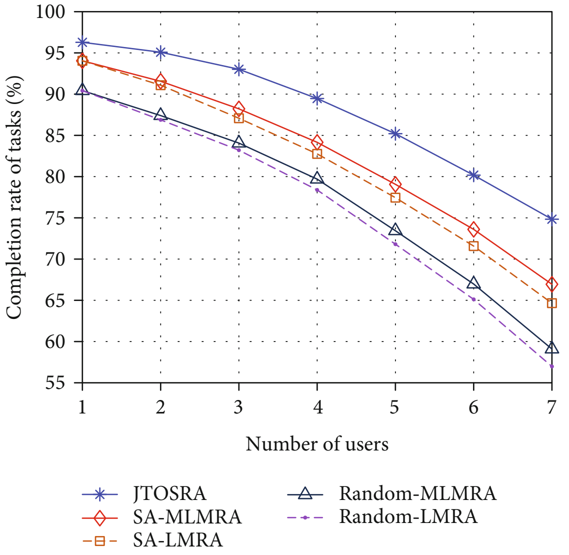


Figure 5: Completion rate of tasks vs. number of users.

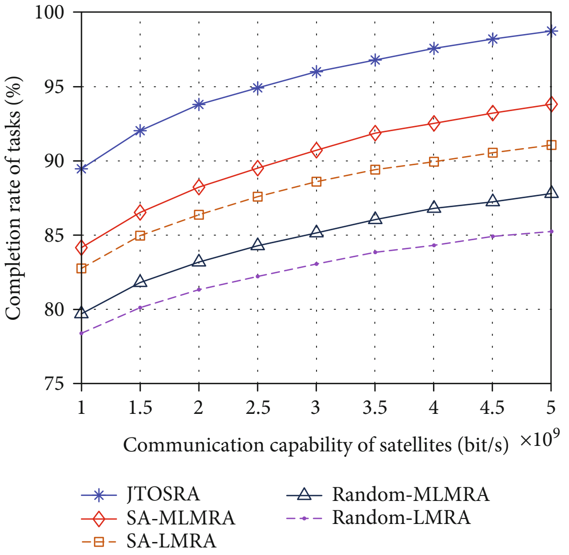


Figure 6: Completion rate of tasks vs. communication capability of satellites.

completion rate of tasks. In addition, MLMRA performs better than LMRA.

In Figure 6, satellite communication capability is adopted as variable to investigate the performance of the proposed algorithm. It can be seen that the increase of satellite communication resources, which means that more communication resource can be allocated to tasks, will lead to the increasing of completion rate. But the increasing rate of the curve decreases with the rise of satellite communication resources. This is because that communication resource is the main factor influencing the latency of tasks when the amount of communication resources is small. When the amount of communication resources of satellites increases to a certain value, the computing resources of satellites will become the dominant factor, which will mainly affect the latency of tasks. In addition, the performance of the JTOSRA algorithm is still better than that of the other two algorithms, and the performance of the MLMRA algorithm is better than that of the LMRA algorithm. Similarly, the completion rate with respect to satellite computing capability is shown in Figure 7. The variation trend of each curve in the figure is close to that in Figure 6.

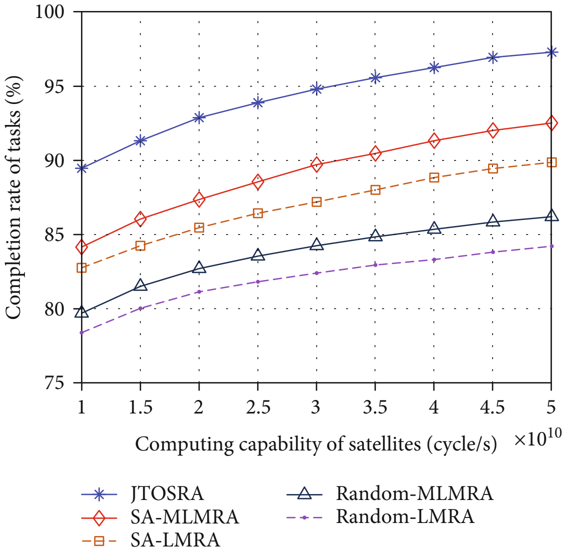


Figure 7: Completion rate of tasks vs. computing capability of satellites.

6. Conclusion

In this paper, collaborative computing and resource allocation for LEO satellite networks are investigated. A framework for collaborative computing among LEO satellites with varying topology is proposed, and the joint task offloading, scheduling, and multidimensional resource allocation problem is divided into two subproblems with low complexity. JTOSRA based on DRL and max-min fairness is proposed to solve the problems, and simulation results demonstrate that the JTOSRA outperforms the referred schemes in terms of task completion rate.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61971054 and Grant 62171052 and Science and Technology on Information Transmission and Dissemination in

Communication Networks Laboratory Foundation under Grant HHX21641X002.

References

[1] O. Kodheli, E. Lagunas, N. Maturo et al., “Satellite communications in the new space era: a survey and future challenges,” *IEEE Communications Surveys Tutorials*, vol. 23, no. 1, pp. 70–109, 2021.

[2] R. Xie, Q. Tang, Q. Wang, X. Liu, F. R. Yu, and T. Huang, “Satellite-terrestrial integrated edge computing networks: architecture, challenges, and open issues,” *IEEE Network*, vol. 34, no. 3, pp. 224–231, 2020.

[3] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, “Mobile edge computing: a survey,” *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2018.

[4] C. Adams, A. Spain, J. Parker, M. Hevert, J. Roach, and D. Cotten, “Towards an integrated GPU accelerated SoC as a flight computer for small satellites,” in *2019 IEEE Aerospace Conference*, pp. 1–7, Big Sky, MT, USA, 2019.

[5] L. Yan, S. Cao, Y. Gong et al., “Satec: a 5G satellite edge computing framework based on microservice architecture,” *Sensors*, vol. 19, no. 4, p. 831, 2019.

[6] Z. Zhang, W. Zhang, and F. Tseng, “Satellite mobile edge computing: improving qos of high-speed satellite-terrestrial networks using edge computing techniques,” *IEEE Network*, vol. 33, no. 1, pp. 70–76, 2019.

[7] I. F. Akyildiz and A. Kak, “The internet of space things/cubesats,” *IEEE Network*, vol. 33, no. 5, pp. 212–218, 2019.

[8] S. Cioni, R. De Gaudenzi, O. Del Rio Herrero, and N. Girault, “On the satellite role in the era of 5G massive machine type communications,” *IEEE Network*, vol. 32, no. 5, pp. 54–61, 2018.

[9] J. Liu, X. Du, J. Cui, M. Pan, and D. Wei, “Task-oriented intelligent networking architecture for the space?air?ground?aqua integrated network,” *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5345–5358, 2020.

[10] C. Jiang and X. Zhu, “Reinforcement learning based capacity management in multi-layer satellite networks,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4685–4699, 2020.

[11] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, “Broadband leo satellite communications: architectures and key technologies,” *IEEE Wireless Communications*, vol. 26, no. 2, pp. 55–61, 2019.

[12] I. Del Portillo, B. G. Cameron, and E. F. Crawley, “A technical comparison of three low earth orbit satellite constellation systems to provide global broadband,” *Acta Astronautica*, vol. 159, pp. 123–135, 2019.

[13] N. U. L. Hassan, C. Huang, C. Yuen, A. Ahmad, and Y. Zhang, “Dense small satellite networks for modern terrestrial communication systems: benefits, infrastructure, and technologies,” *IEEE Wireless Communications*, vol. 27, no. 5, pp. 96–103, 2020.

[14] M. Sheng, D. Zhou, R. Liu, Y. Wang, and J. Li, “Resource mobility in space information networks: opportunities, challenges, and approaches,” *IEEE Network*, vol. 33, no. 1, pp. 128–135, 2019.

[15] A. Papa, T. de Cola, P. Vizarreta, M. He, C. Mas-Machuca, and W. Kellerer, “Design and evaluation of reconfigurable sdn leo constellations,” *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1432–1445, 2020.

[16] N. Cheng, F. Lyu, W. Quan et al., “Space/aerial-assisted computing offloading for IoT applications: a learning-based approach,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 5, pp. 1117–1129, 2019.

[17] B. Cao, J. Zhang, X. Liu et al., “Edge-cloud resource scheduling in space-air-ground integrated networks for internet of vehicles,” *IEEE Internet of Things Journal*, p. 1, 2021.

[18] Y. Wang, J. Yang, X. Guo, and Z. Qu, “Satellite edge computing for the internet of things in aerospace,” *Sensors*, vol. 19, no. 20, p. 4375, 2019.

[19] F. Wang, D. Jiang, S. Qi, C. Qiao, and H. Song, “Fine-grained resource management for edge computing satellite networks,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Waikoloa, HI, USA, 2019.

[20] M. LiWang, S. Dai, Z. Gao, X. Du, M. Guizani, and H. Dai, “A computation offloading incentive mechanism with delay and cost constraints under 5G satellite-ground IoV architecture,” *IEEE Wireless Communications*, vol. 26, no. 4, pp. 124–132, 2019.

[21] J. Jiao, Y. Sun, S. Wu, Y. Wang, and Q. Zhang, “Network utility maximization resource allocation for noma in satellite-based internet of things,” *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3230–3242, 2020.

[22] B. Denby and B. Lucia, “Orbital edge computing: machine inference in space,” *IEEE Computer Architecture Letters*, vol. 18, no. 1, pp. 59–62, 2019.

[23] G. Cui, Y. Long, L. Xu, and W. Wang, “Joint offloading and resource allocation for satellite assisted vehicle-to-vehicle communication,” *IEEE Systems Journal*, vol. 15, no. 3, pp. 3958–3969, 2021.

[24] S. Zhang, G. Cui, Y. Long, and W. Wang, “Optimal resource allocation for satellite-aided collaborative computing among multiple user pairs,” *International Journal of Satellite Communications and Networking*, vol. 39, no. 5, pp. 500–508, 2021.

[25] Z. Song, Y. Hao, Y. Liu, and X. Sun, “Energy efficient multiaccess edge computing for terrestrial-satellite internet of things,” *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 14202–14218, 2021.

[26] H. Liao, Z. Zhou, X. Zhao, and Y. Wang, “Learning-based queue-aware task offloading and resource allocation for space–air–ground-integrated power iot,” *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5250–5263, 2021.

[27] Q. Tang, Z. Fei, B. Li, and Z. Han, “Computation offloading in leo satellite networks with hybrid cloud and edge computing,” *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 9164–9176, 2021.

[28] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, “Deep exploration via bootstrapped dqn,” *Advances in neural information processing systems*, pp. 4026–4034, 2016.

[29] V. Mnih, K. Kavukcuoglu, D. Silver et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[30] X. Huang, S. Leng, S. Maharjan, and Y. Zhang, “Multi-agent deep reinforcement learning for computation offloading and interference coordination in small cell networks,” *IEEE Transactions on Vehicular Technology*, p. 1, 2021.

[31] Y. Wang, J. Zhang, X. Zhang, P. Wang, and L. Liu, “A computation offloading strategy in satellite terrestrial networks with double edge computing,” in *2018 IEEE International*

Conference on Communication Systems (ICCS), pp. 450–455, Chengdu, China, 2018.

[32] S. Yang, S. Cao, J. Wei, Y. Zhao, and L. Yan, “Space-based computing platform based on soc fpga,” in *2019 IEEE World Congress on Services (SERVICES)*, Milan, Italy, 2019.

[33] J. Zhang, H. Guo, J. Liu, and Y. Zhang, “Task offloading in vehicular edge computing networks: a load-balancing solution,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2092–2104, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6673807, 12 pages
https://doi.org/10.1155/2021/6673807

Research Article

Impact of IQ Imbalance on RIS-Assisted SISO Communication Systems

Asma Bouhlel[1] and Anis Sakly[2]

[1]*Laboratory of Electronic and Micro-Electronic, Faculty of Sciences Monastir, University of Monastir, Tunisia*
[2]*Laboratory of Industrial Systems Study and Renewable Energy, National Engineering School of Monastir, Tunisia*

Correspondence should be addressed to Asma Bouhlel; bouhlel_asmaa@yahoo.fr

Received 16 November 2020; Revised 2 July 2021; Accepted 23 July 2021; Published 15 August 2021

Academic Editor: Mumtaz Shahid

Reconfigurable intelligent surface (RIS) for wireless networks has emerged as a promising future transmission technique to create smart radio environments that improve the system performance by turning the wireless channel into an adjustable system block. However, transceivers come with various hardware impairments, such as phase noise and in-phase/quadrature-phase imbalance (IQI). Hence, for robust configuration of RIS-based communication under practical conditions, assuming the identical performance analysis when subject to IQI, will lead to inaccurate analysis. In this paper, the implementation of this novel transmission technique is thoroughly investigated under intensive realistic circumstances. For this purpose, based on the maximum likelihood (ML) detector, a novel analytical expression of average pairwise error probability under IQI is proposed and compared to the standard ML detector. Further, the proposed analytical approaches are confirmed by numerical simulations.

1. Introduction

Many attempts have been done in recent years to deliver new deployment models with high speeds, superior reliability, and negligible latency to meet the requirements of 5G standards. To achieve these goals, several transmission techniques have been used such as millimeter wave (mmWave), orthogonal frequency division multiplexing (OFDM), and massive multiple input multiple output (MIMO) [1]. Although the successful launch of the first 5G service, the introduced technologies suffer from high energy consumption and uncontrolled propagation environments effects. Therefore, for the sixth-generation (6G) mobile communication systems with very-high frequency bands and more power efficiency, researchers are already exploring new methods [2] such as reconfigurable intelligent surface (RIS) [3]. RIS, also known as large intelligent surface (LIS) [4], has attracted a significant amount of attention from researchers as a promising future transmission technique to create a smart propagation environment [5, 6]. Conventionally, only the source and the destination are controlled with coding, encoding, and many processing operations to enhance the quality of the signal. By putting a RIS between the transmitter (Tx) and the receiver (Rx), an additional propagation path appears. Thus, the created channel behavior can be software-controlled in order to achieve a smart programmable wireless environment that provides more freedom degrees and boost the performance [6].

The announced technique has been compared with the massive MIMO [7], amplify-and-forward (AF) relaying [8], backscatter communication [9], mmWave communication [10], and network densification [11]. Although it is similar to other existing technologies, the RIS is based on a large number of thin passive reflectors without buffering and processing any incoming signals [4]. These reflectors are designed based on two-dimensional meta-surfaces [12, 13]. In addition, RIS is equipped with a programmable microcontroller that modifies the phase of the incident electromagnetic waves in a way that can enhance the signal quality at the Rx and improve the network coverage. Hence, RIS improves the quality of the received signal by simply reflecting and adjusting the incident signal phase shifts favorably with low cost and low energy consumption [3]. The novel proposed concept presents important theoretical and electromagnetic design challenges [14, 15].

However, the use of RIS poses several new challenges for the transceiver design, and it is paramount to analyze the system performance under practical conditions [16]. Accordingly, RIS-aided communication performance can be significantly degraded by different types of realistic imperfections, including noise signal, imperfect channel state information (CSI), and transceiver hardware impairments. The effect of the hardware impairments general model on the achievable rate of RIS has been studied in [17]. In [18, 19], an asymptotic analysis of the uplink data rate in a RIS-based communication system affected by channel estimation errors and interference was investigated. The design and the implementation of RIS were also detailed in [20].

To the best of the authors' knowledge, RIS under in-phase and quadrature imbalance (IQI) has not been highlighted yet. However, for robust configuration, modeling the transceiver radio frequency (RF) front-end hardware as perfect will lead to inaccurate analysis [21]. Indeed both in-phase (I) and quadrature (Q) modulator and demodulator at the Tx and Rx may introduce phase and/or amplitude mismatch [22]. Additional harmful effects could degrade the system performance such as crosstalk and frequency interference [23–25]. Hence, taking into account IQI effects is a crucial factor for RIS effective design policies.

Motivated by the aforementioned limitations of the existing literature, this paper explores the design of an optimal Rx detector which is compared to the performance of the classical ML in the presence of IQI. Accordingly, novel error probability analytical expressions are derived and proved with simulation results.

2. System Model and Signal Detection

2.1. System Model. In this section, we have adapted a general RIS-assisted single-input single output (SISO) wireless communication system as presented in Figure 1. The direct signal path between the source and the destination is ignored in the rest of the paper [4], and the RIS is deployed to relay the scattered signal. Indeed, this assumption holds in the case of unfavorable propagation conditions that might be caused by an obstacle or a long distance, for example, [26–28].

First, the information is conveyed from the source to the RIS. Then, the RIS software controls the amplitude and the phase of the received signal to combat the propagation environment's harmful effects and reflect it to the destination.

Although RIS is based on small passive elements and does not need any signal processing power, the font-ends of both Tx and Rx could be affected by the IQI which limits the system performance. Actually in practical conditions, due to the local oscillator (LO), filters, analog components, and up-and-down-conversion steps at Tx and Rx sides, the generated signals present a mismatch between the I and Q parts [29]. Accordingly, the Tx and Rx sides IQI parameters are introduced, respectively, using (G_1, G_2) and (K_1, K_2) which can be expressed while based on the complex LO signals [24, 30] as

$$G_1 = \frac{1}{2}\left(1 + \xi_t e^{j\varphi_t}\right), G_2 = \frac{1}{2}\left(1 - \xi_t e^{-j\varphi_t}\right), \tag{1}$$

$$K_1 = \frac{1}{2}\left(1 + \xi_r e^{-j\varphi_r}\right), K_2 = \frac{1}{2}\left(1 - \xi_r e^{j\varphi_r}\right), \tag{2}$$

where (ξ_t, φ_t) and (ξ_r, φ_r) denote, respectively, Tx and Rx amplitude and phase imbalances. In the ideal case, where the IQ branch is perfectly matching these parameters that are reduced to $\xi_t = \xi_r = 1$ and $\varphi_t = \varphi_r = 0°$. Consequently, we have $G_1 = K_1 = 1$ and $G_2 = K_2 = 0$. The up-converted signal modulated over M-ary mapper at the source affected by IQI can be written as

$$x^{IQ} = G_1 x + G_2 x^*, \tag{3}$$

where $()^*$ denotes the complex conjugate.

Using RIS with N reflectors, the signal is firstly transmitted from the source antenna to the RIS then it is conveyed from the RIS to the destination through, respectively, the flat fading channels h_k and g_k for the k^{th} reflecting meta-surface $(k = 1, 2, \cdots, N)$. Note that h_k, g_k follow zero-mean complex Gaussian distribution with unit variance.

In such case, an intelligent RIS software is deployed to adjust the reflection phases based on the channel phases in order to maximize the received SNR. Hence, the adjustable signal received at the destination is given as

$$y = \sqrt{P}\left[\sum_{k=1}^{N} h_k e^{j\varphi_k} g_k\right] x^{IQ} + n, \tag{4}$$

where φ_k characterizes the adapted phase for the k^{th} RIS reflector, P is the average of the transmitted power symbol, and n is a complex additive white Gaussian noise (AWGN) with zero mean and N_0 variance. Using (3), the received signal becomes

$$y = \sqrt{P}\left[\sum_{k=1}^{N} h_k e^{j\varphi_k} g_k\right] [G_1 x + G_2 x^*] + n. \tag{5}$$

Taking into account the Rx IQI and the adjusted phases, the resulting signal at the destination can be expressed as

$$\begin{aligned} y^{IQ} &= K_1 y + K_2 y^* \\ &= \underbrace{\sqrt{P}\left[K_1 G_1 \sum\nolimits_{k=1}^{N} h_k e^{j\varphi_k} g_k + K_2 G_2^* \sum\nolimits_{k=1}^{N} h_k^* e^{-j\varphi_k} g_k^*\right] x}_{S} \\ &\quad + \underbrace{\sqrt{P}\left[K_1 G_2 \sum\nolimits_{k=1}^{N} h_k e^{j\varphi_k} g_k + K_2 G_1^* \sum\nolimits_{k=1}^{N} h_k^* e^{-j\varphi_k} g_k^*\right] x^*}_{I} \\ &\quad + K_1 n + K_2 n^* = \left(\sqrt{P} S x + \sqrt{P} I x^*\right) + \eta = \sqrt{P}\chi + \eta, \end{aligned} \tag{6}$$

where $\chi = Sx + Ix^*$ and $\eta = K_1 n + K_2 n^*$.

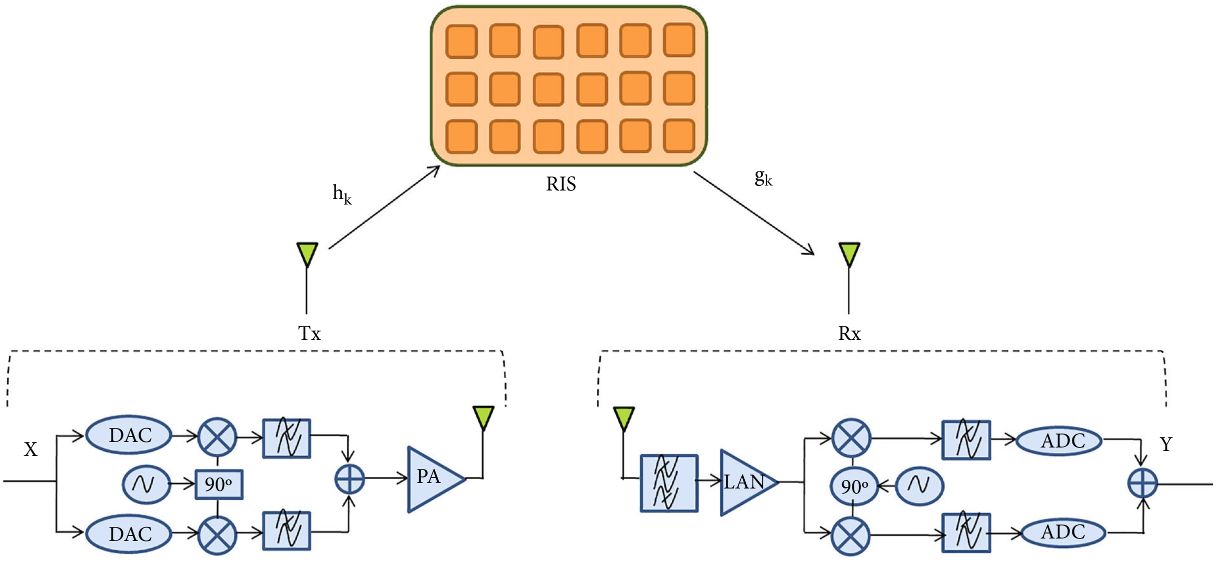


Figure 1: RIS under IQI system model.

Note that due to Tx and Rx IQI, the baseband signal x is interfered by its complex conjugate x^*. Hence, Ix^* represents the self-interferences.

Analogous to [4], with the assistance of a software communication, the following derived expressions are analyzed based on the knowledge of the channel at the RIS in function of amplitudes and phases as $h_k = \varepsilon_k e^{-j\theta_k}$, $g_k = \beta_k e^{-j\psi_k}$. In such case, the RIS adjusts the phases in order to maximize the signal-to-noise ratio (SNR) such as $\varphi_k = \theta_k + \psi_k$.

Moreover, the noise can be expressed as $\eta = n^I + j(K_c n^I + K_d n^Q)$ where n^I and n^Q note, respectively, the real and imaginary parts of n, $K_c = K_1^Q + K_2^Q$, and $K_d = K_1^I - K_2^I$ [23].

Accordingly, stating that $\eta = \eta^I + j\eta^Q$ where η^I and η^Q note, respectively, the real and imaginary parts of η, it is worth noting that η is an improper Gaussian noise with unequal real and imaginary parts variances such as

$$\sigma_{\eta^I}^2 = \frac{\sigma_n^2}{2} = \frac{N_0}{2}, \tag{7}$$

$$\sigma_{\eta^Q}^2 = \left(K_c^2 + K_d^2\right)\frac{\sigma_n^2}{2} = \frac{N_0}{2}\xi_r^2. \tag{8}$$

Further, the correlation factor between the noise components is $\rho = \operatorname{cov}\left(\eta^I, \eta^Q\right)/\sigma_{\eta^I}\sigma_{\eta^Q} = K_c(\sigma_n^2/2)/\xi_r\sigma_n^2/2 = -\sin\left(\varphi_r\right)$ [25]. It is important to emphasize the IQI effects in changing the behavior of the noise from proper to improper. Thus, the received signal has correlated IQ components. Considering the presence of improper noise is a critical factor in analyzing the RIS performance.

The received signal expression in (6) can be analyzed under several scenarios taking into account the possible values of IQ parameters:

(1) *Perfect IQ Matching*. In this scenario, Tx and Rx sides are assumed to be perfect and IQI parameters are defined as $\boldsymbol{G}_1 = \boldsymbol{K}_1 = \mathbf{1}, \boldsymbol{G}_2 = \boldsymbol{K}_2 = \mathbf{0}$. The received signal in (6) can be expressed as the following well known expressed used for all previous studies:

$$y = \sqrt{P}\left[\sum_{k=1}^{N} \varepsilon_k \beta_k\right] x + n. \tag{9}$$

(2) *Tx Impaired by IQI*. Taking into account the destructive effects of only the Tx side with perfect IQI parameters on the destination, i.e., $\boldsymbol{K}_1 = \mathbf{1}$ and $\boldsymbol{K}_2 = \mathbf{0}$, the resulting signal is given as

$$y_{Tx}^{IQ} = \sqrt{P}\left[\sum_{k=1}^{N} \varepsilon_k \beta_k\right]\left[G_1 x + G_2 x^*\right] + n. \tag{10}$$

(3) *Rx Impaired by IQI*. Considering ideal Tx IQI parameters and stating that $\boldsymbol{G1} = \mathbf{1}, \boldsymbol{G2} = \mathbf{0}$, (6) can be rewritten as

$$y_{Rx}^{IQ} = \sqrt{P}\left[\sum_{k=1}^{N} \varepsilon_k \beta_k\right]\left[K_1 x + K_2 x^*\right] + K_1 n + K_2 n^*. \tag{11}$$

It can be observed from the above expressions that Tx impaired with IQI causes self-interference from the conjugate of the transmitted signal. On the other side, Rx suffering from IQI affects both the signal and the noise.

As previously mentioned, with the presence of IQI, a new derived received signal expression is required which shows the presence of self-interference and new noise behavior.

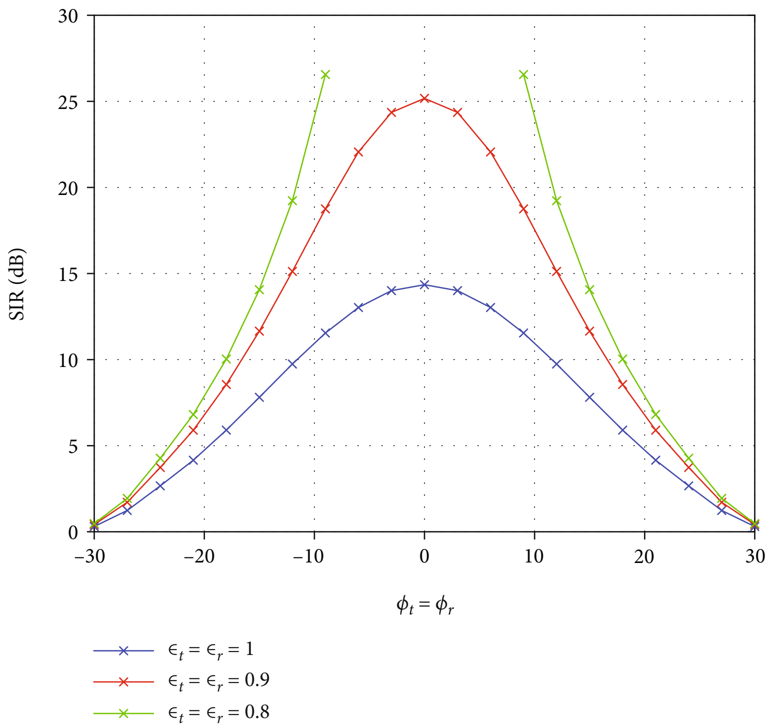


FIGURE 2: SIR values variation of RIS under IQI in function of phases with fixed gains $\varepsilon_t = \varepsilon_r = [0.8, 0.9, 1]$.

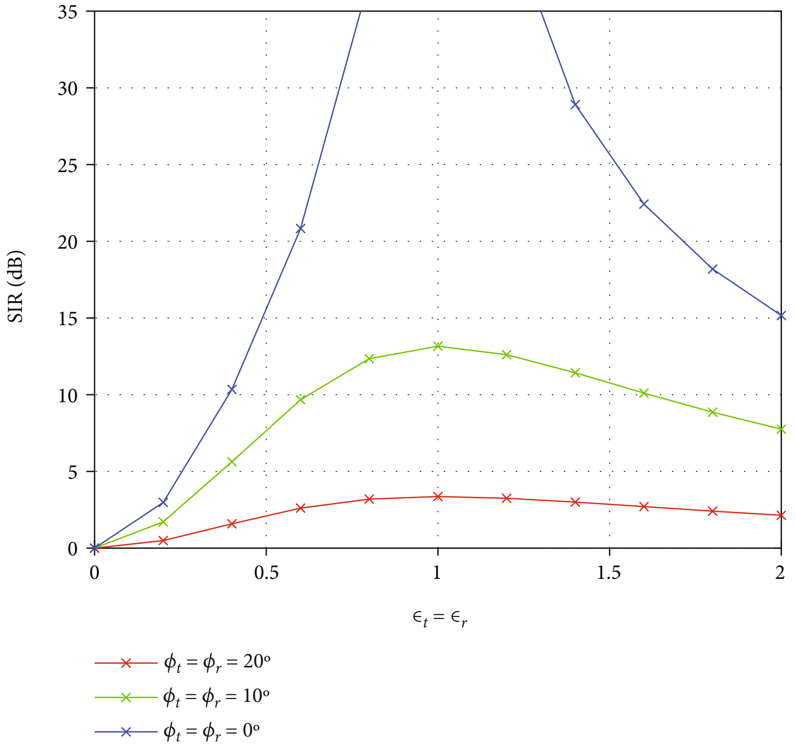


FIGURE 3: SIR values variation of RIS under IQI in function of gains with fixed phases $\phi_t = \phi_r = [0^\circ, 10^\circ, 20^\circ]$.

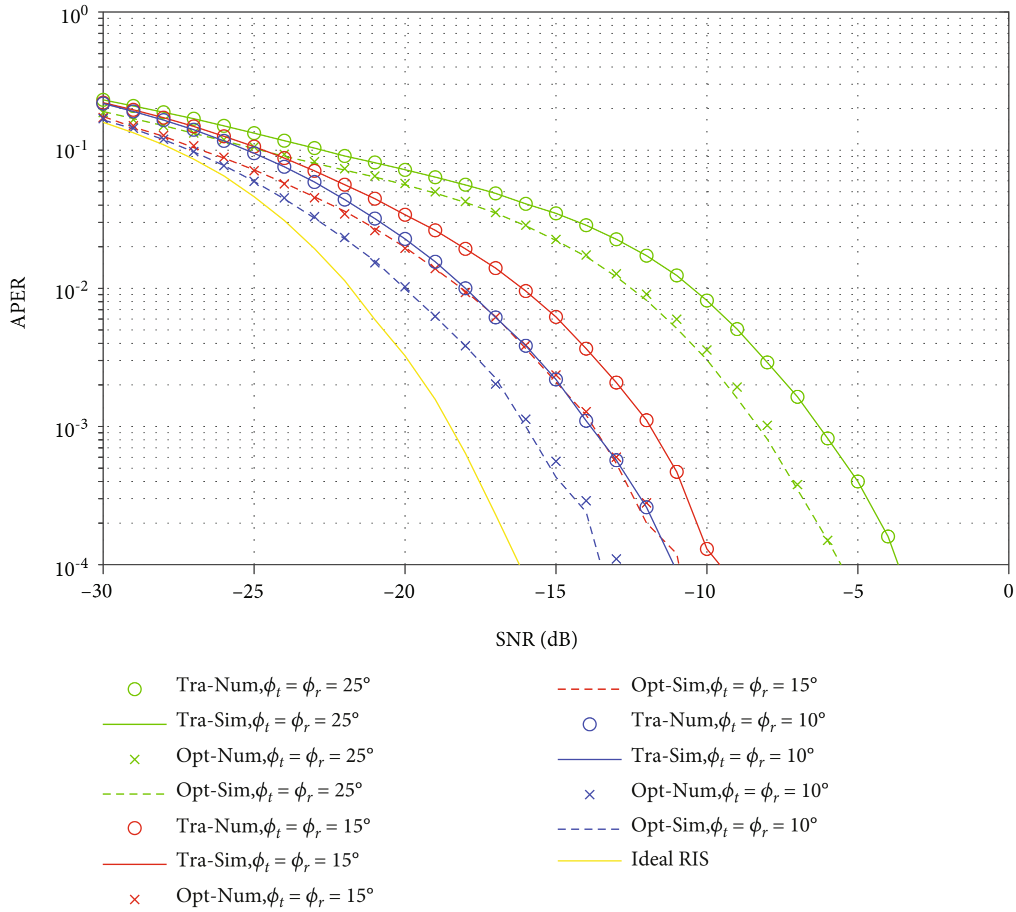


FIGURE 4: RIS APEP of optimal and traditional ML detector under IQI in function of phases with fixed gains $\varepsilon_t = \varepsilon_r = 3$dB using $N = 32$ and 4-QAM.

Hence, for an effective performance study of RIS-based communication systems, a novel design of optimal ML detector that incorporates the IQI effects is obligatory.

2.2. Signal-to-Interference Ratio (SIR). The average SIR is calculated as follows to illustrate the harmful effects of the caused self-interference

$$\mathrm{SIR} = \frac{\mathbb{E}\left\{|S|^2\right\}}{\mathbb{E}\left\{|I|^2\right\}}. \tag{12}$$

Note that the image rejection ratio IRR which denotes the measure of image frequency band attenuation can be defined for Tx and Rx sides, respectively, as $\mathrm{IRR}_{Tx} = |G_1|^2/|G_2|^2$ and $\mathrm{IRR}_{Rx} = |K_1|^2/|K_2|^2$ [31]. IRR has a typical value in the range of 20-40 dB for practical analog RF front-end electronics [32]. Based on (2), the following relation can be defined $K_1 = 1 - K_2^*$. It can be depicted that for high IRR_{Rx} values $K_1 \longrightarrow 1$ and, $K_1 K_2^* = K_1 - K_1^2 \longrightarrow 0$ [33]. Supposing $\gamma_k = \varepsilon_k \beta_k$, an approximate widely used in the literature [31–34] is obtained by assuming

$$\begin{aligned}
&\mathbb{E}\left\{\left[\sum_{k=1}^{N}\gamma_k\right]^2\right\}\left(|K_1|^2|G_1|^2+|K_2|^2|G_2|^2\right)\\
&\quad > > \mathbb{E}\left\{2\Re\left(K_1G_1K_2^*G_2\left[\sum_{k=1}^{N}\gamma_k\right]^2\right)\right\}\mathbb{E}\left\{\left(\sum_{k=1}^{N}\gamma_k\right)^2\right\}\\
&\qquad \times\left(|K_1|^2|G_2|^2+|K_2|^2|G_1|^2\right)\\
&\quad > > \mathbb{E}\left\{2\Re\left(K_1G_2K_2^*G_1\left[\sum_{k=1}^{N}\gamma_k\right]^2\right\}.
\end{aligned} \tag{13}$$

Hence, the averaged SIR can be tightly approximated as

$$\mathrm{SIR} \simeq \frac{|K_1|^2|G_1|^2+|K_2|^2|G_2|^2}{|K_1|^2|G_2|^2+|K_2|^2|G_1|^2}. \tag{14}$$

It is obviously seen that even small values of IQI degrade the system performance. Further, the SIR expression is independent of N. Hence, increasing the reflector number cannot mitigate the interference caused by the IQI. In the ideal case of perfect IQ matching $\mathrm{SIR} = \infty$.

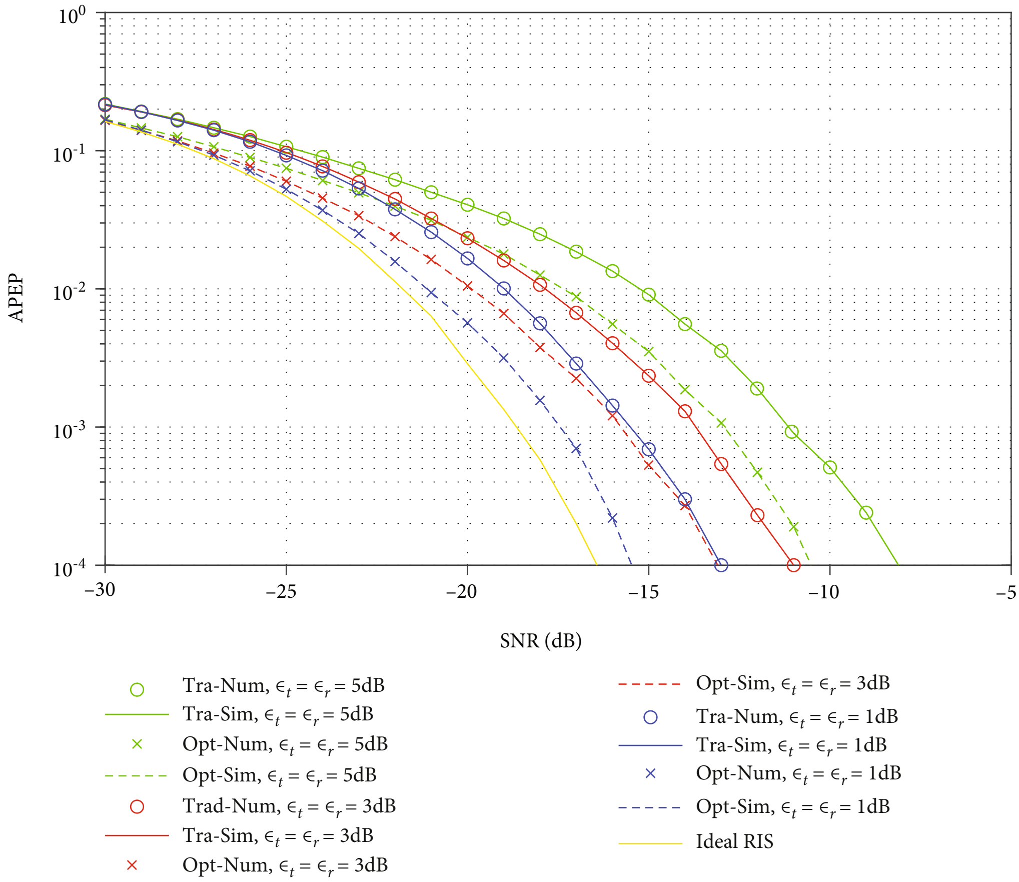


FIGURE 5: RIS APEP of optimal and traditional ML detector under IQI in function of gains with fixed phases $\phi_t = \phi_r = 10^\circ$ using $N = 32$ and 4-QAM.

2.3. Signal Detection

2.3.1. Optimal ML Detector. In the attempt to cover the presence of an improper Gaussian noise, an optimal ML detector is designed based on the received signal expression in (6). Hence, the bivariate Gaussian random variable (RV) distribution of the real, y^I_{IQ}, and imaginary, y^Q_{IQ}, correlated components of the received signal vector y is described as follows:

$$f_{y^I_{IQ},y^Q_{IQ}}\left(y^I_{IQ}, y^Q_{IQ} \mid x\right) = \left(\frac{1}{2\pi\sigma_{\eta^I}\sigma_{\eta^Q}\sqrt{1-\rho^2}}\right) \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{\left\|y^I_{IQ} - \sqrt{P}\chi^I\right\|^2}{\sigma^2_{\eta^I}} + \frac{\left\|y^Q_{IQ} - \sqrt{P}\chi^Q\right\|^2}{\sigma^2_{\eta^Q}} - \frac{2\rho\left(y^I_{IQ} - \sqrt{P}\chi^I\right)\left(y^Q_{IQ} - \sqrt{P}\chi^Q\right)}{\sigma_{\eta^I}\sigma_{\eta^Q}}\right]\right), \quad (15)$$

where χ^I and χ^Q represent, respectively, $\Re\{\chi\}$ and $\Im\{\chi\}$ components.

Regarding that the transmitted symbols are assumed equally distributed, the optimal ML detector is given by maximizing the argument of the conditional joint probability density function (PDF) defined in (15) which is equivalent to the following expression:

$$\{\hat{x}\}_{ML_{op}} = \arg\min_x \left\{\frac{\left\|y^I_{IQ} - \sqrt{P}\chi^I\right\|^2}{\sigma^2_{\eta^I}} + \frac{\left\|y^Q_{IQ} - \sqrt{P}\chi^Q\right\|^2}{\sigma^2_{\eta^Q}} - \frac{2\rho\left(y^I_{IQ} - \sqrt{P}\chi^I\right)\left(y^Q_{IQ} - \sqrt{P}\chi^Q\right)}{\sigma_{\eta^I}\sigma_{\eta^Q}}\right\}. \quad (16)$$

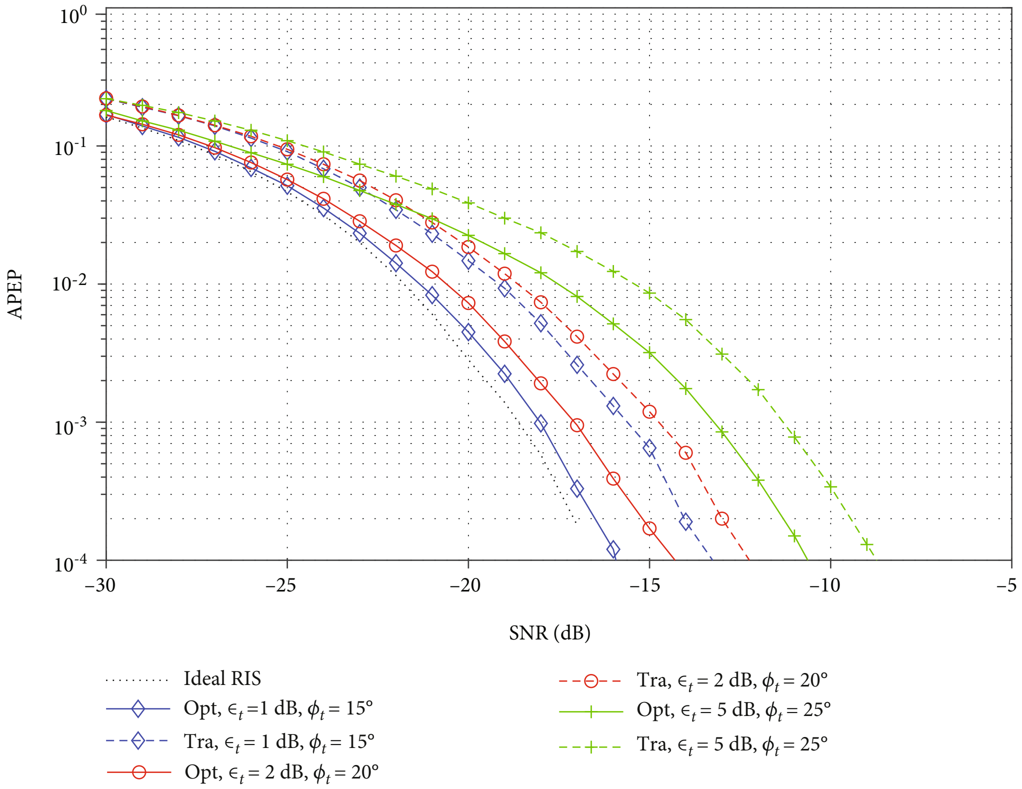


Figure 6: RIS APEP of optimal and traditional ML detector in the presence of Tx IQI with $N = 32$ using 4-QAM.

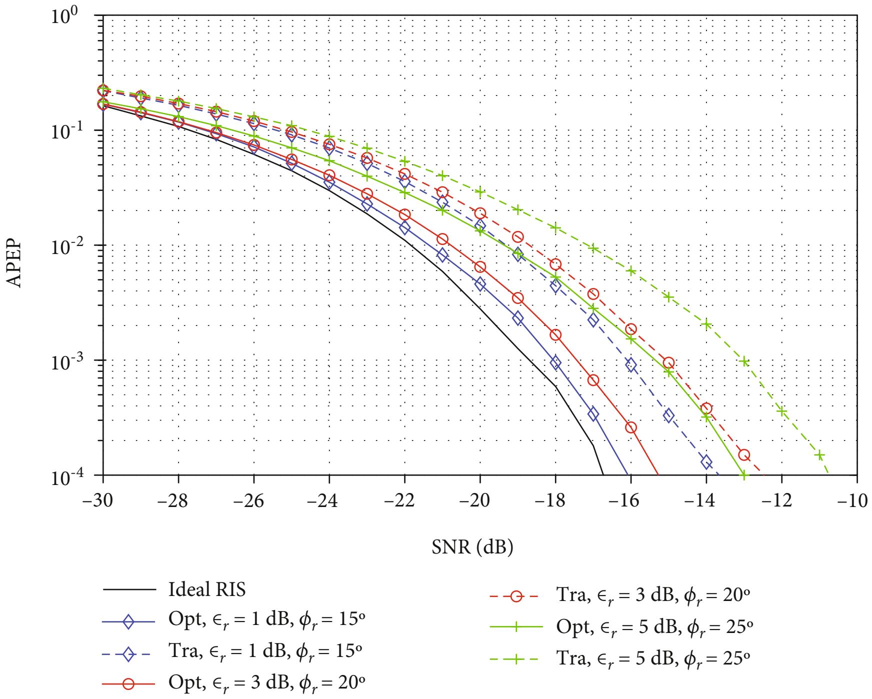


Figure 7: RIS APEP of optimal and traditional ML detector in the presence of Rx IQI with $N = 32$ using 4-QAM.

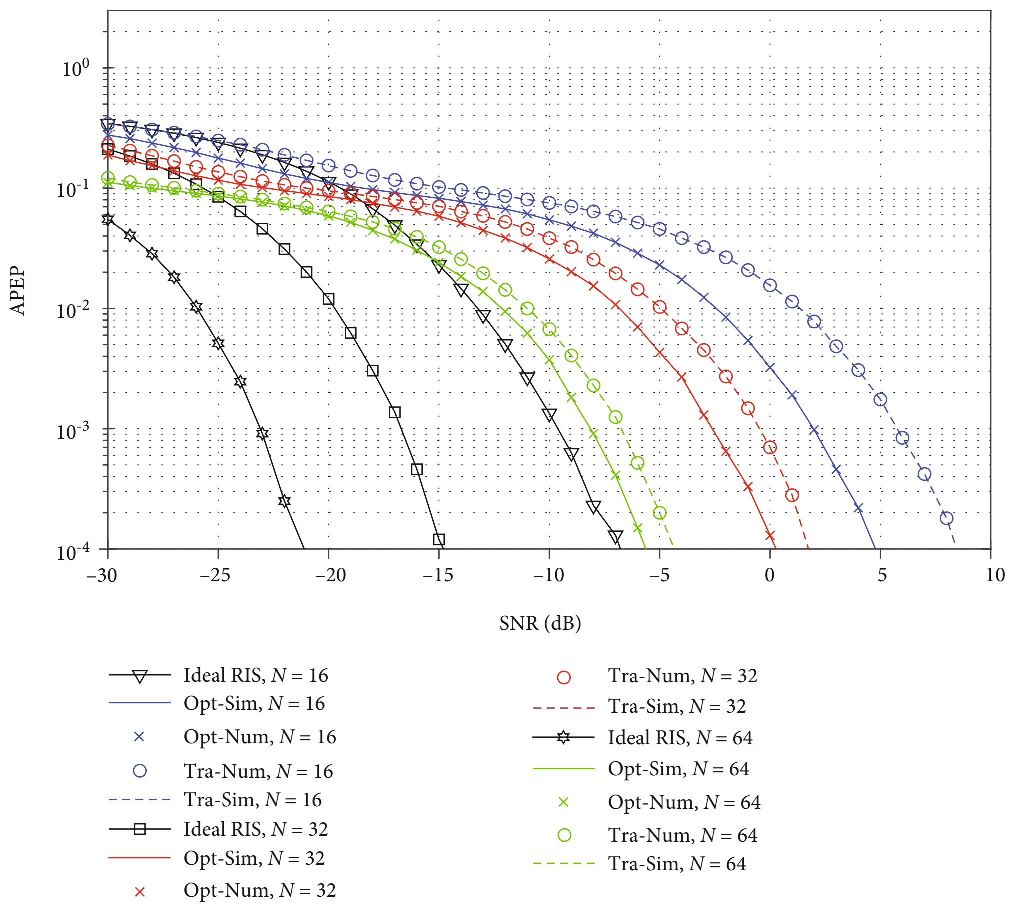


FIGURE 8: RIS APEP of optimal and traditional ML detector in the presence of joint Tx and Rx IQI with $\varepsilon_t = \varepsilon_r = 5\text{dB}$, $\phi_t = \phi_r = 20°$, and $N = [16,32,64]$ using 4-QAM.

2.3.2. Traditional ML Detector. The traditional ML detector is the classical ML detector used in the previous performance analysis of RIS in which the presence of IQI in the received signal expression is neglected. Indeed, it is simply expressed as the following well-known expression used for all previous studies of RIS performance:

$$\{\widehat{x}\}_{\mathrm{ML}_{tra}} = \arg\ \min_{x}\left\{\left\| y^{IQ} - \sqrt{P}\chi \right\|^2\right\}. \tag{17}$$

3. Performance Analysis

3.1. Conditional Error Probability

3.1.1. Optimal ML Detector. Assuming χ is transmitted, the pairwise error probability (PEP) of deciding in favor of $\tilde{\chi}$ is given from the optimal ML detector expression in (16) as

$$\begin{aligned}\mathrm{PEP}_{\mathrm{opt}} &= \mathrm{P}_r\left\{\frac{\left\|y_{IQ}^{I} - \sqrt{P}\chi^{I}\right\|^2}{\sigma_{\eta^I}^2} + \frac{\left\|y_{IQ}^{Q} - \sqrt{P}\chi^{Q}\right\|^2}{\sigma_{\eta^Q}^2} - \frac{2\rho\left(y_{IQ}^{I} - \sqrt{P}\chi^{I}\right)\left(y_{IQ}^{Q} - \sqrt{P}\chi^{Q}\right)}{\sigma_{\eta^I}\sigma_{\eta^Q}} > \frac{\left\|y_{IQ}^{I} - \sqrt{P}\tilde{\chi}^{I}\right\|^2}{\sigma_{\eta^I}^2} + \frac{\left\|y_{IQ}^{Q} - \sqrt{P}\tilde{\chi}^{Q}\right\|^2}{\sigma_{\eta^I}^2} - \frac{2\rho\left(y_{IQ}^{I} - \sqrt{P}\tilde{\chi}^{I}\right)\left(y_{IQ}^{Q} - \sqrt{P}\tilde{\chi}^{Q}\right)}{\sigma_{\eta^I}\sigma_{\eta^Q}}\right\}\\ &= \mathrm{Pr}\ \{\beta > 0\},\end{aligned} \tag{18}$$

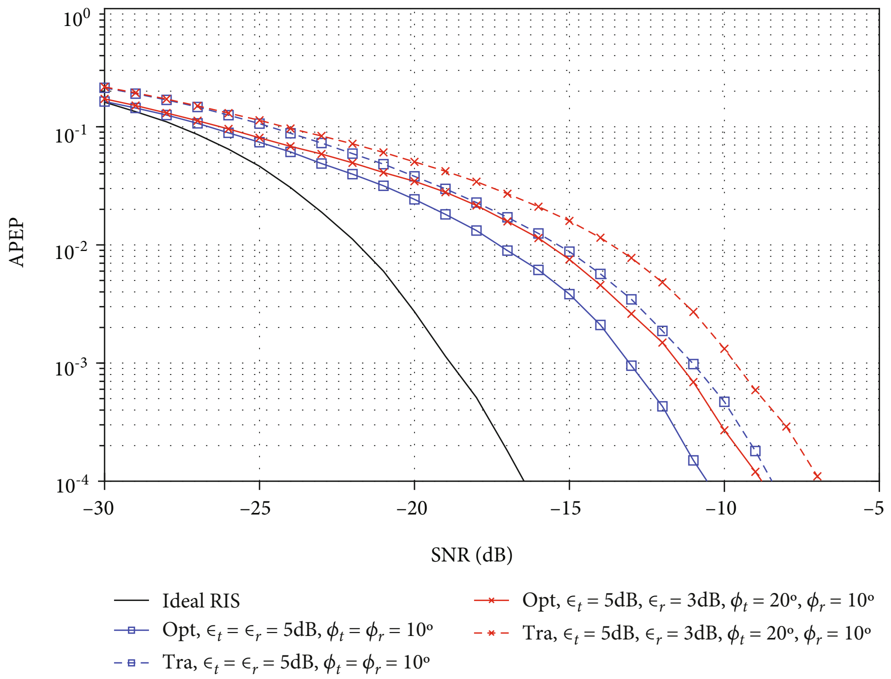


FIGURE 9: RIS APEP of optimal and traditional ML detector in the presence of joint Tx and Rx IQI with $\varepsilon_t = \varepsilon_r = 5$dB, $\phi_t = \phi_r = 10^\circ$ and $\varepsilon_t = 5$dB, $\varepsilon_r = 3$dB, $\phi_t = 20^\circ$, $\phi_r = 10^\circ$, and $N = 32$ using 4-QAM.

where $\tilde{\chi} = S\tilde{x} + I\tilde{x}^*$ and β is obtained after simple mathematical operations as

$$\beta = \frac{2\rho\sqrt{P}\left\{\left(\chi^I - \tilde{\chi}^I\right)\eta^Q + \left(\chi^Q - \tilde{\chi}^Q\right)\eta^I\right\}}{\sigma_{\eta^I}\sigma_{\eta^Q}} - \frac{2\sqrt{P}\left(\chi^I - \tilde{\chi}^I\right)\eta^I}{\sigma_{\eta^I}^2} - \frac{2\sqrt{P}\left(\chi^Q - \tilde{\chi}^Q\right)\eta^Q}{\sigma_{\eta^Q}^2} - \frac{E\left(\chi^I - \tilde{\chi}^I\right)^2}{\sigma_{\eta^I}^2} - \frac{E\left(\chi^Q - \tilde{\chi}^Q\right)^2}{\sigma_{\eta^Q}^2} + \frac{2\rho E\left\{\left(\chi^I - \tilde{\chi}^I\right)\left(\chi^Q - \tilde{\chi}^Q\right)\right\}}{\sigma_{\eta^I}\sigma_{\eta^Q}}. \tag{19}$$

Without loss of generality β, conditioned on χ, is a Gaussian RV with the following mean and variance values

$$\mu_\beta = \frac{E\left\|\chi^I - \tilde{\chi}^I\right\|^2}{\sigma_{\eta^I}^2} + \frac{E\left\|\chi^Q - \tilde{\chi}^Q\right\|^2}{\sigma_{\eta^Q}^2} - \frac{2\rho E\left(\chi^I - \tilde{\chi}^I\right)\left(\chi^Q - \tilde{\chi}^Q\right)}{\sigma_{\eta^I}\sigma_{\eta^Q}}, \tag{20}$$

$$\sigma_\beta^2 = 4E\left(1 - \rho^2\right)\left\{\frac{E\left\|\chi^I - \tilde{\chi}^I\right\|^2}{\sigma_{\eta^I}^2} + \frac{E\left\|\chi^Q - \tilde{\chi}^Q\right\|^2}{\sigma_{\eta^Q}^2} - \frac{2\rho\left(\chi^I - \tilde{\chi}^I\right)\left(\chi^Q - \tilde{\chi}^Q\right)}{\sigma_{\eta^I}\sigma_{\eta^Q}}\right\}. \tag{21}$$

Using (18) and (20), the conditional PEP (CPEP) can be written as in (22) on the top of the page, where $Q(x)$ denotes the Q-function defined as $Q(x) = 1/2\pi\int_x^\infty \exp\left(-u^2/2\right)du$.

$$\mathrm{CPEP}_{\mathrm{opt}} = Q\left(\sqrt{\frac{P}{4(1-\rho^2)}\left[\frac{\left\|\chi^I - \tilde{\chi}^I\right\|^2}{\sigma_{\eta^I}^2} + \frac{\left\|\chi^Q - \tilde{\chi}^Q\right\|^2}{\sigma_{\eta^Q}^2} - \frac{2\rho\left(\chi^I - \tilde{\chi}^I\right)\left(\chi^Q - \tilde{\chi}^Q\right)}{\sigma_{\eta^I}\sigma_{\eta^Q}}\right]}\right) = Q\left(\sqrt{\frac{PY}{2(1-\rho^2)N0}}\right), \tag{22}$$

3.1.2. Traditional ML Detector. Note that the improper behavior of the noise is not considered for the traditional ML detector. Thus, assuming x is transmitted, the PEP of deciding in favor of $\tilde{x}$ is given from the classical ML detector expression in (17) as

$$\begin{aligned}\mathrm{PEP}_{\mathrm{tra}} &= \mathrm{P}_r\left\{\left\|y^{IQ} - \sqrt{P}\chi\right\|^2 > \left\|y^{IQ} - \sqrt{P}\tilde{\chi}\right\|^2\right\} \\ &= \Pr\left\{D > E\left\|\chi - \tilde{\chi}\right\|^2\right\},\end{aligned} \tag{23}$$

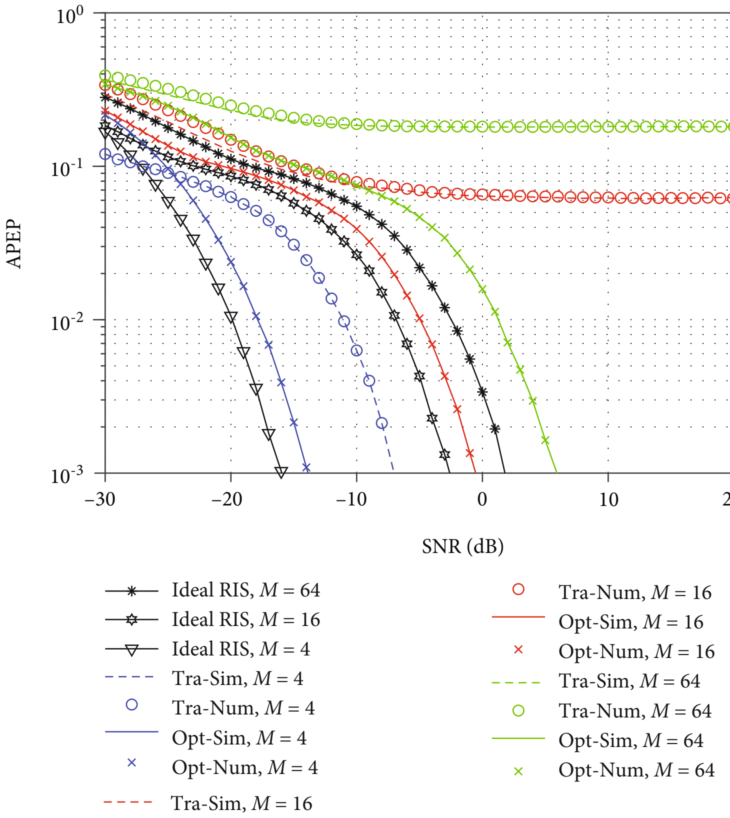


Figure 10: RIS APEP of optimal and traditional ML detector in the presence of joint Tx and Rx IQI with $\varepsilon_t = \varepsilon_r = 3$dB, $\phi_t = \phi_r = 10°$, and $N = 32$ using different modulation order.

where $D = -2R\{\sqrt{P}(\chi - \tilde{\chi})\eta^*\}$ which has zero mean and $\sigma_D^2 = 4P(\chi^I - \tilde{\chi}^I)^2\sigma_{\eta^I}^2 + 4P(\chi^Q - \tilde{\chi}^Q)^2\sigma_{\eta^Q}^2 + 8\rho\sigma_{\eta^I}\sigma_{\eta^Q}P(\chi^I - \tilde{\chi}^I)(\chi^Q - \tilde{\chi}^Q)$ as variance. Hence, the CPEP using Q function is presented in (24).

$$\mathrm{CPEP}_{\mathrm{tra}} = Q\left(\sqrt{\frac{P(\chi^I - \tilde{\chi}^I)^4}{4(\chi^I - \tilde{\chi}^I)^2\sigma_{\eta^I}^2 + 4(\chi^Q - \tilde{\chi}^Q)^2\sigma_{\eta^Q}^2 + 8\rho\sigma_{\eta^I}\sigma_{\eta^Q}(\chi^I - \tilde{\chi}^I)(\chi^Q - \tilde{\chi}^Q)}}\right). \tag{24}$$

3.2. Average Error Probability

3.2.1. Optimal ML Detector. To derive the average PEP (APEP), (22) should be averaged over the PDF of Y which can be depicted as:

$$\mathrm{APEP}_{\mathrm{opt}} = \int_0^\infty Q\left(\sqrt{\frac{PY}{4(1-\rho^2)}}\right) f_Y(Y)dY. \tag{25}$$

However, note that $Y = \Psi_1^2 + \Psi_2^2 - 2\rho\Psi_1\Psi_2$ where $\Psi_1 = (\chi^I - \tilde{\chi}^I)$ and $\Psi_2 = (\chi^Q - \tilde{\chi}^Q)/\xi_r$. Indeed, Ψ_1 and Ψ_2 depend on real and imaginary parts of S and I given in (6). Thus, based on central limit theorem (CLT) and for large number of reflectors N, Y is considered as a combination of correlated noncentral chi-squared random variables. The determination of the APEP using the Y PDF is very hard. Hence, the APEP is numerically computed by averaging PEP over large number of channel realization.

3.2.2. Traditional ML Detector. Traditional ML detector APEP based on (24) is difficult and even impossible to determinate. However, it will be numerically calculated with the same method used for the optimal ML detector.

4. Numerical Results and Discussions

In this part, the APEP performance of the optimal and traditional ML detector for RIS under IQI based on the proposed analytic scenarios will be presented and proved with simulation results.

First, the determined expression of SIR is evaluated in Figures 2 and 3 for different values, respectively, of gains ε_t,

ε_r and phases ϕ_t, ϕ_r of IQI and IRR = 20 dB. It can be depicted that a considerable SIR value degradation occurs even for small imbalance values which highlight the importance of taking into account the IQI in RIS performance analysis.

Figures 4 and 5 are presented to show the effects of, respectively, amplitudes and phases of Tx and Rx IQI on APEP performance. Indeed, in Figure 4, the APEP is analyzed for fixed amplitudes values $\varepsilon_t = \varepsilon_r = 3\text{dB}$ and for variable values of phase imbalance $\phi_t = \phi_r = [10^\circ, 15^\circ, 25^\circ]$. Accordingly, the effects of increasing the amplitude imbalance from 1dB to 5dB with fixed phases values $\phi_t = \phi_r = 25^\circ$ is given in Figure 5. It can be depicted that increasing the phases and amplitudes of IQI cause a considerable performance loss. Further, the analytical analysis matches perfectly the simulation results for both ML detectors, which validates the proposed analysis. Additionally, it is observed that the proposed ML detector outperforms the classical ML detector for all IQI values and a considerable gain is achieved by using the optimal detector.

Figure 6 shows APEP simulation results of standard ML and optimal ML under only Tx imbalance for different values of ε_t, ϕ_t with $N = 32$. In such case, $\varepsilon_r = \phi_r = 0^\circ$. It is worth mentioning that the optimal design improves the system performance and decreases the IQI effects. The destructive effects of Tx imbalance degrade the performance of perfect RIS. Actually, comparing the ideal RIS with the optimal ML detector performance, the system degrades about 6 dB when $\varepsilon_t = 5\text{dB}$ and $\phi_t = 25^\circ$, while the degradation is about 1 dB when $\varepsilon_t = 2\text{dB}$ and $\phi_t = 15^\circ$ at APEP=10^{-2} .

In Figure 7, the impact of Rx IQI on RIS with perfect Tx IQI parameters ($\varepsilon_t = 0, \phi_t = 0^\circ$) is carried for different ε_r, ϕ_r and fixed $N = 32$. For comparison reasons, the case of ideal IQ is also presented. It can be seen the harmful effects of increasing the IQI values on the system performance. Indeed, a 2 dB performance degradation is shown when $\varepsilon_r = 5\text{dB}$ and $\phi_r = 25^\circ$ for optimal ML compared to the ideal case at APEP=10^{-2}. It can be noticed that there is a performance gain of the designed ML detector compared with the traditional one.

In order to put in evidence the impact of RIS reflector number, Figure 8 presents the numerical and simulation results of optimal and traditional ML detectors under joint Tx and Rx IQI when the number of reflectors changes among 16, 32, and 64. This figure shows the accuracy of the presented analysis for various reflectors numbers. A remarkable performance gain is achieved by using the optimal ML detector. However, even if we increase the number of reflectors, IQI will harm the system performance. For instance, Figure 8 illustrates that if the system has 16 reflectors, the optimal performance degrades by 12 dB when $\varepsilon_t = \varepsilon_r = 5\text{dB}$ and $\phi_t = \phi_r = 20^\circ$ compared with the perfect IQ matching system at 10^{-3} APEP. This degradation has less impact when increasing the number of reflectors. Hence, it can be concluded that increasing the SNR or the number of reflectors enhances the system performance.

Figure 9 shows the effects of changing IQI parameters at joint Tx and Rx on optimal and standard ML detector performances. The results are carried for fixed IQI parameters at the Tx and Rx $\varepsilon_t = \varepsilon_r = 5\text{dB}$, $\phi_t = \phi_r = 10^\circ$ and for different values such as $\varepsilon_t = 5\text{dB}, \varepsilon_r = 3\text{dB}$ and $\phi_t = 20^\circ, \phi_r = 10^\circ$ with $N = 32$. It can be depicted that the optimal ML detector under equal Tx and Rx IQI parameters outperforms the optimal ML under different values due to the additional mismatch.

Figure 10 validates the derived expressions of optimal and traditional ML detectors for a variable number of modulation order with fixed RIS-IQI parameters $\varepsilon_t = \varepsilon_r = 3\text{dB}$, $\phi_t = \phi_r = 10^\circ$, and $N = 32$. It can be shown that RIS using conventional ML detector with increased modulation order suffers from an error floor in the high SNR region due to the fact of ignoring the noise behavior. Consequently, in the presence of IQI, RIS with a traditional ML detector could not support high order modulation which is not acceptable in practice. A careful study for high data rate RIS-aided communication systems in the presence of IQI is required. However, the designed optimal ML detector prevents the error floor and decreases the IQI impairments.

5. Conclusion

This paper investigates the harmful effects of IQI on RIS-aided communication. Hence, novel analytical expressions of PEP and the APEP for optimal and traditional ML detectors are derived. Consequently, the behavior of the designed detectors is analyzed in function of Tx and Rx IQ parameters to further characterize the effects of IQI. Finally, numerical results are illustrated to prove the efficiency of the proposed analysis. The carried results highlight RIS performance degradation's with the presence of IQI even in high SNR regions and an increased number of reflectors. It was proved for the different carried cases that optimal ML detector performs far better than conventional detection method. However, it is worth noting that an optimal detector presents a significant computational complexity. Hence, studying IQI compensation algorithms with low complexity for RIS systems could be potential future works. Moreover, we intend to analyze the performance of RIS under IQI with the presence and absence of direct links between the source and the destination.

Data Availability

The data that support the findings of this study are available from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

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

[2] K. David and H. Berndt, "6G vision and requirements: is there any need for beyond 5G?," *IEEE Vehicular Technology Magazine*, vol. 13, no. 3, pp. 72–80, 2018.

[3] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy

efficiency in wireless communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, 2019.

[4] E. Basar, M. di Renzo, J. de Rosny, M. Debbah, M. S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[5] M. Di Renzo, M. Debbah, D. T. Phan-Huy et al., "Smart radio environments empowered by reconfigurable AI meta-surfaces: an idea whose time has come," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, Article ID 129, 2019.

[6] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[7] S. Hu, F. Rusek, and O. Edfors, "Beyond massive MIMO: the potential of data transmission with large intelligent surfaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2746–2758, 2018.

[8] E. Bjornson, O. Ozdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: how large surfaces are needed to beat relaying?," *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 244–248, 2019.

[9] A. N. Parks, A. Liu, S. Gollakota, and J. R. Smith, "Turbocharging ambient backscatter communication," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 619–630, 2015.

[10] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3029–3056, 2015.

[11] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, "Are we approaching the fundamental limits of wireless network densification?," *IEEE Communications Magazine*, vol. 54, no. 10, pp. 184–190, 2016.

[12] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software controlled metasurfaces," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 162–169, 2018.

[13] A. C. Tasolamprou, A. Pitilakis, S. Abadal et al., "Exploration of intercell wireless millimeter-wave communication in the landscape of intelligent metasurfaces," *IEEE Access*, vol. 7, pp. 122931–122948, 2019.

[14] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: principles, challenges, and opportunities," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 990–1002, 2020.

[15] L. Dai, B. Wang, M. Wang et al., "Reconfigurable intelligent surface-based wireless communications: antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45913–45923, 2020.

[16] X. Yuan, Y. J. Zhang, Y. Shi, W. Yan, and H. Liu, "Reconfigurable-intelligent-surface empowered 6G wireless communications: challenges and opportunities," 2020, arXiv preprint.

[17] J. V. Alegría and F. Rusek, "Achievable rate with correlated hardware impairments in large intelligent surfaces," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 559–563, Le gosier, Guadeloupe, December 2019.

[18] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, "Reliability analysis of large intelligent surfaces (LISs): rate distribution and outage probability," *IEEE Wireless Communications Letters*, vol. 8, no. 6, pp. 1662–1666, 2019.

[19] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, "Performance analysis of large intelligence surfaces (LISs): asymptotic data rate and channel hardening effects," 2018, https://arxiv.org/abs/1810.05667.

[20] S. Gong, X. Lu, D. T. Hoang et al., "Toward smart wireless communications via intelligent reflecting surfaces: a contemporary survey," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 4, pp. 2283–2314.

[21] T. Schenk, *RF Imperfections in High-Rate Wireless Systems: Impact and Digital Compensation*, Springer Publishing Company, Incorporated, 1st edition, 2008.

[22] J. Li, M. Matthaiou, and T. Svensson, "I/Q imbalance in AF dual-hop relaying: performance analysis in Nakagami-m fading," *IEEE Transactions on Communications*, vol. 62, no. 3, pp. 836–847, 2014.

[23] A. E. Canbilen, M. M. Alsmadi, E. Basar, S. S. Ikki, S. S. Gultekin, and I. Develi, "Spatial modulation in the presence of I/Q imbalance: optimal detector and performance analysis," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1572–1575, 2018.

[24] A. E. Canbilen, S. S. Ikki, E. Basar, S. S. Gultekin, and I. Develi, "Impact of I/Q imbalance on amplify-and-forward relaying: optimal detector design and error performance," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3154–3166, 2019.

[25] M. M. Alsmadi, A. E. Canbilen, N. Abu Ali, S. S. Ikki, and E. Basar, "Cognitive networks in the presence of I/Q imbalance and imperfect CSI: receiver design and performance analysis," *IEEE Access*, vol. 7, pp. 49765–49777, 2019.

[26] S. Li, L. Yang, D. B. . Costa, M. D. Renzo, and M. S. Alouini, "On the performance of RIS-assisted dual-hop mixed RF-UWOC systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 2, pp. 340–353, 2021.

[27] F. A. P. de Figueiredo, M. S. P. Facina, R. C. Ferreira et al., "Large intelligent surfaces with discrete set of phase-shifts communicating through double-Rayleigh fading channels," *IEEE Access*, vol. 9, pp. 20768–20787, 2021.

[28] I. Trigui, E. K. Agbogla, M. Benjillali, W. Ajib, and W. P. Zhu, "Bit error rate analysis for reconfigurable intelligent surfaces with phase errors," *IEEE Communications Letters*, vol. 25, no. 7, pp. 2176–2180, 2021.

[29] Y. Zou, M. Valkama, and M. Renfors, "Digital compensation of I/Q imbalance effects in space-time coded transmit diversity systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2496–2508, 2008.

[30] T. Schenk, *IQ imbalance. RF Imperfections in High-rate Wireless Systems: Impact and Digital Compensation*Springer, Dordrecht.

[31] B. Selim, S. Muhaidat, P. C. Sofotasios et al., "Performance analysis of non-orthogonal multiple access under I/Q imbalance," *IEEE Access*, vol. 6, pp. 18453–18468, 2018.

[32] B. Razavi, *RF Microelectronics*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

[33] A. A. Boulogeorgos, V. M. Kapinas, R. Schober, and G. K. Karagiannidis, "I/Q-imbalance self-interference coordination," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4157–4170, 2016.

[34] A.-A. A. Boulogeorgos, P. C. Sofotasios, B. Selim, S. Muhaidat, G. K. Karagiannidis, and M. Valkama, "Effects of RF impairments in communications over cascaded fading channels," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 8878–8894, 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9971565, 11 pages
https://doi.org/10.1155/2021/9971565

Research Article

Soft Mobility: Transparent Handover with Zero Handover Failure in User-Centric Networks

Yong Sun,[1] Haoyan Wei,[1] Shusheng Wang,[2] and Hongtao Zhang[1]

[1] *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2] *Space Star Technology Co., Ltd., Beijing 100086, China*

Correspondence should be addressed to Hongtao Zhang; htzhang@bupt.edu.cn

Received 15 March 2021; Accepted 29 June 2021; Published 13 July 2021

Academic Editor: Wei Duan

User-centric network (UCN) is regarded as a promising candidate to approach the challenges of more radio link failures (RLFs) due to the ultradense deployment of small base stations (SBSs) and meet the requirements of ultrahigh throughput, ultrahigh reliability, and ultralow latency for the 6G system. In this paper, soft mobility is proposed for UCN with the split of control and user plane (C/U-plane) and shared physical cell identifier (PCI) to achieve the goal of zero handover failure (HOF) probability, where transparent handover (HO) within a cell is realized with user configuration duplication and measurement enhancement. Specifically, the cell is composed of several SBSs around the user, where one anchor SBS is selected for controlling, and others act as slave SBSs for transmission with duplicated UE configuration from the anchor SBS. Based on the proposed architecture, the user measures downlink channel quality for cells and SBSs distinguishingly, via SS/PBCH Block (SSB) and channel-state information-reference signal (CSI-RS), respectively, and then makes the HO decision. Results show that soft mobility can reduce the number of HOF by about 50% over the current system, and the HOF probability is lower than 1% for TTT = 40 ms and offset = −1 dB.

1. Introduction

Ubiquitous wireless access is one of the prominent features in six-generation (6G) wireless communication networks, so the number of Femto Access Points (FAPs) will grow exponentially in order to address the high traffic demands caused by more and more smart devices [1]. Ultradense networks (UDNs) are regarded as one of the most significant technologies for the fifth-generation and beyond. However, the increasing number of small cells may cause frequent handovers and degraded mobility robustness [2]. So the crucial topic is proposed in the 6G network which is user-centric network (UCN).

The philosophy of UCN is introduced to deal with the strong interference and frequent handover (HO) in UDN [3–5]. UCN breaks through the network-centric cellular architecture and consequently provides senseless movement for user equipment (UE), which is regarded as a promising candidate to meet the requirements of ultrahigh throughput, ultrahigh reliability, and ultralow latency for the 6G system [6–8]. However, the mobility management requires more considerations for UCN when a cluster of cooperating small cells appears to UE as a single cell.

The proposed soft mobility for UCN is aimed at reducing radio link failures (RLFs) due to the ultradense deployment of small base stations (SBSs) and reducing the HOF probability to zero. Through user configuration duplication and measurement enhancement, transparent HO within SBSs groups is realized, where the configuration of physical (PHY) layer parameters is transparent between a set of SBSs; thus, no reconfiguration is needed for HO. In this paper, the architecture of UCN with the split of control and user plane (C/U-plane) is proposed, and the concepts of the anchor SBS and the slave SBS are introduced, and SBS-level mobility procedure and cell-level mobility procedure are also designed correspondingly. The main contributions of this paper can be summarized as follows:

(i) The architecture of UCN with anchor and slave SBSs is proposed for providing transparent handover.

Compared to 5G system where the responsibilities of each SBSs are independent and roughly the same, in this work, a certain number of SBS sets will be formed and divided into anchor SBSs and slave SBSs which have different responsibilities in the process of handover and transmission. Anchor SBS with strong capability acted as a handover anchor and its neighboring SBSs are slave SBSs

(ii) Mobility management procedures are redesigned for SBS-level and cell-level handover. Different from the mobility procedure in the 5G system, it is needed to redesign the handover procedure separately when UE changes the slave SBS and anchor SBS, including SBS-level mobility procedure and cell-level mobility procedure, and transparent handover between slave SBSs is realized with user configuration duplication and measurement enhancement

(iii) UE-controlled mobility management is applied for its scalability instead of network-controlled mobility management. That is because UE can get better aware of the surrounding wireless communication environment, and fewer measurement reports are needed

(iv) The simulation results verify the superiority of the proposed architecture and mobility management procedures. Furthermore, the effect of system parameters (such as UE velocity, BS densities, time-to-trigger, and offset) on handover, handover failure, and throughput is analyzed, which could provide guidance for actual network planning in 6G systems

This paper proposes a soft mobility model for UCN to reduce the number of HOs and handover failures (HOFs) without extra resource occupancy of dual connectivity (DC), where multiple SBSs around UE form a cell for transparent handover. Simulation results showed that better performance on delay, signal overheads, and HOF reduction is achieved in the soft mobility scheme proposed.

The rest of the paper is structured as follows. Section 2 introduces the previous related work and compares the scheme proposed in this paper with the previous work. Section 3 illustrates the network architecture and mobility management events. Section 4 describes the formulation models involved in the handover process. Section 5 explains new designs for soft mobility which are needed to support the procedures. And mobility procedures are detailly presented in Section 6. Numerical results and explanations are shown in Section 7. Finally, Section 8 concludes this paper.

2. Related Works

Considering the dense deployment of BSs, [4, 9, 10] analyze the handover performance theoretically. The negative effect of channel fading and the overhead of handover on mobility performance in UDN are analyzed in [9, 10], respectively. Poor mobility performance under dense SBSs deployment and performance gain of UCN architecture is confirmed in [4, 9] through theoretical results without specific network architecture design. HO management strategies are proposed in [11–14] to reduce HO and HOF in UDN at the cost of spectrum resources [11] or high computational complexity [12–14], which lack implementability. The specific mobility implementation scheme for UCN needs to be designed.

Recent studies investigate the mobility performance with the formation of cooperating SBSs, which shows that the HO reliability is improved. A softer HO scheme is proposed in [15] with DC for cell-edge users, which enables fast HO by duplicating control messages. However, the mobility improvement comes at the expense of more signal overheads between each serving-target cell pair. Further, local anchor-based DC is applied in UCN in [5], which achieves a remarkably decreased HOF rate without increasing control overheads. By synchronizing the multiple SBSs in the same cluster, the selected anchor SBS will manage the HO within the cluster which only requires few procedures. However, DC will occupy extra resources of the SBSs, which reduces resource utilization rate and causes more frequent HO events due to the additional transmission link for robustness. How to design a low signaling overhead mobility management scheme in UCN is still an open problem.

It would be useful to generalize the anchor-based architecture and the method of configuration duplication to mobility enhancement in UCN. However, in UCN, the measurement for SBSs and cells should be distinguished [16]. Especially, SBS-level and cell-level measurements and decisions are implemented with channel-state information-reference signal (CSI-RS) and SS/PBCH Block (SSB), respectively.

3. Network Architecture for Soft Mobility Enhancement

UCN makes the user feel like the network is always following it, and the network intelligently recognizes the user’s wireless communication environments and then flexibly organizes the required cell group and resource to serve the user. Inspired by this, a UCN architecture that supports soft mobility is presented in this section to realize transparent handover.

3.1. Network Architecture. In this section, the architecture of UCN with the split of control and user plane (C/U-plane) is introduced, as shown in Figure 1, which naturally supports the features of soft mobility in terms of transparent handover [16]. Several SBSs around a user form a cell to provide user-centric coverage. The anchor SBS is selected for the C-plane, which in a way operates as a gateway in the system by terminating the signaling and data plane between other SBSs (terms as slave SBS) and the core network. The number of SBSs in a cell is set by the network operator, and the SBS with the largest load capacity in the cell is selected as the anchor SBS because the anchor SBS is expected higher capacity. The other SBSs around the user provide the U-plane as slave SBSs.

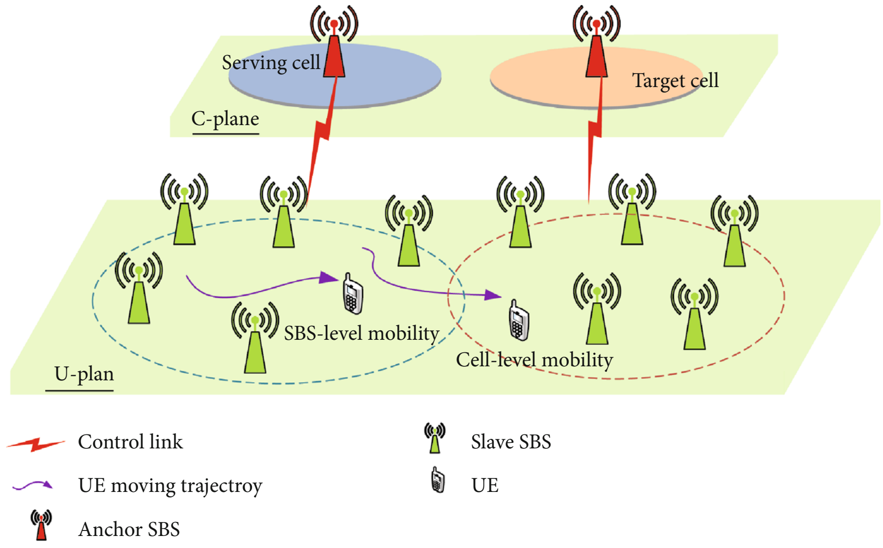


Figure 1: Soft mobility model and network architecture.

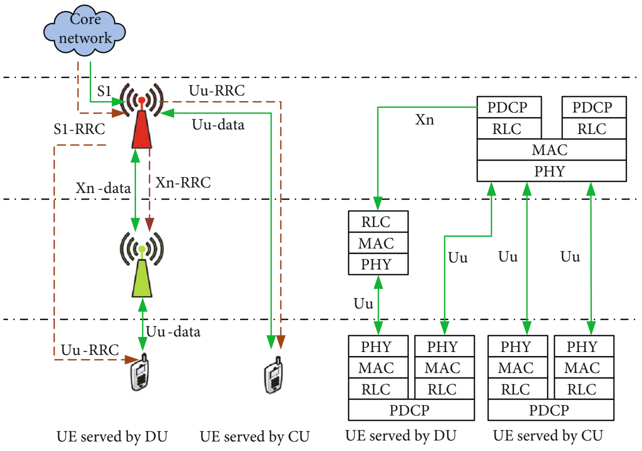


Figure 2: The protocol stack in the UCN.

3.2. Protocol Stack. The design of U-plane protocol stack 3C is applied for the two-layer UCN architecture, shown in Figure 2, where the anchor SBS and its slave SBSs share the protocol data unit (PDU) of packet data convergence protocol (PDCP) layer, while the radio link control (RLC) and medium access control (MAC) are independent as mentioned above. With the downlink measurement result on CSI-RS, the proper SBS for transmission can be selected dynamically.

On the other hand, one of the slave SBSs provides data service for the UE acted as the transmission node and the RLC, MAC, and PHY layers of the slave SBS and the anchor SBS are independent. So when the service transmission node changes, the terminal only needs to reconfigure the parameters of the RLC and MAC layers.

3.3. Mobility Management for Soft Mobility. Under the proposed architecture, mobility management events should be redesigned for soft mobility. As shown in Figure 1, there is an example of the user moving trajectory and 3 types of mobility management events are included.

(i) *Initialization*. In the beginning, the UE is at the beginning of the trajectory; then, the SBS with the largest load capacity among the SBSs around the user is chosen as the anchor SBS, and then, a certain number SBSs are chosen as the slave SBSs, and anchor SBS provides data service at the beginning as it provides maximum reference signal receiving power (RSRP) among the slave SBSs.

(ii) *SBS-Level Handover*. A SBS-level handover is triggered when the RSRP from another SBS is stronger than serving SBS due to the user's movement and the SBS is in the set of slave SBSs.

(iii) *Cell-Level Handover*, As the UE moves, it became farther from the anchor SBS. When the signal from SBSs in the current cell is unable to satisfy the A3 entering condition, cell-level handover is needed.

The proposed soft mobility model is aimed at achieving transparent HO and senseless movement for mobile users. With duplicated UE configuration, the slave SBSs share physical cell identifier (PCI) with the anchor SBS within the same cell, so that there is no sense of changing cells for the user when the slave SBS is changed within the same cell.

4. Formulation Models

This section presents formulaic models for the soft mobility handover (HO), signal to interference and noise ratio (SINR), throughput, spectrum efficiency (SE), and handover failure (HOF).

4.1. Model for HO. As motioned above, several SBSs and a SBS near UE form slave SBS and anchor SBS service UE together, and the soft mobility scheme is different from the traditional scheme in the current system. A novel scheme for the soft mobility architecture is discussed in this subsection.

The UE needs to select a target SBS for SBS-level mobility and a target cell with a target SBS for cell-level mobility. Similar to event A3 [17], the HO decision should consider the RSRP and is made at t_0 if the following condition is fulfilled.

$$P^{\Theta}_{t,u} > P^{\Theta}_{s,u} + \mathrm{Offset}^{\Theta} \text{ for } t_0 - T^{\Theta} < t < t_0, \tag{1}$$

where $\Theta \in \{\mathrm{SBS}, \mathrm{cell}\}$ denotes the SBS-individual or cell-individual parameters, $P^{\Theta}_{t,u}$ and $P^{\Theta}_{s,u}$ are the measured RSRPs of a neighboring SBS (or cell) and the serving SBS (or cell), respectively, Offset^{Θ} is the offset, and T^{Θ} is the time-to-trigger (TTT). It is worth noting that the $P^{\mathrm{SBS}}_{s,u}$ is only measured from the slave SBS in the cell in order to achieve senseless movement.

Similar to the 5G system, the offset and TTT are designed to improve mobility robustness and reduce unnecessary handover and ping-pong (PP) effect. At the same time, the offset and TTT should not be configured too big, which may lead to not timely handover trigger and cause RLF afterward.

4.2. Model for SINR, Throughput, and SE. The RSRP from the neighbor SBS is measured periodically, in order to make handover decisions in time. A universal path loss-plus-fading model is used to describe the received signal power. So the RSRP received by UE u from SBS t is given by

$$P^{n}_{t,u}(d,t) = p^{n}_{t} g_t(d) h_t(t), \tag{2}$$

where p^n_t denotes the transmit power of SBS t at the subchannel n, $g_t(d)$ denotes the pathloss gain that only depends on the distance d between UE and SBS t, and $h_t(t)$ is the multiplicative channel gain at time t modeling the multipath fading effect.

Although slave SBSs logically serve UE together, the bandwidth is reused among the slave SBS in order to achieve high spectrum efficiency. So the UE will still be interfered by the other slave SBS. Then, the corresponding SINR of the UE u which connects with SBS s can be calculated as

$$\mathrm{SINR}^{n}_{s,u}(t) = \frac{P^{n}_{s,u}(d,t)}{\sum_{j\in\Psi\setminus\{s\}} P^{n}_{j,u}(d,t) + \eta}, \tag{3}$$

where $P^n_{s,u}(d,t)$ denotes the signal received by the UE u from the serving slave SBS at the subchannel n, $P^n_{j,u}(d,t)$ denotes the interference received by UE u from other SBS at the subchannel n, and η is the white noise power.

By using (3) and Shannon capacity theory, the achievable data rate of UE u at the subchannel n is shown as follows:

$$C^{n}_{u} = W \log_2\left(1 + \mathrm{SINR}^{n}_{s,u}\right), \tag{4}$$

where $\mathrm{SINR}^n_{s,u}$ is the SINR of UE u at the subchannel n calculated by (4) and W is the bandwidth at the subchannel n. Therefore, the SE of UE u is calculated as follows:

$$\mathrm{SE}_u = \frac{\sum_n C^n_u}{\mathrm{BW}_u}, \tag{5}$$

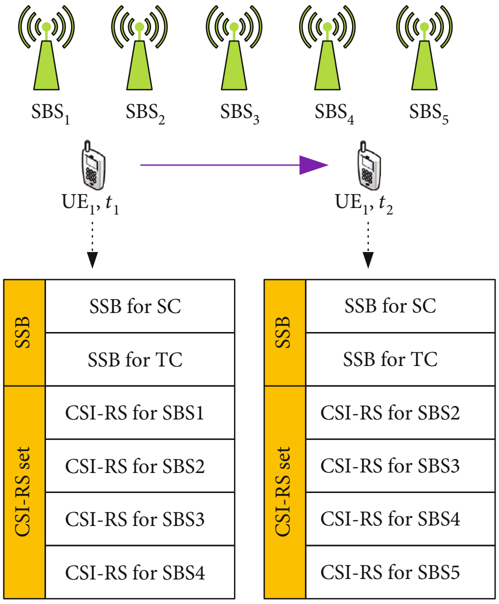


Figure 3: An example for the maintenance of CSI-RS set.

where BW_u is the total bandwidth allocated to UE u. And after averaging the spectrum efficiency on all of the subchannel, we can get the SE of UE u.

4.3. Model for HOF. The SINR can evaluate the link quality, so it can be used to judge the HOF. According to the 3rd Generation Partnership Project (3GPP) specification, the handover failure model in state 2 can be summarized as follows:

(i) The link quality becomes worse after the handover is triggered because of the user's moving, and the SINR decreases at the same time

(ii) When the SINR is below a certain threshold, that is, Q_{out} during the TTT, handover failure is caused

So the HOF model can be expressed by

$$\mathrm{SINR}_{s,u}(t) < Q_{\mathrm{out}} \text{ for } t_0 - T^{\Theta} < t < t_0, \tag{6}$$

where handover is triggered at $t_0 - T^{\Theta}$, and T^{Θ} is the TTT.

5. Design for Soft Mobility

Given the above architecture, in order to support the procedures to be mentioned later, we have introduced some special designs for soft mobility. In this section, we explain the main three points of the design in detail.

5.1. UE Configuration Duplication. The process of the UE configuration can be expressed as follows:

(i) The anchor SBS transfers both UE static configuration and UE running-time configuration to the slave SBS, so that slave SBS could act as a transmission point of the anchor SBS from the UE point of view.

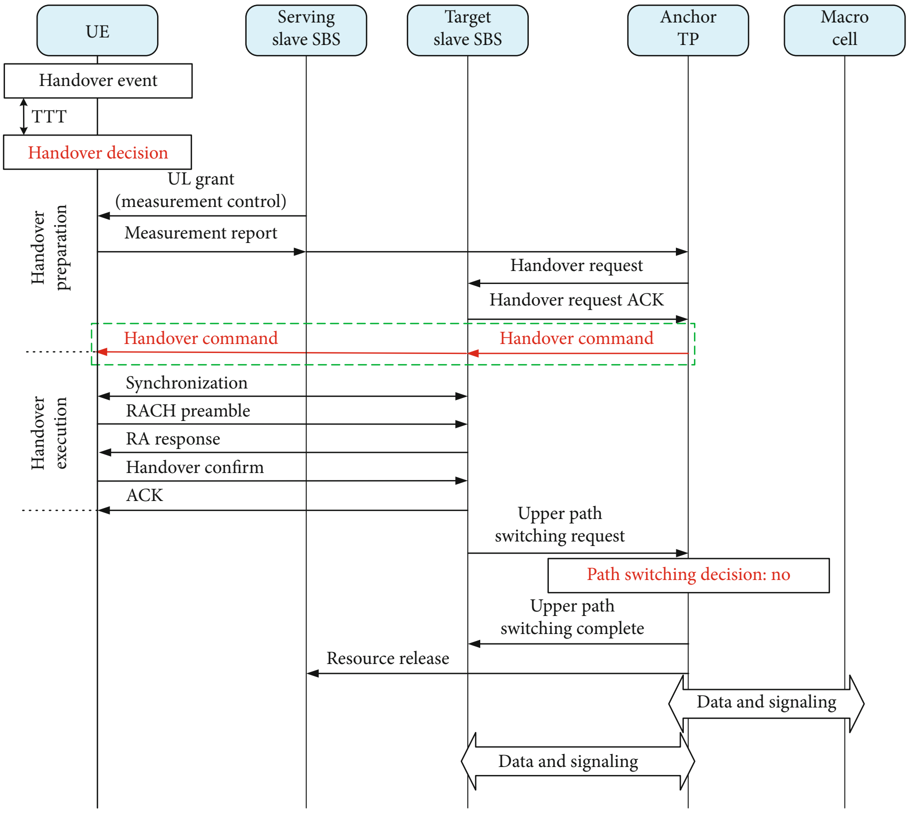


Figure 4: SBS-level mobility management procedure.

That is because the anchor SBS and the slave SBS have the same PCI

(ii) Both the anchor SBS and the slave SBS transfer the same L2 data to UE in a redundant way

(iii) UE performs combination in L2 RLC layer

5.2. Measurement and Reference Signal. In order to make the decision on target SBS and target cell for intracell and intercell mobility, measurement configurations and related signaling are needed. Since a cell is composed of densely deployed SBSs, the set of SBSs within one cell and the set of neighboring cells to be measured should be different. The UE needs to obtain measurement results by measuring candidate cells and candidate SBSs wherein each cell includes multiple SBSs that share a common cell ID (see Figure 3).

For the proposed soft mobility model, the downlink measurements can be classified into two types:

(i) *SBS-Level Measurement*. The UE measures downlink channel quality via CSI-RS and later reports to the anchor SBS after the UE makes the HO decision with proper SBS for SBS-level mobility.

(ii) *Cell-Level Measurement*. It is based on the SSB which is sent to UE only when the serving SBS is changed to an edge SBS, or the UE is moving between edge SBSs.

In addition, adjacent UE transmits the UE-specific orthogonal uplink sounding reference signal (SRS) number, for the uplink channel estimation which can be measured by arbitrary slave SBSs. The slave SBSs are informed by the anchor SBS to monitor the SRS and then send the feedback to the anchor SBS. Based on the uplink measurement, in order to handle the SRS conflict, the slave SBS chooses its serving UE for SRS reconfiguration before the HO process.

5.3. Uplink and Downlink Channel. Different from the current system, the key characteristics of channel design are described as follows:

(i) Enhanced physical downlink/uplink control channel (ePDCCH/ePUCCH) is applied instead of PDCCH/PUCCH for frame control/uplink control

(ii) For the uplink channel, anchor-assisted physical random access channel (PRACH) takes over PRACH

(iii) For downlink channel, anchor-assisted physical broadcast channel is applied for system information bearer

6. Procedures

Base on the proposed architecture, the handover management would be different from the current system. As both anchor SBS and slave SBS exist in the proposed architecture, two types of mobility management are correspondingly designed, which are SBS-level mobility procedure and cell-

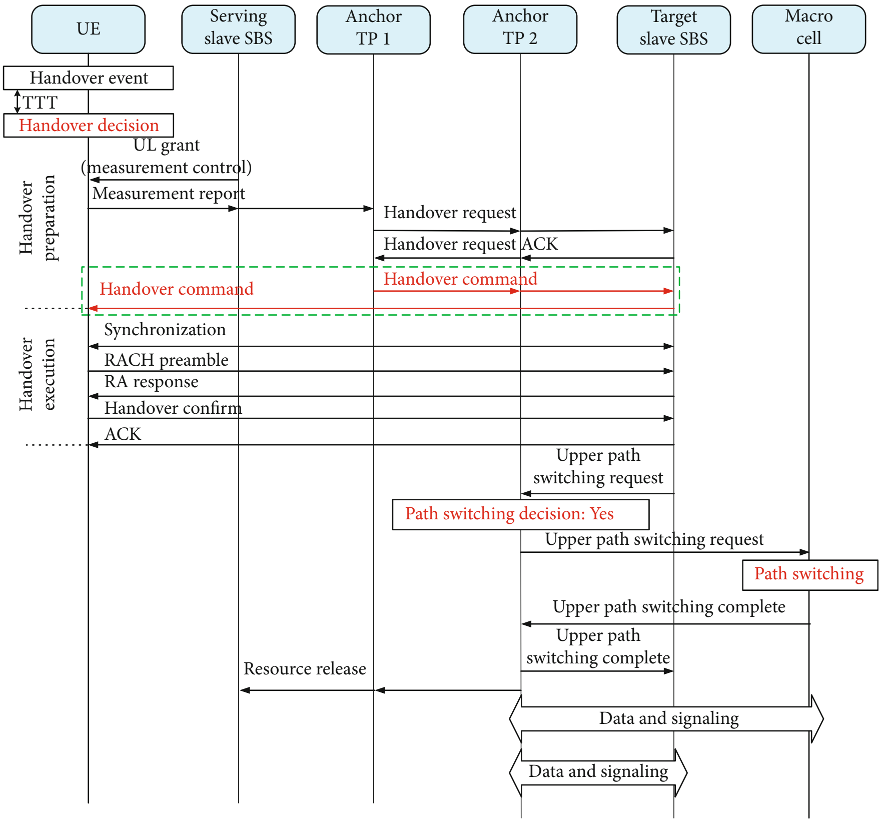


FIGURE 5: Cell-level mobility management procedure.

level mobility procedure. This section gives detailed explanations of the procedures, which are described in Figures 4 and 5, respectively.

There are two places that can decide to change the serving SBS which are the UE and the network and are termed UE-controlled mobility management and network-controlled mobility management.

(i) In UE-controlled mobility management, the UE estimates the channel quality from the neighboring SBSs or cells and determines mobility events

(ii) In network-controlled mobility management, the SBSs decide and initiate the processes based on the measurement feedback information assisted by UE

This paper is concerned about UE-controlled mobility management for its scalability. When UE moves across the adjacent cells, cell-level mobility is needed, where not only a target cell but also a specific SBS in the cell needs to be identified.

6.1. SBS-Level Mobility Procedure. The key procedure of the SBS-level mobility procedure is shown in Figure 4. In the preparing phase, before the HO procedures, the UE must already have a radio resource control connection with the anchor SBS in C-plane and a SBS in U-plane which could be either the anchor SBS or slave SBS. The main procedure of the SBS-level mobility procedure is described as follows:

TABLE 1: Simulation parameters.

Parameters	Values
Carrier frequency	2 GHz
Bandwidth	10 MHz
SBS transmit power	30 dBm
Area of simulation	$500 * 500\,\text{m}^2$
UE antenna height	1.5 m
Variance of white Gaussian noise	-174 dBm/Hz
UE density	1000 users/km^2
SBS density	180, 320, 500, 720, 980, 1280, 1620 SBSs/km^2
Number of SBSs in a cell	5
TTT	$T^{\text{SBS}} = T^{\text{cell}} = 320$ ms
Offset	Offset$^{\text{SBS}}$ = 1 dB, Offset$^{\text{cell}}$ = 3 dB
Q_{in}	-8 dB
Q_{out}	-6 dB
T_{RLF}	200 ms
T_{recovery}	100 ms

(i) First, the RSRPs of the neighboring SBSs are obtained periodically by UE, and UE makes the handover decision when condition (1) is satisfied

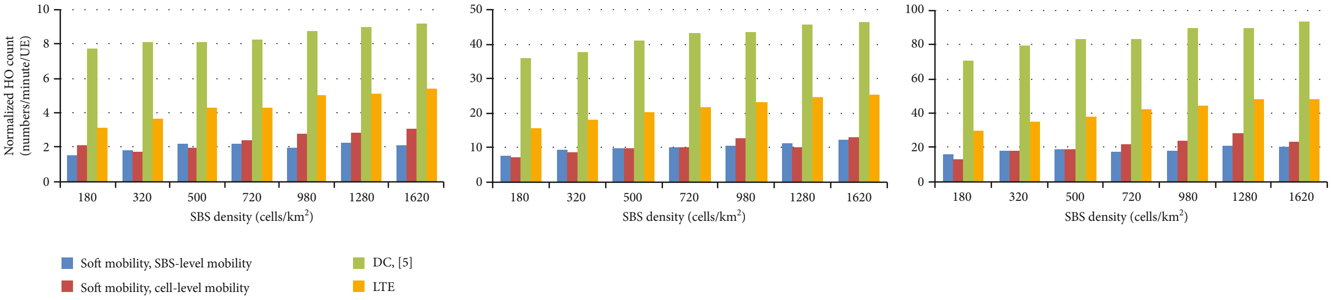


FIGURE 6: The normalized HO count versus SBS density under the soft mobility model, the DC scheme, and the LTE system. *User velocity* = 3 km/h, 15 km/h, and 30 km/h.

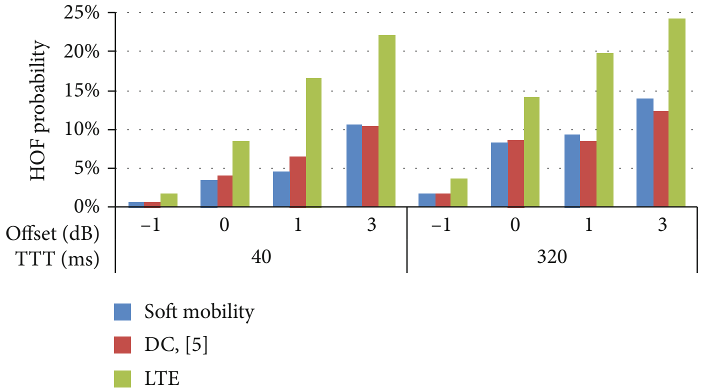


FIGURE 7: Effect of different TTT/offset parameters on the HOF probability (user velocity = 3 km/h and SBS density = 1280 count/km^2).

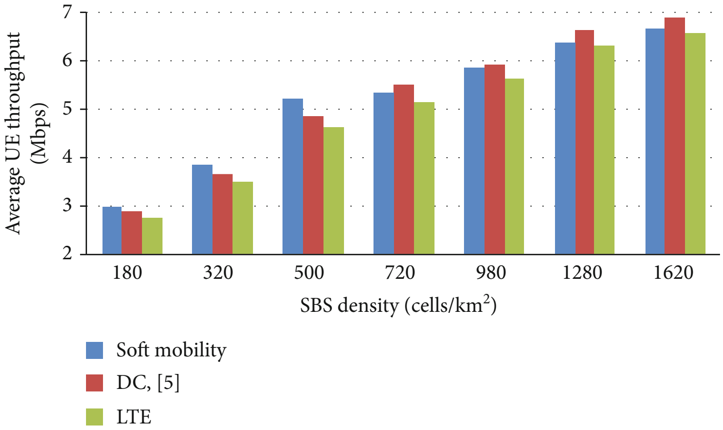


FIGURE 8: The average UE throughput under different SBS density (user velocity = 3 km/h and SBS density = 1280 count/km^2).

and holds for a certain time TTT. The target slave SBS is selected from the SBSs which are under the same anchor SBS's control as the serving slave SBS

(ii) Second, the serving slave SBS sends an uplink grant (UL grant) to the UE periodically in order to obtain the measurement report. Then, the UE sends the measurement report to serving slave SBS and anchor SBS to inform the anchor SBS that SBS-level handover is needed

(iii) Third, the anchor SBS sends the handover request to the target slave SBS, and the handover request acknowledge character (ACK) is sent by target slave SBS to the anchor SBS. Then, the anchor SBS sends a handover command to both target slave SBS and UE in order to prepare and execute the handover

(iv) Finally, handover execution is carried out, and the UE is synchronized with the target slave SBS. After

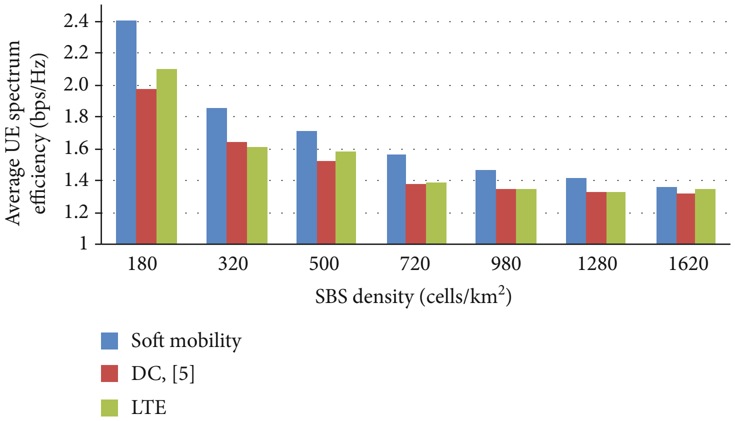


FIGURE 9: The average UE spectrum efficiency (user velocity = 3 km/h and SBS density = 1280 count/km^2).

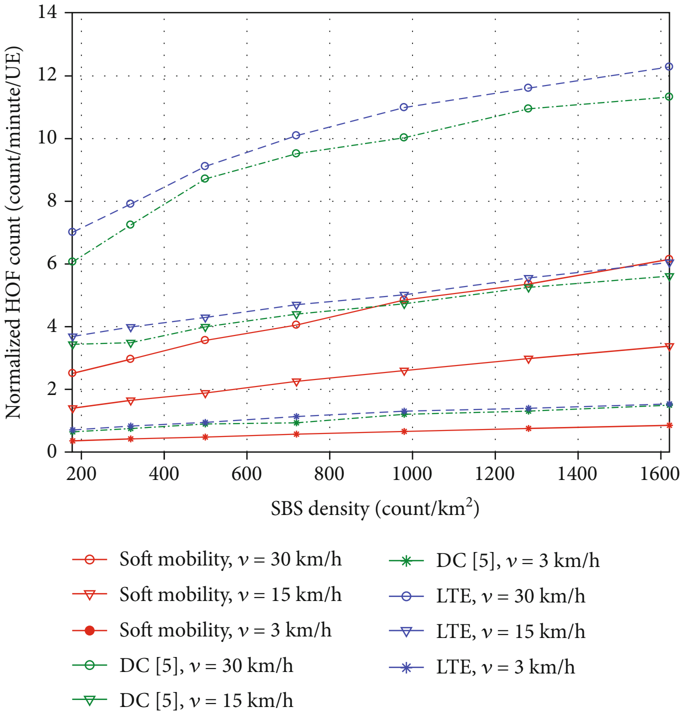


FIGURE 10: The normalized HOF count versus SBS density under the soft mobility model, the DC scheme, and the LTE system.

that, the anchor SBS informs the serving slave SBS to release the resource

It is important to note that compared with the handover procedure in the current system, the main differences of the SBS-level mobility management of soft mobility can be summarized as follows:

(i) The SSBs are shared by SBSs within the same cell, while CSI-RS patterns are used for downlink measurement and target SBS selection

(ii) The UE makes the HO decision, which achieves better flexibility than the current system by selecting the optimal SBS based on the user-centric decision

(iii) The RSRP from SBSs, the load condition of SBSs, and the context information of UE can be taken into consideration of HO decision under the UE-controlled mobility

6.2. Cell-Level Mobility Procedure. When a UE moves to the edge of the cell, and the signal from SBSs in the current cell is unable to satisfy the A3 entering condition, the cell-level mobility from the current cell to another will be executed. The anchor SBS will trigger the downlink measurement, and UE will search for a new cell with a new SBS for providing high-quality continuous service. The main procedure of the cell-level mobility procedure is described as follows:

(i) First, the RSRP of the neighboring SBSs are obtained periodically by UE, and UE makes the handover decision when condition (1) is satisfied and holds for a certain time TTT

(ii) Second, the serving slave SBS sends an UL grant to the UE periodically in order to obtain the measurement report. Then, the UE sends the measurement report to serving slave SBS and original anchor SBS to inform the anchor SBS that cell-level handover is needed

(iii) Third, the anchor SBS sends the handover request to the target slave SBS and the target anchor SBS; then, the handover request ACK is sent by target slave SBS to the original anchor SBS and target anchor SBS. Then, the original anchor SBS sends a handover command to both target slave SBS and target anchor SBS, and the target anchor SBS sends a handover command to the UE

(iv) Finally, handover execution is carried out, and the UE is synchronized with the target slave SBS and target anchor SBS. After that, the target anchor SBS informs the serving slave SBS and original anchor SBS to release the resource

And the main differences between SBS-level and cell-level mobility procedures can be concluded as follows:

(i) Cell-level mobility will be executed between SBSs when slave SBSs in the current cell are unable to provide high-quality service, and slave SBSs in the adjacent cell may be more appropriate to provide data service instead

(ii) For cell-level mobility, other than the SBS-level measurement based on CSI-RS, an extra layer of measurement which is the cell-level measurement based on SSB should be triggered by the connection to edge SBSs

(iii) The context of the UE is stored in the anchor SBS, and additional mobility procedures, e.g., path switch, are executed among the core network, the serving anchor SBS, and the target anchor SBS

(iv) In the cell-level mobility scenario, not only the U-plane packet transmission path but also the C-plane RRC connection is reconfigured

(v) The target SBS sends the *HO command* to the UE instead of the serving SBS for a better radio link

7. Performance Evaluation

The simulation parameters are configured according to the 3GPP [18, 19], as shown in Table 1. It is assumed that the initial position of SBSs and UEs obeys the Poisson point process (PPP), and the Random Waypoint (RWP) model proposed in [20] is adopted to simulate user's random motions.

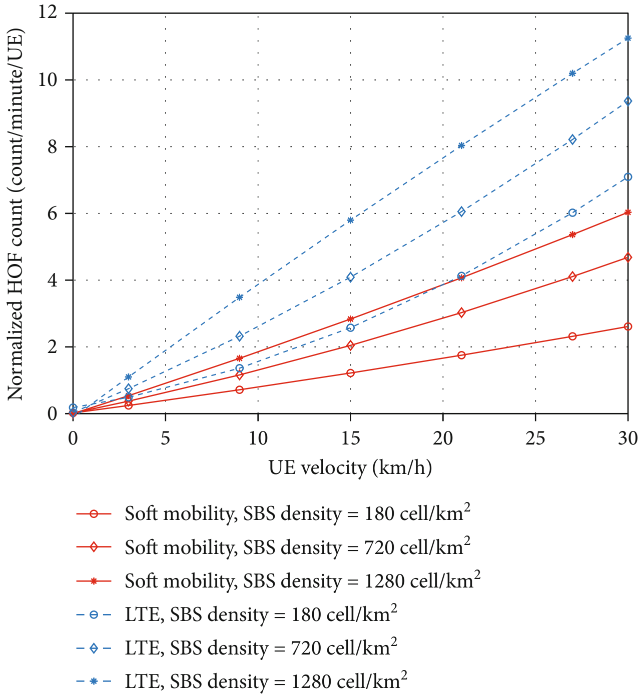


Figure 11: Effect of UE velocity on HOF rate comparison between the proposed scheme and conventional scheme.

The normalized HO count for SBS density is shown in Figure 6. Since path switching is not required for SBS-level mobility in the soft mobility model, the HO delay and signal overheads are reduced. Because of an extra connection, the DC scheme consequently results in a more HO number. Compared with the number of HO in the LTE system, the number of cell-level mobility achieves a decrease of about 50%.

Figure 7 describes the effect of different TTT/offset parameters on the HOF probability where HOF probability = the number of HOF events/the total number of HO attempts .

Using Shannon capacity theory, the average user throughput and the spectrum efficiency (SE) are given by (4) and (5), shown in Figures 8 and 9, respectively. The SE decreases with the SBS density as the utilized resource blocks increase with the shorter UE-to-SBS distance. The average throughput and average user SE in the soft mobility model are better than those in the LTE system and the DC scheme. This is because, in the soft mobility model, the wasted resources of the SBS brought by the extra connection in the DC scheme are utilized.

Figure 10 compares the normalized HOF number in the soft mobility model with that in the LTE system and DC scheme. The HOF number for DC is detected and counted during the HO process of any of the two serving SBSs. It is shown that the soft mobility model reduces the normalized HOF count; thanks to the design for transparent handover. Specifically, we can see a significant improvement in the HOF number of soft mobility with a maximum decrease of more than 50% over the LTE system and a decrease of 45% over the DC scheme.

Figure 11 compares the HOF rate between the proposed scheme and the conventional scheme as a function of UE velocity. It is shown in Figure 11 that the HOF rate linearly increases with the UE density, approximately. This can be explained by that the HO rate linearly increases with UE velocity, thus the HOF rate increase with the HO rate correspondingly. What is more, the simulation results showed that the HOF rate under our proposed scheme keeps decreasing by about 50% at different UE velocity compared with the LTE system.

Setting lower TTT and lower offset can decrease HOF probability; however, it will introduce high PP rate on the other hand. To trade off the HOF probability and PP rate, offset = 1 dB and TTT = 320 ms are the preference parameter configurations in the soft mobility model. Based on above figures, although the HOF probability under the DC scheme almost performs the same as that of soft mobility, it has a larger HO number.

8. Conclusion

This paper proposes and evaluates the soft mobility model as a solution to tackle mobility challenges in future dense networks. Soft mobility is aimed at minimizing the number of HOFs by achieving transparent mobility. The UCN is introduced to eliminate the cell edges where multiple SBSs in UE's vicinity serve the UE in the form of a single cell. Further, feasible mobility procedures are provided, where UE makes the decision and target SBS sends the HO command message instead of serving SBS. Simulation results showed that the soft mobility scheme gives better performance on HOF reduction than the LTE system.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation under Grant 4202048 and the National Natural Science Foundation of China under Grant 61971064 and Grant 61901049.

References

[1] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: the future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133995–134030, 2020.

[2] H. Zhang, Y. Chen, Z. Yang, and X. Zhang, "Flexible coverage for backhaul-limited ultradense heterogeneous networks: throughput analysis and η-optimal biasing," *IEEE transactions on vehicular communications*, vol. 67, no. 5, pp. 4161–4172, 2018.

[3] H. Zhang, Y. Chen, and Z. Han, "Explicit modelling and performance analysis of cell group selection with backhaul-aware biasing," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 273–276, 2019.

[4] H. Zhang, W. Huang, and Y. Liu, "Handover probability analysis of anchor-based multi-connectivity in 5G user-centric network," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 396–399, 2019.

[5] H. Zhang, N. Meng, Y. Liu, and X. Zhang, "Performance evaluation for local anchor-based dual connectivity in 5G user-centric network," *IEEE Access*, vol. 4, pp. 5721–5729, 2016.

[6] Y. He, L. Dai, and H. Zhang, "Multi-branch deep residual learning for clustering and beamforming in user-centric network," *IEEE Communications Letters*, vol. 24, no. 10, pp. 2221–2225, 2020.

[7] W. Huang, J. Peng, and H. Zhang, "User-centric intelligent UAV swarm networks: performance analysis and design insight," *IEEE Access*, vol. 7, pp. 181469–181478, 2019.

[8] L. Yang, H. Zhang, and Y. He, "Temporal correlation and long-term average performance analysis of multiple UAV-aided networks," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8854–8864, 2021.

[9] Y. He, W. Huang, H. Wei, and H. Zhang, "Effect of channel fading and time-to-trigger duration on handover performance in UAV networks," *IEEE Communications Letters*, vol. 25, no. 1, pp. 308–312, 2021.

[10] S. Choi, J. Choi, and S. Bahk, "Mobility-aware analysis of millimeter wave communication systems with blockages," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 5901–5912, 2020.

[11] O. Semiari, W. Saad, M. Bennis, and M. Debbah, "Integrated millimeter wave and sub-6 GHz wireless networks: a roadmap for joint mobile broadband and ultra-reliable low-latency communications," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 109–115, 2019.

[12] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving sustainable ultra-dense heterogeneous networks for 5G," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

[13] M. Alhabo, L. Zhang, and N. Nawaz, "GRA-based handover for dense small cells heterogeneous networks," *IET Communications*, vol. 13, no. 13, pp. 1928–1935, 2019.

[14] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 779–793, 2018.

[15] I. Viering, H. Martikainen, A. Lobinger, and B. Wegmann, "Zero-zero mobility: intra-frequency handovers with zero interruption and zero failures," *IEEE Network*, vol. 32, no. 2, pp. 48–54, 2018.

[16] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: challenges, methodologies, and directions," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 78–85, 2016.

[17] 3GPP TS 38.331, *Technical Specification Group Radio Access Network; NR; Radio Resource Control (RRC); Protocol Specification (Release 15)*, 3rd Generation Partnership Project, 2020.

[18] 3GPP TR 36.839 V11.1.0, *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility Enhancements in Heterogeneous Networks (Release 11)*, 3rd Generation Partnership Project, 2012.

[19] 3GPP TR 36.872 V12.1.0, *Technical Specification Group Radio Access Network; Small Cell Enhancements for E-UTRA and E-UTRAN-Physical Layer Aspects (Release 12)*, 3rd Generation Partnership Project, 2013.

[20] D. B. Johnson and D. A. Maltz, "Dynamic source routing in *ad hoc* wireless networks," *Mobile Computing*, vol. 353, pp. 153–181, 1996.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5563338, 9 pages
https://doi.org/10.1155/2021/5563338

Research Article

Real Time Arrhythmia Monitoring and Classification Based on Edge Computing and DNN

Mingxin Liu,[1] **Ningning Shao**,[2] **Chaoxuan Zheng**,[3] **and Ji Wang**[1]

[1]*College of Electrical and Information Engineering, Guangdong Ocean University, Zhanjiang 524088, China*
[2]*College of Information Science and Engineering, Yanshan University, Qinhuangdao 006004, China*
[3]*School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China*

Correspondence should be addressed to Ji Wang; 13902576499@163.com

Received 26 January 2021; Revised 26 February 2021; Accepted 31 March 2021; Published 17 May 2021

Academic Editor: Wei Duan

In this paper, we investigate how to incorporate intelligence into the human-centric IoT edges to detect arrhythmia, a heart condition often associated with morbidity and even mortality. We propose a classification algorithm based on the intrapatient convolutional neural network model and the interpatient attention residual network model to automatically identify the type of arrhythmia in the edges. As the imbalance categories in the MIT-BIH arrhythmia database which needs to be used in the algorithm, we slice and overlap the original ECG signal to homogenize the heartbeat sets of different types, and then the preprocessed data was used to train the two proposed network models; the results reached an overall accuracy rate of 99.03% and an F1 value of 0.87, respectively. The proposed algorithm model can be used as a real-time diagnostic tool for the remote E-health system in next generation wireless communication networks.

1. Introduction

It is reported by The World Health Organization that cardiovascular diseases are the primary cause of the world's highest mortality, and arrhythmias are the most common [1]. Arrhythmias are caused by abnormalities in the conduction system of the heart. They can be slowly, rapidly, or irregular heartbeats and can be life-threatening or nonlife-threatening. Nonlife-threatening arrhythmias need to be tested for a long period of time to ensure that the pathologic causes of the arrhythmia can be detected early. ECG signal is a kind of physiological signal that can record and reflect the condition of the heart, and its simple and noninvasive advantages are widely used in the diagnosis of arrhythmia. In the early stage of cardiovascular diseases, arrhythmias are often accompanied by the appearance of arrhythmia. Therefore, early diagnosis and prevention are very important for the intervention treatment of patients.

Traditionally, the diagnosis of arrhythmias relies on a cardiologist's ability to identify specific types of arrhythmias by analyzing the waveform of electrical signals collected from the heart. However, due to the complexity and suddenness characteristics of arrhythmia, the ECG signals detected in a short time may not accurately reflect the real cardiac activity of the patient. For the ECG signal recording that needs to be monitored for a long time, the identification of arrhythmia type by artificial means is time-consuming and laborious, and it is easy to miss detection. To improve the reading efficiency of arrhythmia, real-time monitoring through automatic analysis technology can play a great auxiliary role in the diagnosis of arrhythmia. 6G wireless communication networks are believed that it should be ubiquitous, human-centric, full band, strongly secure, and intelligent [2, 3], which offers distributed, low latency, and reliable machine learning at the wireless network edge [4, 5].

In the machine learning method, the process of the arrhythmia diagnosis algorithm usually includes three main steps: preprocessing, feature extraction, and classification. Feature extraction has dominated the diagnosis field of arrhythmia for decades, including feature extraction based on wavelet, morphology, and statistics [6]. Wavelet transform decomposes the signal into components of different scales [7], and the time positioning of spectral components can be obtained through wavelet analysis. Some studies on the analysis of ECG signal using extracted by wavelet feature [8]. In literature [9], a random matrix was selected to extract the morphological features of heartbeat, in which each column was normalized, and each row was extracted by discrete cosine transform as the projection matrix. For the statistical characteristics of ECG, signal analysis usually is the use of ECG signal of time-domain characteristic value calculation, such as energy, mean, kurtosis, skewness, maximum, and minimum [10, 11], these features provide an effective method to analyze the complexity, and different types of the time series of ECG signal can help distinguish the types of arrhythmia patients, to obtain better classification performance. However, the advantages and disadvantages of such arrhythmia diagnosis algorithms usually depend on the feature extraction stage, and the robustness of the diagnosis model is still limited due to the complex feature extraction process.

In recent years, the end-to-end deep learning method has shown outstanding performance in automatic feature extraction, and the trend of using convolutional neural network models for arrhythmia diagnosis has become more and more obvious. Paper [12] proposes a 34-layer neural network model that does not take any complex preprocessing or feature extraction steps. The data set used is 500 times that of the open data set. The classification of arrhythmias has a high diagnostic performance similar to that of cardiologists. Although existing work has laid a solid foundation for this field, due to the long recording time of ECG signals, low signal quality, diversity of pathological reasons, and extremely scarce data sets, how to improve the robustness of arrhythmia diagnosis results remains is a challenge.

One of the most effective tools for arrhythmia diagnosis is the detection of ECG signal, and the morphological characteristics and frequency spectrum of a single heartbeat can provide meaningful clinical information about the automatic identification of ECG. However, the shape and time characteristics of the ECG signal between different patients are very different under different physical environments, which leads to the problem of ECG signal classification that has not been fully solved. The main problem of using ECG to diagnose arrhythmia is that different patients have different ECG shapes although they suffer from the same disease, and two different diseases may have roughly the same characteristics in the ECG signal. These problems bring some difficulties to the diagnosis of heart disease [13]. Most of the algorithms in the literature are evaluated based on intrapatient paradigms rather than interpatient programs. Although these algorithms can obtain good accuracy by evaluating intrapatient programs [14, 15], due to individual differences, sexuality exists objectively, and the result is not particularly reliable. So, it is the most consistent with the actual application scenario to avoid the training data and the test data coming from the same sample. In addition, when the amount of sample data in the arrhythmia database is scarce and the number of categories is unbalanced, the existing arrhythmia diagnosis algorithms show poor performance when identifying categories with relatively small amounts of data and whose sensitivity and accuracy are both very low [16]; so, the automatic classification of ECG signals is still a difficult problem.

In order to cope with the above challenges, we use the data in the MIT-BIH arrhythmia database [17] to propose based on intrapatient with the convolution neural network model to simulate the ECG records within the normal beat (N), ventricular premature beat (V), right bundle branch block (R), left bundle branch block (L), and based on the interpatient with attention residual network model [18, 19] for normal (N), ventricular ectopic (S), ventricular ectopic (V), the fusion (F), and unknown beat (Q) five types of classification. Compared with the results in the existing literature, our method can obtain relatively good results. The main contributions of this work are listed as follows:

(a) We propose a one-dimensional convolutional neural network model to classify the heartbeat intrapatient in four categories

(b) We propose to combine the residual network module with the attention mechanism with interpatient ablation study on the proposed network model

(c) By adopting slice and overlap processing to enhance the original ECG signal, the amount of data of various types can be balanced

The rest of this paper is arranged as follows: Section 2 introduces the work related to the study of arrhythmia. Section 3 describes the data set used and the preprocessing of the data. Section 4 describes the proposed two network model structures. Section 5 introduces the evaluation indicators of training network model and analyzes the results. Section 6 summarizes the whole thesis and prospects.

2. Related Work

The algorithm process of arrhythmia diagnosis based on the deep neural network can be roughly divided into the following steps: ECG data preprocessing and arrhythmia classification. The algorithm flow chart is shown in Figure 1.

2.1. Preprocessing. The ECG signal is usually a low-frequency weak signal collected by an electrocardiogram machine with electrodes attached to the surface of the human body. The signal frequency is usually between 0.05 and 100HZ, which is extremely susceptible to external noise. The purpose of data preprocessing steps is to reduce these noises. Typical noise types are as follows:

(a) Baseline drift: it belongs to low-frequency noise (0.15-0.3 Hz), which is the noise caused by the change of electrodeposition caused by movement artifact or the patient's respiration

(b) Power interference: it is mainly a noise signal with a frequency of 50/60HZ generated by the interference of the power system, and its bandwidth is lower than 1HZ

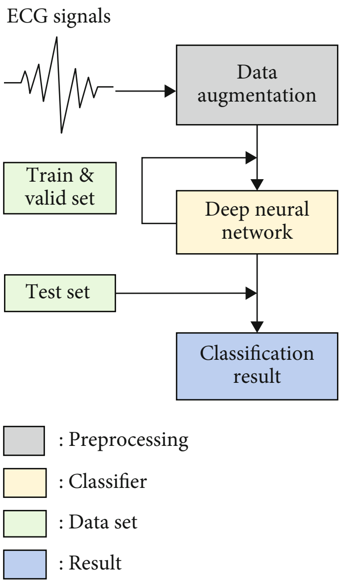


Figure 1: ECG arrhythmia algorithm flow chart.

(c) EMG interference: high-frequency noise signals (30-300 Hz) generated by muscle contractions other than the heart

In the classification process of arrhythmia, noise signals of different degrees will have a great impact on the diagnosis of patients and reduce the accuracy of diagnosis. Therefore, it is necessary to select an appropriate preprocessing method for noise removal [18].

2.2. Based on Existing Methods of Deep Learning. In the deep neural network method, a classifier that can automatically extract features is needed to identify the types of arrhythmia after the ECG data preprocessing step. At present, probabilistic neural network, fuzzy clustering neural network, and recursive neural network are used to classify arrhythmia. The probabilistic neural network is a feed forward network, which is derived from the Bayesian network and Fisher discriminant analysis. Literature [19] believes that the probabilistic neural network model is more robust and effective in calculation than the traditional model. In the structure of the fuzzy clustering neural network, the neural network layer composed of a fuzzy clustering layer and a multilayer perceptron works in turn. When the fuzzy layer performs the initial operation of the classification task, the neural network layer serves as the final classifier, and finally the fuzzy clustering is used to improve the performance of the neural network classifier [20]. In recent years, the fuzzy clustering neural network has been applied in some studies [21, 22]. However, in literature [23], a hybrid fuzzy neural network method is proposed to minimize the problem of multilayer perceptron, improve its generalization ability, and reduce training time. The recurrent neural network is a neural network structure with closed-loop connections between neurons [24]. This neural network can achieve highly nonlinear dynamic mapping and has been used in some ECG signal classification studies [25, 26].

Table 1: The number of data in each category after slice interception and overlap processing.

Dataset	DS1	DS2	DS1′
N	5907	5874	5907
S	344	477	5847
V	1633	1468	5884
F	54	118	6060
Q	4	5	3303
Total	7942	7942	27001

3. Arrhythmia Data Preprocessing

3.1. Arrhythmia Dataset. This study uses the MIT-BIH arrhythmia data set [27]. The benchmark database was created by the Massachusetts Institute of Technology and Beth Israel Hospital in Boston, Massachusetts, USA, in 1980 and started to release. It is the first data set used to evaluate the performance of arrhythmia detectors and has been widely used in some famous studies [28, 29]. Each record in the data set is independently annotated and confirmed by two or more cardiologists, and the R wave peak value or local extreme value of the heartbeat is indicated.

3.2. Segmentation of Intrapatient Heartbeat. The QRS wave in the ECG signal data is located, and then a single heartbeat beat is extracted. First, use the 15-25HZ band-pass filter to obtain the QRS band and then perform the double-slope processing [30] to become a signal composed of single-mode peaks. After the preprocessing is completed, the ECG signal is located by the QRS wave through an adaptive threshold. Taking the QRS wave as the central reference, 100 and 150 sampling points are selected forward and backward, respectively, for the rough interception. At this time, the length of the heartbeat beat obtained by interception is 250 sampling points. The ECG signal is divided into normal heartbeat (74962), premature ventricular contraction (7034), right bundle branch block (7254), and left bundle branch block (8068). Since the number of four heartbeat types is extremely unbalanced, each type selects only 7000 heartbeats after segmentation and interception during model training and then randomly divides them into a training set and test set in half. Finally, the samples corresponding to the first 14,000 indexes in the cut sample are the training set, and the rest are used as the test set.

3.3. ECG Sequence Processing for Interpatient. In this paper, four records containing rhythmic ECG signals have been deleted from the MIT-BIH arrhythmia data set. The division method adopts interpatient, and the remaining ECG signal data is divided into training set DS1 and test set DS2. DS1 and DS2 contain 22 records of mixed conventional and complex arrhythmia, each data set with about 50,000 heartbeats [6, 31].

Due to the serious class imbalance in the data set used in this paper, especially the imbalance of training set data is very likely to cause the network model to learn invalid or even fail

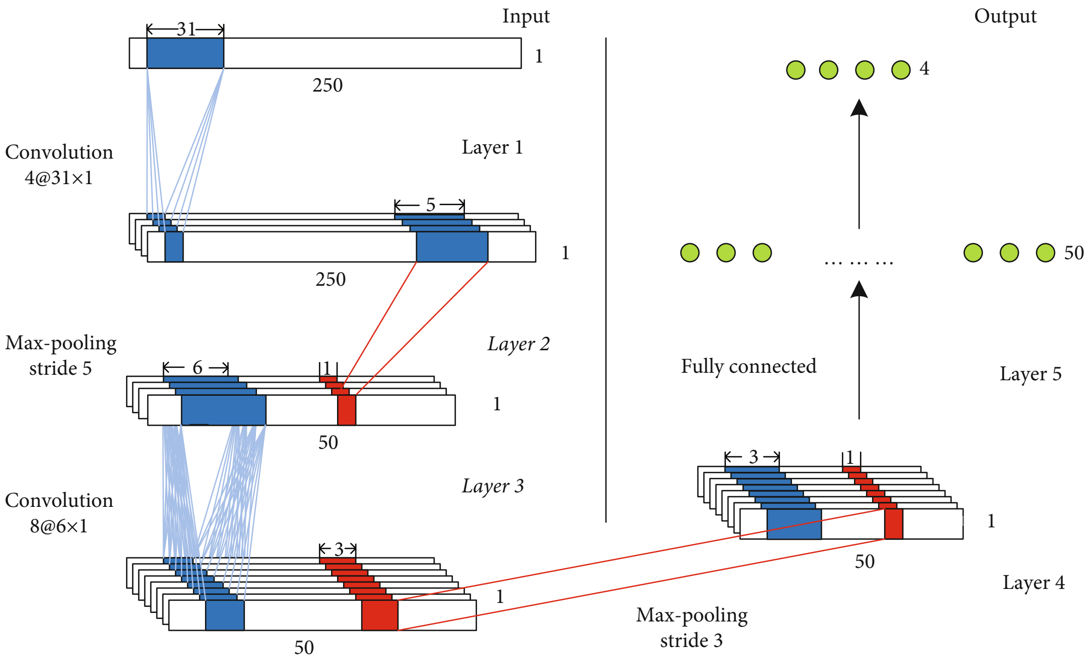


Figure 2: Schematic diagram of the convolutional network model.

Table 2: Convolutional neural network model parameter settings.

Parameter	Value	Parameter	Value
Input	250	Learning rate	0.01
Number of convolution kernels	2	Optimizer	SGD
Kernel size	31 × 1, 8 × 1	Iteration	50
Activation function	RELU	Batch	12
Pooling method	Average pooling	Output	4

to converge. To solve this problem, this paper adopts the method of slice and interception of ECG signal data, with the length of each interception segment being 5 s, and the amount of data is increased by overlapping between segments to alleviate the impact of category imbalance. Taking the most nonoverlapping category as the benchmark, the overlapping length of the remaining slices can be estimated by the following formula:

$$ol = \text{round}\left(L * \left(1 - \frac{n}{N}\right)\right), \quad (1)$$

where ol represents the length of the ECG signal overlap, round represents rounding, L represents the length of each slice, n represents the number of samples in the current category, and N represents the number of samples in the largest category. After training set, DS1 is processed as DS1$'$ according to formula (1). The number of different types of data in DS1$'$ basically reaches a balance with the N types. The number of each type after slice interception and overlap processing is shown in Table 1. Finally, the intercepted ECG signal is subjected to wavelet transform based on the db6 wavelet system [30] and Z-score standardization before resampling. The test set DS2 does not do overlap processing.

4. Architecture of the Deep Learning Network Model

4.1. Structure of the Convolutional Network Model for Intrapatient Paradigm. This paper proposes a five-layer convolutional neural network model for the classification of arrhythmia with normal beat (N), premature ventricular beat (V), right bundle branch block (R), and left bundle branch block (L) based on intrapatient. The local connection and weight sharing of convolutional neural networks reduce the number of network parameters, decrease the complexity of the model, and alleviate the problem of model overfitting, which has achieved great success in many fields such as computer vision. The structure of the network model is shown in Figure 2.

This paper proposed that the one-dimensional convolutional neural network model is composed of 2 convolutional layers, 2 pooling layers, and 1 fully connected layer, in which the convolutional layer of each layer will pass through a RELU activation function after the convolutional operation. The first layer of the convolution layer performs convolution operation on the input single heartbeat beat to extract local features. The size of the convolution kernel is set as 31 × 1, and the number of feature maps starts from 4. The size of

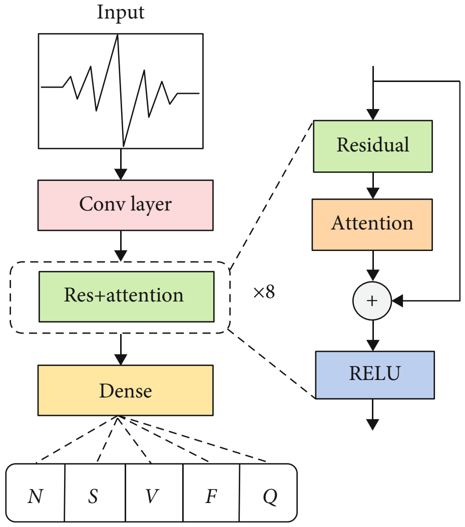


Figure 3: Schematic diagram of the attention residual network model.

the convolution kernel of the third convolution layer is set to 6 × 1, and the number of feature maps is 8. When the convolution operation is performed in the convolution layer, the movement step of the convolution kernel is set to 1. In the second and fourth layers, the average pooling operation is performed, the key feature information is extracted from the local features, and redundant features are discarded. The pooling step size is set to 5 and 3, respectively. The specific parameters of the network model are shown in Table 2.

4.2. Structure of the Attention Residual Network Model for Interpatient Paradigm. In this paper, the residual network module and attention mechanism are combined to form the attention residual unit, and the attention residual network model stacked by the attention residual unit is used to conduct the ablation study for interpatient paradigm. The structure of the network model is shown in Figure 3.

This paper uses the deep learning framework Keras and TensorFlow as the backend to build the model [32]. The size of the convolution kernel in the residual network is 32 × 32, the number of feature maps starts from 12, the weight of each layer is added with L2 regularization, the dropout probability value is set to 0.5 [33], and the value for small batch processing set to is 128, the initial value of the learning rate is set to 0.1, and the subsequent stepwise changes. Because the momentum optimizer has strong generalization ability in the ECG signal classification problem, to optimize the loss function, this paper uses stochastic gradient descent and momentum optimizer. The size of the convolution kernel of the attention module introduced in the network model is 7 × 7. The network model is optimized by adjusting the number of convolutional layers in the residual network and the number of convolution kernels in the attention module.

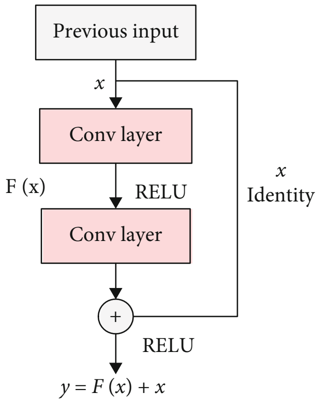


Figure 4: Schematic diagram of residual module.

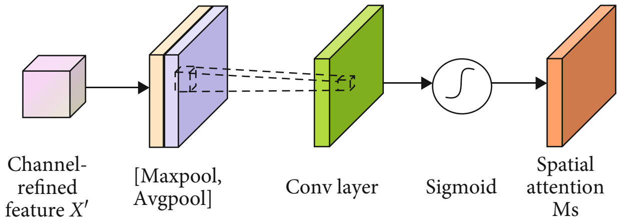


Figure 5: Spatial attention module.

4.2.1. Residual Network Model. The residual network is composed of the stack of residual module, and shows superior performance in the application of computer vision and other fields. The schematic diagram of the residual module is shown in Figure 4. The process can perform the following mathematical calculation:

$$Y = F(X, W) + X, \tag{2}$$

where X is the input of the residual module, $F(X, W)$ is the residual function, W is the weight parameter corresponding to the residual function, and Y is the output of the residual module.

For the arrhythmia diagnosis algorithm, a traditional neural network will more or less have information loss and waste when transmitting the information. It also maybe causes gradients to disappear or explode that making the deep network models unable to train. And the residual block inside the residual network uses the jump connection to directly pass the input information to the output by bypassing, protecting the integrity of information, and optimizing the problem of gradient disappearance caused by increasing the network depth in the neural network.

4.2.2. Attention Mechanism. The attention mechanism is a method of data processing in machine learning. It can be understood as a mechanism to redistribute resources based on the importance of the attention object to the originally allocated resources. The core idea is to find data based on the original data and then focus on some important features that inhibit unnecessarily. Because of the advantages of the attention mechanism, this paper proposes to introduce a spatial attention module into the residual network.

The spatial attention module [34, 35] uses the spatial relationship between features to generate a spatial attention map. The focus of attention is on the “where” of the feature

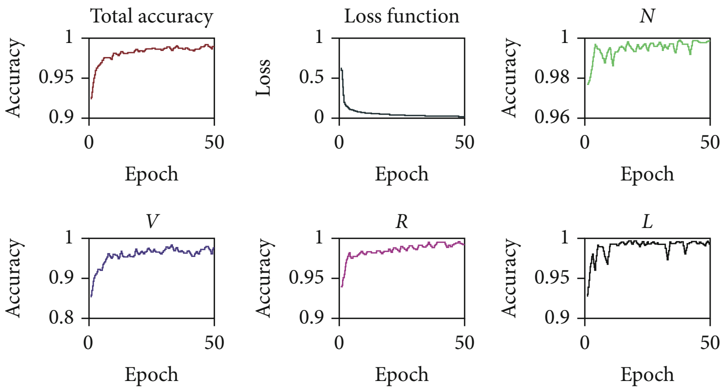


Figure 6: CNN model.

map that is the information part. The schematic diagram of the module is shown in Figure 5. The feature map X' is used as the input of the spatial attention module, and after Figure 5, the two-dimensional spatial attention map Ms can be obtained. The process can be summarized as

$$\mathrm{Ms}(X') = \sigma\Big(f^{7\times7}\Big(\Big[\mathrm{AveragePool}(X')\,;\mathrm{MaxPool}(X')\Big]\Big), \tag{3}$$

$$X'' = \mathrm{Ms}(X') \otimes X' \tag{4}$$

where AveragePool(X') means average pooling, MaxPool(X') means maximum pooling, $f^{7\times7}$ means convolution operation with a convolution kernel size of 7×7, σ is the sigmoid activation function, ⊗ represents element-wise multiplication, and X'' is the precise output obtained after passing through the spatial attention module.

4.2.3. Ablation Study. In this section, we conduct ablation experiments to better understand the effect of adding an attention module [36]. This paper uses the 18-layer residual network as the backbone architecture. By adding the attention module to the residual module, the network can more efficiently concentrate on the important information part of the ECG signal. Finally, analyze and compare the results of ablation experiments, and all experiments are performed on the same machine with the same parameter settings.

5. Experiments and Result Analysis

5.1. Model Evaluation Metrics. In this paper, we follow the classification standards of the American Association for the Advancement of Medical Instrumentation (AAMI) [37] and refers to other literature on the evaluation methods of the ECG signal classification in the MIT-BIH arrhythmia database, using accuracy (accuracy) and sensitivity (sensitivity), prediction rate (precision+), recall rate (recall), F1 value, and confusion matrix to evaluate the network model [35].

Table 3: Test results of model ablation study.

Type	Resnet		Resnet+attention	
	Se	P+	Se	P+
N	0.91	0.83	0.91	*0.93*
V	0.92	0.79	0.85	0.79
S	0.27	0.31	0.43	0.46
F	0.35	0.36	0.66	0.55
Q	0	0	0	0
Accuracy	0.86		0.87	
F1	0.84		0.87	

The final evaluation indicators are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{5}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$\text{Precision+} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

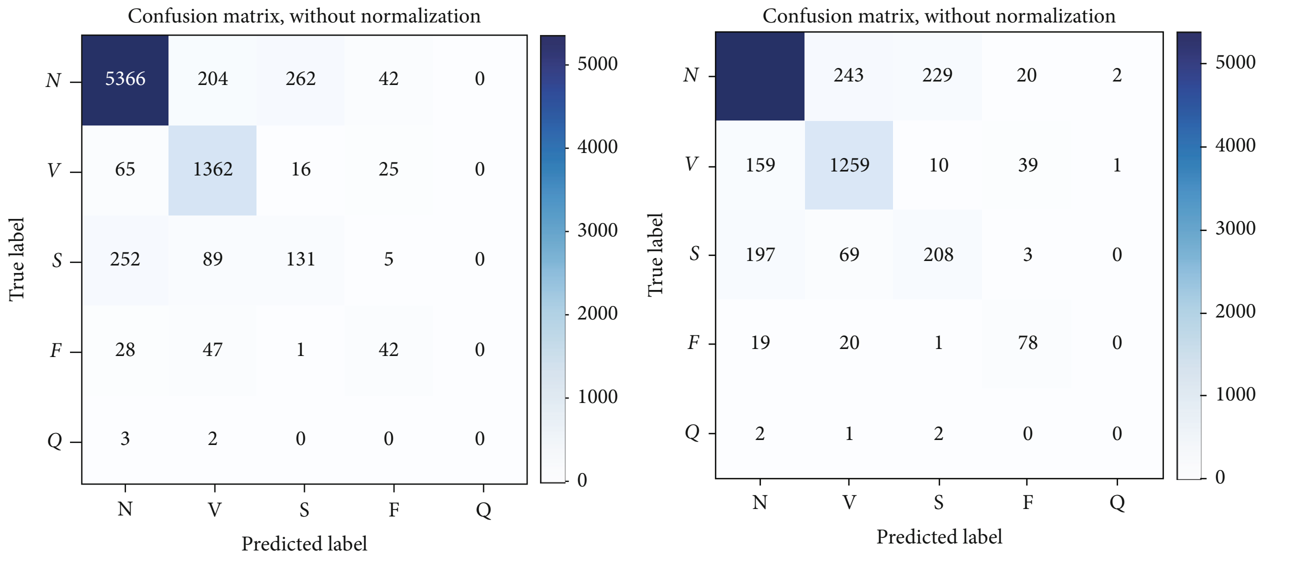
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$F1 = 2 \times \frac{\text{Presision} \times \text{Recall}}{\text{Precison} + \text{Recall}}, \tag{9}$$

where TP is a true positive sample, TN is a true negative sample, FP is a false positive sample, FN is a false negative sample, and $TP + TN + FP + FN$ is the total number of samples.

6. Result Analysis

6.1. Intrapatient Model Performance. The network modeled by the end-to-end deep learning method avoids manual extraction of data features, and the network can perform automatic feature extraction for classification. The center beat of the ECG signal is passed through a 5-layer convolutional neural network model. After continuous hyperparameter adjustments, the overall accuracy of the four types of classification is 99.03%, and the normal beat of the specific



(a) Resnet

(b) Resnet+attention

Figure 7: Confusion matrix of model ablation study.

four types (N) is as follows: 99.88%, premature ventricular beats (V): 97.83%, right bundle branch block (R): 99.12%, and left bundle branch block (L): 99.29%; the results are shown in Figure 6:

The model algorithm puts all the extracted heartbeats together and then divides the training set and the test set, without considering the differences between individuals, but this seems to be somewhat inconsistent with the actual scene. The labeled data that has been obtained in the actual scene comes from some old patients, and the model algorithm needs to predict new patients based on the rule of these data. At this time, the influence of individual differences will be reflected, making it difficult for the model we trained on old patient data to effectively generalize to the data of new patients, and individual differences will cause the deterioration of model performance.

6.2. Model Performance Interpatient. Through the ablation study on the MIT-BIH arrhythmia data set, the results of the two network models are shown in Table 3. Table 3 shows that the model that introduces the attention mechanism in the residual network has higher accuracy than the residual network model, and the F1 value has increased by 3%. In the residual network model, the sensitivity of the classification results for the two categories of normal heartbeat and supraventricular ectopic heartbeat is both above 90%. Comparing the attention residual network model with the residual network model, the prediction rate in the classification results of normal heartbeats has increased by 10%, but the sensitivity of the classification of supraventricular ectopic heartbeats has decreased. The sensitivity and prediction rate of the classification results of ventricular ectopic heartbeat and fusion heartbeat increased by 16%, 31%, 15%, and 19%, respectively, but neither model can classify unknown heartbeats. From the results of the ablation study, it can be seen that the introduction of the attention mechanism into the residual network greatly improves the diagnostic results of a normal heartbeat, supraventricular ectopic heartbeat, and fusion heartbeat and also increases the robustness of the network model.

The confusion matrix of the two models after the ablation study is shown in Figure 7. Observation shows that after the introduction of the attention mechanism in the residual network model, the predictions of N, S, and F categories in arrhythmia have been greatly improved. Among them, there are numerous mutual wrong predictions among N, V, and S types. So, it can be guessed that the waveform graphs of these three categories of the heartbeat are more similar.

Since the amount of original data in the Q category in the MIT-BIH data set used in this paper is very small, the Q category cannot be distinguished. In many literatures that use this dataset, it is basically indistinguishable. The accuracy rate in some literatures exceeds 90%, but the balance of the categories is not good. For example, in the literature [34], the accuracy rate reached 94.61%, but the Se of the S category was only 20%, the P+ was only 0.16%, and the P+ of the F category was only 0.52%. In the literature [9], the accuracy rate reached 93%, but in the five categories, F and Q categories could not be distinguished, the V category Se was only 70.8%, and the S category Se and P+ were 29.5% and 38.4%, respectively.

7. Conclusion

Due to the data set used in this paper is extremely unevenly distributed among categories, overlap processing is used when preprocessing the data, which increases the amount of data in each class and optimizes the overfitting problem of the proposed network model. In addition, compared with a single heartbeat, the sample obtained after arbitrary segment interception of a limited amount of data is much more complex, which enables the network model to get rid of the coupling problem with the QRS detection algorithm and makes the ECG signal diagnosis process more simple and generalized. The attention residual network model proposed in this paper greatly improves the *N*, *S*, and *F* types of arrhythmia, optimizes the performance of the network model, and increases the robustness of the model. The research results presented in this paper have positive significance for improving the accuracy of arrhythmia diagnosis, but limited by the small amount of data, it has a certain impact on the research. It is necessary to increase the number of data sets and try to combine other neural network structures to explore the arrhythmia diagnosis algorithm. Note that we will consider our further work into the device-to-device (D2D) and index modulation systems [38–40], which might automatically and adaptively monitor the arrhythmia situation.

Data Availability

All data, models, or code generated or used during the study are available from the corresponding author by request. (Ji Wang, email: 13902576499@163.com).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (61871645, J2024023). Innovation Training Program for College Students of Guangzhou University (Provincial) under grant S202011078027.

References

[1] National Center for Cardiovascular Diseases, *China Cardiovascular Health and Disease Report 2019*, Science Press, Beijing, 2020.

[2] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, “Wireless network intelligence at the edge,” *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

[3] J. Li, S. Dang, M. Wen, X.-Q. Jiang, Y. Peng, and H. Hai, “Layered orthogonal frequency division multiplexing with index modulation,” *IEEE Systems Journal*, vol. 13, no. 4, pp. 3793–3802, 2019.

[4] W. Duan, Y. Ji, J. Hou, B. Zhuo, M. Wen, and G. Zhang, “Partial-DF full-duplex D2D-NOMA systems for IoT with/without an eavesdropper,” *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6154–6166, 2019.

[5] X. Pei, W. Duan, M. Wen, Y. C. Wu, H. Yu, and V. Monteiro, “Socially-aware joint resource allocation and computation offloading in NOMA-aided energy harvesting massive IoT,” *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5240–5249, 2021.

[6] P. de Chazal, “Detection of supraventricular and ventricular ectopic beats using a single lead ECG,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*no. 1, pp. 45–48, Osaka, Japan, July 2013.

[7] S. Banerjee and M. Mitra, “Application of cross wavelet transform for ECG pattern analysis and classification,” *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 2, pp. 326–333, 2014.

[8] F. A. Elhaj, N. Salim, A. R. Harris, T. T. Swee, and T. Ahmed, “Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals,” *Computer Methods and Programs in Biomedicine*, vol. 127, no. 4, pp. 52–63, 2016.

[9] S. Chen, W. Hua, Z. Li, J. Li, and X. Gao, “Heartbeat classification using projected and dynamic features of ECG signal,” *Biomedical Signal Processing and Control*, vol. 31, no. 1, pp. 165–173, 2017.

[10] M. Javadi, S. A. A. A. Arani, A. Sajedin, and R. Ebrahimpour, “Classification of ECG arrhythmia by a modular neural network based on mixture of experts and negatively correlated learning,” *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 289–296, 2013.

[11] A. R. Hassan and M. A. Haque, “An expert system for automated identification of obstructive sleep apnea from single-lead ECG using random under sampling boosting,” *Neurocomputing*, vol. 235, no. 4, pp. 122–130, 2017.

[12] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi et al., “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[13] S. Shadmand and B. Mashoufi, “A new personalized ECG signal classification algorithm using block-based neural network and particle swarm optimization,” *Biomedical Signal Processing and Control*, vol. 25, no. 3, pp. 12–23, 2016.

[14] V. Moskalenko, N. Zolotykh, and G. Osipov, “Deep learning for ECG segmentation,” in *international conference on Neuroinformatics*, vol. 856, no. 9pp. 246–254, Springer, Cham, 2019.

[15] U. R. Acharya, S. L. Oh, Y. Hagiwara et al., “A deep convolutional neural network model to classify heartbeats,” *Computers in Biology and Medicine*, vol. 89, no. 10, pp. 389–396, 2017.

[16] T. Li and M. Zhou, “ECG classification using wavelet packet entropy and random forests,” *Entropy*, vol. 18, no. 8, p. 285, 2016.

[17] S. Xiguo and Q. Deng, “Reading and application of MIT-BIH arrhythmia database,” *Chinese Journal of Medical Physics*, vol. 21, no. 4, pp. 230–232, 2004.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[19] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” *Advances in neural information processing systems.*, pp. 5998–6008, 2017.

[20] M. A. Awal, S. S. Mostafa, M. Ahmad, and M. A. Rashid, “An adaptive level dependent wavelet thresholding for ECG denoising,” *Biocybernetics and Biomedical Engineering*, vol. 34, no. 4, pp. 238–249, 2014.

[21] S. N. Yu and Y. H. Chen, “Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network,” *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1142–1150, 2007.

[22] R. Ceylan, Y. Özbay, and B. Karlik, “A novel approach for classification of ECG arrhythmias: Type-2 fuzzy clustering neural network,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 6721–6726, 2009.

[23] Y. P. Meau, F. Ibrahim, S. A. L. Narainasamy, and R. Omar, “Intelligent classification of electrocardiogram (ECG) signal using extended Kalman filter (EKF) based neuro fuzzy

system," *Computer Programs in Biomedicine*, vol. 82, no. 2, pp. 157–168, 2006.

[24] L. V. Fausett, *Pearson. Fundamentals of neural networks: architectures, algorithms, and applications*, Prentice-Hall, 1994.

[25] E. Derya Übeyli, "Recurrent neural networks employing Lyapunov exponents for analysis of ECG signals," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1192–1199, 2010.

[26] S. Dutta, A. Chatterjee, and S. Munshi, "Identification of ECG beats from cross-spectrum information aided learning vector quantization," *Measurement*, vol. 44, no. 10, pp. 2020–2027, 2011.

[27] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[28] Y. H. Hu, S. Palreddy, and W. J. Tompkins, "A patient-adaptable ECG beat classifier using a mixture of experts approach," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 9, pp. 891–900, 1997.

[29] C. Ye, B. V. K. V. Kumar, and M. T. Coimbra, "An automatic subject-adaptable heartbeat classifier based on multi-view learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 6, pp. 1485–1492, 2016.

[30] Y. Wang, C. J. Deepu, and Y. Lian, "A computationally efficient QRS detection algorithm for wearable ECG sensors," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5641–5644, Boston, MA, USA, August 2011.

[31] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 205–220, 1992.

[32] F. Chollet, *Keras: The Python Deep Learning library*, Astrophysics Source Code Library, 2018.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, , pp. 3–19, Springer, 2018.

[35] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[36] E. J. D. S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: a survey," *Computer Methods and Programs in Biomedicine*, vol. 127, no. 4, pp. 144–164, 2016.

[37] A. Swami and R. Jain, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 10, pp. 2825–2830, 2013.

[38] Y. Ji, W. Duan, M. Wen et al., "Spectral efficiency enhanced cooperative device-to-device systems with NOMA," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2020.

[39] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz, "Emerging technologies for 5G-IoV networks: applications, Trends and Opportunities," *IEEE Network*, vol. 34, no. 5, pp. 283–289, 2020.

[40] M. Wen, X. Chen, Q. Li, E. Basar, Y.-C. Wu, and W. Zhang, "Index modulation aided subcarrier mapping for dual-hop OFDM relaying," *IEEE Transactions on Communications*, vol. 67, no. 9, pp. 6012–6024, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5576356, 9 pages
https://doi.org/10.1155/2021/5576356

Research Article

Resource Allocation for SWIPT Systems with Nonlinear Energy Harvesting Model

Yifan Hu, Mingang Liu, and Yizhi Feng

School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China

Correspondence should be addressed to Yizhi Feng; yzfeng@scut.edu.cn

Received 27 January 2021; Revised 5 March 2021; Accepted 19 March 2021; Published 7 April 2021

Academic Editor: Wei Duan

In this paper, we study the resource allocation for simultaneous wireless information and power transfer (SWIPT) systems with the nonlinear energy harvesting (EH) model. A simple optimal resource allocation scheme based on the time slot switching is proposed to maximize the average achievable rate for the SWIPT systems. The optimal resource allocation is formulated as a nonconvex optimization problem, which is the combination of a series of nonconvex problems due to the binary feature of the time slot-switching ratio. The optimal problem is then solved by using the time-sharing strong duality theorem and Lagrange dual method. It is found that with the proposed optimal resource allocation scheme, the receiver should perform EH in the region of medium signal-to-noise ratio (SNR), whereas switching to information decoding (ID) is performed when the SNR is larger or smaller. The proposed resource allocation scheme is compared with the traditional time switching (TS) resource allocation scheme for the SWIPT systems with the nonlinear EH model. Numerical results show that the proposed resource allocation scheme significantly improves the system performance in energy efficiency.

1. Introduction

In traditional energy-constrained wireless networks, the wireless devices normally use batteries as energy source and require periodically recharging or replacing the batteries, which is difficult for a large number of wireless devices and even hazardous or impossible in some circumstances [1], resulting limited lifetime of the wireless devices and the networks.

Energy harvesting (EH) that allows the energy-limited wireless devices to harvest energy from the ambient environment is a promising solution for extending the lifetime of energy-constrained wireless networks. Among the EH technologies, simultaneous wireless information and power transfer (SWIPT) takes advantage of the radio frequency (RF) signal's ability of carrying both information and energy at the same time, providing great convenience of recharging to energy-constrained devices by harvesting energy from the RF signals. SWIPT is especially suitable for the wireless terminals with low-power consumption whereas hard to access.

The SWIPT technique has gained wide attention from both researchers and engineers since Varshney proposed the idea in 2008 [1]. In [2], the trade-off between the amount of the harvested energy and the achievable rate is studied for the SWIPT systems in the frequency selection channel with additive white Gaussian noise (AWGN). In [3], two kinds of SWIPT receivers, namely, time switching (TS) and power splitting (PS) receivers, are, respectively, proposed. Since they were proposed, the TS and PS receivers have attracted a lot of interest due to the simplicity of realization [4–8]. Specifically, the trade-off between the information and energy transmission is studied for the point-to-point single-input single-output (SISO) systems with a PS receiver in [4], which is extended to the point-to-point multiple-input single-output (MISO) systems with a TS receiver in [5]. In [6], the system throughput maximization is proposed for the MISO systems with TS and PS receivers. The TS and PS ratios are optimized to maximize the weighted sum rate of all receivers for the multiuser SISO orthogonal frequency-division multiplexing (OFDM) systems in [7]. Aiming at the minimization for the transmission power, the power allocation problem for the multiuser MISO downlink system is studied, and the optimal PS ratio is obtained in [8].

Besides the receiver design and the trade-off between the energy harvesting and information decoding, one of the other key issues in the implementation of SWIPT is efficient resource allocation [9–15]. In [9], the authors study the power and subcarrier allocation schemes for energy-efficient SWIPT in multicarrier systems. In [10], the secrecy rate maximization is studied in an OFDM secrecy communication system. A multiuser OFDM system for maximizing the sum rate with a minimum transmit power constraint is designed in [11]. In [12], an optimal resource allocation policy is derived in a generalized WPCN where the devices can harvest energy from both multiple-antenna power station and ambient energy harvesting. In [13], the energy efficiency maximization is considered in large-scale multiple-antenna SWIPT systems. In [14], the authors propose an energy efficiency maximization optimization scheme for the multiuser multicarrier energy-constrained amplify-and-forward (AF) multirelay network. In [15], a robust energy-efficient optimization is designed for MIMO two-way relay networks with SWIPT.

Most of the aforementioned works about SWIPT systems consider the linear EH model, where the power conversion efficiency factor of the EH receiver is assumed to be a constant. However, it is found that the power conversion efficiency of the practical RF to direct current (DC) converter is affected by the input power, i.e., when the input power is greater than a certain threshold, the output power changes nonlinearly with the input power and shows a saturation characteristic [16]. Hence, the linear EH model cannot properly model the practical EH implementations and may lead to the resource mismatch in the resource allocation or the overestimation in the system performance evaluation [17]. In [16, 18], the parametric and logistic function-based nonlinear EH model and the piece-wise linear EH model are, respectively, proposed to capture the nonlinear saturation input-output characteristic of the practical EH circuit, which are further exploited for the various scenarios [19–23]. Specifically, in [19], the authors study the joint transmit power allocation and receive power splitting for SWIPT systems with the realistic nonlinear EH model. Considering a Nakagami-m channel, the authors investigate the outage probability and reliable throughput of a multiuser wireless-powered SWIPT system in [20]. In [21], the analytical results on outage probability performance are presented for a cooperative relay-aid network with spectrum sensing and energy harvesting. In [22], the power split factor is optimized to minimize the outage probability for the AF relay system with PS receiver and the nonlinear EH model. In [23], considering imperfect channel state information (CSI) conditions, the authors analyse the outage probability for the multirelay SWIPT systems with PS receivers and the nonlinear EH model.

In this paper, we consider the single-input single-output (SISO) point-to-point SWIPT communication systems with TS receiver. The piece-wise linear EH model is considered to model the nonlinear saturation input-output characteristic for the EH circuit. We propose a simple optimal resource allocation scheme based on the time slot-switching strategy to maximize the average achievable rate for the systems. The information transmission block time T is divided into N time slots. Each time slot is used for information decoding (ID) or EH according to the optimal scheme. The optimal problem is formulated as a nonconvex optimization, which is first proved to satisfy the time-sharing condition and then solved by using the time-sharing strong duality theorem and Lagrange dual method.

Compared with work [24] where the linear EH model is considered, the major contribution of our work is that we consider the effect of the saturation characteristic of the practical nonlinear EH model and derive a more realistic optimal resource allocation scheme.

The rest of this paper is organized as follows. “System Model” introduces the system model. The optimal resource allocation design is proposed and the optimal problem is solved in “Design of Resource Allocation Optimization Algorithm.” In “Numerical Results,” the numerical results and discussion are presented. “Conclusions” concludes this paper.

2. System Model

We consider a SISO point-to-point SWIPT communication system as shown in Figure 1, where both the receiver R_x and the transmitter T_x have single antenna and R_x is assumed to be energy-limited and could harvest energy from the received signals with a TS scheme. We assume that the channel between the transmitter and receiver is subjected to frequency flat and the block Rayleigh fading. The channel coefficient is denoted as h, which is a random variable following the complex Gaussian distribution with zero mean and variance σ^2. Without loss of generality, we assume that each time slot is normalized transmission time. In each time slot, the TS ratio β is equal to 0 or 1, indicating that the receiver implements EH or ID operations, respectively. The received signal can be expressed as

$$y = \sqrt{\theta P} h x + n_a, \tag{1}$$

where P is the transmit power, x is the data symbol with unity power, i.e., $E[|x|^2] = 1$ where $E[\,]$ means the mathematical expectation, $n_a \sim CN(0, \sigma^2)$ is the channel noise, $\theta = d^{-m}$ represents the path loss where d is the distance between the source and the destination nodes, and m represents the path-loss exponent. The achievable rate of the system based on the TS scheme can be expressed as

$$R(\beta) = \beta \log_2 \left(1 + \frac{\theta P H}{\sigma^2} \right), \tag{2}$$

where $\mathrm{H} = |h|^2$ is the channel power gain.

We use the piece-wise linear function to model the nonlinear saturation input-output characteristic of the EH

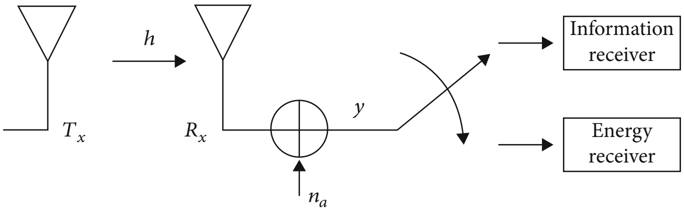


FIGURE 1: The SISO point-to-point SWIPT system.

receiver. The harvested power at the EH receiver is then given as [17, 18]

$$P_o = \begin{cases} \eta\theta PH, \eta\theta PH < P_s \\ P_s, \eta\theta PH \geq P_s \end{cases} \tag{3}$$

where $\eta (0 < \eta < 1)$ is the energy conversion efficiency of the energy harvester in the linear region, P_s is the maximum saturation harvested power of the EH receiver. As shown in (3), when the conversion power of the energy receiver $\eta\theta PH$ exceeds the saturation output power P_s, the output power of the energy receiver remains unchanged and some of the power is wasted, which means that in such case, the time slot should be switched to the ID receiver rather than EH receiver to avoid the waste of the power. Hence, the resource allocation scheme should be redesigned for the practical SWIPT system considering the nonlinear input-output characteristic of the EH circuit.

3. Design of Resource Allocation Optimization Algorithm

In this section, we propose an optimal resource allocation scheme based on the simple time slot-switching strategy to achieve the balance between the maximum average achievable rate and the maximum average harvested energy.

From (3), the harvested energy can be expressed as

$$Q(\beta) = (1 - \beta)P_o. \tag{4}$$

We consider maximizing the average achievable rate for the SISO SWIPT systems as shown in Figure 1. The optimization problem is formulated as

$$\max_{\beta} E[R(\beta)], \tag{5}$$

$$\begin{aligned} \text{s.t.} \quad & E[Q(\beta)] \geq \bar{Q} \\ & \beta \in \{0, 1\}, \end{aligned} \tag{6}$$

where $\bar{Q}$ is the minimum amount of the harvested energy required to maintain the normal operation of the EH receiver.

In general, it is difficult to solve the optimal problem (5) directly since it is the combination of a series of nonconvex problems due to the binary feature of β. In addition, the complexity of solving the optimization problem using a numerical calculation method increases exponentially with the number of the time slots. In this section, to solve the optimization problem (5), we use the time-sharing strong duality theorem proposed in [25], which is given as follows.

We first prove that the optimization problem (5) satisfies the time-sharing condition (for more detail about the definition of time-sharing condition, the reader is referred to Ref. [25]).

Proposition 1. *Let β_x and β_y be the optimal solutions of (5) with $\bar{Q} = Q_x$ and $\bar{Q} = Q_y$, respectively. Then, for any $\gamma \in [0, 1]$, there exists a feasible solution β_z to the optimization problem (5) such that*

$$E[R(\beta_z)] \geq \gamma E[R(\beta_x)] + (1 - \gamma)E\left[R\left(\beta_y\right)\right], \tag{7}$$

$$E[Q(\beta_z)] \geq \gamma Q_x + (1 - \gamma)Q_y, \tag{8}$$

and hence, the optimization problem (5) satisfies the time-sharing condition.

Proof. Let β_x be the feasible solution to the optimization problem (5).

(1) When $\gamma = 0$, let $\beta_z = \beta_y$. Since β_y is the optimal solution of the problem (5) when $\bar{Q} = Q_y$, the feasible solution β_z obviously satisfies Equations (7) and (8)

(2) When $\gamma = 1$, let $\beta_z = \beta_x$. Since β_x is the optimal solution of the problem (5) when $\bar{Q} = Q_x$, the feasible solution β_z also satisfies Equations (7) and (8)

(3) When $0 < \gamma < 1$, let H_k denote the channel power gain of the kth$(1 \leq k \leq N)$ time slot. The average achievable rate and the average harvested energy can be, respectively, expressed as

$$\begin{aligned} E\left[R\left(\beta_j\right)\right] &= \lim_{N \longrightarrow \infty} \frac{1}{N} \sum_{k=1}^{N} R_k\left(\beta_{j,k}\right), \\ E\left[Q\left(\beta_j\right)\right] &= \lim_{N \longrightarrow \infty} \frac{1}{N} \sum_{k=1}^{N} Q_k\left(\beta_{j,k}\right), \end{aligned} \tag{9}$$

where $j \in \{x, y, z\}$, $\beta_{j,k}$ is the value of β_j in the kth time slot. Define an integer M such that $M = \lceil \gamma N \rceil$, where $\lceil \cdot \rceil$ means ceiling round operation. Let

$$\beta_{z,k} = \begin{cases} \beta_{x,k}, k = 1, 2, \cdots, M \\ \beta_{y,k}, k = M + 1, M + 2, \cdots, N. \end{cases} \tag{10}$$

When $N \longrightarrow \infty$, it can be obtained that $M \longrightarrow \infty$ and $(N - M) \longrightarrow \infty$. Then, for any $\gamma \in (0, 1)$, it can be derived that

$$\begin{aligned} E[R(\beta_z)] &= \lim_{N \longrightarrow \infty} \frac{1}{N} \sum_{k=1}^{N} R_k(\beta_{z,k}) \\ &= \lim_{N \longrightarrow \infty} \left[\frac{\gamma}{M} \sum_{k=1}^{M} R_k(\beta_{x,k}) + \frac{1-\gamma}{N-M} \sum_{k=M+1}^{N} R_k\left(\beta_{y,k}\right) \right], \\ &= \lim_{M \longrightarrow \infty} \frac{\gamma}{M} \sum_{k=1}^{M} R_k(\beta_{x,k}) + \lim_{N-M \longrightarrow \infty} \frac{1-\gamma}{N-M} \sum_{k=M+1}^{N} R_k\left(\beta_{y,k}\right) \\ &= \gamma E[R(\beta_x)] + (1-\gamma) E\left[R\left(\beta_y\right)\right], \end{aligned} \tag{11}$$

which means that the feasible solution β_z satisfies the equation (7). Similarly, it can be derived that

$$\begin{aligned} \mathrm{E}[\mathrm{Q}(\beta_z)] &= \lim_{\mathrm{N} \longrightarrow \infty} \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=1}^{\mathrm{N}} \mathrm{Q}_{\mathrm{k}}(\beta_{\mathrm{z,k}}) \\ &= \lim_{\mathrm{N} \longrightarrow \infty} \left[\frac{\gamma}{\mathrm{M}} \sum_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{Q}_{\mathrm{k}}(\beta_{\mathrm{x,k}}) + \frac{1-\gamma}{\mathrm{N}-\mathrm{M}} \sum_{\mathrm{k}=\mathrm{M}+1}^{\mathrm{N}} \mathrm{Q}_{\mathrm{k}}\left(\beta_{\mathrm{y,k}}\right) \right] \\ &= \lim_{\mathrm{M} \longrightarrow \infty} \frac{\gamma}{\mathrm{M}} \sum_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{Q}_{\mathrm{k}}(\beta_{\mathrm{x,k}}) + \lim_{\mathrm{N}-\mathrm{M} \longrightarrow \infty} \frac{1-\gamma}{\mathrm{N}-\mathrm{M}} \sum_{\mathrm{k}=\mathrm{M}+1}^{\mathrm{N}} \mathrm{Q}_{\mathrm{k}}\left(\beta_{\mathrm{y,k}}\right), \\ &= \gamma \mathrm{E}[\mathrm{Q}(\beta_{\mathrm{x}})] + (1-\gamma) \mathrm{E}\left[\mathrm{Q}\left(\beta_{\mathrm{y}}\right)\right] \geq \gamma \mathrm{Q}_{\mathrm{x}} + (1-\gamma) \mathrm{Q}_{\mathrm{y}} \end{aligned} \tag{12}$$

which means that the feasible solution β_z satisfies Equation (8). From (11) and (12), Proposition 1 is proved.

Since the optimization problem (5) satisfies the time-sharing condition, according to the *Time-Sharing Strong Duality Theorem*, the primal problem (5) has the same optimal solution as its dual problem and can be solved by the Lagrange dual method.

The Lagrange function of (5) can be expressed as

$$L(\beta, \lambda) = E[R(\beta)] + \lambda(E[Q(\beta)] - Q), \tag{13}$$

where $\lambda \geq 0$ is the Lagrange multiplier associated with $E[R(\beta)] \geq Q$. Accordingly, the Lagrange dual function can be expressed as

$$g(\lambda) = \max_{\beta \in \{0,1\}} L(\beta, \lambda). \tag{14}$$

The dual problem is then given as

$$\begin{aligned} \min \quad & g(\lambda) \\ \text{s.t.} \quad & \lambda \geq 0, \end{aligned} \tag{15}$$

In order to effectively solve the above dual problem, we first decouple the optimization problem (14) into parallel N subproblems that has the same structure as (14). The kth ($k = 1, 2, \cdots N$) subproblem can be expressed as

$$\max_{\beta \in \{0,1\}} L_k(\beta), \tag{16}$$

where $L_k(\beta) = R(\beta) + \lambda Q(\beta)$. In order to solve the optimization problem (16), we compare the value of the objective function $L_k(\beta)$ when $\beta = 1$ or $\beta = 0$, which can be expressed as

$$\begin{aligned} L_k(\beta = 1) &= \log_2\left(1 + \frac{\theta P H}{\sigma^2}\right), \\ L_k(\beta = 0) &= \lambda P_o, \end{aligned} \tag{17}$$

respectively. Hence, in the optimal solution of problem (14), β^* can be expressed as

$$\beta^* = \begin{cases} 1, \log_2\left(1 + \dfrac{\theta P H}{\sigma^2}\right) > \lambda P_o \\ 0, \text{else} \end{cases}. \tag{18}$$

Then, for a given value λ, β^* can be obtained from (18) according to the channel state in each time slot. Let λ^* be the optimal dual variable, which is associated with the required minimum harvested energy value $\bar{Q}$ in the inequality constraint term in (5). The optimal dual variable λ^* can be obtained by iterative search and updating until the average energy collection meets the minimum energy constraint, i.e., $E[Q(\beta)] = \bar{Q}$, for which the detailed iterative search algorithm will be discussed later.

The proposed resource allocation scheme is based on the optimal TS strategy according to the channel state in each time slot. To describe the optimal switching strategy (18) more clearly, from (3) and (18), we define two functions G_1 and G_2 with respect to the channel power gain

$$G_1(H) = \log_2\left(1 + \frac{\theta P H}{\sigma^2}\right) - \lambda^* \eta \theta P H, \tag{19}$$

$$G_2(H) = \log_2\left(1 + \frac{\theta P H}{\sigma^2}\right) - \lambda^* P_s. \tag{20}$$

Equation (19) is not easy to solve because the first and second terms for $G_1(H)$ are, respectively, logarithmic and linear functions of H. In this paper, we solve it by traversing the value of H from 0, when the difference between the value of the first and second terms is close (we set it as 10^{-5}), we consider the values of the two terms to be equal, i.e., $G_1(H) = 0$; then, we can obtain an approximate nonzero real root H_1. Next, it can be found by derivation that $G_1(H)$ is a function that increases first and then decreases, and the position of the point that changes its monotonicity is related to the value of λ. Obviously, we can find a λ that takes the position in $(0, H_1)$. At this time, since $G_1(0) = 0$, $G_1(H_1) = 0$, when $H \in (0, H_1)$, $G_1(H) > 0$, from (3) and (18), $\log_2(1 + (\theta PH/\sigma^2)) > \lambda^* P_0 = \lambda^* \eta\theta PH$, and hence, $\beta^* = 1$. Similarly, it can be obtained by derivation that $G_2(H)$ is an increasing function, when $H \in (H_2, \infty)$, $\log_2(1 + (\theta PH/\sigma^2)) > \lambda^* P_0 = \lambda^* P_s$, $\beta^* = 1$, where H_2 is the nonzero real root of equation $G_2(H) = 0$. Then, it can be deduced that when $H \in [H_1, H_2]$, $\beta^* = 0$. Therefore, the optimal TS strategy can be expressed as

$$\beta^*(H) = \begin{cases} 1, & H < H_1 \text{ or } H > H_2 \\ 0, & H_1 \le H \le H_2. \end{cases} \tag{21}$$

In (21), the optimal TS thresholds H_1 and H_2 depend on the optimal dual variables λ^*, which is determined by the inequality constraints $\bar{Q}$ in (5) and should be chosen so that $E[Q(\beta(H))] = \bar{Q}$. The average energy collection can be expressed as

$$E[Q(\beta(H))] = \int_{H_1}^{H_{th}} \eta\theta Px f_H(x) dx + \int_{H_{th}}^{H_2} P_s * f_H(x) dx, \tag{22}$$

where $f_H(x)$ is the probability density function (pdf) of H, $H_{th} = P_s/\eta\theta P$. The iterative search algorithm for the optimal dual variables λ^* is summarized as follows in Algorithm 1. The initial value of λ_0 is set as 1.0, and $\Delta\lambda$ is set as 0.01.

The optimal resource allocation scheme can then be described as follows. Firstly, the information transmission block time T is divided into N time slots. Then, for each time slot, say, the kth$(1 \le k \le N)$ time slot, the optimal TS thresholds H_1 and H_2 are determined by using Algorithm 1. Finally, the channel power gain H_k is compared with H_1 and H_2 in the kth$(1 \le k \le N)$ time slot; if $H_k < H_1$ or $H_k \ge H_2$, the receiver switches to information decoding or else switches to energy harvesting. The optimal resource allocation scheme is summarized as follows in Algorithm 2.

Remark. It is shown in (21) and Algorithm 2 that for the SWIPT systems with nonlinear EH model, the optimal resource allocation scheme based on the time slot-switching strategy requires that the receiver switches to information decoding when the signal-to-noise ratio (SNR) is larger or smaller, whereas switching to energy harvesting is performed in the region of medium SNR.

4. Numerical Results

In this section, we present simulation results for the proposed optimal resource allocation scheme for the SWIPT systems with nonlinear EH model. In order to validate the proposed scheme, we compare the proposed resource allocation scheme with the traditional TS resource allocation scheme [26] in the energy efficiency performance for the SWIPT systems with nonlinear EH model, where the energy efficiency for the traditional TS scheme is obtained by sweeps, the time switching factor from 0 to 1 with a step 0.01, and the energy efficiency is defined as

$$\eta_E = R(\beta)/(P - E[Q(\beta)]). \tag{23}$$

Unless otherwise specified, the transmission power is set to $P = 1W$, and the required minimum harvested energy is set to $\bar{Q} = 5\,\text{mW}$. The energy conversion efficiency in the linear region and the saturation output power of the nonlinear EH receiver are set to $\eta = 0.8$ and $P_s = 24\,\text{mW}$ [17], respectively. The variance of the additive white Gaussian noise is $\sigma^2 = N_0$, and the SNR is defined as $\text{SNR} = P/N_0$. The information transmission block time T is divided into $N = 10^5$ subslots, and in each subslot, the channel obeys the Rayleigh distribution and is independent of each other. The distance between the source node and the destination node is set to $d = 5\,\text{m}$, and the pathloss exponent m is 2.0.

Figure 2 shows the energy efficiency versus the transmit power P for the SISO SWIPT systems with nonlinear EH model under various SNRs. It is shown that when the transmit power increases, the energy efficiency decreases in the region of middle and higher SNRs (i.e., $\text{SNR} = 15 \text{ or } 25\,\text{dB}$), whereas it keeps almost invariant when the SNR is smaller, i.e., $\text{SNR} = 5\,\text{dB}$. It can be observed that compared with traditional TS resource allocation scheme, the proposed resource allocation scheme significantly improves the system performance in energy efficiency. Furthermore, it can be also observed that the gap of system performance between the proposed scheme and traditional TS scheme becomes more obvious as the SNR increases, indicating that the proposed performs better in the region of higher SNR. The reason is that, when the SNR increases, the EH receiver is more likely to work in the nonlinear region. Due to the saturation characteristic of the nonlinear EH model, the traditional TS scheme is more likely to waste the received power in such case, thus resulting in the larger performance gap.

In Figure 3, the energy efficiency of the SISO SWIPT systems with nonlinear EH model under various SNRs is plotted against the distance d between the source and the destination nodes. It can be observed that the energy efficiency decreases when the distance d increases. It shows similar results in Figure 2 that the proposed resource allocation scheme substantially outperforms the traditional TS resource allocation scheme and that the performance gap gets larger when SNR increases. Moreover, it can be observed that as the distance d increases, the energy efficiency of both schemes and the performance gap between the two schemes tend to be zero in the lower SNR region, since as the distance d increases, the received power of the signal becomes very weak; thus, the achievable rate is very small in the lower SNR region for both the schemes.

Figure 4 shows the energy efficiency versus the pathloss exponent m for the SISO SWIPT systems with the nonlinear EH model under various SNRs. Similar results can be observed in Figure 3 that the energy efficiency decreases when the pathloss exponent m increases. The reason is that a larger value of m means more transmit loss of the power of the signal, which brings a lower received SNR and smaller achievable rate for the system. It can be also observed that compared with the distance d, the system performance is more susceptible to the variation of the pathloss exponent m, since the pathloss exponent m has a greater impact on pathloss than the source-destination distance d. Also, it is shown that the proposed resource allocation scheme substantially outperforms the traditional TS resource allocation

```
Initialize: $i \longleftarrow 0, \lambda_0, \Delta\lambda, \delta, i_{\max}$
repeat
    Obtain $G_1(H)$ and $G_2(H)$ using (19) and (20), respectively
    Obtain $H_{1_i}$ and $H_{2_i}$ by solving $G_1(H) = 0$ and $G_2(H) = 0$, respectively
    Obtain $E[Q(\beta(H))]$ using (22), then $mut \longleftarrow E[Q(\beta(H))] - \bar{Q}$
    if$|mut| \le \delta$then
        return Optimal dual variable $\lambda^* = \lambda_i$ and optimal TS threshold $H_1^* = H_{1_i}, H_2^* = H_{2_i}$
    else
        $i \longleftarrow i + 1, \lambda_{i+1} \longleftarrow \lambda_i + \Delta\lambda$
    end if
until$|mut| \le \delta$ or $i = i_{\max}$
```

ALGORITHM 1: Iterative search algorithm for $\lambda^{*\cdot}$.

```
Initialize:$N, k \longleftarrow 0$
repeat
    if$H_1^* \le H_k \le H_2^*$then
        $\beta^* = 0$, and let the $k$th subslot do energy harvesting
    else
        $\beta^* = 1$, and let the $k$th subslot do information decoding
    end if
    $k \longleftarrow k + 1$
until$k = N$
```

ALGORITHM 2: Resource allocation algorithm for the SWIPT systems with nonlinear EH model.

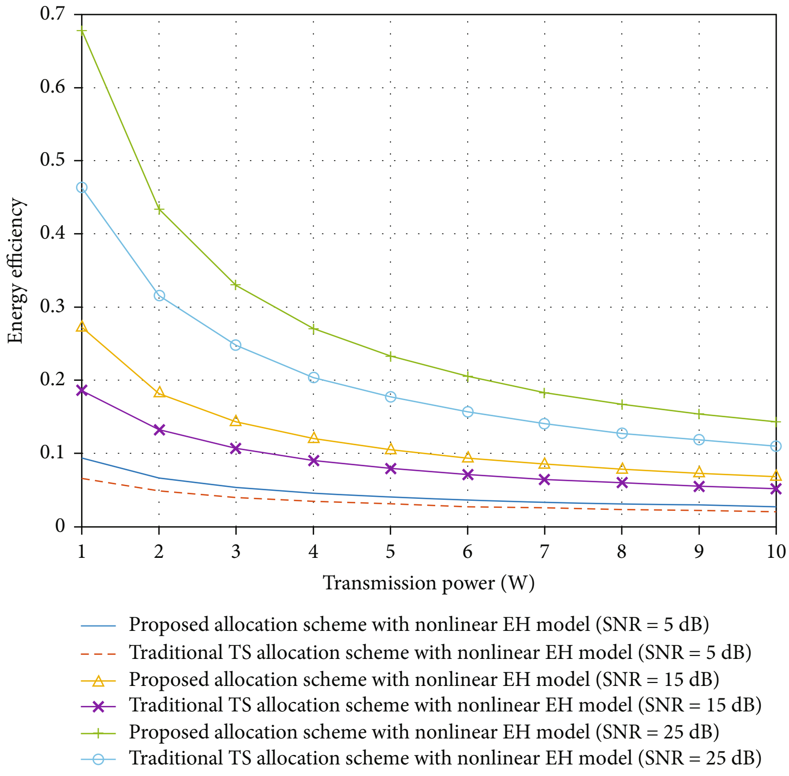


FIGURE 2: Energy efficiency versus transmission power for the SISO SWIPT systems with various signal-to-noise ratios (SNRs).

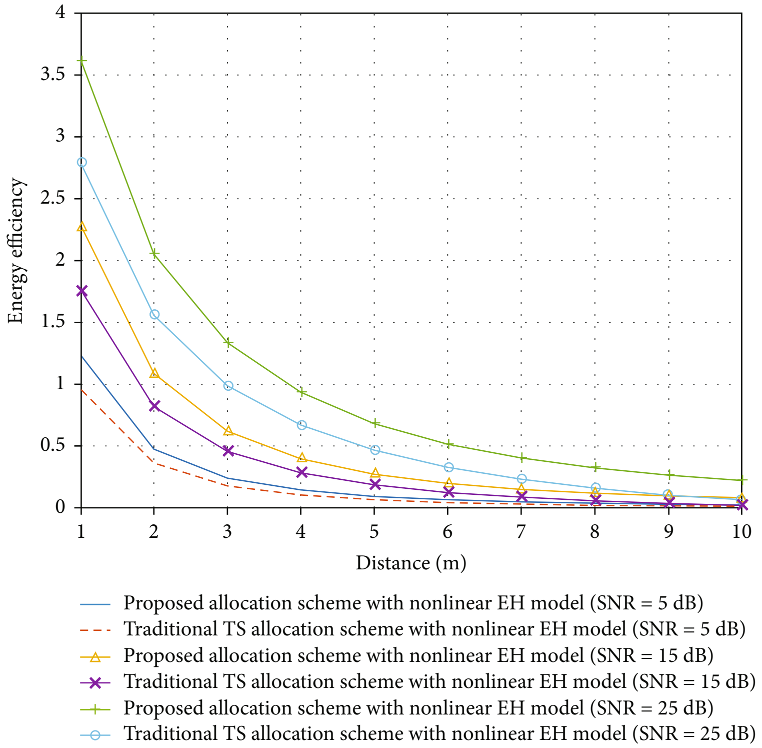


Figure 3: Energy efficiency versus distance for the SISO SWIPT systems with various signal-to-noise ratios (SNRs).

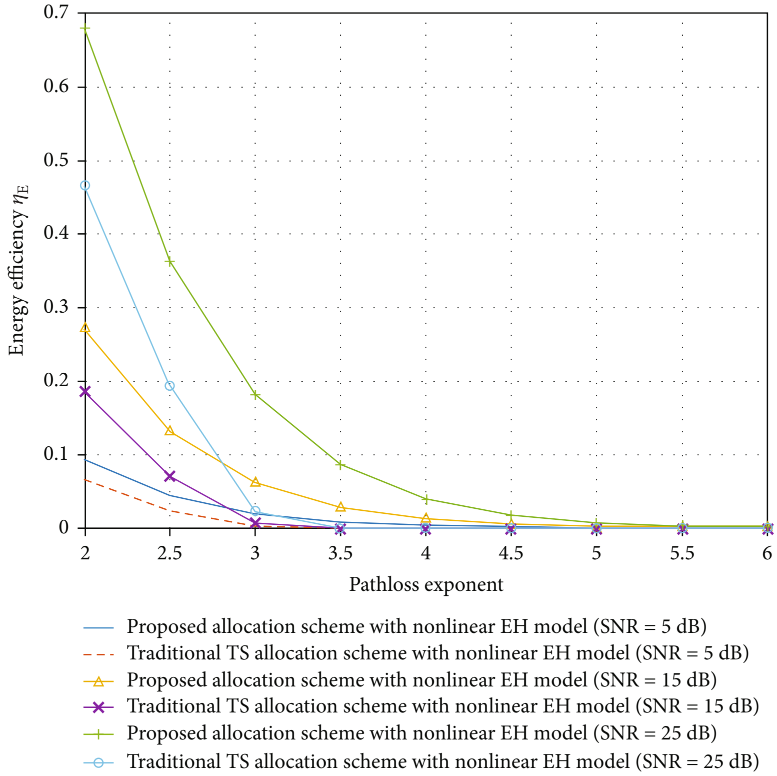


Figure 4: Energy efficiency versus pathloss exponent for the SISO SWIPT systems with various signal-to-noise ratios (SNRs).

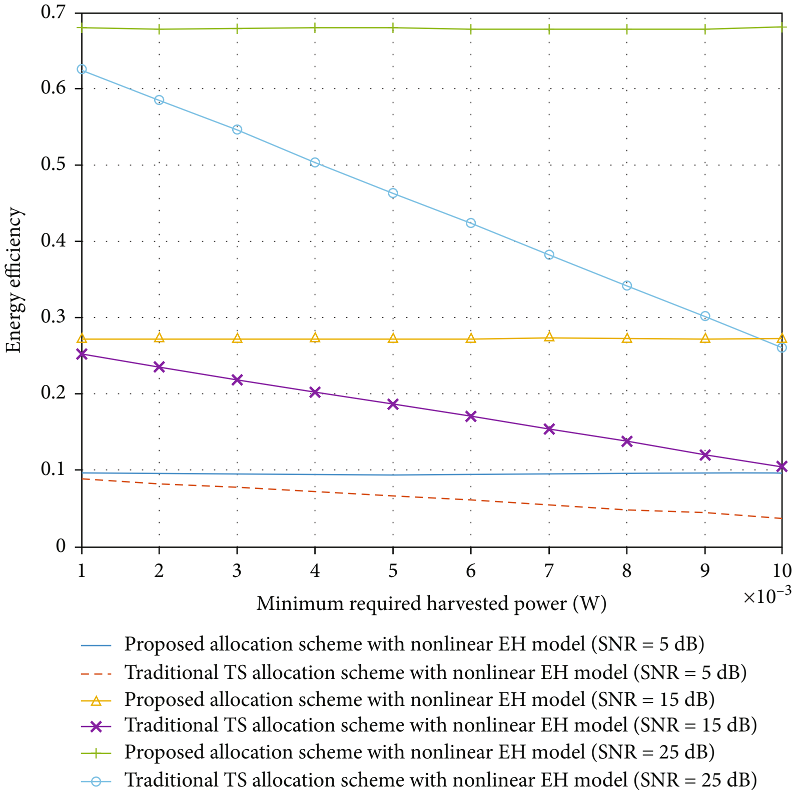


Figure 5: Energy efficiency versus minimum required harvested energy for the SISO SWIPT systems with various signal-to-noise ratios (SNRs).

scheme for smaller pathloss exponent m, whereas the energy efficiency of both schemes and the performance gap between the two schemes tend to be zero in lower SNR region when the pathloss exponent m increases.

In Figure 5, the energy efficiency versus the minimum required harvested energy $\bar{Q}$ is shown for the SISO SWIPT systems with the nonlinear EH model under various SNRs. Not surprisingly, the proposed scheme substantially outperforms the traditional TS scheme despite the variation of the minimum required harvested energy. In fact, the energy efficiency of the system with the traditional TS scheme decreases with the increases of the minimum required harvested energy $\bar{Q}$, whereas for the proposed time slot-switching scheme, the system energy efficiency keeps almost unchangeable when the minimum required harvested energy $\bar{Q}$ increases from 1 mW to 10 mW.

5. Conclusions

In this paper, we have studied the resource allocation scheme for the point-to-point SISO SWIPT systems with the nonlinear EH model. We have proposed an optimal resource allocation scheme based on the time slot switching to maximize the average information rate for the systems, with which the receiver performs information decoding in the region of higher or lower SNRs, whereas switching to energy harvesting is performed in the region of medium SNR. Compared to the traditional TS resource allocation scheme, the proposed scheme significantly improves the system performance in energy efficiency, and the system performance improvements gets larger when the SNR is higher. We have investigated the impacts of the source-destination distance d and the pathloss exponent m on the energy efficiency performance. Results have demonstrated that the system performance is more susceptible to the variation of the pathloss exponent m than the distance d. We have also investigated the impacts of the minimum required harvested energy $\bar{Q}$ on the energy efficiency performance. It is demonstrated that for the traditional TS scheme, the system energy efficiency decreases with the increases of the minimum required harvested energy $\bar{Q}$, whereas for the proposed time slot-switching scheme, the system energy efficiency is hardly affected by the variation of the minimum required harvested energy $\bar{Q}$. In our setup, we consider the SISO SWIPT systems and Rayleigh channels for the proposed optimal scheme. MIMO systems and other more complex channel models for the proposed scheme can be further studied in future work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant numbers 61871192 and 61871191) and the Science and Technology Program of Guangzhou (grant number 201904010373).

References

[1] L. R. Varshney, "Transporting information and energy simultaneously," in *2008 IEEE International Symposium on Information Theory*, pp. 1612–1616, Toronto, ON, Canada, July 2008.

[2] P. Grover and A. Sahai, "Shannon meets Tesla: wireless information and power transfer," in *2010 IEEE International Symposium on Information Theory*, pp. 2363–2367, Austin, TX, USA, June 2010.

[3] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 1989–2001, 2013.

[4] L. Liu, R. Zhang, and K. C. Chua, "Wireless information and power transfer: a dynamic power splitting approach," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3990–4001, 2013.

[5] H. Ju and R. Zhang, "A novel mode switching scheme utilizing random beamforming for opportunistic energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2150–2162, 2014.

[6] C. Shen, W.-C. Li, and T.-H. Chang, "Wireless information and energy transfer in multi-antenna interference channel," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6249–6264, 2014.

[7] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer in multiuser OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2282–2294, 2014.

[8] Q. Shi, L. Liu, W. Xu, and R. Zhang, "Joint transmit beamforming and receive power splitting for MISO SWIPT systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 6, pp. 3269–3280, 2014.

[9] D. W. K. Ng, E. S. Lo, and R. Schober, "Wireless information and power transfer: energy efficiency optimization in OFDMA systems," *IEEE Transactions on Wireless Communications*, vol. 12, no. 12, pp. 6352–6370, 2013.

[10] G. Zhang, J. Xu, Q. Wu, M. Cui, X. Li, and F. Lin, "Wireless powered cooperative jamming for secure OFDM system," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1331–1346, 2018.

[11] S. Yin and Z. Qu, "Resource allocation in multiuser OFDM systems with wireless information and power transfer," *IEEE Communications Letters*, vol. 20, no. 3, pp. 594–597, 2016.

[12] Q. Wu, G. Zhang, D. W. K. Ng, W. Chen, and R. Schober, "Generalized wireless-powered communications: when to activate wireless power transfer?," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8243–8248, 2019.

[13] X. Chen, X. Wang, and X. Chen, "Energy-efficient optimization for wireless information and power transfer in large-scale MIMO systems employing energy beamforming," *IEEE Wireless Communications Letters*, vol. 2, no. 6, pp. 667–670, 2013.

[14] A. Gupta, K. Singh, and M. Sellathurai, "Time-switching EH-based joint relay selection and resource allocation algorithms for multi-user multi-carrier AF relay networks," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 2, pp. 505–522, 2019.

[15] Q. Li and L. Yang, "Robust optimization for energy efficiency in MIMO two-way relay networks with SWIPT," *IEEE Systems Journal*, vol. 14, no. 1, pp. 196–207, 2020.

[16] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Communications Letters*, vol. 19, no. 12, pp. 2082–2085, 2015.

[17] Y. Feng, M. Wen, F. Ji, and V. C. M. Leung, "Performance analysis for BDPSK modulated SWIPT cooperative systems with nonlinear energy harvesting model," *IEEE Access*, vol. 6, pp. 42373–42383, 2018.

[18] Y. Dong, M. J. Hossain, and J. Cheng, "Performance of wireless powered amplify and forward relaying over *Nakagami-m* fading channels with nonlinear energy harvester," *IEEE Communications Letters*, vol. 20, no. 4, pp. 672–675, 2016.

[19] J. M. Kang, I. M. Kim, and D. I. Kim, "Joint Tx power allocation and Rx power splitting for SWIPT system with multiple nonlinear energy harvesting circuits," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 53–56, 2019.

[20] S. Gao, K. Xiong, R. Jiang, L. Zhou, and H. Tang, "Outage performance of wireless-powered SWIPT networks with nonlinear EH model in Nakagami-m fading," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 668–671, Beijing, China, August 2018.

[21] T. L. N. Nguyen and Y. Shin, "Outage probability analysis for SWIPT systems with nonlinear energy harvesting model," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 196–199, Jeju Island, Korea (South), October 2019.

[22] K. Wang, Y. Li, Y. Ye, and H. Zhang, "Dynamic power splitting schemes for non-linear EH relaying networks: perfect and imperfect CSI," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, Toronto, ON, Canada, September 2017.

[23] J. Zhang and G. Pan, "Outage analysis of wireless-powered relaying MIMO systems with non-linear energy harvesters and imperfect CSI," *IEEE Access*, vol. 4, pp. 7046–7053, 2016.

[24] L. Liu, R. Zhang, and K. C. Chua, "Wireless information transfer with opportunistic energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 288–300, 2013.

[25] W. Yu and R. Lui, "Dual methods for nonconvex Spectrum optimization of multicarrier systems," *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1310–1322, 2006.

[26] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer: architecture design and rate-energy tradeoff," *IEEE Transactions on Communications*, vol. 61, no. 11, pp. 4754–4767, 2013.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6618126, 11 pages
https://doi.org/10.1155/2021/6618126

Research Article

A Semidynamic Bidirectional Clustering Algorithm for Downlink Cell-Free Massive Distributed Antenna System

Panpan Qian,[1] Huan Zhao,[1] Yanmin Zhu,[1] and Qiang Sun[1,2]

[1]*School of Information Science and Technology, Nantong University, Nantong 226019, China*
[2]*Nantong Research Institute for Advanced Communication Technologies, Nantong 226019, China*

Correspondence should be addressed to Qiang Sun; sunqiang@ntu.edu.cn

Received 19 October 2020; Revised 3 January 2021; Accepted 8 January 2021; Published 21 January 2021

Academic Editor: Miaowen Wen

Cell-free massive distributed antenna system (CF-MDAS) can further reduce the access distance between mobile stations (MSs) and remote access points (RAPs), which brings a lower propagation loss and higher multiplexing gain. However, the interference caused by the overlapping coverage areas of distributed RAPs will severely degrade the system performance in terms of the sum-rate. Since that clustering RAPs can mitigate the interference, in this paper, we investigate a novel clustering algorithm for a downlink CF-MDAS with the limited-capacity backhaul. To reduce the backhaul burden and mitigate interference effectively, a semidynamic bidirectional clustering algorithm based on the long-term channel state information (CSI) is proposed, which has a low computational complexity. Simulation results show that the proposed algorithm can efficiently achieve a higher sum-rate than that of the static clustering one, which is close to the curve obtained by dynamic clustering algorithm using the short-term CSI. Furthermore, the proposed algorithm always reveals a significant performance gain regardless of the size of the networks.

1. Introduction

1.1. Background and Related Work. Recently, with the widespread adoption of smartphones and the popularity of multimedia services, mobile data traffic is exploding. As current cellular networks are reaching their breaking point, there is an urgent need to develop new innovative solutions [1]. In cell-free massive distributed antenna systems (CF-MDASs), the antennas are distributed over the intended coverage area. Meanwhile, it has a very large number of remote access points (RAPs) which can use a direct measurement of channel characteristics to serve all mobile stations (MSs) in the same frequency band [2]. It is expected to be a key technology enabler of the sixth generation (6G) mobile communication systems [3]. In CF-MDAS, a large number of MSs in the whole area will be served simultaneously by a large number of separately distributed RAPs, which coordinate with the central processing unit (CPU) [4].

In contrast to the traditional DAS [5], CF-MDAS can further reduce the access distance between MSs and RAPs, which brings low path loss and high spatial multiplexing gain. However, CF-MDAS brings more serious inter-RAP interference, especially in the overlapping area than that of the conventional DAS. Due to the collaboration among RAPs, the system performance based on the sum-rate can be optimized effectively in this way. Nevertheless, it requires the complete channel state information (CSI) of all RAPs processed jointly, a strict synchronization across RAPs, and strong information exchange backhaul capability. Thanks to RAP clustering, which is rated as one of the promising techniques to combat inter-RAP interference for CF-MDAS [6, 7], the scale of collaboration can be reduced and backhaul burden by sharing full CSI between a limited number of RAPs can be diminished as well.

Generally, the clustering algorithms are classified into static clustering and dynamic clustering. Static clustering is formed according to the geographic locations of the BSs without any CSI [8, 9]. The number of the selected BSs in the cluster is fixed, which does not change over time. Though static clustering is simple enough and does not rely on fast backhaul, MSs at the edge of clusters still suffer from serious intercluster influence. Several studies on dynamic clustering have investigated to overcome the above mentioned problems, Shi et al. proposed a dynamic user-centric cell clustering algorithm [10], which can not only cancel the joint intracluster interference but also effectively alleviate the overall and per-BS cooperation cost. However, it can only count on short-term CSI. In [11], the authors proposed a clustering algorithm based on maximum coordination gain which focuses on minimizing the intercell interference to the cell-edge MS. Nevertheless, the clustering algorithm ignores the bidirectional cooperation gain between RAPs. The authors in [12] proposed a bidirectional dynamic network (BDN) to improve efficiency in achieving better spectral efficiency (SE) performance. It is worth noting that even dynamic clustering can be exploited to achieve higher cooperative gains than static clustering, but its complexity is very high.

It is also noted that most of the existing clustering algorithms are unidirectional. One cluster chooses the best cluster freely that can bring high channel gain to itself. At the same time, the dynamic forming cooperative clusters will result in frequent changes of clusters and lead to a large signaling overhead, which is based on the short-term CSI. Furthermore, in the CF-MDAS, owing to the limited-capacity backhaul, sharing the short-term CSI and data information of all RAPs is difficult.

1.2. Motivation and Contributions. In the existing literature, we investigate the clustering problem of the CF-MDAS with limited-capacity backhaul, aimed at maximizing the system sum-rate. Since traditional dynamic clustering algorithms cannot be applied to the CF-MDAS directly, it can only consider unidirectional clustering and depend on the short-term CSI. To this end, we propose a semidynamic bidirectional clustering algorithm using long-term CSI. The main idea of the algorithm is to cluster RAPs according to the bidirectional average rate gain among clusters. Our novelties and contributions can be summarized as follows:

(i) We develop a network model where each MS is randomly placed into the network. Compared to most bibliographies where the number of MSs in each RAP is fixed, we consider the number of MSs is different in each cluster

(ii) We derive a closed-form expression for the average rate per MS based on some approximation techniques, which can be computed with a low computational complexity

(iii) Based on the derived expression, we proposed that the process of the semidynamic bidirectional clustering algorithm can be approximately equivalent to combining two clusters with the maximum bidirectional average rate gain per MS in each iteration

(iv) We propose a semidynamic bidirectional clustering algorithm for the downlink cell-free CF-MDAS. The proposed algorithm can reduce the backhaul burden and obtain a higher sum-rate with long-term CSI. Simulation results show that our proposed algorithm can achieve a higher sum-rate than the static clustering. Furthermore, our proposed algorithm achieves a performance very close to the optimum curve obtained by dynamic clustering algorithm with the short-term CSI

The remainder of this paper is organized as follows. In Section 2, the system model used in this study is described. A semidynamic bidirectional clustering algorithm is proposed in Section 3. Then, we discuss the simulation assumptions and compare the performance of different RAP clustering algorithms in Section 4. Finally, the paper is concluded in Section 5.

For notations, matrices and column vectors are denoted by bold capital letters $\mathbf{X}$ and bold letters $\mathbf{x}$, respectively. The transpose and Hermitian transpose are denoted by $(\cdot)^T$ and $(\cdot)^H$, respectively. The $F \times F$ identity matrix is denoted by I_F. The vector 2-norm of x is represented by $\|x\|$. The space of all $M \times N$ matrices with complex entries is represented by $\mathbb{C}^{M\times N}$. A combination of k elements taken from n different elements is presented by $\binom{n}{k}$. A complex Gaussian distribution function with mean 0 and variance σ^2 is given by $\mathscr{C}\mathscr{N}(0, \sigma^2)$. The Gamma distribution function with the shape parameter μ and the scale parameter θ is given by $\Gamma(\mu, \theta)$. The cardinality of a set $\mathscr{U}$ is denoted by $|\mathscr{U}|$. The expectation operation is denoted by $\mathbb{E}(\cdot)$.

2. System Model

In this section, the general system model for the CF-MDAS is introduced including the network model, channel model, and signal model descriptions, respectively. Then, the ergodic achievable sum-rate is given.

2.1. Network Model. We consider a 2-dimension downlink CF-MDAS which consists of $M = \{1, 2, \cdots, M\}$ RAPs and $K = \{1, 2, \cdots, K\}$ MSs as shown in Figure 1. RAPs are located at the center of each hexagon, where each RAP is equipped with N_t antennas. Define R_1 as the distance between one RAP and any vertex of its hexagon. Therefore, the distance between two nearest RAPs is $\sqrt{3}R_1$. MSs are distributed randomly in the network, and each MS is equipped with a single antenna. We define the number of RAPs along each dimension is the size of network, and it can be changed. A simple illustration of clustering is given in Figure 1, the RAPs in the same color hexagons form a cluster. If a RAP is associated with no MS, it is assumed to be sleeping.

Let $V = \{V_1, V_2, \cdots, V_L\}$, $\forall V_i \cap V_j = \phi$ be the set of clusters, where L is the number of clusters. All MSs choose the best RAP with the maximal large-scale fading. Denote the set of MSs in cluster i as U_i, where $U_i = \{1, 2, \cdots, |U_i|\}$. Then, the set of MSs can be finally defined as $U = \{U_1, U_2, \cdots, U_L\}$.

2.2. Channel Model. The channel vector between RAP m in cluster i and MS k in cluster i is noted as

$$h_{miki} = \sqrt{\beta_{miki}} g_{miki}, \tag{1}$$

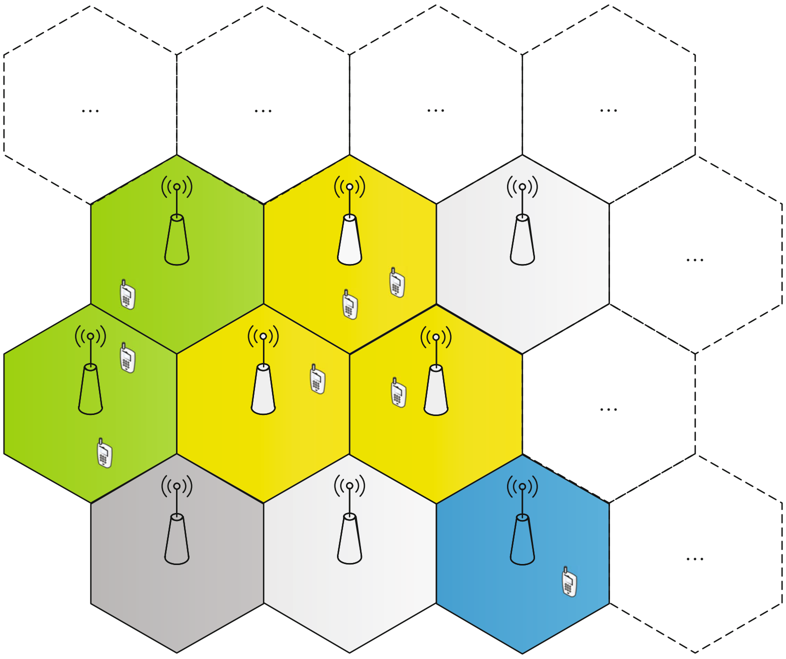


Figure 1: A 2-dimension CF-MDAS.

where $\mathbf{g}_{miki}$ denotes the small-scale fading with $\mathscr{CN}(0,1)$ independent and identically distributed (i.i.d.) elements, and β_{miki} denotes the large-scale fading, which can be modeled as

$$\beta_{miki} = f_{miki} d_{miki}^{-\eta}, \tag{2}$$

where f_{miki} is a log-normal shadow fading variable between RAP m and MS k, η is the path loss exponent, and d_{miki} is the distance between RAP m and MS k.

2.3. Signal Model. The received signal vector of MS k in cluster i is

$$y_{ki} = \underbrace{\sqrt{P} h_{ki}^H w_{ki} s_{ki}}_{\text{usefulsignal}} + \underbrace{\sum\nolimits_{m \in U_i \backslash k} \sqrt{P} h_{ki}^H w_{mi} s_{mi}}_{\text{intraclusterinterference}} + \underbrace{\sum\nolimits_{j=1, j \neq i}^{L} \sum\nolimits_{b \in U_j} \sqrt{P} h_{jki}^H w_{bj} s_{bj}}_{\text{interclusterinterference}} + \underbrace{z_{ki}}_{\text{noise}}, \tag{3}$$

where h_{ki} is the composite channel vector from all RAPs in cluster i to MS k in cluster i noted as $h_{ki} = [h_{1iki}, h_{2iki}, \cdots, h_{|V_i|iki}]$, h_{jki} is the composite channel vector from all RAPs in cluster j to MS k in cluster i, w_{ki} is the beamforming vector assigned for MS k in cluster i defined as $h_{jki} = [h_{1jki}^{\mathrm{T}}, h_{2jki}^{\mathrm{T}}, \cdots, h_{|V_j|jki}^{\mathrm{T}}]^{\mathrm{T}}$, and s_k is the data symbol with unit variance destined to for MS k. z_{ki} is the noise following the distribution $\mathscr{CN}(0, \sigma^2)$. P is the average transmit power of each RAP.

2.4. Ergodic Sum-Rate. The intracluster interference can be cancelled by using zero forcing (ZF) beamforming, that is,

$$\mathbf{w}_{ki} = \frac{\tilde{\mathbf{w}}_{ki}}{\|\tilde{\mathbf{w}}_{ki}\|}, \tag{4}$$

where $\tilde{\mathbf{w}}_{ki} \in \mathbb{C}^{|\mathscr{V}_i| N_t \times 1}$ is the kth column of $\mathbf{H}_i (\mathbf{H}_i^H \mathbf{H}_i)^{-1}$, and $\mathbf{H}_i = [\mathbf{h}_{1i}, \mathbf{h}_{2i}, \cdots, \mathbf{h}_{|\mathscr{U}_i| i}] \in \mathbb{C}^{|\mathscr{V}_i| N_t \times |\mathscr{U}_i|}$ is the compound channel matrix between RAPs in cluster i and MSs within the cluster.

Therefore, the signal-to-interference-plus-noise ratio (SINR) of MS k is

$$\gamma_{ki} = \frac{P \left| h_{ki}^H w_{ki} \right|^2}{P \sum_{j=1, j \neq i}^{L} \sum_{b \in U_j} \left| f_{jki}^H w_{bj} \right|^2 + \sigma^2}. \tag{5}$$

Then, the downlink rate of MS k can be expressed as

$$R_{ki} = \mathbb{E}[\log_2(1 + \gamma_{ki})]. \tag{6}$$

With above observations, the ergodic achievable sum-rate of the system can be presented as

$$R = \sum_{i=1}^{L} \sum_{k \in \mathscr{U}_i} \mathbb{E}[\log_2(1 + \gamma_{ki})]. \tag{7}$$

3. Semidynamic Bidirectional Clustering Algorithm

As stated previously, dynamic clustering based on short-term CSI will lead to a large signaling overhead among RAPs and MSs, making it infeasible in practical systems. Therefore, we propose to form clusters based on long-term CSI. In this section, we first derive the asymptotical average rate per MS associated with long-term CSI, which will be used in the following clustering algorithm design. In what following, we analyze the bidirectional cooperation willingness and the complexity of the optimal clustering by exhaustive search (ES) algorithm. Finally, a semidynamic bidirectional clustering algorithm is proposed.

3.1. Average Rate per MS.

In this subsection, we first employ a Gamma approximation technique pioneered in [13, 14] to obtain the distributions of both the signal and the interference terms. Based on the distributions, we derive the asymptotical average rate per MS in the high-SINR regime.

The useful channel strength can be denoted as

$$\mathbf{h}_{ki}^{H}\mathbf{h}_{ki} = \sum_{m\in\mathscr{V}_i} \beta_{miki}\mathbf{g}_{miki}^{H}\mathbf{g}_{miki}, \tag{8}$$

where $\mathbf{g}_{miki}^{H}\mathbf{g}_{miki} = \|\mathbf{g}_{miki}\|^2$ is distributed as a chi-square random variable (RV) with $2N_t$ degrees of freedom scaled with 1/2 [14], thus $\beta_{miki}\mathbf{g}_{miki}^{H}\mathbf{g}_{miki} \sim \Gamma(N_t, \beta_{miki})$. Therefore, $\mathbf{h}_{ki}^{H}\mathbf{h}_{ki}$ is a sum of independent Gamma RVs which does not yield a mathematically tractable expression. Fortunately, the sum of independent nonidentically distributed Gamma RVs can be well approximated by employing the second-order matching technique shown in the following lemma.

Lemma 1 (see [13]). *Assume $\{x_i\}$ are independent Gamma RVs with shape and scale parameters μ_i and θ_i, the sum $\sum_i x_i$ can be approximated as another Gamma distributed RV Y which has the same first- and second-order moments, $Y \sim (\mu, \theta)$, where $\mu = (\sum_i \mu_i\theta_i)^2/(\sum_i \mu_i\theta_i^2)$ and $\theta = \sum_i \mu_i\theta_i^2/\sum_i \mu_i\theta_i$.*

As the consequence of Lemma 1, the approximate distribution of the useful channel strength can be presented as the $\Gamma(\mu_{ki}, \theta_{ki})$ distribution, wherein

$$\begin{aligned} \mu_{ki} &= N_t \frac{\left(\sum_{m\in\mathscr{V}_i}\beta_{miki}\right)^2}{\sum_{m\in\mathscr{V}_i}\beta_{miki}^2}, \\ \theta_{ki} &= \frac{\sum_{m\in\mathscr{V}_i}\beta_{miki}^2}{\sum_{m\in\mathscr{V}_i}\beta_{miki}}. \end{aligned} \tag{9}$$

Similarly, the interference channel strength can be noted as

$$\mathbf{h}_{jki}^{H}\mathbf{h}_{jki} = \sum_{t\in\mathscr{V}_j} \beta_{tjki}\mathbf{g}_{tjki}^{H}\mathbf{g}_{tjki}, \tag{10}$$

and its approximate distribution is $\Gamma(\mu_{kj}, \theta_{kj})$, where

$$\begin{aligned} \mu_{kj} &= N_t \frac{\left(\sum_{t\in V_j}\beta_{tjki}\right)^2}{\sum_{t\in V_j}\beta_{tjki}^2}, \\ \theta_{kj} &= \frac{\sum_{t\in V_j}\beta_{tjki}^2}{\sum_{t\in V_j}\beta_{tjki}}. \end{aligned} \tag{11}$$

From (11), it is easy to see that $\mu_{ki} \le |\mathscr{V}_i|N_t$, where the upper bound becomes exact, when $\beta_{1iki} = \beta_{2iki} = \cdots\beta_{|\mathscr{V}_i|iki}$. In order to get tractable distributions, in [13], the authors used (9) and (11) to obtain an equivalent i.i.d. channel vector of MS k, i.e., approximate the nonisotropic channel vector $\mathbf{h}_{ki}$ as an isotropic vector with i.i.d. $\mathscr{CN}(0, \theta_{ki})$ elements and the nonisotropic channel vector $\mathbf{h}_{jki}$ as an isotropic vector with i.i.d. $\mathscr{CN}(0, \theta_{kj})$ elements. Besides, the authors in [12] proposed that each spatial dimension contributes $\mu_{ki}/|\mathscr{V}_i|N_t$ and $\mu_{kj}/|\mathscr{V}_i|N_t$ to the shape parameters of the distribution of the signal term and interference term, respectively. Noting that each signal beam lies in a $(|V_i|N_t - |U_i| + 1)$ dimensional space and each interference beam spans a one-dimensional [14–16], the shape parameter associated with the distribution of the signal term $|h_{ki}^{H}w_{ki}|^2$ becomes $(|V_i|N_t - |U_i| + 1)(\mu_{ki}/|V_i|N_t)$, and the shape parameter associated with the distribution of the interference term $|h_{jki}^{H}w_{bj}|^2$ turns into $\mu_{kj}/|\mathscr{V}_j|N_t$. Therefore, the distributions of the signal and interference terms can be written as, respectively,

$$\left|h_{ki}^{H}w_{ki}\right|^2 \sim \Gamma\left(\frac{\mu_{ki}(|V_i|N_t - |U_i| + 1)}{|V_i|N_t}, \theta_{ki}\right), \tag{12}$$

$$\left|h_{jki}^{H}w_{bj}\right|^2 \sim \Gamma\left(\frac{\mu_{kj}}{\left|V_j\right|N_t}, \theta_{kj}\right), \tag{13}$$

with μ_{kl}, θ_{kl}, μ_{kj}, and θ_{kj} as the ones defined in (9) and (11).

In [14], the authors assumed the ZF beams designed at each interfering cluster are orthogonal and verifies the accuracy of this approximation. Based on this, $\sum_{b\in U_j}|h_{jki}^{H}w_{bj}|^2$ is a sum of $|U_j|$ independent Gamma RVs which have the same scale parameters. Therefore, the total interference power produced by cluster j follows that

$$\sum_{b\in U_j}\left|h_{jki}^{H}w_{bj}\right|^2 \sim \Gamma\left(\frac{\left|U_j\right|\mu_{kj}}{\left|V_j\right|N_t}, \theta_{kj}\right). \tag{14}$$

Proposition 2. *Based on (12) and (14) and in the high-SINR regime, the average rate of MS k in cluster i can be approximated as*

$$R_{ki} \approx \log_2\left(1 + \frac{\bar{\beta}_{iki}(|V_i|N_t - |U_i| + 1)}{\sum_{j\ne l}^{L}\left|U_j\right|\bar{\beta}_{jki}}\right), \tag{15}$$

where $\bar{\beta}_{iki} = (\sum_{m\in\mathscr{V}_i}\beta_{miki})/|\mathscr{V}_i|$, and $\bar{\beta}_{jki} = (\sum_{t\in\mathscr{V}_j}\beta_{tjki})/|\mathscr{V}_j|$.

For proof, see Appendix A.

3.2. Bidirectional Cooperation. As shown in Figure 2(a), there are two RAPs and two MSs, and each of the RAPs belongs to one cluster. The solid line denotes the useful signal, and the dashed line represents the interference signal. For cluster i, whether it wants to cooperate with cluster j depends on the SINR of MS k. For cluster j, whether it wants to cooperate with cluster i depends on the SINR of MS b. We define γ_{ki} as the SINR of MS k in cluster i, γ_{bj} is the SINR of MS b in cluster j. Due to the intracluster, interference can be eliminated by ZF beamforming. Thus, γ_{ki} and γ_{bj} are, respectively, given by,

$$\gamma_{ki} = \frac{P\left|h_{ki}^{H} w_{ki}\right|^2}{P\left(\sum_{b\in U_j}\left|f_{jki}^{H} w_{bj}\right|^2 + \sum_{r\neq i,j}^{L}\sum_{c\in U_r}\left|h_{rki}^{H} w_{cr}\right|^2\right) + \sigma^2},$$
$$\gamma_{bj} = \frac{P\left|h_{bj}^{H} w_{bj}\right|^2}{P\left(\sum_{k\in U_i}\left|f_{ibj}^{H} w_{ki}\right|^2 + \sum_{r\neq i,j}^{L}\sum_{c\in U_r}\left|h_{rki}^{H} w_{cr}\right|^2\right) + \sigma^2}. \qquad (16)$$

γ_{ki} and γ_{bj} are all small, both cluster i and j want to cooperate with each other. γ_{ki} is small and γ_{bj} is large, cluster i wants to cooperate with cluster j but cluster j does not want to cooperate with cluster i. γ_{ki} is large and γ_{bj} is small, cluster i does not want to cooperate with cluster j but cluster j wants to cooperate with cluster i. γ_{ki} and γ_{bj} are all large, neither cluster i nor j wants to cooperate with each other.

In general, $\gamma_{ki} \neq \gamma_{bj}$, which means cooperation in different directions making a huge difference. As shown in Figures 2(a)–2(d), there are four different cooperation scenarios according to the locations of MSs.

In Figures 2(b) and 2(c), the cooperation desire of one side is strong, without loss of generality, the other side is correspondingly weak. If only considering the unidirectional cooperation desire of one side, it will not be able to obtain optimal clustering results. To this end, we should consider the bidirectional cooperation between cluster i and j.

3.3. The Problem of Clustering. Due to the limited-capacity backhaul, the maximum number of RAPs in a cluster is defined as Q. By restricting the value of Q, the information exchange of intracluster RAPs can be reduced. To maximize the system sum-rate, the objective problem of clustering can be described as

$$\max \sum_{i=1}^{L}\sum_{k\in U_i} R_{ki}, \text{ s.t.}\begin{cases} \left|V_i\right| + \left|V_j\right| \leq Q \forall V_i \subseteq V \\ V_i \cap V_j = \phi \forall i, j \end{cases}. \qquad (17)$$

The above problem is a combinatorial optimization problem, and the optimal solution can be obtained by exhaustive search (ES) algorithm. All possible clustering results are defined as a set G that satisfies the cluster size no more than Q. Then, $|G|$ can be written as

$$|G| = \begin{pmatrix} M \\ |\mathscr{V}_1| \end{pmatrix} \cdot \begin{pmatrix} M - |\mathscr{V}_1| \\ |\mathscr{V}_2| \end{pmatrix} \cdots \begin{pmatrix} |\mathscr{V}_L| \\ |\mathscr{V}_L| \end{pmatrix}, \text{s.t.}\begin{cases} \forall V_i \subseteq V, \\ \sum_{i=1}^{L} |\mathscr{V}_i| = M. \end{cases} \qquad (18)$$

Clustering RAPs requires two steps. The first step is to determine the cluster size $|\mathscr{V}_i|, i = 1, 2, \cdots, L$. The second step is to calculate the number of RAP combination $|G|$. Assume the size of each cluster is same, i.e., $|\mathscr{V}_1| = |\mathscr{V}_2| = \cdots = |\mathscr{V}_L| = [M/L]$, the possible number of RAP cluster combination scheme is $\begin{pmatrix} M \\ |\mathscr{V}_1| \end{pmatrix} \cdot \begin{pmatrix} M - |\mathscr{V}_1| \\ |\mathscr{V}_1| \end{pmatrix} \cdots \begin{pmatrix} |\mathscr{V}_1| \\ |\mathscr{V}_1| \end{pmatrix}$. Therefore, the complexity of clustering by ES algorithm is $O(M \cdot M!)$. For a large M, this method is not feasible. Alternatively, we propose a low-complexity semidynamic bidirectional clustering algorithm using greedy algorithm which will be discussed in the next subsection to find a good suboptimal solution.

3.4. Semidynamic Bidirectional Clustering Algorithm. Consider the bidirectional cooperation, we use the rate gain per MS of the cluster i after cooperating with cluster j to measure its unidirectional cooperation desire to cooperate with cluster j. Assume cluster i and j cooperate as a new cluster l, then we define $\alpha(i, j)$ and $\alpha(j, i)$ as the rate gain per MS of cluster i and j after cooperating, where

$$\alpha(i, j) = \frac{\bar{R}_l}{\bar{R}_i}, \alpha(j, i) = \frac{\bar{R}_l}{\bar{R}_j}, \qquad (19)$$

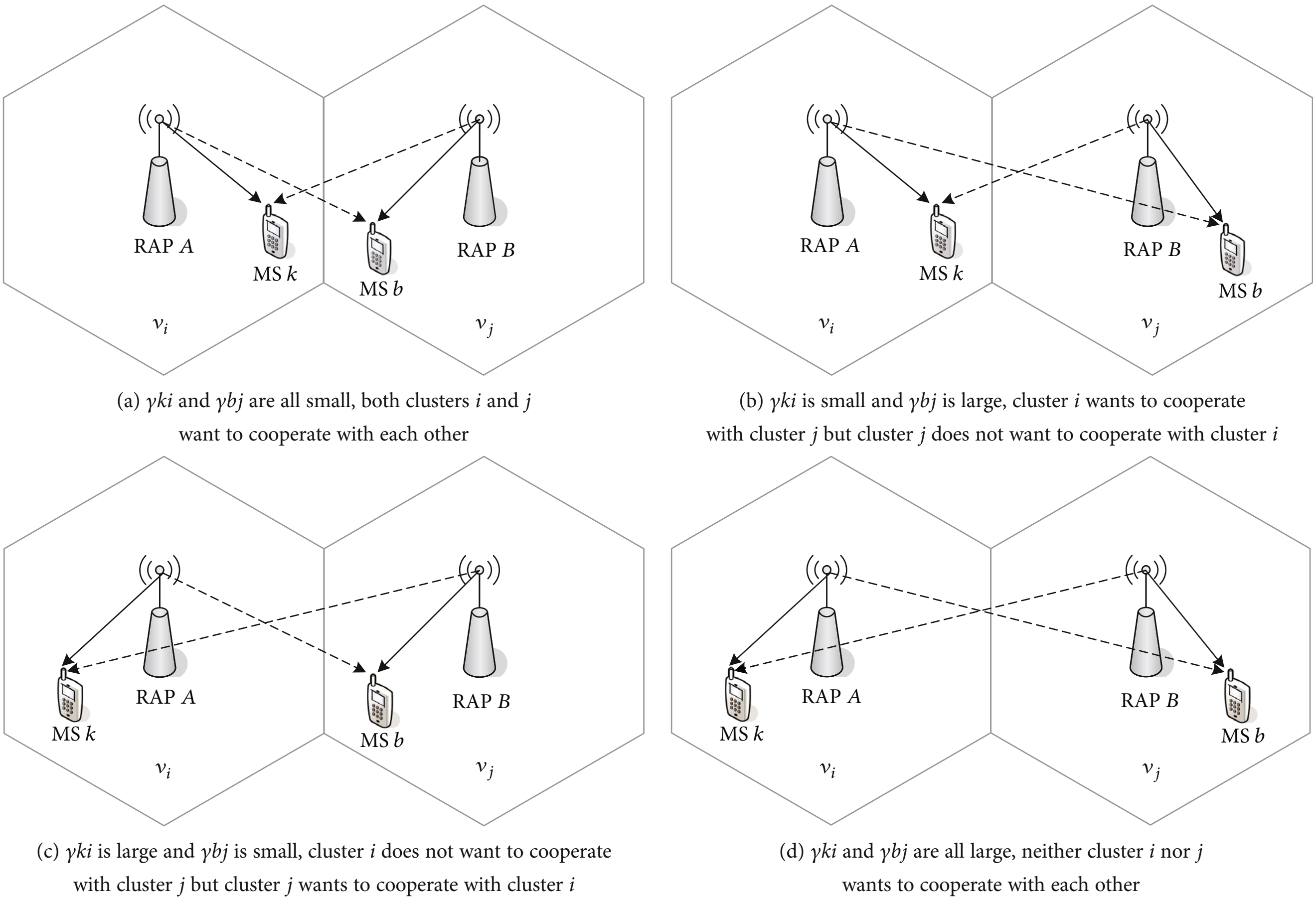
where $\bar{R}_i = (\sum_{k\in\mathscr{U}_i} R_{ki})/|\mathscr{U}_i|$ is the average rate per MS of cluster i. Analogously, $\bar{R}_j$ and $\bar{R}_l$ are the average rate of each MS of cluster j and l.

Using the greedy algorithm, the problem of clustering to maximize the system sum-rate while considering the bidirectional cooperation can be translated into combine two clusters with the maximum bidirectional average rate gain per MS in each iteration as

$$\max \frac{\alpha(i, j) + \alpha(j, i)}{2}, \quad \text{s.t.}\begin{cases} \left|V_i\right| + \left|V_j\right| \leq Q \forall V_i \subseteq V, \\ V_i \cap V_j = \phi \forall i, j. \end{cases} \qquad (20)$$

According to the analysis above, the proposed semidynamic clustering algorithm (the large-scale fading coefficients change slowly compared to the small-scale fading and can be easily tracked, e.g. in a few minutes. Thus the period of updating the cluster depends on the change of large-scale fading.) is summarized in Algorithm 1.

In step 1, each MS is associated with one RAP, which can be determined by the large-scale fading channel factors. Step



(a) γki and γbj are all small, both clusters i and j want to cooperate with each other

(b) γki is small and γbj is large, cluster i wants to cooperate with cluster j but cluster j does not want to cooperate with cluster i

(c) γki is large and γbj is small, cluster i does not want to cooperate with cluster j but cluster j wants to cooperate with cluster i

(d) γki and γbj are all large, neither cluster i nor j wants to cooperate with each other

Figure 2: Four different cooperation scenarios.

```
Initialization: RAPs set $M = \{1, 2, \cdots, M\}$, MSs set $K = \{1, 2, \cdots, K\}$, cluster set $V = \{V_1, V_2, \cdots, V_L\}$, $\forall V_i \cap V_j = \phi$. MSs set $U = \{U_1, U_2, \cdots, U_L\}$.
Step 1: associate MSs with their respective RAP according to the maximum large-scale fading.
  for $k = 1, 2, \cdots, K$ do
    $m^* = \text{argmax}\beta_{m\_k\_}$, for $m = 1, 2, \cdots, M$.
  end for
Step 2: calculate the average rate gain per MS of cooperation cluster between any two clusters.
  for $i = 1 ; i < L ; i + +$ do
    for $j = i + 1 ; \mathrm{j} < = L ; j + +$ do
      calculate $\alpha(i, j)$ and $\alpha(j, i)$ based on (19),
      $\bar{\alpha}(i, j) = \alpha(i, j) + \alpha(j, i)/2$.
    end for
  end for
Step 3: merge cooperation cluster.
  while $\bar{\alpha}(i, j) > 1$ do
    $(i*, j*) = \arg \max_{i,j} \bar{\alpha}(i, j)$,
    $c = i^*, d = j^*$.
    if $|V_c| + |V_d| \le Q$ then
      update $V$ and $L$, $V_c = V_c \cup V_d$, $V_d = \phi$, $U_c = U_c \cup U_d$, $U_d = \phi$, $L = L - 1$, go to step 2.
    end if
    if $|V_c| + |V_d| > Q$ then
      update $\bar{\alpha}(c, d) = 0$, go to step 3.
    end if
  end while
```

Algorithm 1: Semidynamic Bidirectional Clustering Algorithm.

2 calculates the approximate average rate gain per MS of cooperation cluster between any two clusters using long-term CSI, which is based on formula (19). Step 3 describes the process of merging cluster, which uses greedy algorithm to find the suboptimal cooperative clusters under the limitation of cluster size, aimed at reducing the computational complexity. The number of available cluster combinations is $\sum_{j=2}^{M}\binom{j}{2}$. Thus, the complexity of our proposed algorithm is $O(M^3)$, which is lower than clustering by ES algorithm.

4. Numerical Results

In this section, we give numerical simulations to compare different clustering algorithms. In the simulations, we use (1) to generate the channels and set path loss exponent $\eta = 4$; each RAP is equipped with 2 antennas and each MS is equipped with single antenna. In addition, the height of transmit antennas is 20 m. We define the maximum number of RAPs in a cluster as $Q = 4$. We assume the noise power is -102 dBm and cell-edge received power is set from -10 dBm to 10dBm. The detail simulation parameters are listed in Table 1. To prevent the contingency of the experiment, the system sum-rate is obtained by averaging 50 drops of MS locations, each of which consists of 10,000 realizations of i.i.d. small-scale channels.

In order to compare the proposed algorithm, we simulate the following seven algorithms:

(1) Clustering all RAPs: all RAPs are clustered to serve the whole MSs

(2) Semidynamic bidirectional clustering algorithm (with legend "SBCA"): the algorithm is proposed in Section 3

(3) Dynamic bidirectional clustering algorithm (with legend "DBCA"): the algorithm uses short-term CSI and considers bidirectional cooperation proposed in Section 3.2

(4) Dynamic unidirectional clustering algorithm (with legend "DUCA") [11]: the algorithm uses short-term CSI but only considers unidirectional selection

(5) No clustering: each MS is only served by the RAP with the strongest massive channel gain and suffers from interference from all the other RAPs

(6) Static clustering (2 RAPs): we give a 4 × 4 network topology for example as shown in Figure 3(a), where each MS is served by a cluster consisting of 2 RAPs. It can be applied to other network topologies

(7) Static clustering (4 RAPs): we give a 4 × 4 network topology for example as shown in Figure 3(b), where each MS is served by a cluster consisting of 4 RAPs. It can be applied to other network topologies

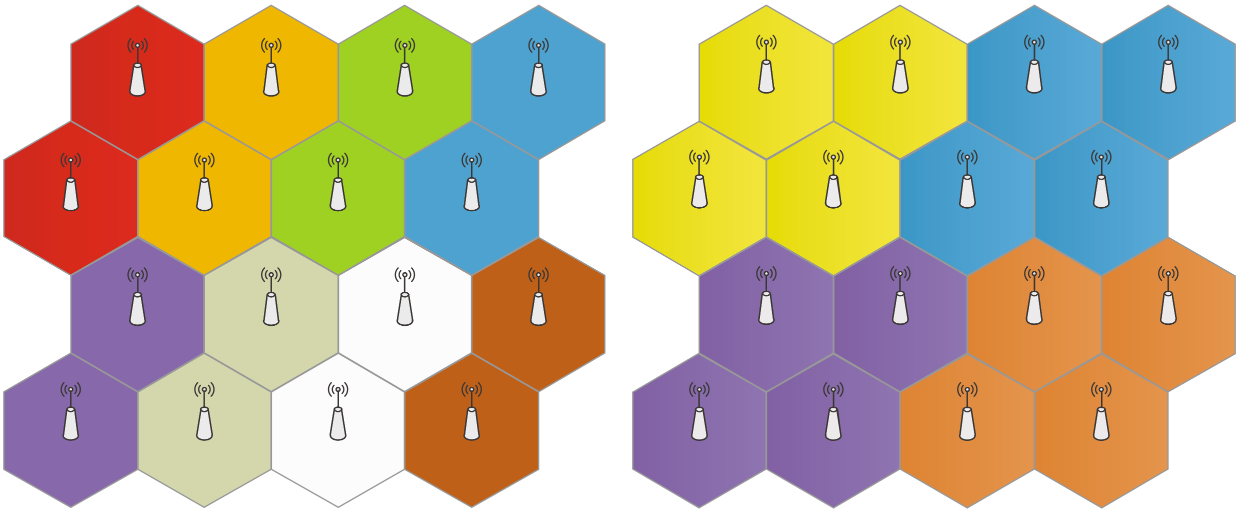
Table 1: Simulation parameters.

Parameters	Value
Cell model	Square grid (1 km^2)
Number of RAPs	100
Minimum distance between two RAPs	125 m
Number of MSs	$5 \le K \le 25$
Number of transmit antennas	$N_t = 2$
Number of receive antennas	$N_r = 1$
Transmit antenna height	$h = 20$ m
Path loss exponent	$\eta = 4$
Shadow fading σ_{sh}	8 dBm
Maximum number of RAPs in a cluster	$Q = 4$
Noise power	-102 dBm
Cell-edge received power	[-10 dBm,10 dBm]

Figure 4 shows the system sum-rates achieved by different clustering algorithms, when K=10, versus cell-edge received power. It can be seen that the system sum-rates of all algorithms increase with the increasing of the cell-edge received power, especially at the low cell-edge received power segment. This is because the interference is small when the cell-edge received power is low, and the signal power increases faster than the interference power. Clustering all RAPs can achieve the highest system sum-rate, but it is an ideal condition which is impossible to implement in reality. If no clustering, MSs will suffer from serious inter-RAP interference, the system sum-rate is the lowest. Even that the static clustering algorithm can improve the sum-rate, however, it cannot adapt to the changes of MS locations; the clusters will not change once formed. The proposed semidynamic bidirectional clustering algorithm exhibits a higher sum-rate than all the static clustering and dynamic unidirectional clustering algorithms. This is because the proposed algorithm uses the long-term CSI and considers the bidirectional cooperation. Though the dynamic bidirectional clustering algorithm performs better than the proposed algorithm, it will cause more signaling overhead. Considering the limited-capacity backhaul and computational complexity, the proposed algorithm is more practical.

Figure 5 depicts the system sum-rates for different number of MSs when the cell-edge received power is 10 dBm. As the number of MSs increases, the proposed algorithm is always superior to the static clustering and dynamic unidirectional clustering algorithms. When the number of MSs is 25, the proposed algorithm can increase the sum-rate about 20% over the static algorithm (4 RAPs). At the same time, the sum-rate of the proposed algorithm is only 1% lower than the dynamic bidirectional clustering algorithm.

Figure 6 describes the average rate per MS for different sizes of network when K=10 and the cell-edge receive power is 10 dBm. When the network size increases, the MS average rate will also increase. Since that the distance between MS and its served RAP is more close when the network size increases, the large-scale fading of MSs will become small. From Figure 6, it is easy to see that the proposed algorithm always yields a significant performance.



(a) Static clustering (2 RAPs)

(b) Static clustering (4 RAPs)

FIGURE 3: Two static clustering.

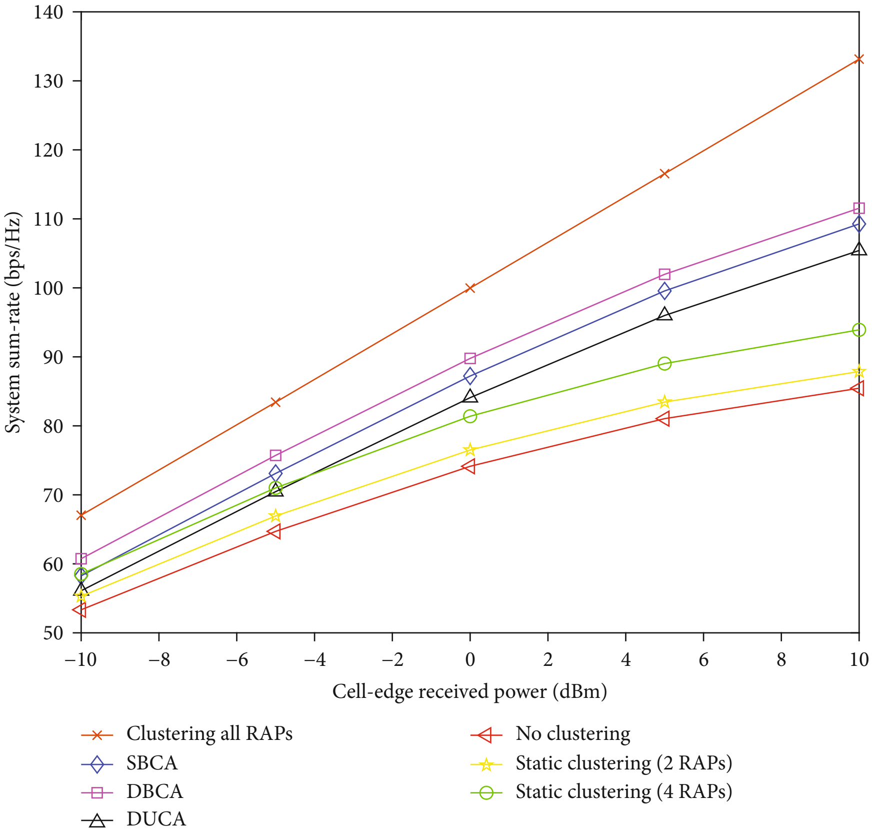


FIGURE 4: System sum-rate for different clustering algorithms.

The computational time in MATLAB of different algorithms is presented in Table 2. Since the static cluster algorithms rely on the geographic location of RAPs without any CSI, it is no computational time. The dynamic clustering algorithms forms clusters in each time slot, while the semidynamic clustering algorithm utilizes the long-term CSI, so the complexity of dynamic clustering is much higher than that of the proposed semidynamic bidirectional clustering algorithm. The dynamic bidirectional clustering algorithm considers bidirectional cooperation among clusters. It needs to calculate the average rate gain per MS of two clusters while the dynamic unidirectional clustering algorithm only calculates the average rate gain per MS of one cluster in each iteration. Therefore, the simulation time of dynamic

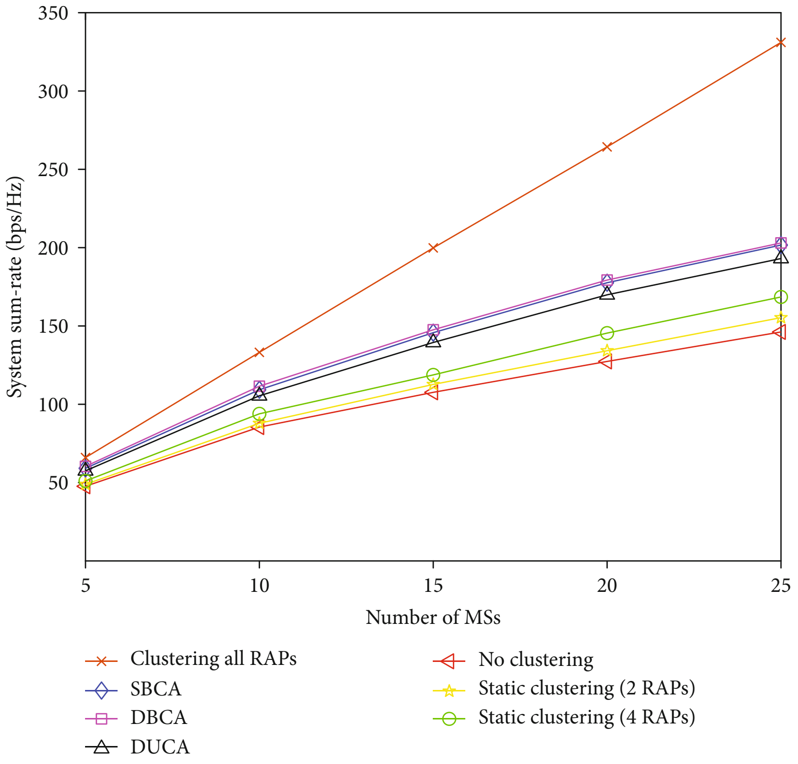


Figure 5: System sum-rate for different numbers of MSs.

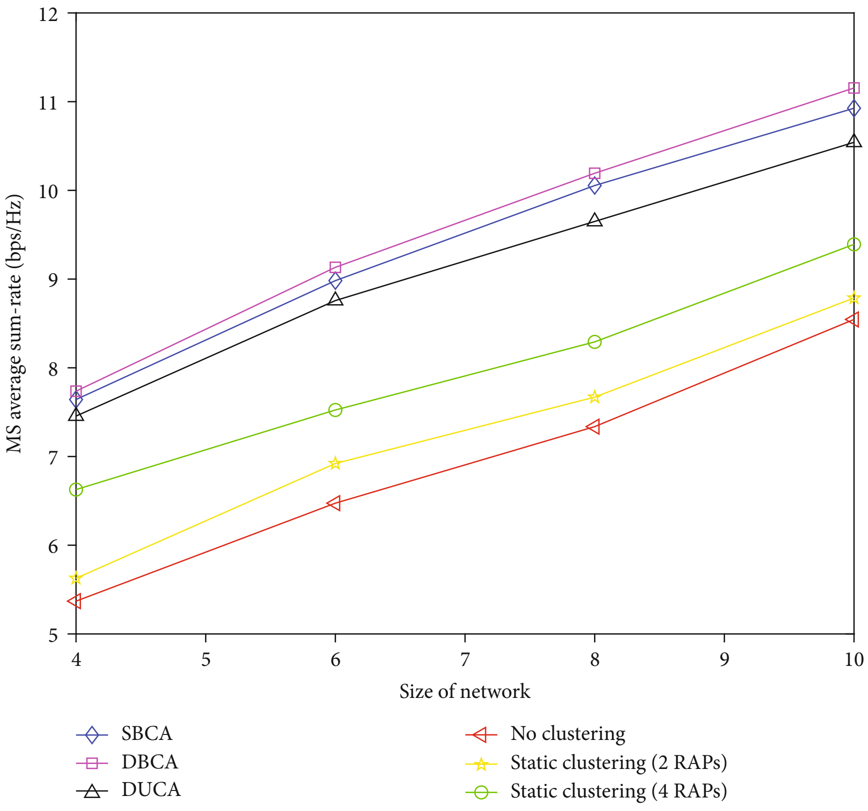


Figure 6: Average rate per MS for different sizes of network.

Table 2: Simulation time of different clustering algorithms.

Clustering algorithms	Simulation time (s)
SBCA	179.437
DBCA	129.324
DUCA	0.041

bidirectional clustering algorithm is longer than dynamic unidirectional clustering algorithm. It can be found from Table 2 that the complexity of our proposed algorithm is low.

5. Conclusion

In this paper, we have studied the RAP clustering of downlink CF-MDAS and proposed a semidynamic bidirectional clustering algorithm with a lower computational complexity. We derived an approximate expression of the average rate per MS associated with the long-term CSI. Considering the limited-capacity backhaul of CF-MDAS, it is impractical to share the short-term CSI and data information of all RAPs. Based on which, the proposed clustering algorithm only relies on long-term CSI. On the other hand, the proposed algorithm also considers bidirectional cooperation among clusters. The results showed that the proposed algorithm exhibits a higher sum-rate than all the static clustering and dynamic unidirectional clustering algorithms. Meanwhile, the sum-rate provided by the proposed algorithm is closed to the dynamic bidirectional clustering algorithm using short-term CSI. Moreover, the proposed algorithm presents a good performance regardless of the network size.

Appendix

Proof of Proposition 2

Lemma 3 (see [17]). *If X and Y are independent positive random variables, then* $\mathbb{E}[log_2(1+(X/Y))] \approx \log_2(1+(\mathbb{E}(X)/\mathbb{E}(Y)))$.

Based on Lemma 3, then, we have

$$R_{ki} = \mathbb{E}[log_2(1+\gamma_{ki})] \approx \log_2\left(1+\frac{P\mathbb{E}\left\{\left|\mathbf{h}_{ki}^H\mathbf{w}_{ki}\right|^2\right\}}{P\sum_{j\neq i,j}^{L}\mathbb{E}\left\{\sum_{b\in\mathscr{U}_j}\left|\mathbf{f}_{jki}^H\mathbf{w}_{bj}\right|^2\right\}+\sigma^2}\right). \tag{A.1}$$

From (13), (15), and the properties of gamma distribution,

$$\mathbb{E}\left\{\left|\mathbf{h}_{ki}^H\mathbf{w}_{ki}\right|^2\right\} = \frac{(|\mathscr{V}_i|N_t - |\mathscr{U}_i| + 1)}{|\mathscr{V}_i|}\sum_{m\in\mathscr{V}_i}\beta_{miki}, \tag{A.2}$$

$$\mathbb{E}\left\{\sum_{b\in\mathscr{U}_j}\left|\mathbf{f}_{jki}^H\mathbf{w}_{bj}\right|^2\right\} = \frac{|\mathscr{U}_i|}{|\mathscr{V}_j|}\sum_{t\in\mathscr{V}_j}\beta_{tjki}. \tag{A.3}$$

To simplify the expression, defining $\bar{\beta}_{iki}$ as the average large-scale fading from MS k to its own cluster i, defining $\bar{\beta}_{jki}$ as the average large-scale fading from MS k to its interference cluster j, where

$$\begin{aligned}\bar{\beta}_{iki} &= \frac{\sum_{m\in\mathscr{V}_i}\beta_{miki}}{|\mathscr{V}_i|},\\ \bar{\beta}_{jki} &= \frac{\sum_{t\in\mathscr{V}_j}\beta_{tjki}}{|\mathscr{V}_j|}.\end{aligned} \tag{A.4}$$

When in the high-SINR regime, σ^2 can be ignored, then by substituting (A.2) and (A.3) into (A.1), we can obtain

$$R_{ki} \approx \log_2\left(1+\frac{\bar{\beta}_{iki}(|\mathscr{V}_i|N_t - |\mathscr{U}_i| + 1)}{\sum_{j\neq i}^{L}|\mathscr{U}_j|\bar{\beta}_{jki}}\right). \tag{A.5}$$

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61971467).

References

[1] Y. Zhang and L. Dai, "A closed-form approximation for uplink average ergodic sum capacity of large-scale multi-user distributed antenna systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1745–1756, 2019.

[2] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: a new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99878–99888, 2019.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[4] Y. Al-Eryani, M. Akrout, and E. Hossain, "Multiple access in cell-free networks: outage performance, dynamic clustering, and deep reinforcement learning-based design," *IEEE Journal on Selected Areas in Communications*, 2020.

[5] Q. Sun, S. Jin, J. Wang, Y. Zhang, X. Q. Gao, and K.-K. Wong, "Downlink massive distributed antenna systems scheduling," *IET Communications*, vol. 9, no. 7, pp. 1006–1016, 2015.

[6] Z. Jiang, S. Zhou, and Z. Niu, "Optimal antenna cluster size in cell-free large-scale distributed antenna systems with imperfect CSI and intercluster interference," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 2834–2845, 2015.

[7] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: a survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 743–764, 2017.

[8] M. A. Wahdan, M. F. Al-Mistarihi, and M. Shurman, "Static cluster and dynamic cluster head (SCDCH) adaptive prediction-based algorithm for target tracking in wireless sensor networks," in *Proceedings 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 596–600, Opatija, 2015.

[9] Z. Zhang, N. Wang, J. Zhang, and X. Mu, "Dynamic user-centric clustering for uplink cooperation in multi-cell wireless networks," *IEEE Access*, vol. 6, pp. 8526–8538, 2018.

[10] J. Shi, M. Chen, W. Zhang, Z. Yang, and H. Xu, "Dynamic AP clustering and precoding for user-centric virtual cell networks," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2504–2516, 2019.

[11] M. Yoon, M. S. Kim, and C. Lee, "A dynamic cell clustering algorithm for maximization of coordination gain in uplink coordinated system," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1752–1760, 2016.

[12] Y. Xin, R. Zhang, D. Wang, J. Li, L. Yang, and X. You, "Antenna clustering for bidirectional dynamic network with large-scale distributed antenna systems," *IEEE Access.*, vol. 5, pp. 4037–4047, 2017.

[13] K. Hosseini, W. Yu, and R. S. Adve, "A stochastic analysis of network MIMO systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 16, pp. 4113–4126, 2016.

[14] H. B. Almelah and K. A. Hamdi, "Spectral efficiency of distributed large-scale MIMO systems with ZF receivers," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 4834–4844, 2017.

[15] W. A. Mahyiddin, N. A. B. Zakaria, K. Dimyati, and A. L. A. Mazuki, "Downlink rate analysis of training-based massive MIMO systems with wireless backhaul networks," *IEEE Access*, vol. 6, pp. 45086–45099, 2018.

[16] Y. Dhungana and C. Tellambura, "Performance analysis of SDMA with inter-tier interference nulling in HetNets," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2153–2167, 2017.

[17] Q. Zhang, S. Jin, M. McKay, D. Morales-Jimenez, and H. Zhu, "Power allocation schemes for multicell massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 5941–5955, 2015.