tech.zte.com.cn

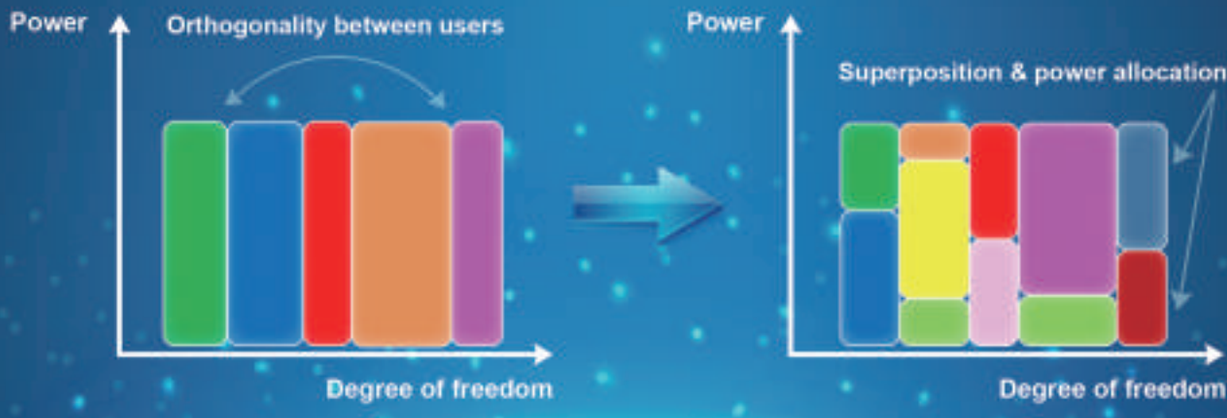
ISSN 1673-5188
CN 34-1294/ TN
CODEN ZCTOAK

ZTE COMMUNICATIONS

An International ICT R&D Journal Sponsored by ZTE Corporation

October 2016, Vol. 14 No. 4

SPECIAL TOPIC:
Multiple Access Techniques for 5G

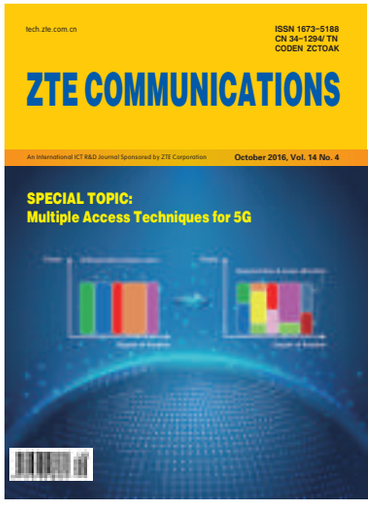


9 771673 518161 10>

ZTE中兴 ZTE COMMUNICATIONS
VOLUME 14 NUMBER 4 OCTOBER 2016

ZTE Communications Editorial Board

CONTENTS



Special Topic: Multiple Access Techniques for 5G

Guest Editorial 01

YUAN Jinhong, XIANG Jiying, DING Zhiguo, and YUAN Zhifeng

Evaluation of Preamble Based Channel Estimation for MIMO-FBMC Systems 03

Sohail Taheri, Mir Ghoraishi, XIAO Pei, CAO Aijun, and GAO Yonghong

Non-Orthogonal Multiple Access Schemes for 5G 11

YAN Chunlin, YUAN Zhifeng, LI Weimin, and YUAN Yifei

A Survey of Downlink Non-Orthogonal Multiple Access for 5G Wireless Communication Networks 17

WEI Zhiqiang, YUAN Jinhong, Derrick Wing Kwan Ng, Maged Elkashlan, and DING Zhiguo

Unified Framework Towards Flexible Multiple Access Schemes for 5G 26

SUN Qi, WANG Sen, HAN Shuangfeng, and Chih-Lin I

CONTENTS

Multiple Access Rateless Network Coding for Machine-to-Machine Communications 35

JIAO Jian, Rana Abbas, LI Yonghui, and ZHANG Qinyu

Multiple Access Technologies for Cellular M2M Communications 42

Mahyar Shirvanimoghaddam and Sarah J. Johnson

Review

Software Defined Optical Networks and Its Innovation Environment 50

LI Yajie, ZHAO Yongli, ZHANG Jie, WANG Dajiang, and WANG Jiayu

Research Paper

Depth Enhancement Methods for Centralized Texture-Depth Packing Formats 58

YANG Jar-Ferr, WANG Hung-Ming, and LIAO Wei-Chen

Roundup

New Members of *ZTE Communications* Editorial Board 57

ZTE COMMUNICATIONS
Vol. 14 No. 4 (Issue 53)
Quarterly
First English Issue Published in 2003

Supervised by:
Anhui Science and Technology Department

Sponsored by:
Anhui Science and Technology Information Research Institute and ZTE Corporation

Staff Members:
Editor-in-Chief: CHEN Jie
Executive Associate Editor-in-Chief: HUANG Xinming
Editor-in-Charge: ZHU Li
Editors: XU Ye, LU Dan, ZHAO Lu
Producer: YU Gang
Circulation Executive: WANG Pingping
Assistant: WANG Kun

Editorial Correspondence:
Add: 12F Kaixuan Building,
329 Jinzhai Road,
Hefei 230061, P. R. China
Tel: +86-551-65533356
Fax: +86-551-65850139
Email: magazine@zte.com.cn

Published and Circulated (Home and Abroad) by:
Editorial Office of
ZTE Communications

Printed by:
Hefei Tiancai Color Printing Company

Publication Date:
October 25, 2016

Publication Licenses:

ISSN 1673-5188
CN 34-1294/ TN

Advertising License:
皖合工商广字0058号

Annual Subscription:
RMB 80

Guest Editorial
YUAN Jinhong, XIANG Jiying, DING Zhiguo, and YUAN Zhifeng

Multiple Access Techniques for 5G

▶ YUAN Jinhong

YUAN Jinhong received his BE and PhD degrees in electronics engineering from Beijing Institute of Technology in 1991 and 1997. From 1997 to 1999, he was a research fellow at the School of Electrical Engineering, University of Sydney, Australia. In 2000, he joined the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia, and is currently a professor of telecommunications there. Dr. Yuan has authored two books, three book chapters, and more than 200 papers for telecom journals and conferences. He has also authored 40 industry reports. He is a co-inventor of one patent on MIMO systems and two patents on low-density parity-check (LDPC) codes. He has co-authored three papers that won Best Paper Awards or Best Poster Awards. Dr. Yuan served as the NSW Chair of the joint Communications/Signal Processions/Ocean Engineering Chapter of IEEE during 2011–2014. He is an IEEE fellow and an associate editor for *IEEE Transactions on Communications*. His research interests include error-control coding and information theory, communication theory, and wireless communications.

▶ DING Zhiguo

DING Zhiguo received his BEng in electrical engineering from Beijing University of Posts and Telecommunications, China in 2000, and the PhD degree in electrical engineering from Imperial College London, UK in 2005. From Jul. 2005 to Aug. 2014, he worked in Queen's University Belfast, Imperial College and Newcastle University, UK. Since Sept. 2014, he has been with Lancaster University, UK as a chair professor. From Oct. 2012 to Sept. 2017, he has also been an academic visitor in Princeton University, USA. His research interests are 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He is serving as an editor for *IEEE Transactions on Communications*, *IEEE Transactions on Vehicular Technology*, *IEEE Wireless Communication Letters*, *IEEE Communication Letters*, and *Journal of Wireless Communications and Mobile Computing*. He received the best paper award in IET Comm. Conf. on Wireless, Mobile and Computing, 2009, IEEE Communication Letter Exemplary Reviewer 2012, and the EU Marie Curie Fellowship 2012–2014.

▶ XIANG Jiying

XIANG Jiying, PhD, is the Chief Scientist of ZTE Corporation. His research is focused on 3G, 4G, 5G, and multi-mode wireless infrastructure technologies. He led the development of the first commercial SDR base station in the industry in 2007. He proposed the first solution that support COMP on non-ideal backhaul (also called Cloud Radio) in 2012. In 2014, he proposed the "pre-5G" conception, which includes massive MIMO, D-MIMO, MUSA, and UDN. Pre-5G allows 5G-like user experience on legacy 4G handsets.

▶ YUAN Zhifeng

YUAN Zhifeng received his MS degree in signal and information processing from Nanjing University of Post and Telecommunications, China in 2005. He has been working at the Wireless Technology Advance Research Department, ZTE Corporation since 2006 and as the leader of the New Multi-Access (NMA) for 5G Wireless System Team since 2012. His research interests include wireless communications, MIMO systems, information theory, multiple access, error control coding, adaptive algorithm, and high-speed VLSI design.

Over the past few decades, wireless communications have advanced tremendously and have become an indispensable part of our lives. Wireless networks have become more and more pervasive in order to guarantee global digital connectivity. Wireless devices have quickly evolved into multimedia smartphones running applications that demand high-speed and high-quality data connections. The upcoming fifth generation (5G) mobile cellular networks are required to provide significant increase in network throughput, cell-edge data rates, massive connectivity, superior spectrum efficiency, high energy efficiency and low latency, compared with the currently deployed long-term evolution (LTE) and LTE-advanced networks. To meet these demanding challenges of 5G networks, innovative technologies on radio air-interface and radio access network (RAN) are of great importance in PHY designs. Recently non-orthogonal multiple access (NOMA) has attracted increasing research interests from both academic and industrial fields as a potential radio access technique. A few examples include multiuser shared access (MUSA), sparse code multiple access (SCMA), resource spread multiple access (RSMA) and pattern division multiple access (PDMA) proposed by ZTE, Huawei, Qualcomm, DTmobile, etc. In the mean time, multicarrier (MC) technologies that divide frequency spectrum into many narrow subchannels, such as filter bank multicarrier (FBMC) and generalized frequency division multiplexing (GFDM), become attractive and new concepts for dynamic access spectrum management and cognitive radio applications.

With these new developments, this special issue is dedicated to multiple access transmission technologies and

Guest Editorial
YUAN Jinhong, XIANG Jiying, DING Zhiguo, and YUAN Zhifeng

related for 5G cellular mobile communications. The main focus is on the cutting-edge research, review and application on non-orthogonal multiple access and related signal processing and coding methods for the air-interface of 5G enhanced mobile broadband (eMBB), mMTC, and ultra reliable and low latency communication (URLLC). Papers for this issue were invited, and after peer review, six were selected for publication. The selected papers cover reviews of various uplink and downlink NOMA schemes, novel designs for MIMO-FBMC systems, review and new designs on multiple access technologies for cellular M2M communications and IoT applications. This issue is intended to be a timely, high-quality forum for scientists and engineers.

In "Evaluation of Preamble Based Channel Estimation for MIMO-FBMC Systems" by Taheri, Ghoraishi, XIAO, CAO and GAO, the authors discuss a candidate waveform design for future wireless communications based on MIMO-FBMC and tackle the challenging problem of channel estimation facing the waveform design. Specifically, they propose a novel channel estimation method which employs intrinsic interference cancellation at the transmitter side. Their research results demonstrate that the proposed novel technique incurs less pilot-overhead compared to the well-known intrinsic approximation methods (IAM). In addition, it also has a better PAPR, BER and MSE performance.

In "Non-Orthogonal Multiple Access Schemes for 5G," YAN, YUAN, LI, and YUAN provide a comprehensive review of six potential multiple access schemes for 5G, including MUSA, RSMA, SCMA, PDMA, interleaver-division multiple access (IDMA) and NOMA. The principles, advantages and disadvantages of these multiple access schemes are discussed. More importantly, this review offers a comprehensive comparison of these solutions from the perspective of user overload, receiver type, receiver complexity, performance and grant-free transmission.

In "A Survey of Downlink Non-Orthogonal Multiple Access for 5G Wireless Communication Networks" by WEI, YUAN, Ng, Elkashlan and DING, the authors use a simple downlink model with two users served by a single-carrier to illustrate the basic principles of NOMA and its performance. The related questions and designs for a more general model with an arbitrary number of users and multiple carriers are discussed. In addition, an overview of existing works on performance analysis, resource allocation, and multiple-input multiple-output NOMA are summarized and discussed. The key features of NOMA and its potential research challenges in future networks are raised.

In "Unified Framework Towards Flexible Multiple Access Schemes for 5G", SUN, WANG, HAN and I provide a comprehensive overview for the multiple access schemes proposed for 5G networks. The authors distinguish three types of multiple access techniques in power, code and interleaver based solutions, respectively. The key features of these multiple access techniques are highlighted, and the authors also provide comparison among these multiple access techniques. Another important contribution of this paper is that a unified framework of the aforementioned multiple access techniques is provided.

In "Multiple Access Rateless Network Coding for Machine-to-Machine Communications" by JIAO, Abbas, LI and ZHANG, the authors propose a novel multiple access rateless network coding scheme for machine-to-machine (M2M) communications. The scheme is capable of increasing transmission efficiency by reducing occupied time slots yet with high decoding success rates. In addition, in contrast to existing state-of-the-art coding schemes, the novel rateless network coding is able to dynamically recode, making it suitable for M2M multicast networks with heterogeneous erasure features.

In "Multiple Access Technologies for Cellular M2M Communications", Shirvanimoghaddam and Johnson provide a comprehensive survey of the multiple access techniques for machine-to-machine (M2M) communications in future wireless cellular networks. In particular, the overview highlights the multiple access strategies and explains their limitations when used for M2M communications. The throughput efficiency of different multiple access techniques when used in coordinated and uncoordinated scenarios are illustrated. The authors demonstrate that in uncoordinated scenarios, NOMA can support a larger number of devices compared to orthogonal multiple access techniques.

We thank all authors for their valuable contributions and all reviewers for their timely and constructive comments on the submitted papers. We hope the content of this issue is informative and helpful to all readers.

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 001
http://www.cnki.net/kcms/detail/34.1294.TN.20161014.0955.002.html, published online October 14, 2016

Evaluation of Preamble Based Channel Estimation for MIMO-FBMC Systems

Sohail Taheri[1], Mir Ghoraishi[1], XIAO Pei[1], CAO Aijun[2], and GAO Yonghong[2]
(1. 5G Innovation Centre, Institute for Communication Systems (ICS), University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom;
2. ZTE Wistron Telecom AB, Kista, Stockholm 164 51, Sweden)

Abstract

Filter-bank multicarrier (FBMC) with offset quadrature amplitude modulation (OQAM) is a candidate waveform for future wireless communications due to its advantages over orthogonal frequency division multiplexing (OFDM) systems. However, because of orthogonality in real field and the presence of imaginary intrinsic interference, channel estimation in FBMC is not as straightforward as OFDM systems especially in multiple antenna scenarios. In this paper, we propose a channel estimation method which employs intrinsic interference cancellation at the transmitter side. The simulation results show that this method has less pilot overhead, less peak to average power ratio (PAPR), better bit error rate (BER), and better mean square error (MSE) performance compared to the well-known intrinsic approximation methods (IAM).

Keywords

channel estimation; filter-bank multicarrier (FBMC); multiple-input multiple-output (MIMO); offset quadrature amplitude modulation (OQAM); wireless communication

1 Introduction

Orthogonal frequency division multiplexing (OFDM) has been widely used in communication systems in the last decade. This is because of its immunity to multipath fading and simplicity of channel estimation and data recovery with a low complexity single-tap equalization, and also suitability for multiple-input multiple-output (MIMO) systems [1]. However, it suffers from disadvantages such as sensitivity to carrier frequency offset (CFO), significant out-of-band radiation, and cyclic prefix overhead. In the presence of CFO, there is loss of orthogonality between subcarriers leading to inter carrier interference (ICI). Moreover, to efficiently use the available spectrum, a waveform with very low spectral leakage is needed.

Because of the OFDM shortcomings, filter-bank multicarrier (FBMC) modulation combined with offset quadrature amplitude modulation (OQAM) has drawn attention in the last decade [2], [3]. Regardless of the higher complexity compared to OFDM, FBMC (known as OFDM/OQAM and FBMC/OQAM in the literature) provides significantly reduced out-of-band emissions, robustness against CFO [4], and under certain conditions, better spectral efficiency as there is no need to use cyclic prefix (CP) [5]. These advantages come from well localized prototype filters in time and frequency domain for pulse shaping. Accordingly, FBMC can be a promising alternative to conventional radio access techniques to improve wireless access capacity.

On the other hand, as orthogonality in FBMC systems only holds in the real field, received symbols are contaminated with an imaginary intrinsic interference term coming from the neighbouring real symbols. The interference becomes a source of problem in channel estimation and equalization processes, especially in MIMO systems. The pilot symbols used for channel estimation should be protected from interference as the receiver has no knowledge about their neighbours to estimate the amount of interference. These protections cause overheads when designing a transmission frame. In a preamble-based approach, the preamble should be protected from the subsequent data transmission and the previous frame by inserting null symbols, which causes longer preamble and thus more overhead compared to OFDM. This is also true for scattered pilots where the neighbouring data symbols contribute to the interference on the pilots [6]. In this scenario, typically one or two time-frequency points adjacent to the pilots are used to cancel the interference on the pilots [7]–[10].

Interference Approximation Method (IAM) for preamble-

This work is supported by ZTE Industry-Academia-Research Cooperation Funds under Grant No. Surrey-Ref-9953.

based channel estimation in single-input, single-output (SISO) systems was first introduced in [11]. The preamble was named IAM-R in the literature, where R denotes real-valued pilots. Alternatively, IAM-I and IAM-C were introduced in [12], [13], where I and C stand for imaginary and complex pilots. Those preamble based channel estimation schemes were extended to FBMC-MIMO systems in [14]. In IAM-I and IAM-C, pilots on each subcarrier interfere with their adjacent subcarriers in a constructive way. That is, these methods use the intrinsic interference to enhance amplitude of the pilots. As a result, better performance of channel estimation is achieved. Despite good performance, IAM methods suffer from increased pilot overhead, i.e., a number of zero symbols are required to protect the pilot symbols from the interference of their adjacent symbols. While the number of pilot symbols is equal to the number of antennas, the total number of symbols in the preamble will be more than twice the number of transmit antennas.

This paper proposes a channel estimation method with reduced preamble overhead compared to the IAM family. The idea was first introduced in [15] for MIMO-OFDM. Applying this method to MIMO-FBMC with spatial multiplexing needs further consideration to cancel intrinsic interference. By using basic idea of zero forcing from single antenna, this method has modest computation complexity, while it can outperform IAM methods in terms of peak to average power ratio (PAPR), bit error rate (BER), and mean square error (MSE) under perfect synchronization conditions and in presence of carrier frequency offset.

The rest of this paper is organized as follows: Section 2 reviews the MIMO-FBMC systems, the effect of intrinsic interference, and the conventional channel estimation methods. In Section 3, the new method for channel estimation is proposed and Section 4 shows the results and comparisons with IAM methods. Finally, conclusions are drawn in Section 5.

2 MIMO-FBMC System

2.1 System Model

FBMC systems are implemented by a prototype filter $g(t)$ and synthesis and analysis filter-banks in transmitter and receiver side respectively. The real and imaginary parts of complex symbols are separated in two different branches where they are modulated in FBMC modulators as real symbols. Therefore, at a specific time, each subcarrier in this system carries a real-valued symbol. Denoting T_0 as symbol duration and F_0 as subcarrier spacing in OFDM systems, duration and subcarrier spacing in FBMC are either $\tau_0 = \frac{T_0}{2}$, $\nu_0 = F_0$ or $\tau_0 = T_0$, $\nu_0 = \frac{F_0}{2}$ [16]. For the system model in this paper, the former approach is adopted. That is, subcarrier spacing remains the same as OFDM, while symbol duration is reduced by half.

Assuming a multiple antenna scenario with P transmit antennas, Q receive antennas, and M subcarriers, the baseband signal to be transmitted over the pth branch in general form is expressed as

$$s^{(p)}(t) = \sum_{n=-\infty}^{+\infty} \sum_{m=0}^{M-1} a_{m,n}^{(p)} g_{m,n}(t), \tag{1}$$

where $a_{m,n}^{(p)}$ is the real-valued symbol, and $g_{m,n}(t)$ is the shifted version of the prototype filter on the mth subcarrier and at nth symbol duration:

$$g_{m,n}(t) = j^{m+n} e^{j2\pi m\nu_0 t} g(t - n\tau_0). \tag{2}$$

The prototype filter $g(t)$ is designed to keep its shifted versions are orthogonal only in the real field [17], i.e.,

$$R\left(\int g_{m,n}(t) g^*_{m_0,n_0}(t)\mathrm{d}t\right) = \delta_{m,m_0}\delta_{n,n_0}, \tag{3}$$

where $R(.)$ denotes the real-part of a complex number. As a consequence, the outputs of the analysis filter-bank have a so-called intrinsic interference term which is pure imaginary. The demodulated signal on the qth receive antenna at a particular subcarrier and symbol point (m_0,n_0) is given by

$$y_{m_0,n_0}^{(q)} = \sum_{p=1}^{P} h_{m_0,n_0}^{q,p} a_{m_0,n_0}^{(p)} + jI_{m_0,n_0}^{(q)} + \eta_{m_0,n_0}^{(q)}, \tag{4}$$

where $h_{m_0,n_0}^{q,p}$ is channel frequency response at (m_0,n_0) between q^{th} receive and p^{th} transmit antenna, $\eta_{m_0,n_0}^{(q)}$ is the noise component at q^{th} receive antenna, and the interference term $I_{m_0,n_0}^{(q)}$ is formed as

$$jI_{m_0,n_0}^{(q)} = \sum_{p=1}^{P} \sum^{(m,n)\neq(m_0,n_0)} h_{m,n}^{q,p} a_{m,n}^{(p)} \langle g \rangle_{m,n}^{m_0,n_0}. \tag{5}$$

In (5), $\langle g \rangle_{m,n}^{m_0,n_0}$ is expressed as

$$\langle g \rangle_{m,n}^{m_0,n_0} = \int g_{m,n}(t) g^*_{m_0,n_0}(t)\mathrm{d}t. \tag{6}$$

Having the prototype filter $g(t)$ well localized in time and frequency, it can be assumed that the intrinsic interference is mostly due to the first-order neighbouring points. That is, (m,n) in (5) can take the values of Ω^* as follows [6]:

$$\Omega^* = \{(m_0,n_0\pm1),(m_0\pm1,n_0),(m_0\pm1,n_0\pm1)\}, \tag{7}$$

which covers the (m_0,n_0) point first-order neighbours. By assuming constant channel frequency response over (m_0,n_0) and Ω^*, we can simplify (5) as

$$jI_{m_0,n_0}^{(q)} = \sum_{p=1}^{P} h_{m_0,n_0}^{p,q} \sum^{(m,n)\in\Omega^*} a_{m,n}^{(p)} \langle g \rangle_{m,n}^{m_0,n_0}. \tag{8}$$

Consequently, (4) can be written as

$$y_{m_0,n_0}^{(q)}=\sum_{p=1}^{P}h_{m_0,n_0}^{p,q}\left(\underbrace{\underbrace{a_{m_0,n_0}^{(p)}+ju_{m_0,n_0}^{(p)}}_{c_{m_0,n_0}^{(p)}}}\right)+\eta_{m_0,n_0}^{(p)}, \tag{9}$$

where

$$ju_{m_0,n_0}^{(p)}=\sum^{(m,n)\in\Omega^*}a_{m,n}^{(p)}\left\langle g\right\rangle_{m,n}^{m_0,n_0}. \tag{10}$$

Table 1 shows the number of $\left\langle g\right\rangle_{m,n}^{m_0,n_0}$ coefficients on the first-order neighbours of the point (m_0,n_0). The weights of interference, β, γ, and δ, depend on the prototype filter and have been derived in [18]. In this work, the isotropic orthogonal transform algorithm (IOTA) [19] filter is employed. It exploits the symmetrical property of Gaussian function in time and frequency. Therefore, the amount of interference out of first-order neighbouring points is negligible. The weights of interference for this filter are $\beta=0.2486$, $\gamma=0.5755$, and $\delta=0.1898$ (Table 1).

The MIMO-FBMC signal model can be represented as

$$\begin{pmatrix}y_{m_0,n_0}^{(1)}\\ \vdots\\ y_{m_0,n_0}^{(Q)}\end{pmatrix}=\begin{pmatrix}h_{m_0,n_0}^{1,1}&\cdots&h_{m_0,n_0}^{1,P}\\ \vdots&\ddots&\vdots\\ h_{m_0,n_0}^{Q,1}&\cdots&h_{m_0,n_0}^{Q,P}\end{pmatrix}\begin{pmatrix}c_{m_0,n_0}^{(1)}\\ \vdots\\ c_{m_0,n_0}^{(Q)}\end{pmatrix}+\begin{pmatrix}\eta_{m_0,n_0}^{(1)}\\ \vdots\\ \eta_{m_0,n_0}^{(Q)}\end{pmatrix} \tag{11}$$

where $c_{m_0,n_0}^{(p)}$ is defined in (9). To retrieve the transmitted symbols from the system above, it is necessary to have an evaluation of the channel coefficients, which are used to detect the linearly combined demodulated complex symbols $c_{m_0,n_0}^{(p)}$ at each receiver branch using zero forcing (ZF), minimum mean square error (MMSE), or maximum likelihood (ML). In $c_{m_0,n_0}^{(p)}$, the imaginary parts are intrinsic interference terms. By taking $R\{.\}$ operation, the transmitted symbols $a_{m_0,n_0}^{(p)}=R\left\{c_{m_0,n_0}^{(p)}\right\}$ are recovered.

2.2 Channel Estimation

To obtain the channel information over one frame duration on each receive antenna, we need to know the transmitted pilot symbols. The number of these pilot symbols should be equal to P to form a linear equation system with the least square estimation method. For simplicity, let us consider a 2-by-2 antenna scenario. By allocating two pilot symbols at times $n=n_0$ and $n=n_1$ on each antenna, the equation set of the system on subcarrier m is given by

$$\begin{pmatrix}y_{m,n_0}^{(1)}&y_{m,n_1}^{(1)}\\ y_{m,n_0}^{(2)}&y_{m,n_1}^{(2)}\end{pmatrix}=\begin{pmatrix}h_{m,n_0}^{1,1}&h_{m,n_0}^{1,2}\\ h_{m,n_0}^{2,1}&h_{m,n_0}^{2,2}\end{pmatrix}\begin{pmatrix}x_{m,n_0}^{(1)}&x_{m,n_1}^{(1)}\\ x_{m,n_0}^{(2)}&x_{m,n_1}^{(2)}\end{pmatrix}+\begin{pmatrix}\eta_{m,n_0}^{(1)}&\eta_{m,n_1}^{(1)}\\ \eta_{m,n_0}^{(2)}&\eta_{m,n_1}^{(2)}\end{pmatrix}. \tag{12}$$

In (12), $x_{m,n}^{(p)}$ are pilot symbols. We have assumed that there is no significant variations in the channel between time slots n_0 and n_1. Hence, we can drop the time subscript and express (12) as

$$\mathbf{Y}_m=\mathbf{H}_m\mathbf{X}_m+\boldsymbol{\eta}_m. \tag{13}$$

Thus, channel coefficients can be calculated by the least square estimation method:

$$\hat{\mathbf{H}}_m=\mathbf{Y}_m\left(\mathbf{X}_m^H\mathbf{X}_m\right)^{-1}\mathbf{X}_m^H=\mathbf{H}_m+\boldsymbol{\eta}_m\left(\mathbf{X}_m^H\mathbf{X}_m\right)^{-1}\mathbf{X}_m^H, \tag{14}$$

or in a special case with the equal number of transmit and receive antenna:

$$\hat{\mathbf{H}}_m=\mathbf{Y}_m\mathbf{X}_m^{-1}=\mathbf{H}_m+\boldsymbol{\eta}_m\mathbf{X}_m^{-1}. \tag{15}$$

The preamble in the IAM methods is composed of $2P+1$ symbols. That is, the length of the preamble grows linearly with P. The symbols with even time indices are pilots, while other symbols are all zeros to protect pilots from intrinsic interference. Based on the values of pilot symbols, i.e. real, imaginary, or complex valued pilots, IAM-R, IAM-I and IAM-C were proposed. In these approaches, the channel coefficients can be obtained using (12). For P=2, pilot symbols in (12) are set as $x_{m,n_0}^{(1)}=x_{m,n_1}^{(1)}=x_{m,n_0}^{(2)}=-x_{m,n_1}^{(2)}=x_m$. Hence, they form a system based on (12) as

$$\mathbf{Y}_m=x_m\mathbf{H}_m\begin{pmatrix}1&1\\ 1&-1\end{pmatrix}+\boldsymbol{\eta}_m=x_m\mathbf{H}_m\mathbf{A}_2+\boldsymbol{\eta}_m, \tag{16}$$

where $\mathbf{A}_2=\mathbf{A}_2^{-1}$ is an orthogonal matrix if omitting the constant coefficient of the inverse [14]. Finally, the channel coefficients are obtained as follows:

$$\hat{\mathbf{H}}_m=\frac{1}{x_m}\mathbf{Y}_m\mathbf{A}_2=\mathbf{H}_m+\frac{1}{x_m}\boldsymbol{\eta}_m\mathbf{A}_2. \tag{17}$$

The length of the preamble in this method is $2P+1=5$ with just two pilot symbols. As a result, this approach suffers from significant pilot overhead which reduces the spectral efficiency. Furthermore, the periodic nature of the pilots in these preambles results in high PAPR at the output of the synthesis filter-

▼**Table 1. Weights of interference on the first-order neighbours**

	n_0-1	n_0	n_0+1
m_0-1	$(-1)^{m_0}\delta$	$-\beta$	$(-1)^{m_0}\delta$
m_0	$-(-1)^{m_0}\gamma$	1	$(-1)^{m_0}\gamma$
m_0+1	$(-1)^{m_0}\delta$	β	$(-1)^{m_0}\delta$

bank [14].

3 Proposed Method

In order to reduce the preamble overhead and accordingly increase the spectral efficiency, a novel channel estimation approach with modest computation complexity is proposed. Since there is no need to have an estimation of the channel on each subcarrier, we can reduce the number of pilot symbols to one. In this way, each subcarrier is allocated to only one branch to transmit pilot. That is, while a branch is transmitting pilot on a subcarrier, the other branches remain silent. Therefore, the channel parameters between the receive branch and the pilot transmitting branch on that specific subcarrier can be obtained. This method enables the increase of transmit branches with a constant length of the preamble.

To elaborate the system more precisely, we assume a 2x2 MIMO system where preambles for branches 1 and 2 are shown in **Fig. 1**. It can be seen that the first and third symbols are all zero to protect the preamble from intrinsic interference from data section and previous frame. In the middle symbol for branch 1, complex pilots are placed on odd subcarriers, while the other subcarriers carry zeros. On branch 2, orthogonal pilots to branch 1 are sent, i.e., even subcarriers carry complex pilots and the rest are zero valued. On a particular subcarrier $m = m_0$, the system equations is written as follows:

$$\begin{pmatrix} y_{m_0}^{(1)} \\ y_{m_0}^{(2)} \end{pmatrix} = \begin{pmatrix} h_{m_0}^{1,1} & h_{m_0}^{1,2} \\ h_{m_0}^{2,1} & h_{m_0}^{2,2} \end{pmatrix} \begin{pmatrix} x_{m_0}^{(1)} \\ x_{m_0}^{(2)} \end{pmatrix} + \begin{pmatrix} \eta_{m_0}^{(1)} \\ \eta_{m_0}^{(2)} \end{pmatrix}. \tag{18}$$

On odd subcarriers $m_0 = 2k+1$, we have $x_{m_0}^{(1)} = X_{m_0}$, while $x_{m_0}^{(2)} = 0$. Then, the channel coefficients $h_{m_0}^{1,1}$ and $h_{m_0}^{2,1}$ are obtained as

$$\begin{aligned} h_{m_0}^{1,1} &= \frac{y_{m_0}^{(1)}}{X_{m_0}}\Big|_{x_{m_0}^{(2)}=0} \\ h_{m_0}^{2,1} &= \frac{y_{m_0}^{(2)}}{X_{m_0}}\Big|_{x_{m_0}^{(2)}=0}. \end{aligned} \tag{19}$$

Likewise, on even subcarriers the channel coefficients of $h_{m_0}^{1,2}$ and $h_{m_0}^{2,2}$ are given by

$$\begin{aligned} h_{m_0}^{1,2} &= \frac{y_{m_0}^{(1)}}{X_{m_0}}\Big|_{x_{m_0}^{(1)}=0} \\ h_{m_0}^{2,2} &= \frac{y_{m_0}^{(2)}}{X_{m_0}}\Big|_{x_{m_0}^{(1)}=0}. \end{aligned} \tag{20}$$

Hence, we have calculated the channel parameters between each pair of antennas on alternative subcarriers. Channel Coefficients on the rest of subcarriers can be obtained by interpolation. Due to short distance between pilots in this system, linear interpolation provides enough accuracy with the advantage of low complexity.

The technique works perfectly for MIMO - OFDM systems [15]. When applying this method to MIMO-FBMC, intrinsic interference degrades the channel estimation performance, i.e., transmitted pilots from one branch interfere with the received pilots on other branch. Consequently, the conditions in (19) and (20) no longer hold. To tackle this problem, we propose a precoding approach in which the interference is calculated at the transmitter side. Then, the zero points in pilot symbols are replaced by $I_{m,n}$, so that there are no interference on the corresponding points at the receiver side. That is, the pilots are received without any interference from other branches.

Fig. 2 shows the precoded pilots. The value of cancelling interference on subcarrier m is calculated by using (10) as

$$I_{m,n} = -\sum_{(m',n)\in\Omega^*} a_{m,n}^{(p)} \langle g \rangle_{m,n}^{m',n'}. \tag{21}$$

Moreover, the adjacent points of the pilot X_m are filled with pre-calculated values to maximize the received signal energy, thereby to enhance the estimation accuracy [18]. Defining $X_m = X_m^R + jX_m^I$, These values would be

$$\begin{aligned} X_m' &= -jX_m^I \\ X_m'' &= -X_m^R. \end{aligned} \tag{22}$$

Consequently, the amplitude of the real and imaginary parts

0	X_{m-3}	0	0	0	0
0	0	0	0	X_{m-2}	0
0	X_{m-1}	0	0	0	0
0	0	0	0	X_m	0
0	X_{m+1}	0	0	0	0
0	0	0	0	X_(m+2)	0
	Branch 1			Branch 2	

▲**Figure 1. The basic preamble for two antennas.**

X_{m-3}'	X_{m-3}	X_{m-3}''	0	$I_{m-3,1}$	0
0	$I_{m-2,1}$	0	X_{m-2}'	X_{m-2}	X_{m-2}''
X_{m-1}'	X_{m-1}	X_{m-1}''	0	$I_{m-1,1}$	0
0	$I_{m,1}$	0	X_m'	X_m	X_m''
X_{m+1}'	X_{m+1}	X_{m+1}''	0	$I_{m+1,1}$	0
0	$I_{m+2,1}$	0	X_{m+2}'	X_{m+2}	X_{m+2}''
	Branch 1			Branch 2	

▲**Figure 2. The preambles for two antennas after interference cancellation of the first and third time symbols that helps the pilots become stronger.**

of the received pilots becomes

$$\begin{aligned}\left|\hat{X}_m^R\right| &= \left|X_m^R\right| + \gamma\left|X_m^{'}\right| + \gamma\left|X_m^I\right| \\ \left|\hat{X}_m^I\right| &= \left|X_m^I\right| + \gamma\left|X_m^{''}\right| + \gamma\left|X_m^R\right|\end{aligned}, \tag{23}$$

where γ is the interference weight shown in Table 1. The complete design of the preambles is displayed in Fig. 2. The pilots can take arbitrary values. In this work, the maximum amplitude of the used QAM modulation is used so that $X_m^R = X_m^I$. In order to avoid high PAPR, the sign of the pilots should be changed alternatively after a number of repetitions. The final value of the received pilots in (23) with $X_m^R = X_m^I$ is

$$\hat{X}_m = (1+2\gamma)X_m. \tag{24}$$

The extension to P-branch MIMO system is straightforward. In this case, one subcarrier of every P subcarriers carries a pilot (non-zero), while each branch's pilot symbol is orthogonal to other branches. The more transmit branches, the more distance between pilot subcarriers. Consequently, for larger number of branches, the quality of channel estimation degrades.

4 Simulation Results

In this section, different preamble-based channel estimators for a 2x2 MIMO-FBMC system are simulated and compared. The simulations are performed using 7-tap EPA-5Hz and 9-tap ETU-70Hz channel models with low spatial correlations. Perfect synchronization is assumed for BER and MSE comparison, i.e., there is no timing or frequency offset errors. In order to detect symbols, MMSE equalizer is used. **Table 2** summarizes the simulation parameters.

The results are compared with IAM-R and IAM-C methods introduced in [14]. For fair comparison, the transmission power is kept equal for all methods. In this system, $\frac{E_b}{N_0}$ is defined by

$$\frac{E_b}{N_0} = Q\frac{SNR}{\alpha \times \log_2(M)}, \tag{25}$$

where $M = 16$ is the modulation order, SNR is signal-to-noise ratio, and $\alpha = N_s - \frac{N_p}{N_s}$ with the frame length $N_s = 14$ and the preamble length N_p. The length of preamble N_p in the proposed method is three symbols resulting in 40% overhead reduction compared to IAMs. As a result, a performance gain is expected due to shorter preamble. The extra symbols generated by the synthesis filter-banks can be dropped before transmission, but one of them with the most power should be kept to avoid filtering errors after demodulation, i.e., $N_s + 1$ symbols are transmitted. To consider this extra symbol, α can be changed to $\alpha = N_s - \frac{N_p}{N_s} + 1$.

▼**Table 2. Simulation parameters**

Modulation type	M-QAM, M =16
FFT size	256
Used subcarriers	144
Sampling frequency	3.84 MHz
Symbols per frame	14
Channel	EPA 5 Hz, ETU 70 Hz

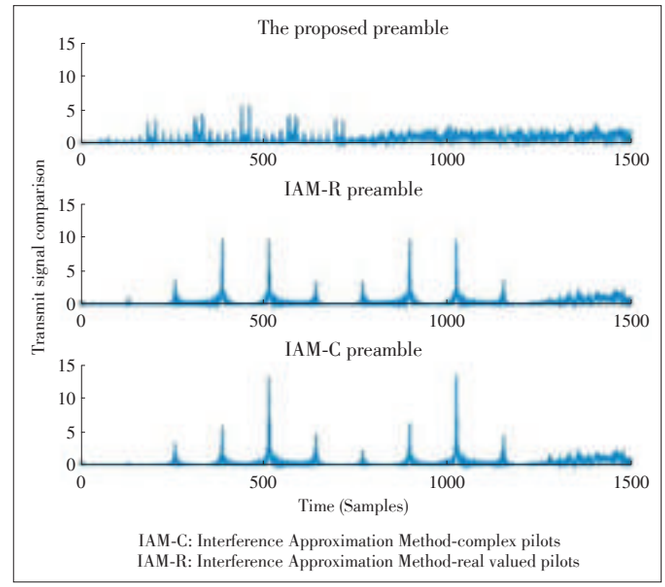
EPA: Extended Pedestrian A model
ETU: Extended Typical Urban model
FFT: fast Fourier transform
QAM: quadrature amplitude modulation

4.1 PAPR Comparison

Fig. 3 shows the comparison between the proposed method and IAMs in terms of PAPR. The plots show the squared magnitude of the preambles at the output of the synthesis filter-bank on branch 1. Evidently, from the point of practical implementations, the proposed method is preferable. Whereas in the others, the signal level should be kept very low to avoid A/D saturations. The PAPR levels for the pilot symbols are compared in **Table 3** for the three methods.

4.2 Channel Estimation Performance Comparison

Fig. 4 shows the MSE comparison of the channel estimation methods. To calculate MSE, the channel tap on the second symbol in frame is considered as reference and it is assumed constant during the symbol duration. Then, the MSE is calcu-

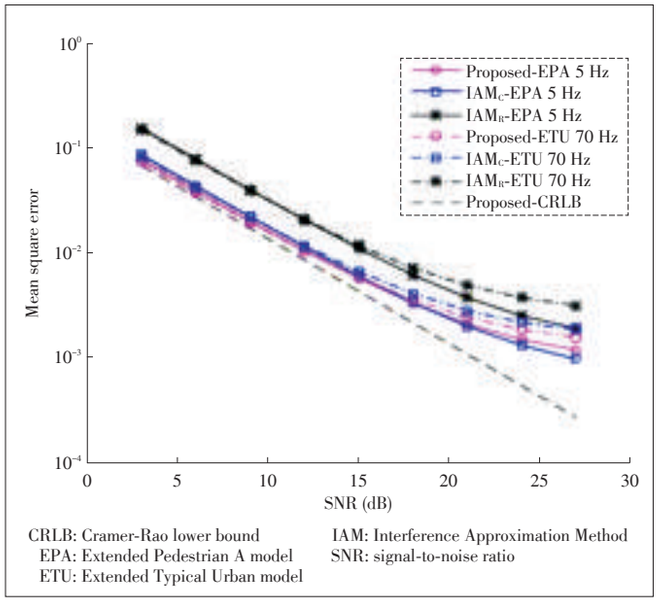


▲**Figure 3. Squared magnitude of the preambles on output of the branch 1.**

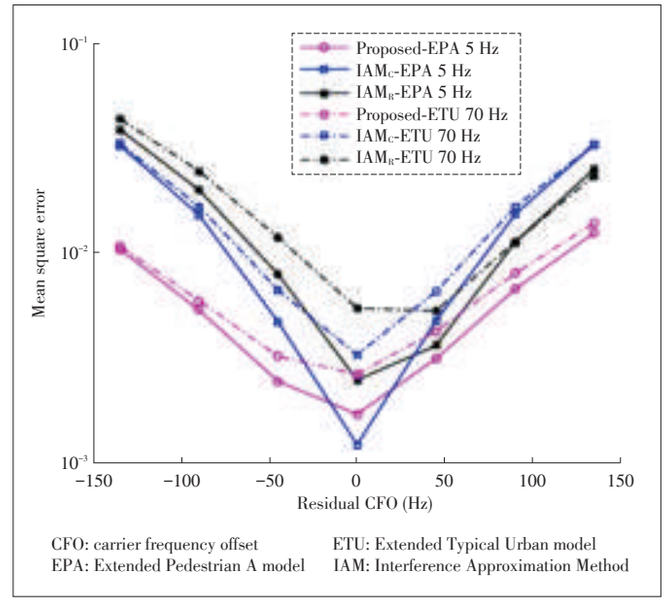
▼**Table 3. PAPR comparison for the three methods**

	IAM-C	IAM-R	Proposed
PAPR	17.5	9.3	7.2

IAM-C: Interference Approximation Method-complex pilots
IAM-R: Interference Approximation Method-real valued pilots
PAPR: peak to average power ratio



▲Figure 4. MSE performance of the channel estimation methods.



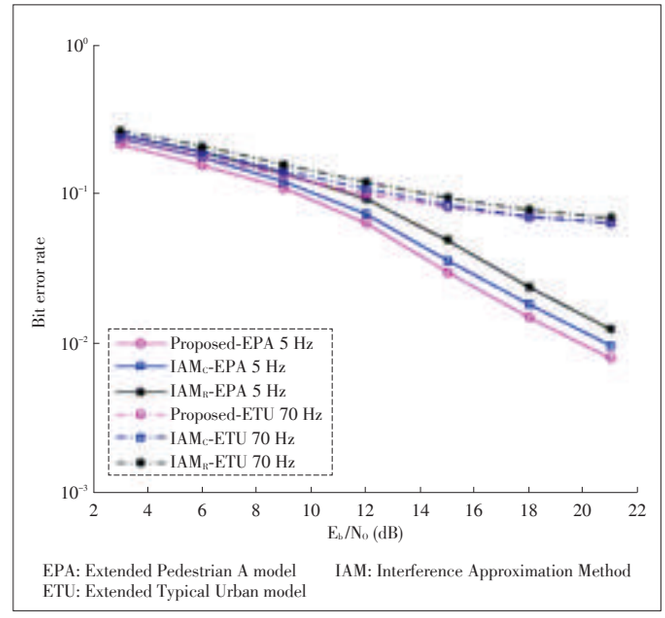
▲Figure 5. MSE performance of the channel estimation methods in presence of residual CFO.

lated using the estimated channel $\hat{\mathbf{H}}$ as $\mathrm{E}\left(\left(\mathbf{H}-\hat{\mathbf{H}}\right)^{H}\left(\mathbf{H}-\hat{\mathbf{H}}\right)\right)$. It can be seen that the proposed preamble outperforms IAM-R and has approximately the same performance as IAM-C in both channel models. In the EPA-5Hz scenario, the proposed method gradually reaches an error floor. This is due to domination of errors from ISI and interference cancellation residual. However, the performance is still as good as IAM-C. In the ETU-70Hz scenario, because of rapid variation of the channel taps, the assumption of constant channel over Ω^{*} in (8) is invalid. Consequently, the performance of all the methods degrades and reaches an error floor in higher SNRs. This is a general problem in channel estimation for FBMC systems where the receiver should necessarily have an estimation of intrinsic interferences for channel estimation. However, the degradation on IAMs is more significant as the channel is estimated using two symbols with one zero symbol in between. Therefore, as the channel is not constant over the two pilot symbols, degradation is higher than the proposed method with only one symbol for channel estimation. The Cramer-Rao lower bound (CRLB) for the proposed method, derived in Appendix A has also been plotted in the figure for benchmark comparison. It can be seen that the proposed scheme achieves closest performance to the theoretical lower bound in comparison to the other schemes.

Fig. 5 shows the MSE comparisons in terms of residual CFO. It is assumed that the CFO has been estimated and compensated before channel estimation. As the estimated CFO is not perfect, the residual CFO affects the quality of channel estimation. Therefore, the methods are compared in presence of residual CFO in the two channel scenarios without added white Gaussian noise. When the CFO is zero, the MSEs show the error floor of the methods in Fig. 4 at very high SNRs. It can be seen that in EPA channel, the error floor of the proposed method is higher than IAM-C, while it has the best performance under ETU channel. This is also true for the other values of CFO, where the degradation of MSE in the proposed method is lower than the other two in both channels.

4.3 Bit Error Rate Performance Comparison

The BER performance comparison with respect to $\frac{E_b}{N_0}$ is illustrated in **Fig. 6**. Evidently, the proposed method performs



▲Figure 6. BER performance of the channel estimation methods.

better compared to the others in low mobility EPA-5Hz scenario. In the high mobility ETU-70Hz channel, the performance deteriorates as the channel varies significantly during the frame time. Consequently, the preamble-based channel estimation is not a proper choice for high mobility applications and there is an error floor for all the curves showing around six percent bit error rate.

5 Conclusions

In this paper, we proposed a novel channel estimation algorithm with much reduced pilot overhead compared to the existing IAM based approaches. Our results show that the proposed method has better PAPR property. The system performance under low mobility and high mobility channels, as well as in the presence of CFO, has been simulated and compared. According to the results, the proposed method achieves comparable channel estimation performance to the IAM methods, and better BER performance due to shorter preamble.

Appendix A

Cramer-Rao Lower Bound for the Proposed Channel Estimation

In this section, a lower bound for the proposed channel estimator is derived. We simplify the system using equations (13), (18), (19), and (20) as

$$\mathbf{Y} = X\mathbf{H} + \boldsymbol{\eta}, \tag{26}$$

where $\mathbf{Y} = [y_1\ y_2]$ is the received signal vector, $\boldsymbol{\eta} = [\eta_1\ \eta_2]$ is the noise vector, $\mathbf{H} = [h_1\ h_2]$ is the channel vector to be estimated, X is the pilot symbol. The subcarrier index has also been dropped for simplicity.

The CRLB is a bound on the smallest covariance matrix that can be achieved by an unbiased estimator, $\hat{\mathbf{H}}$, of a parameter vector $\mathbf{H}$ as

$$\begin{aligned} &\mathbf{J}^{-1} \leqslant \mathbf{C}_{\hat{\mathbf{H}}} = \mathrm{E}\left\{(\mathbf{H}-\hat{\mathbf{H}})(\mathbf{H}-\hat{\mathbf{H}})^*\right\}; \\ &\mathbf{J} = \mathrm{E}\left\{\left(\frac{\partial \ln p(\mathbf{Y};\mathbf{H})}{\partial \mathbf{H}}\right)\left(\frac{\partial \ln p(\mathbf{Y};\mathbf{H})}{\partial \mathbf{H}}\right)^*\right\}, \end{aligned} \tag{27}$$

where $(\cdot)^*$ denotes conjuagate transpose operation, $\mathbf{J}$ is the Fisher information matrix and $\ln p(\mathbf{Y};\mathbf{H})$ is the log-likelihood function of the observed vector $\mathbf{Y}$. The vector $\mathbf{Y}$ is a complex Gaussian random vector, i.e., $\mathbf{Y} \sim CN(X\mathbf{H}, N_0\mathbf{I})$ with likelihood function and log-likelihood function as

$$\begin{aligned} &(\mathbf{Y};\mathbf{H}) = \frac{1}{(\pi N_0)^2}\exp\left[-\frac{(\mathbf{Y}-X\mathbf{H})^*(\mathbf{Y}-X\mathbf{H})}{N_0}\right] = \\ &\frac{1}{(\pi N_0)^2}\exp\left[-\frac{\|\mathbf{Y}\|^2 - \mathbf{H}^*X^*\mathbf{Y} - \mathbf{Y}^*X\mathbf{H} + \mathbf{H}^*X^*X\mathbf{H}}{N_0}\right]; \\ &\ln p(\mathbf{Y};\mathbf{H}) = K - \frac{\|\mathbf{Y}\|^2 - \mathbf{H}^*X^*\mathbf{Y} - \mathbf{Y}^*X\mathbf{H} + \mathbf{H}^*X^*X\mathbf{H}}{N_0}, \end{aligned} \tag{28}$$

where K is a constant. Taking the complex gradient [20] of $\ln p(\mathbf{Y};\mathbf{H})$ with respect to $\mathbf{H}$ yields

$$\frac{\partial \ln p(\mathbf{Y};\mathbf{H})}{\partial \mathbf{H}} = -\frac{1}{N_0}\left[X^*X\mathbf{H} - X^*\mathbf{Y}\right]^*. \tag{29}$$

The above equality holds since

$$\begin{aligned} &\frac{\partial \|\mathbf{Y}\|^2}{\partial \mathbf{H}} = 0; \qquad \frac{\partial \mathbf{H}^*X^*\mathbf{Y}}{\partial \mathbf{H}} = 0; \\ &\frac{\partial \mathbf{Y}^*X\mathbf{H}}{\partial \mathbf{H}} = \left(X^*\mathbf{Y}\right)^*; \ \frac{\partial \mathbf{H}^*X^*X\mathbf{H}}{\partial \mathbf{H}} = \left(X^*X\mathbf{H}\right)^*. \end{aligned} \tag{30}$$

Thus we can derive,

$$\begin{aligned} &\frac{\partial \ln p(\mathbf{Y};\mathbf{H})}{\partial \mathbf{H}^*} = \left(\frac{\partial \ln p(\mathbf{Y};\mathbf{H})}{\partial \mathbf{H}}\right)^* = \frac{X^*\mathbf{Y} - X^*X\mathbf{H}}{N_0} = \\ &\frac{X^*X}{N_0}\left\{\left(X^*X\right)^{-1}X^*\mathbf{Y} - \mathbf{H}\right\} = \mathbf{J}(\mathbf{H})\left[\hat{\mathbf{H}} - \mathbf{H}\right]. \end{aligned} \tag{31}$$

This proves that the minimum variance unbiased estimator of $\mathbf{H}$ is

$$\hat{\mathbf{H}} = \left(X^*X\right)^{-1}X^*\mathbf{Y} = \frac{\mathbf{Y}}{X}. \tag{32}$$

It is efficient in that it attains the CRLB. The Fisher information matrix $\mathbf{J}(\mathbf{H})$ and covariance matrix $\mathbf{C}_{\hat{\mathbf{H}}}$ of this unbiased estimator are

$$\begin{aligned} &\mathbf{J}(\mathbf{H}) = \mathrm{E}\left[\frac{X^*X\mathbf{I}_2}{N_0}\right] = \frac{\mathrm{E}\left[X^*X\right]\mathbf{I}_2}{N_0} = \frac{E_x}{N_0}\mathbf{I}_2 \\ &\mathbf{C}_{\hat{\mathbf{H}}} = \mathbf{J}^{-1}(\mathbf{H}) = \frac{N_0}{\mathrm{E}_x}\mathbf{I}_2. \end{aligned} \tag{33}$$

In (33), E_x is the pilot energy. The CRLB for each diagonal element of $\mathbf{J}^{-1}(\mathbf{H})$ is

$$\mathrm{var}\left(\hat{h}_1\right) = \mathrm{var}\left(\hat{h}_2\right) = \mathrm{diag}\left[\mathbf{C}_{\hat{\mathbf{H}}}\right]_i = \frac{N_0}{E_x}. \tag{34}$$

As the pilots in this system are amplified exploiting intrinsic interference by the factor of $1+2\gamma$, E_x should be replaced by $E_x' = (1+2\gamma)^2 E_x$. Assuming $\frac{E_x}{N_0}$ is approximately equal to SNR and considering (25), (34) becomes

$$\mathrm{var}\left(\hat{h}_1\right) = \mathrm{var}\left(\hat{h}_2\right) = \frac{N_0}{E_x}\frac{1}{(1+2\gamma)^2}. \tag{35}$$

References

[1] A. Sahin, I. Guvenc, and H. Arslan, "A survey on multicarrier communications: Prototype filters, lattice structures, and implementation aspects," *IEEE Communications Surveys Tutorials*, vol. 16, no. 3, pp. 1312–1338, Mar. 2014.

[2] B. Farhang-Boroujeny, "OFDM versus filter bank multicarrier," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 92–112, May 2011.
[3] F. Schaich and T. Wild, "Waveform contenders for 5G—OFDM vs. FBMC vs. UFMC," in *6th International Symposium on Communications, Control and Signal Processing*, Athens, Greece, 2014, pp. 457 – 460. doi: 10.1109/ISCCSP.2014.6877912.
[4] Q. Bai and J. Nossek, "On the effects of carrier frequency offset on cyclic prefix based OFDM and filter bank based multicarrier systems," in *IEEE Eleventh International Workshop on Signal Processing Advances in Wireless Communications*, Marrakech, Morocco, Jun. 2010, pp. 1–5. doi: 10.1109/SPAWC.2010.5670999.
[5] M. Sriyananda and N. Rajatheva, "Analysis of self interference in a basic FBMC system," in *IEEE 78th Vehicular Technology Conference*, Las Vegas, USA, Sept. 2013, pp. 1–5. doi: 10.1109/VTCFall.2013.6692102.
[6] J. Javaudin and Y. Jiang, "Channel estimation in MIMO OFDM/OQAM," in *IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, Recife, Brazil, Jul. 2008, pp. 266–270. doi: 10.1109/SPAWC.2008.4641611.
[7] J. Javaudin, D. Lacroix, and A. Rouxel, "Pilot-aided channel estimation for OFDM/OQAM," in *57th IEEE Semiannual Vehicular Technology Conference*, Jeju, South Korea, Apr. 2003, pp. 1581–1585. doi: 10.1109/VETECS.2003.1207088.
[8] C. Lele, R. Legouable, and P. Siohan, "Channel estimation with scattered pilots in OFDM/OQAM," in *IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, Recife, Brazil, Jul. 2008, pp. 286–290. doi: 10.1109/SPAWC.2008.4641615.
[9] Z. Zhao, N. Vucic, and M. Schellmann, "A simplified scattered pilot for FBMC/OQAM in highly frequency selective channels," in *11th international symposium on Wireless communications systems*, Barcelona, Spain, Oct. 2014, pp. 819–823. doi: 10.1109/ISWCS.2014.6933466.
[10] J. Bazzi, P. Weitkemper, and K. Kusume, "Power efficient scattered pilot channel estimation for FBMC/OQAM," in *10th International ITG Conference on Systems, Communications and Coding*, Hamburg, Germany, Feb. 2015, pp. 1–6.
[11] C. Lélé, J. Javaudin, R. Legouable, A. Skrzypczak, and P. Siohan, "Channel estimation methods for preamble-based OFDM/OQAM modulations," *Transactions on Emerging Telecommunications Technologies*, pp. 741–750, Sept. 2008. doi: 10.1002/ett.1332.
[12] C. Lélé, P. Siohan, and R. Legouable, "2 dB better than CP-OFDM with OFDM/OQAM for preamble-based channel estimation," in *IEEE International Conference on Communications*, Beijing, China, 2008, pp. 1302–1306. doi: 10.1109/ICC.2008.253.
[13] J. Du and S. Signell, "Novel preamble-based channel estimation for OFDM/OQAM systems," in *IEEE International Conference on Communications*, Dresden, Germany, 2009, pp. 1–6. doi: 10.1109/ICC.2009.5199226.
[14] E. Kofidis and D. Katselis, "Preamble-based channel estimation in MIMO-OFDM/OQAM systems," in *IEEE International Conference on Signal and Image Processing Applications*, Kuala Lumpur, Malaysia, 2011, pp. 579–584. doi: 10.1109/ICSIPA.2011.6144161.
[15] J. Siew, R. Piechocki, A. Nix, and S. Armour. (2002). "A channel estimation method for MIMO-OFDM systems," *London Communicaitons Symposium (LCS)* [Online]. Available: http://www.ee.ucl.ac.uk/lcs/previous/LCS2002/LCS087.pdf
[16] J. Du, P. Xiao, J. Wu, and Q. Chen, "Design of isotropic orthogonal transform algorithm-based multicarrier systems with blind channel estimation," *IET communications*, vol. 6, no. 16, pp. 2695–2704, Nov. 2012. doi: 10.1049/iet-com.2012.0029.
[17] P. Siohan, C. Siclet, and N. Lacaille, "Analysis and design of OFDM/OQAM systems based on filterbank theory," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1170–1183, May 2002.
[18] E. Kofidis and D. Katselis, "Improved interference approximation method for preamble-based channel estimation in FBMC/OQAM," in *19th European signal processing conference (EUSIPCO-2011)*, Barcelona, Spain, 2011. pp. 1603–1607.
[19] J. Du and S. Signell, "Time frequency localization of pulse shaping filters in OFD/OQAM systems," in *6th International Conference on Information, Communications Signal Processing*, Singapore, 2007, pp. 1–5.
[20] S. Kay, *Fundamentals of Statistical Signal Processing*. Upper Saddle River, USA: Prentice Hall, 1998.

Manuscript received: 2016-04-04

Biographies

Sohail Taheri (s.taheri@surrey.ac.uk) received his BS degree in electronic engineering and MSc degree in digital electronics from Amirkabir University of Technology, Iran in 2010 and 2012 respectively. He is currently working towards his PhD degree from the Institute for Communication Systems (ICS), University of Surrey, United Kingdom. His current research interests include signal processing for wireless communications, waveform design for 5G air interface and physical layer for 5G networks.

Mir Ghoraishi (m.ghoraishi@surrey.ac.uk) is a senior research fellow in the Institute for Communication Systems (ICS), University of Surrey. He joined the Institute in 2012 and is currently leading 5GIC testbed and proof-of-concept projects. This work area includes several implementation and proof-of-concept projects, e.g. 5G air-interface proof-of-concept, distributed massive MIMO implementation, wireless in-band full-duplex, millimeter wave hybrid beamforming system, and millimeter wave wireless channel analysis and modelling. He was involved in EU FP7 DUPLO project as work package leader. He has previously worked in Tokyo Institute of Technology as assistant professor and senior researcher from 2004 to 2012, after getting his PhD from the same institute. In Tokyo Tech he was involved in several national and small scale projects in planning, performing, implementation, analysis and modelling different aspect of wireless systems in physical layer, propagation channel and signal processing. He has co-authored 100 publications including refereed journals, conference proceedings and three book chapters.

XIAO Pei (p.xiao@surrey.ac.uk) received the BEng, MSc and PhD degrees from Huazhong University of Science & Technology, Tampere University of Technology, Chalmers University of Technology, respectively. Prior to joining the University of Surrey in 2011, he worked as a research fellow at Queen's University Belfast and had held positions at Nokia Networks in Finland. He is a Reader at University of Surrey and also the technical manager of 5G Innovation Centre (5GIC), leading and coordinating research activities in all the work areas in 5GIC. Dr Xiao's research interests and expertise span a wide range of areas in communications theory and signal processing for wireless communications. He has published 160 papers in refereed journals and international conferences, and has been awarded research funding from various sources including Royal Society, Royal Academy of Engineering, EU FP7, Engineering and Physical Sciences Research Council as well as industry.

CAO Aijun (cao.aijun@zte.com.cn) is a principal architect in ZTE R&D Center, Sweden (ZTE Wistron Telecom AB). He has over 17 years of experience in wireless communications research and development from baseband processing to network architecture, including design and optimization of commercial UMTS/LTE base-station and handset products, HetNet and small cell enhancement, etc. He has also been involved in standardization works and contributed to several 3GPP technical reports. He is also active in academic and industrial workshops and conferences related to the future wireless networks being as panelists or (co-)authors of published papers in refereed journals and international conferences. In addition, he holds more than 50 granted or pending patents. His current focus is 5G technologies related to the new energy-efficient unified air-interface and network architecture, e.g., new waveform design, non-orthogonal multiple access schemes, random access challenges and innovative signaling architecture for 5G networks.

GAO Yonghong (gao.yonghong@zte.com.cn) received his BEng degree in electronic engineering from Tsinghua University, China in 1989, and PhD degree in electronic systems from Royal Institute of Technology, Sweden in 2001. In 1996, he was a visiting scientist at Royal Institute of Technology and Ericsson Sweden. In 1999, he joined Ericsson Sweden to develop 3G base stations, baseband algorithms, and baseband ASICs. He joined ZTE European Research Institute (ZTE Wistron Telecom AB, Sweden) in 2002 and has been the CTO of ZTE European Research Institute till now, leading and participating the development of 3G/4G commercial base stations, baseband/RRM algorithms, and baseband ASICs, 3GPP small cell enhancement, and from 3 years ago focusing on 5G pre-study, 5G standardization, and 5G research projects in Europe. He has filed 40+ patents as a main author or co-author. His research interests include mobile communication standards/systems, and solutions and algorithms for commercial wireless products.

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 002
http://www.cnki.net/kcms/detail/34.1294.TN.20161008.1038.002.html, published online October 8, 2016

Non-Orthogonal Multiple Access Schemes for 5G

YAN Chunlin, YUAN Zhifeng, LI Weimin, and YUAN Yifei
(ZTE Corporation, Shengzhen 518057, China)

Abstract

Multiple access scheme is one of the key techniques in wireless communication systems. Each generation of wireless communication is featured by a new multiple access scheme from 1G to 4G. In this article we review several non-orthogonal multiple access schemes for 5G. Their principles, advantages and disadvantages are discussed, and followed by a comprehensive comparison of these solutions from the perspective of user overload, receiver type, receiver complexity and so on. We also discuss the application challenges of non-orthogonal multiple access schemes in 5G.

Keywords

5G; non-orthogonal multiple access; mMTC

1 Introduction

Multiple access scheme is the key technique of wireless communications. In 3rd generation (3G) code division multiple access is applied. In 4G orthogonal frequency division multiplexing access (OFDMA) is employed. In the coming 5G, non-orthogonal multiple access schemes are hot topics because they can achieve high system capacity. Moreover, massive machine type communication (mMTC) is one of the key scenarios for 5G in which massive connection is required. In this paper, we mainly focus on the non-orthogonal multiple access schemes supporting mMTC which has the rapidest growing speed and the urgent deploy demand.

Several non-orthogonal multiple access schemes are proposed for 5G, which include multi-user shared multiple access (MUSA) [1]–[4], resource spread multiple access (RSMA) [5], sparse code multiple access (SCMA) [6]–[8], pattern division multiple access (PDMA) [9]–[11], interleaver-division multiple access (IDMA) [12], [13], and non-orthogonal multiple access (NOMA) by power domain [14]. In this paper, the principles, merits and demerits of these schemes are discussed to let readers have a full overview on that.

2 Features of 5G

5G has three main technical features, including enhanced mobile broadband (eMBB), mMTC and ultra reliable and low latency communication (URLLC). The eMBB is the evolution of MBB targeting for high data rate and can support high mobility The mMTC is characterized by massive connection with low cost terminals. High reliability and ultra-low latency are the goals of URLLC.

With the development of Internet of things, a large number of terminals will have access to the network. Therefore, mMTC needs to support one million of connections per square kilometer. The mMTC, which has the fastest growing speed and the most urgent deployment demand, will create new chances in 5G. The non-orthogonal multiple access should support at least mMTC where high user overload is the key requirement.

In LTE there are several interactive processes between base station and terminal before the data is transmitted from terminal to the base station. This makes sense for long time and continuous data transmission because signaling overhead is small by averaging over a long time. In mMTC each terminal only transmits small data and massive terminals would sporadically transmit their data to the base station. When the same access procedure like in LTE-A is applied, the signaling overhead will be comparably large and the access efficiency is very low, thus grant-free for mMTC is needed in which multiple terminals can send their data on the same resource block without multi-step negotiations with base station.

3 Non-Orthogonal Multiple Access Schemes for 5G

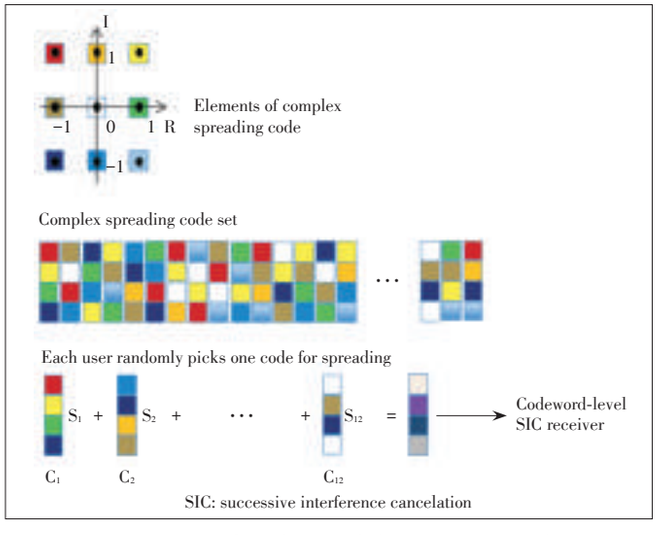
Several non-orthogonal multiple access schemes have been proposed for 5G. Based on their properties, they can be categorized to different types. Most non-orthogonal multiple access schemes use spreading codes. When such schemes have other

predominant properties, such as SCMA and PDMA use code matrix to illustrate how multiple users share the same resource block, and IDMA uses interleaver for user separation, we categorize them as other kind of schemes. In the following joint detection denotes message passing algorithm (MPA) based schemes.

3.1 Non-Orthogonal Multiple Access Schemes Based on Spreading Sequences

3.1.1 MUSA

MUSA is a non-orthogonal multiple access scheme operating in code domain and power domain. Spreading code with short length is applied in MUSA to support a large number of users that share the same resource block. When the number of users is large and the length of the spreading code is small, it is difficult to design large number of spreading code with low correlation when binary element of the spreading code is assumed. For binary spreading code the element of the spreading code belongs to the set {1, -1}. Only two values are employed in the spreading code. To overcome this drawback, non-binary and complex-value spreading code is proposed in MUSA. Either the real or the image element of the non-binary spreading code belongs to the set {1, 0, -1}, there are nine values for selection. This provides much more flexibility of spreading code design. Because the real and image elements of the spreading code are 1, 0 or -1, the multiplication operation can be implemented by addition operation which will reduce the implementation complexity. **Fig. 1** shows the basic features of MUSA, where multiple users could transmit data on the same resources by using randomly selected non-orthogonal complex spreading codes with short length (e.g. 4). In this example 12 users share 4 resource blocks, and the user overload is 300%. MUSA is always modeled by multiple spreading codes superposed on the same resource block. It can also be modeled by a code



▲**Figure 1. An example of MUSA with 300% user overload [4].**

matrix. The code matrix of MUSA with 300% overload is given by

$$G=\begin{bmatrix} 1+i & 1-i & -1+i & i & -i & -1-i & 1 & -1 & 1+i & 1 & 1-i & 0 \\ 1+i & 1+i & -1+i & -i & -i & -1+i & -1 & 1 & -i & -1+i & 1 & 0 \\ 1+i & i & -1 & 1 & 1+i & 1 & 1+i & 1 & 0 & 0 & 0 & 0 \\ 1 & -i & i & 1+i & 1-i & i & -1-i & -1 & 1 & 0 & 0 & 1+i \end{bmatrix}$$

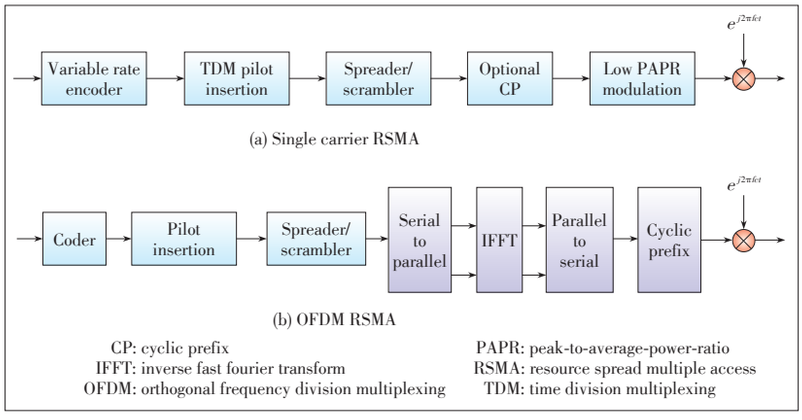
In 5G, mMTC is one important application scenario. In this scenario MUSA is preferred since grant-free transmission can be readily supported. A device terminal autonomously accesses the communication system without base station (BS) scheduling. Blind detection is applied at BS for MUSA in which active user, user spreading code and user channel would not be known before hand. Because the spreading code length is relative short and its elements have limited values, BS can generate numerous local spreading codes with low correlation. By using these local spreading codes and the received signal, we can closely approximate the optimal performance of MMSE estimator. Then the user signal with the highest signal-to-interference-plus-noise ratio (SINR) can be detected and decoded. After that user's signal is successfully decoded, it can be employed for channel estimation. After interference cancellation, the user signal with the second highest SINR is detected and decoded. During this process no pilots or preamble are needed for channel estimation, which facilitates MUSA application in mMTC because most other schemes rely on additional overhead for channel estimation. The blind detection for MUSA is verified over flat fading channel and multi-path fading channel [3], [15].

The main advantages of MUSA are reflected by high overloading factor, robust blind detection and true sense of grant-free transmission. Due to frequency-diversity gain achieved, 700% user overload can be achieved by MUSA over multi-path fading channel [15]. User detection can be carried out without the knowledge of the spreading code. User transmitted signal can be applied for enhanced channel estimation once it has been correctly decoded. Users can transmit their signals according to their demand. The possibility of collision due to the same spreading code applied is small since large number of the spreading codes can be accommodated.

Successive interference cancelation (SIC)-based receiver is applied for MUSA. It works well when there is SINR difference among the received signals. However, when the difference is small, there would be certain performance loss due to error propagation. While there is inherent SINR different in mMTC due to free power control, it is not a so serious problem for the signal detection of MUSA. The SINR difference is small, so it can be solved by using more advanced receiver, such as joint detection and decoding scheme.

3.1.2 RSMA

In RSMA (**Fig. 2**), a group of users' signals are superposed on the same resource blocks, and each user's signal is spread over the entire frequency/time resource blocks. Different users'



▲**Figure 2. RSMA block diagrams [5].**

signals within the resource blocks may be not orthogonal. Low code rate channel codes are employed to achieve large coding gain. Relative long spreading codes with good correlation property are applied to reduce the multi-user interference. Scramblers can be employed with the same purpose as the spreading codes. Interleaver is optional for RSMA according to the system requirements.

Depending on the application scenarios, it includes single carrier RSMA and multi-carrier RSMA. For the former it is optimized for battery power consumption and coverage extension for small data transactions by utilizing single carrier waveforms, very low peak-to-average-power-ratio (PAPR) modulations. It allows grant-less transmission and potentially allow asynchronous access. While for the latter it is optimized for low latency access for radio resource connection (RRC) connected users (i.e., timing with eNB already acquired) and allows for grant-less transmission.

The advantage of RSMA is that it supports asynchronous and grant-less transmission, so the signaling overhead is reduced. The disadvantage is that its user overload is limited when rake receiver is applied. By using advanced receiver, such as SIC based receiver, the overload can be enhanced.

3.2 Non-Orthogonal Multiple Schemes Based on Structured Coding Matrix

3.2.1 SCMA

Sparse codebook is applied at SCMA to reduce the detection complexity. At the same time joint detection is employed for SCMA to achieve excellent performance. The codewords are composed of multi-dimensional complex symbols, and the codewords in the same codebook have the same sparse pattern. Sparse codeword mapping utilizes low density spreading and could be referred to as sparse spreading. At the receiver, iterative multi-user detection based on MPA is used. **Fig. 3** shows an example of SCMA, where the coded bits of a data stream are directly mapped to a codeword with sparse non-zero elements from a codebook. With 6 sparse codewords transmitted over 4 orthogonal resources, the user overload is 150%. The coding matrix of Fig. 3 is given by

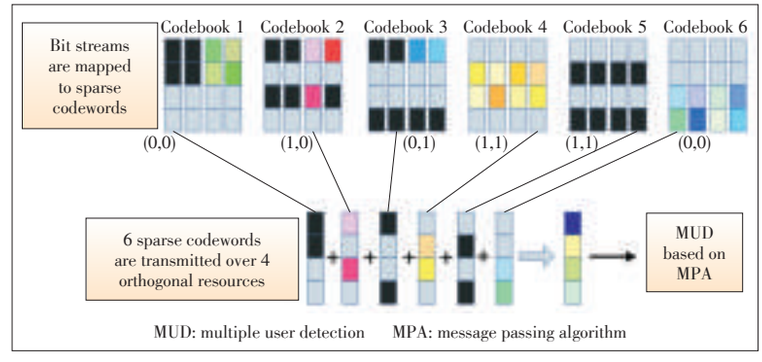
$$G=\begin{bmatrix}1&1&1&0&0&0\\1&0&0&1&1&0\\0&1&0&1&0&1\\0&0&1&0&1&1\end{bmatrix}$$

To reduce the multi-user interference and the detection complexity, sparse signature sequence is applied in SCMA for spreading. User signal is modulated by a codebook in which multidimensional modulation maps of the input coded bits to the points in the multiple complex dimensions [6]. By such operation shaping gain is achieved, which is claimed as one major property of SCMA.

The main disadvantage of SCMA is its high detection and decoding complexity even sparse signature sequence is applied. The detection and decoding complexity is even higher when large size constellation and a large number of users are employed. And additional pilots or preambles are needed for multi-user channel estimation, which may reduce system spectral efficiency. Because the size of the codebook is limited, if two users choose the same codeword, collision will happen. Collision is a serious problem for SCMA, which limits its overload capability. For example, with 6 users transmitted over 4 units, the user overload is only 150%. Although the overloading factor can be enhanced by using longer spreading codes, the detection complexity will increase significantly since the size of the codebook and the searching space is enlarged.

3.2.2 PDMA

For PDMA, the code in a code matrix is used to define mapping from data to a group of resources. Each element in the code corresponds to a resource in the resource group. PDMA can be detected by SIC type receiver. It also can be detected by MPA based scheme in the receiver. PDMA is designed for SIC-based receiver originally. The different diversity orders of different users by carefully design the code matrix facilitate the multi-user signal detection. The user with the largest diver-



▲**Figure 3. An example of SCMA with 150% user overload [8].**

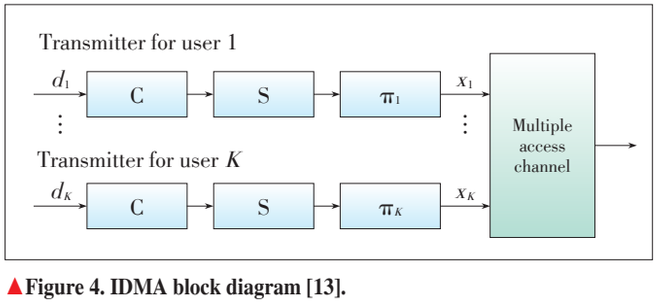
sity order is detected first, and then the user with the largest diversity order among the remaining users is detected; in this way, all users' signals will be detected.

To further improve the performance of PDMA, joint detection based scheme is proposed. In this case the unbalance weight of each column is interpreted as the irregular code weight. As we know irregular low density parity check (LDPC) code has better performance than that of the regular one. By carefully designing the code matrix with joint detection, even better performance can be obtained by PDMA compared with regular code matrix (for example non-orthogonal multiple access with low density signatures can be regards as regular code).

The main disadvantage of PDMA is its low user overload (user overload is defined by the number of user over the resource block that all users share). It is difficult to achieve overload of 400% with the 4-row code matrix (when the row of the code matrix is K, the largest user number it supported is 2^K-1 [10]).The complexity is high for high order modulation when joint-detection scheme is applied. Additional pilots or preamble are needed for channel estimation. Because the number of patterns is limited, there is high probability of collision when users are allowed to randomly select the patterns.

3.3 Non-Orthogonal Multiple Schemes Based on Interleaver

IDMA was proposed by [12], [13], in which users are separated by different interleavers. Low-rate channel decoding is applied and the coded bits are repeated multiple times to increase the SINR after accumulating the received signals. After channel coding and repetition, interleaver is employed to make the transmission bits randomly distributed. A block diagram of IDMA is shown in **Fig. 4** where C represents channel encoding, S denotes repetition and π is the interleaver. The strategy of user separation for IDMA is different from other non-orthogonal multiple access schemes. Interleaver is used for user separation and the length of the interleaver is very large (the length of the interleaver equals to the number of the bits after channel coding and repetition), thus this provides good base for a large number of users access by using IDMA. It is reported that 64 users can be supported by IDMA which share the same resource block [12]. This goal can never be achieved by other non-orthogonal multiple access schemes at present.



▲**Figure 4. IDMA block diagram [13].**

At the receiver side each user's signal is detected, demodulated and de-interleaved according to its own interleaver patterns. The soft information of decoded bits is input to elementary signal estimator (ESE) for soft information updating. After soft information updating new soft information is input to the decoder for channel decoding again. Several iterative detections between ESE and channel decoder are needed to achieve the best performance. The detection and decoding complexity does not increase exponentially with the user number and total spectral efficiency. The complexity increases linearly, which is also different from other non-orthogonal multiple access schemes which use joint detection and decoding scheme.

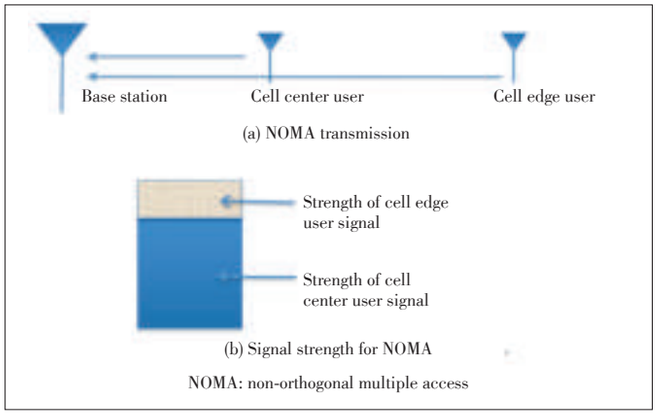
The main advantages of IDMA are its high user overload and excellent performance. And high spectral efficiency can be achieved by IDMA (as high as 8 b/s/Hz). The performance gap between IDMA simulation result and the system capacity bound is almost the same from the spectral efficiency 1 b/s/Hz to 8 b/s/Hz (this means the detection and decoding scheme is very robustness) [12]. These two merits are seldom achieved by other non-orthogonal multiple schemes simultaneously.

The main disadvantage of IDMA may be its large decoding complexity and decoding latency, especially when a large number of users are supported. The reason is that when large number of iterative detection and decoding are needed with the increasing of user number. For example, tens of channel decoder procedures are needed in the signal detection and tens of interactive actions between channel decoder and ESE detector are required. Thus high convergence algorithm is needed in the signal detection for IDMA in future. To solve the problem of large decoding complexity and decoding latency, interleaver patterns can be pre-allocate to small number of users, i.e., the relatively small pool size, so that the complexity of blind decoding and channel decoding latency can be maintained below certain level. Another disadvantage is that additional pilots or long preamble is needed to estimate the users' channels.

3.4 Non-Orthogonal Multiple Access (NOMA) Scheme Based on Power-Domain Division

Multi-user signals can be superposed together in NOMA. In NOMA, capacity or throughput improvement can be expected by sharing the same radio resources among multiple user equipments (UEs) as shown in **Fig. 5a** and **Fig. 5b**. A typical application scenario of NOMA is that a cell-center user and a cell-edge user are serviced by NOMA. Due to small path loss of cell center user, in the signal detection it is detected first and the signal of cell edge user is treated as interference. In the signal detection of cell edge user, the signal of cell center user is detected and decoded first. Then the signal of the cell center user is cancelled from the received signal and signal of cell edge user is detected and decoded.

The main advantage of NOMA is that excellent performance can be achieved when a cell center user and cell edge user are scheduled with moderate computational complexity (SIC detec-



▲Figure 5. NOMA block diagram.

tor is always applied). And a user overload of 200% is easily achieved. The main disadvantage of NOMA is that there is restriction on the scheduled users. Usually a cell center user and a cell edge user should be scheduled on the same resource block. When two cell center users or two cell edge users are scheduled and SIC-type receiver is applied, there is performance loss because one user always has low SINR due to interference from another user's signal. The NOMA is designed for eMBB originally. Thus when it is applied for mMTC, the received SINR would not be high and the number of supported users is very limited (two or three users are supported on the same resource block, which is much smaller than other non-orthogonal multiple access schemes). And additional pilots or long preamble is needed to estimate the users' channels.

A summary of these non-orthogonal multiple schemes are shown on **Table 1**. They are compared in terms of multiplexing domain, user overload, receiver type, receiver complexity and so on. Among these schemes MUSA achieves a good balance between performance and complexity, such as high user overload, low implementation complexity and flexible in grant-free transmission.

4 Application Challenges of Non-Orthogonal Multiple Access Schemes in 5G

Followings are the requirements for the non-orthogonal multiple access schemes. These factors should be considered when we design the non-orthogonal multiple access schemes.

4.1 Coverage

Coverage is an important issue for mMTC since terminals may distribute over a large area, thus it is crucial for non-orthogonal multiple access schemes to support terminals with different received power due to path loss. And the non-orthogonal multiple access schemes should have the ability of robustness to the high interference. To increase the coverage, low code rate channel coding or large spreading factor could be considered. High efficiency power amplifier is appealing for coverage extension, which requires transmit signals with low PAPR.

▼Table 1. Summary of different non-orthogonal multiple access schemes

	MUSA	RSMA	SCMA	PDMA	IDMA	NOMA
Multiplexing domain	Spreading	Spreading/ scramble	Codebooks	Pattern	Interleaver	Power
User overload	High	Low	Middle	Middle	High	Low
Receiver type	SIC	Raker or SIC	Joint detection	SIC or joint detection	Iterative detection and decoding	SIC
Receiver complexity	Low	Low	High	Low for SIC High for joint detection	High*	Low
Grant-free transmission	Users can randomly pick up spreading sequence	Power control needed	Codeword for each user is predefined and known at BS. Codeword collision is a problem due to limited number of codewords	Pattern is predefined and known at BS. User collision is a problem due to limited number of patterns	Interleaver patterns are known at BS	Grant-based

* Unlike joint detection scheme whose complexity increases exponentially as the number of the users and spectral efficiency increases, the complexity of IDMA only linear increases with the number of users and the spectral efficiency. The high complexity is due to large number of iterative detection and decoding.

MUSA: multi-user shared multiple access
RSMA: resource spread multiple access
SCMA: sparse code multiple access
PDMA: pattern division multiple access
IDMA: interleaver-division multiple access
NOMA: non-orthogonal multiple access
SIC: successive interference cancelation
BS: base station

4.2 PAPR

When the non-orthogonal multiple access scheme is applied for uplink, PAPR should be considered to increase the transmission efficiency and reduce the transmission power thus save the battery life. The battery life is desired to be 10 years for mMTC, so it puts a big challenge on the non-orthogonal multiple access scheme. The signal of the non-orthogonal multiple access schemes which have low PAPR will be preferred in practical implementation. Filtered $\pi/2$-binary phase shift keying (BPSK) and Gaussian filtered minimum shift keying (GMSK) have good property of low PAPR and are employed for PAPR reduction in RSMA [16].

4.3 Implementation Complexity

The implementation complexity includes two parts: transmitter implementation complexity and receiver implementation complexity. Because multi-user detection is carried out at receiver side, which has the highest complexity over the entire signal processing chain, the main implementation complexity is at the receiver side. Two types of receivers are always applied for non-orthogonal multiple access schemes: SIC-based receiver and joint-detection-based receiver. The former can achieve a good balance between performance and complexity. As the number of user increases, the complexity only increases linearly. While it suffers performance loss in some cases, such as the path-losses among different users are the same. Joint-detection-based receiver achieves excellent performance at the

cost of high computational complexity. Although by some designs, such as sparse coding matrix, the decoding complexity is reduced significantly, however, as the constellation size and the number of users increase, the decoding complexity grows exponentially. This bottleneck should be solved before such scheme is employed in practical systems.

4.4 Combination with Multiple-Input Multiple-Output (MIMO)

By applying MIMO technique large system capacity or high transmission/receiver reliability can be achieved. It had been proved that MIMO is a very effective technique in wireless communication systems. The non-orthogonal multiple access schemes should be amiable for MIMO. As the first step, SISO is assumed in the research of the new non-orthogonal multiple access schemes. However, compatibility with MIMO should be considered in the next research step.

4.5 Flexibility

The non-orthogonal multiple access schemes should have flexibility. It can change its parameters to support different use scenarios. For example, in some cases high user overload is the system design target, while in other cases coverage is the most important factor. This imposes requirements on the non-orthogonal multiple access scheme design. By changing the parameter of the non-orthogonal multiple access schemes, different targets can be achieved. Another example is that non-orthogonal multiple access schemes should support both multi-carrier system and single-carrier systems to facilitate its application scenarios.

5 Conclusion

This article reviews the main non-orthogonal multiple access schemes for 5G. Their principles and unique properties are discussed. MUSA can support high user overload with low implementation complexity and is more suitable for grant-free transmission. RSMA is suitable for single-carrier system and multi-carrier system. It has good property of large coverage. SCMA can achieve additional shaping gain and PDMA has the flexibility in the patterns design. IDMA can accommodate very high user overload and support high spectral efficiency at the cost of large decoding complexity and decoding latency. NOMA works well for large SINR difference among the non-orthogonal multiple users. At the same time they have their own disadvantages. It is important to integrate the advantages of different schemes to make the final designed scheme fulfill the challenging requirements of coming 5G.

References

[1] *Discussion on Multiple Access for New Radio Interface*, 3GPP R1-162226, Apr. 2016.
[2] Z. Yuan, G. Yu, W. Li, Y. Yuan, and X. Wang, "Multi-user shared access for internet of things," in *IEEE Vehicular Technology Conference*, Nanjing, China, May 2016, pp 1–5. doi: 10.1109/VTCSpring.2016.7504361.
[3] *Receiver Implementation for MUSA*, 3GPP R1-164270, May 2016.
[4] *Contention-Based Non-Orthogonal Multiple Access for UL mMTC*, 3GPP R1-164269, May 2016.
[5] *Resource Spread Multiple Access*, 3GPP R1-164688, May 2016.
[6] M. Taherzadeh, H. Nikopour, A. Bayesteh, H. Baligh, "SCMA codebook design", in *IEEE Vehicular Technology Conference*, Vancouver, Canada, Sept. 2014, pp.1–5, doi: 10.1109/VTCFall.2014.6966170.
[7] H. Nikopour and H. Baligh, "Sparse code multiple access," in *IEEE International Symposium On Personal, Indoor And Mobile Radio Communications*, London, UK, Sept. 2013, pp. 332–336. doi: 10.1109/PIMRC.2013.6666156.
[8] Future Mobile Communication Forum. (2016, Jul. 7). *5G white paper v2.0, part d—alternative multiple access v1* [Online]. Available: http://www.future-forum.org/dl/151106/whitepaper.rar
[9] *Candidate Solution for New Multiple Access*, 3GPP R1-163383, Apr. 2016.
[10] X. Dai, S. Chen, S. Sun, et al., "Successive interference cancelation amenable multiple access (SAMA) for future wireless communications," in *Proc. IEEE International Conference on Communication Systems*, Macau, China, Nov. 2014, pp. 222–226. doi: 10.1109/ICCS.2014.7024798.
[11] X. Dai, "Successive interference cancellation amenable space-time codes with good multiplexing-diversity tradeoff," *Wireless Personal Communications*, vol. 55, no. 4, pp. 645–654, Dec. 2010. doi: 10.1007/s11277-009-9826-9.
[12] P. Li, L. Liu, K. Wu, and W. K. Leung, "On interleave-division multiple-access," in *IEEE International Conference on Communications*, Paris, France, Jun. 2004, pp. 2869–2873. doi: 10.1109/ICC.2004.1313053.
[13] P. Li, L. Liu, K. Wu, and W. K. Leung, "Interleave division multiple-access," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 938–947, Apr. 2006. doi: 10.1109/TWC.2006.1618943.
[14] Y. Saito, Y. Kishiyama, A. Benjebbour, et al., "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *IEEE Vehicular Technology Conference*, Dresden, Germany, Jun. 2013, pp. 1–5. doi: 10.1109/VTC Spring.2013.6692652.
[15] *Receiver Details and Link Performance for MUSA*, 3GPP R1-166404, Aug. 2016.
[16] *Resource Spread Multiple Access*, 3GPP R1-166359, Aug. 2016.

Manuscript received: 2016-07-07

Biographies

YAN Chunlin (yan.chunlin@zte.com.cn) received his PhD degree from University of Electronic Science and Technology of China (UESTC), China in 2004. He worked at DOCOMO Beijing communications lab from 2005 to 2016. Since 2016 he has been with ZTE Corporation. He has published tens of papers in IEEE ICC, Globecom, VTC, PIMRC and other international conferences. His main research interests include synchronization, binary and non-binary channel coding, MIMO detection and non-orthogonal multiple access technique for 5G.

YUAN Zhifeng (yuan.zhifeng@zte.com.cn) received his MS degree in signal and information processing from Nanjing University of Post and Telecommunications (NUPT), China in 2005. He has been worked with the Wireless Technology Advance Research Department of ZTE Corporation since 2006 and the leader of the team for new multi-access (NMA) for 5G wireless systems since 2012. His research interests include wireless communication, MIMO systems, information theory, multiple access, error control coding, adaptive algorithm, and high-speed VLSI design.

LI Weimin (li.weimin6@zte.com.cn) received his master degree from NUPT, China. He joined in ZTE Corporation in 2010, and is responsible for technology research of power control and interference control in wireless communications. His current research focuses on multiple access technology for 5G system.

YUAN Yifei (yuan.yifei@zte.com.cn) received his master degree from Tsinghua University, China, and PhD from Carnegie Mellon University, USA. He was with Alcatel-Lucent from 2000 to 2008, working on 3G/4G key technologies. Since 2008, he has been with ZTE as the technical director of standards research on LTE-advanced physical layer and 5G new radio. His research interests include MIMO, channel coding, resource scheduling, multiple access, and NB-IoT. He was admitted to Thousand Talent Plan Program of China in 2010. He has extensive publications, including two books on LTE-Advanced.

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 003
http://www.cnki.net/kcms/detail/34.1294.TN.20161019.0829.002.html, published online October 19, 2016

A Survey of Downlink Non-Orthogonal Multiple Access for 5G Wireless Communication Networks

WEI Zhiqiang[1], YUAN Jinhong[1], Derrick Wing Kwan Ng[1], Maged Elkashlan[2], and DING Zhiguo[3]
(1. The University of New South Wales, Sydney, NSW 2052, Australia;
2. Queen Mary University of London, London E1 4NS, UK;
3. Lancaster University, Lancaster LA1 4YW, UK)

Abstract

Non-orthogonal multiple access (NOMA) has been recognized as a promising multiple access technique for the next generation cellular communication networks. In this paper, we first discuss a simple NOMA model with two users served by a single-carrier simultaneously to illustrate its basic principles. Then, a more general model with multicarrier serving an arbitrary number of users on each subcarrier is also discussed. An overview of existing works on performance analysis, resource allocation, and multiple-input multiple-output NOMA are summarized and discussed. Furthermore, we discuss the key features of NOMA and its potential research challenges.

Keywords

non-orthogonal multiple access (NOMA); successive interference cancellation (SIC); resource allocation; multiple-input multiple-output (MIMO)

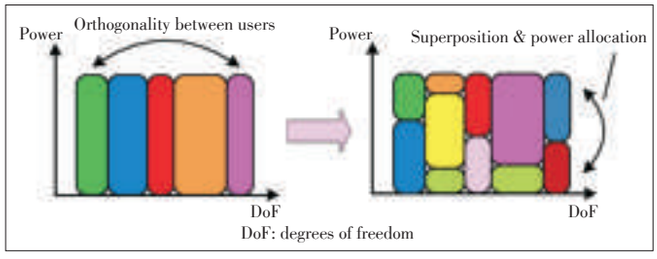
1 Introduction and Background

The fifth generation (5G) communication system is on its way. It is widely believed that 5G is not just an incremental version of the fourth generation (4G) communication systems [1] due to the increasing demand of data traffic and the expected new services and functionalities, such as internet-of-things (IoT) and cloud-based architectural applications [2]. These envisioned services pose challenging requirements for 5G wireless communication systems, such as much higher data rates (100–1000 times faster than current 4G technology), lower latency (1 ms for a roundtrip latency), massive connectivity and support of diverse quality of service (QoS) (10^6 devices/km^2 with diverse QoS requirements) [1]. From a technical perspective, to meet the aforementioned challenges, some potential technologies, such as massive multiple-input multiple-output (MIMO) [3], [4], millimeter wave [5], [6], and ultra densification and offloading [7]–[9], have been discussed extensively. Besides, it is expected to employ a future radio access technology for 5G, which is flexible, reliable [10], and efficient in terms of energy and spectrum [11], [12]. Radio access technologies for cellular communications are characterized by multiple access schemes, such as frequency-division multiple access (FDMA) for the first generation (1G), time-division multiple access (TDMA) for the second generation (2G), code-division multiple access (CDMA) used by both 2G and the third generation (3G), and orthogonal frequency division multiple access (OFDMA) for 4G. All these conventional multiple access schemes are categorized as orthogonal multiple access (OMA) technologies, where different users are allocated to orthogonal resources in either time, frequency, or code domain in order to mitigate multiple access interference (MAI). However, OMA schemes are not sufficient to support the massive connectivity with diverse QoS requirements. In fact, due to the limited degrees of freedom (DoF), some users with better channel quality have a higher priority to be served while other users with poor channel quality have to wait to access, which leads to high unfairness and large latency. Besides, it is inefficient when allocating DoF to users with poor channel quality. In this survey, we focus on one promising technology, non-orthogonal multiple access (NOMA), which in our opinion will contribute to disruptive design changes on radio access and address the aforementioned challenges of 5G.

In contrast to conventional OMA, NOMA transmission techniques intend to share DoF among users via superposition and consequently need to employ multiple user detection (MUD) to separate interfered users sharing the same DoF, as illustrated in **Fig. 1**. NOMA is beneficial to enlarge the number of connections by introducing controllable symbol collision in the same DoF. Therefore, NOMA can support high overloading transmis-

A Survey of Downlink Non-Orthogonal Multiple Access for 5G Wireless Communication Networks

WEI Zhiqiang, YUAN Jinhong, Derrick Wing Kwan Ng, Maged Elkashlan, and DING Zhiguo



▲Figure 1. From OMA to NOMA via power domain multiplexing.

sion and further improve the system capacity given limited resource (spectrum or antennas). In addition, multiple users with different types of traffic request can be multiplexed to transmit concurrently on the same DoF to improve the latency and fairness. The comparison of OMA and NOMA is summarized in **Table 1**. As a result, NOMA has been recognized as a promising multiple access technique for the 5G wireless networks due to its high spectral efficiency, massive connectivity, low latency, and high user fairness [13]. For example, multiuser superposition transmission (MUST) has been proposed for the third generation partnership project long-term evolution advanced (3GPP-LTEA)

networks [14]. Three kinds of non-orthogonal transmission schemes have been proposed and studied in the MUST study item. Through the system-level performance evaluation, it has been shown that the MUST can increase system capacity as well as improve user experience.

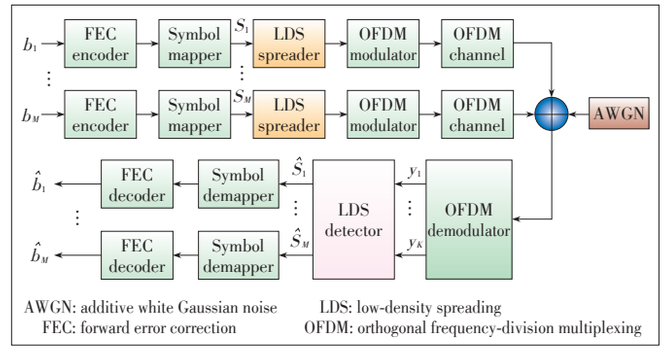
Recently, several NOMA schemes have been proposed and received significant attention. According to the domain of multiplexing, the authors in [13] divided the existing NOMA techniques into two categories, i.e., code domain multiplexing (CDM) and power domain multiplexing (PDM). The CDM-NOMA techniques, including low-density spreading (LDS) [15]–[17], sparse code multiple access (SCMA) [18], pattern division multiple access (PDMA) [19], etc, introduce redundancy via coding/spreading to facilitate the users separation at the receiver.

For instance, LDS-CDMA [15] intentionally arranges each user to spread its data over a small number of chips and then interleave uniquely, which makes optimal MUD affordable at receiver and exploits the intrinsic interference diversity. LDS-OFDM [16], [17], as shown in **Fig. 2**, can be interpreted as a system which applies LDS for multiple access and OFDM for multicarrier modulation. Besides, SCMA is a generalization of LDS methods where the modulator and LDS spreader are merged. On the other hand, PDM-NOMA exploits the power domain to serve multiple users in the same DoF, and performs successive interference cancellation (SIC) at users with better channel conditions. In fact, the non-orthogonal feature can be

▼Table 1. Comparison of OMA and NOMA

	Advantages	Disadvantages
OMA	Simpler receiver detection	• Lower spectral efficiency • Limited number of users • Unfairness for users
NOMA	• Higher spectral efficiency • Higher connection density • Enhanced user fairness • Lower latency • Supporting diverse QoS	• Increased complexity of receivers • Higher sensitivity to channel uncertainty

NOMA: non-orthogonal multiple access
OMA: orthogonal multiple access
QoS: quality of service



▲Figure 2. Block diagram of an uplink LDS-OFDM system.

introduced either in the power domain only or in the hybrid code and power domain. Although DM-NOMA has the potential code gain to improve spectral efficiency, PDM-NOMA has a simpler implement since there is almost no big change in the physical layer procedures at the transmitter side compared to current 4G technologies. In addition, PDM-NOMA paves the way for flexible resource allocation via relaxing the orthogonality requirement to improve the performance of NOMA, such as spectral efficiency [20], [21], energy efficiency [22], and user fairness [23]. Therefore, this paper will focus on the PDM-NOMA, including its basic concepts, key features, existing works, and future research challenges.

2 Fundamentals of NOMA

This section presents the basic model and concepts of single-antenna downlink NOMA. The first subsection presents a simple downlink single-carrier NOMA (SC-NOMA) system serving two users simultaneously, while the second subsection presents a more general multi-carrier NOMA (MC-NOMA) model for serving an arbitrary number of users in each subcarrier.

2.1 Two-User SC-NOMA[1]

Benjebbour, Saito et al. [24], [25] proposed the system model of downlink NOMA with superposition transmission at the base station (BS) and successive interference cancellation

[1] In this paper, a two-user NOMA system means that two users are multiplexed on each subcarrier simultaneously. Similarly, a multiuser NOMA system means that an arbitrary number of users are multiplexed on each subcarrier simultaneously.

(SIC) at the user terminals, which is illustrated in **Fig. 3** in case of one BS and two users. The BS transmits the messages of both user 1 and user 2, i.e., s_1 and s_2, with different transmit powers p_1 and p_2, on the same subcarrier, respectively. The corresponding transmitted signal is represented by

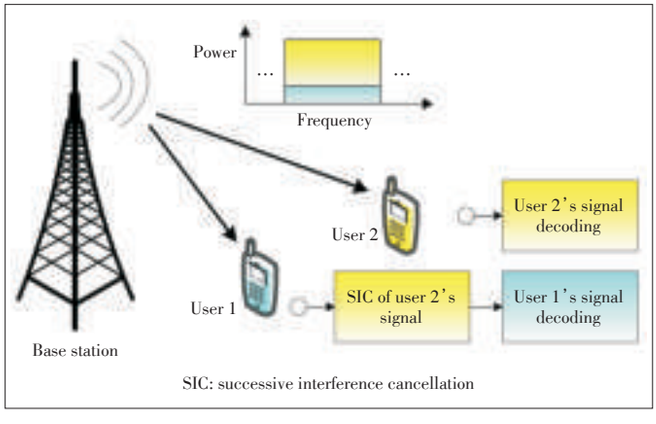
$$x=\sqrt{p_1}s_1+\sqrt{p_2}s_2, \quad (1)$$

where transmit power is constrained by $p_1+p_2=1$. The received signal at user i is given by

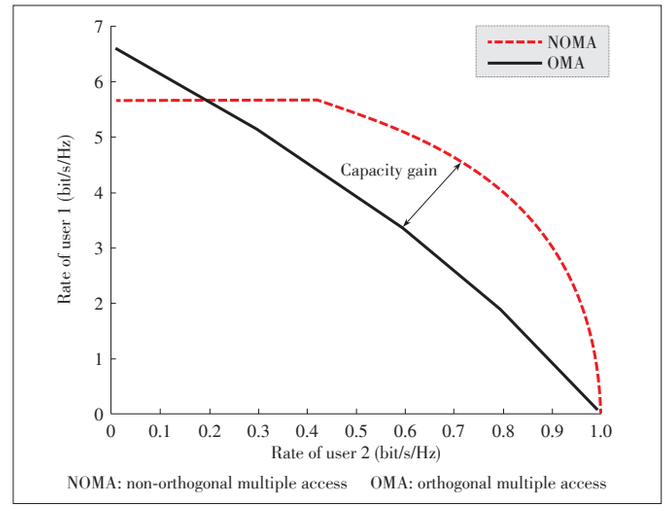
$$y_i=h_i x+v_i, \quad (2)$$

where h_i denotes the complex channel coefficient including the joint effect of large scale fading and small scale fading. Variable v_i denotes the additive white Gaussian noise (AWGN), and $v_i\sim CN(0,\sigma_i^2)$, where $CN(0,\sigma_i^2)$ denotes the circularly symmetric complex Gaussian distribution with mean zero and variance σ_i^2. We assume that user 1 is the cell-center user with a better channel quality (strong user), while user 2 is the cell-edge user with a worse channel quality (weak user), i.e., $(|h_1|^2/\sigma_1^2)\geqslant(|h_2|^2/\sigma_2^2)$. According to the NOMA protocol [26], the BS will allocate more power to the weak user to provide fairness and facilitate the SIC process, i.e., $p_1\leqslant p_2$.

In downlink SC-NOMA, the SIC process is implemented at the receiver side. The optimal SIC decoding order is in the descending order of channel gains normalized by noise. It means that user 1 will decode s_2 first and remove the inter-user interference of user 2 by subtracting s_2 from the received signal y_i before decoding its own message s_1. On the other hand, user 2 does not perform interference cancellation and directly decodes its own message s_2 with interference from user 1. Fortunately, the power allocated to user 2 is larger than that of user 1 in the aggregate received signal y_2, which will not introduce much performance degradation compared to allocating user 2 on this subcarrier exclusively. The rate region of SC-NOMA is illustrated in **Fig. 4** in comparison with that of OMA, where it has been proved that NOMA schemes are very likely to outperform OMA schemes in [27]. It is noted that the rate region of NOMA only covers a part of the capacity region of broadcast channel with SIC receiver [28] due to the power constraint $p_1\leqslant p_2$.



▲**Figure 3. A downlink NOMA model with one base station and two users.**



▲**Figure 4. The rate region of two-user SC-NOMA in comparison with that of OMA. User 1 is a strong user with $(|h_1|^2/\sigma_1^2)=100$, while user 2 is a weak user with $(|h_2|^2/\sigma_2^2)=1$.**

2.2 Multiuser MC-NOMA

For a downlink MC-NOMA system with one BS serving an arbitrary number of users, such as N, the available bandwidth is divided into a set of K subcarriers, where $N>K$, i.e., an overloading scenario that OFDMA cannot afford. The channel between user n and the BS on subcarrier k is denoted by $h_{k,n}$, and is assumed to be perfectly known at both the transmitter and receiver side. The BS schedules all users across all subcarriers by ξ_k and ζ_n, where ξ_k denotes a user set allocated on subcarrier k and ζ_n denotes a subcarrier set occupied by user n. Without loss of generality, the channel gains of all users allocated on subcarrier k are sorted as $|h_{k,b(1)}|^2\geqslant|h_{k,b(2)}|^2\geqslant\cdots\geqslant|h_{k,b(|\xi_k|)}|^2$, where $|\xi_k|$ denotes the cardinality of the user set ξ_k and $b(\cdot)$ indicates the mapping between the sorted channel gain order and the original one. For instance, for subcarrier k occupied by three users $\xi_k=\{1,2,3\}$ and $|h_{k,2}|^2\geqslant|h_{k,3}|^2\geqslant|h_{k,1}|^2$, we will have $b(1)=2$, $b(2)=3$, and $b(3)=1$, respectively. It is noted that the mapping functions are various on different subcarriers due to users' different frequency selective fading patterns.

According to NOMA protocol [26], all users in ξ_k share subcarrier k by different transmission power $p_{k,b(l)}$ based on the given channel gain, where $l=1,2,\cdots,|\xi_k|$ and $p_{k,b(1)}\leqslant p_{k,b(2)}\leqslant\cdots\leqslant p_{k,b(|\xi_k|)}$. The sharing strategy saves the subcarriers those might be wasted by only transmitting the messages of the weak users and accommodates more users with di-

verse QoS requirements, which is favorable to massive connectivity and IoT in 5G networks.

All messages of users in ξ_k are superimposed on subcarrier k, where the transmitted signal is given by

$$x_k = \sum_{l=1}^{|\xi_k|} \sqrt{p_{k,b(l)}} S_{k,b(l)}, \quad (3)$$

where $S_{k,b(l)}$ and $p_{k,b(l)}$ denote the message and allocated power of user $n(l)$ on subcarrier k, respectively.

Assuming the independent and identically distributed (IID) AWGN over all subcarriers and all users for simplicity, the user scheduling, power allocation, and the SIC decoding order only depend on the channel gain order. At the receiver side, the received signal of user $b(l)$ on subcarrier k can be represented by

$$y_{k,b(l)} = h_{k,b(l)} x_k + v \quad (4)$$

$$= h_{k,b(l)} \sum_{l'=1}^{|\xi_k|} \sqrt{p_{k,b(l')}} S_{k,b(l')} + v, \quad \forall l \in \{1,2,\cdots,|\xi_k|\}, \quad (5)$$

where v denotes the AWGN, i.e., $v_i \sim CN(0,\sigma^2)$, and σ^2 denotes the noise power.

On subcarrier k, the scheduled users in ξ_k perform SIC to eliminate inter-user interference. Similar to the case of two-user NOMA, the optimal SIC decoding order is in the descending channel gain order, i.e., $\{b(1),b(2),\cdots,b(|\xi_k|)\}$. It means that the user $b(l)$ first decodes and subtracts the message $s_{k,b(l')}$, $\forall l' > l$, in descending order from $|\xi_k|$ to $l+1$, and then decodes its own message $s_{k,n(l)}$ by treating $s_{k,n(l')}$, $\forall l' > l$, as interference.

3 Performance and Key Features of NOMA

In this section, we present the performance characteristics of NOMA in existing works, and then discuss the pros and cons of NOMA schemes.

3.1 Performance of NOMA

It has been shown that NOMA offers considerable performance gain over OMA in terms of spectral efficiency and outage probability [25]–[27], [29]–[31]. Initially, the performance of NOMA was evaluated through simulations given perfect CSI by utilizing the proportional fairness scheduler [25], [29], fractional transmission power allocation (FTPA) [25], and tree-search based transmission power allocation (TTPA) [30]. These works showed that the overall cell throughput, cell-edge user throughput, and the degrees of proportional fairness achieved by NOMA are all superior to those of OMA. In [27], the author analyzed a two-user SC-NOMA system under statistical CSI from an information theoretic perspective, where it was proved that NOMA outperforms native TDMA with high probability in terms of both the sum rate and individual rates. In [26], for a fixed power allocation, the performance of a multiuser SC-NOMA system in terms of outage probability and ergodic sum rates under statistical CSI was investigated in a cellular downlink scenario with randomly deployed users. With the proposed asymptotic analysis, it showed that user n experiences a diversity gain of n and NOMA is asymptotically equivalent to the opportunistic multiple access technique. Furthermore, the authors in [32] analyzed the performance degradation of a multiuser SC-NOMA system on outage probability and average sum rates due to partial CSI. It showed that NOMA based on second order statistical CSI always achieves a better performance than that of NOMA based on imperfect CSI, while it can achieve similar performance to the NOMA with perfect CSI in the low SNR region.

In summary, most of the existing works on performance analysis of NOMA focused on a SC-NOMA system since the user scheduling in MC-NOMA complicates the analysis due to its combinatorial nature. A remarkable work in [31] characterized the impact of user pairing on the performance of a two-user SC-NOMA system with fixed power allocation and cognitive radio inspired power allocation, respectively. The authors proved that, for fixed power allocation, the performance gain of NOMA over OMA increases when the difference in channel gains between the paired users becomes larger. However, further exploration on performance analysis of MC-NOMA system should be carried out in the future since user scheduling is critical for performance of NOMA.

3.2 Pros

1) Higher spectral efficiency

By exploiting the power domain for user multiplexing, NOMA systems are able to accommodate more users to cope with system overload. In contrast to allocate a subcarrier exclusively to a single user in OMA scheme, NOMA can utilize the spectrum more efficiently by admitting strong users into the subcarriers occupied by weak users without compromising much their performance via utilizing appropriate power allocation and SIC techniques.

2) Better utilization of heterogeneity of channel conditions

As we mentioned before, NOMA schemes intentionally multiplex strong users with weak users to exploit the heterogeneity of channel condition. Therefore, the performance gain of NOMA over OMA is larger when channel gains of the multiplexed users become more distinctive [31].

3) Enhanced user fairness

By relaxing the orthogonal constraint of OMA, NOMA enables a more flexible management of radio resources and offers an efficient way to enhance user fairness via appropriate resource allocation [23].

4) Applicability to diverse QoS requirements

NOMA is able to accommodate more users with different types of QoS requests on the same subcarrier. Therefore, NOMA is a good candidate to support IoT which connects a great

number of devices and sensors requiring distinctive targeted rates.

3.3 Cons

1) The BS needs to know the perfect channel state information (CSI) to arrange the SIC decoding order, which increases the CSI feedback overhead.

2) The SIC process introduces a higher computational complexity and delay at the receiver side, especially for multicarrier and multiuser systems.

3) The strong users have to know the power allocation of the weaker users in order to perform SIC, which also increases the system signalling overhead.

4) Allocating more power to the weak users, who are generally in the cell-edge, will introduce more inter-cell interferences into the whole system.

4 Design of NOMA Schemes

Due to the remarkable performance gain of NOMA over conventional OMA, a lot of works on design of NOMA schemes have been proposed in literatures. In this section, we present the existing works on resource allocation of NOMA and MIMO-NOMA, and then briefly introduce other works associated with NOMA.

4.1 Resource Allocation

Resource allocation has received significant attention since it is critical to improve the performance of NOMA. However, optimal resource allocation is very challenging for MC-NOMA systems, since user scheduling and power allocation couple with each other severely. Some initial works on resource allocation in [25], [29], [30] have been reported, but they are far from optimal. In [33], [34], the authors studied a two-user MC-NOMA system by minimizing the number of subcarriers assigned under the constraints of maximum allowed transmit power and requested data rates, and further introduced a hybrid orthogonal-nonorthogonal scheme. Furthermore, the authors in [21] studied a joint power and subcarrier allocation problem for a two-user MC-NOMA system. They proposed an optimal scheme and a suboptimal scheme with close-to-optimal performance based on monotonic optimization and difference of convex function programming, respectively.

Besides, there are also several works on resource allocation for multiuser MC-NOMA systems. In [35], the authors formulated the resource allocation problem to maximize the sum rate, which is a non-convex optimization problem due to the binary constraint and the existence of the interference term in the objective function. Interestingly, they proposed a suboptimal solution by employing matching theory and water-filling power allocation. In [36], the authors presented a systematic approach for NOMA resource allocation from a mathematical optimization point of view. They formulated the joint power and channel allocation problem of a downlink multiuser MC-NOMA system, and proved its NP-hardness based on [37] via defining a special user. Furthermore, they proposed a competitive suboptimal algorithm based on Lagrangian duality and dynamic programming, which significantly outperforms OFDMA as well as NOMA with FTPA.

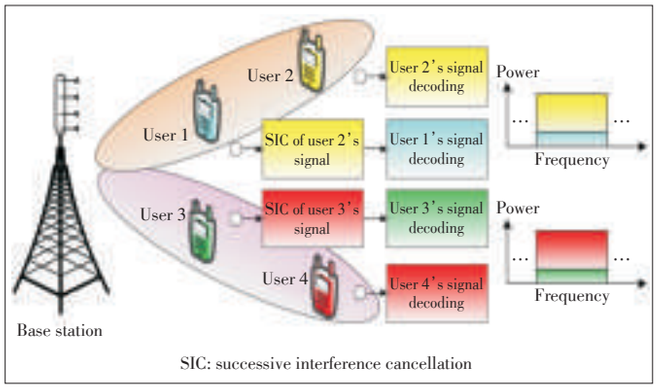
Most of works aforementioned focus on the optimal resource allocation for maximizing the sum rate. However, fairness is another objective to optimize for resource allocation of NOMA. Proportional fairness (PF) has been adopted as a metric to balance the transmission efficiency and user fairness in many works [38], [39]. In [40], the authors proposed a user pairing and power allocation scheme for downlink two-user MC-NOMA based on the PF objective. A prerequisite for user pairing was given and a closed-form optimal solution for power allocation was derived. Apart from PF, max-min or min-max methods are usually adopted to achieve user fairness.

Given a preset user group, the authors in [23] studied the power allocation problem from a fairness standpoint by maximizing the minimum achievable user rate with instantaneous CSI and minimizing the maximum outage probability with average CSI. Although the resulting problems are non-convex, simple low-complexity algorithms were developed to provide close-to-optimal solutions. Similarly, another paper [41] studied the outage balancing problem of a downlink multiuser MC-NOMA system to maximize the minimum weighted success probability with and without user grouping. Joint power allocation and decoding order selection solutions were given, and the inter-group power and resource allocation solutions were also provided in the paper.

In summary, many existing works focus on the resource allocation for NOMA systems under perfect CSI at the transmitter side. However, there are only few works on the joint user scheduling and power allocation problem for MC-NOMA systems under imperfect CSI, not to mention the SIC decoding order selection problem. In fact, under imperfect CSI, the SIC decoding order cannot be determined by channel gain order, and some other metrics, such as distance, priority, and target rates, are potential criteria to decide the SIC decoding order.

4.2 MIMO-NOMA

The application of MIMO techniques to NOMA systems is important for enhancing the performance gains of NOMA. Therefore, MIMO-NOMA is another hot topic that has been researched, where the BS and users are equipped with multiple antennas, and multiple users in the same beam are multiplexed on power domain. **Fig. 5** illustrates a simple MIMO-NOMA system with one base station and four users. Initially, the concept of MIMO-NOMA was proposed in [30], [42], [43], which demonstrated that MIMO-NOMA outperforms conventional MIMO OMA. The authors in [44] proposed a two-user MIMO-NOMA scheme with a clustering and power allocation algorithm, where the correlation and channel gain difference



▲**Figure 5. A downlink MIMO-NOMA model with one base station and four users.**

were taken into consideration to reduce intra-beam interference and inter-beam interference simultaneously. In [45], the authors proposed a minimum power multicast beamforming scheme and applied to two-user NOMA systems for multi-resolution broadcasting. The proposed two-stage beamforming method outperforms the zero-forcing beamforming scheme in [44].

The design of precoding and detection algorithms also received considerable attention since they are the key to eliminate or reduce inter-cluster interference. The authors in [46] studied the ergodic sum capacity maximization problem of a two-user MIMO-NOMA system under statistical CSI with the total power constraint and minimum rate constraint for the weak user. This paper derived the optimal input covariance matrix, and proposed the optimal power allocation scheme as well as a low complexity suboptimal solution. Furthermore, in [47], the authors studied the sum rate optimization problem of two-user MIMO-NOMA under perfect CSI with the same constraints, while different precoders were assigned to different users. The optimal precode covariance matrix was derived by utilizing the duality between uplink and downlink, and a low complexity suboptimal solution based on singular value decomposition (SVD) was also provided. In [48], the authors proposed a new design of precoding and detection matrices for a downlink multiuser MIMO-NOMA system, then analyzed the impact of user pairing as well as power allocation on the sum rate and outage probability of MIMO-NOMA system. Furthermore, in [49], a transmission framework based on signal alignment was proposed for downlink and uplink two-user MIMO-NOMA systems. The authors in [20] studied the sum rate maximization problem of a downlink multiuser multiple-input single-output (MISO) NOMA system. The MISO NOMA transmission outperforms conventional OMA schemes, particularly when the transmit SNR is low, and the number of users is greater than the number of BS antennas. Recently, a multiuser MIMO-NOMA scheme based on limited feedback was proposed and analyzed in [50].

In summary, most of the existing works on MIMO-NOMA focused on design of precoding and detection algorithms, and their performance analyses. However, user scheduling and power allocation were rarely discussed in the spatial domain, which play important roles in improving the spatial efficiency of MIMO-NOMA.

4.3 Other Works on NOMA

In addition to the above two aspects, there are many other works associated NOMA. We will not discuss further in detail due to the limited space. Compared to downlink NOMA, uplink NOMA was also studied in several works [51]–[57]. Moreover, asynchronous NOMA has also been investigated in uplink scenarios [58], [59]. Cooperative NOMA, where strong users serve as relays for weak users, was studied in [60], [61]. In addition, several works on NOMA combined with other techniques were also reported, such as energy harvesting [62], [63], cognitive radio networks [64], visible light communication [65], and physical layer security [66].

5 Research Challenges

As discussed above, NOMA can be employed to improve the spectral efficiency, user fairness, as well as to support massive connections with diverse QoS requirements. Based on our overview of existing works on NOMA and its potential applications in practical systems, we present the research challenges of NOMA in the following three aspects.

5.1 Resource Allocation under Imperfect CSI

Most of existing works on resource allocation of NOMA are based on the assumption of perfect CSI at the transmitter side, which is difficult to obtain in practice due to either the estimation error or the feedback delay. Therefore, it is nature to investigate how CSI error affects the performance of NOMA and to consider robust resource allocation under imperfect CSI. Since NOMA is expected to offer lower latency in order to support delay-sensitive applications in 5G, one promising solution is the outage-based robust approach for designing the resource allocation of NOMA. In this direction, the SIC decoding order under imperfect CSI is still an open problem. Furthermore, it is important to study the joint optimization of power allocation, user scheduling and SIC decoding order selection of NOMA under imperfect CSI.

5.2 Cooperative NOMA

A key feature of NOMA is that the strong users have prior information of the weak users, which has not been fully exploited in existing works. In cooperative NOMA, the strong users can serve as relays for the weak users, which has the potential to utilize the spatial DoF even for users with a single antenna. Some preliminary works showed that cooperative NOMA can achieve the maximum diversity gain for all the users [60], [61]. It is important to study the optimal resource allocation for coop-

erative NOMA. Besides, distributed beamforming can be employed in cooperative NOMA to harvest the spatial DoF without much signalling overhead. Considering that cooperative NOMA will introduce more complexity and extra delay into systems, it is important to investigate the tradeoffs among the system performance, complexity, and delay.

5.3 QoS-Based NOMA

As we mentioned before, NOMA has great potential to support diverse QoS requirements. The heterogeneity of QoS requirements might in turn facilitate the power allocation and user scheduling of NOMA, which is also an interesting topic to explore in the future. For example, users in NOMA systems can be categorized according to their QoS requirements, instead of their channel conditions, which offers two following benefits. One is that the SIC decoding order, power allocation, and user scheduling can be designed more appropriately to meet the users QoS requests. The other is to make NOMA communications more general, e.g., applicable to scenarios in which users channel conditions are the same.

6 Conclusions

In this article, a promising multiple access technology for 5G networks, NOMA, is discussed. A two-user SC-NOMA scheme and a multiuser MC-NOMA scheme were presented and discussed to illustrate the basic concepts and principles of NOMA. A literature review about performance analyses of NOMA, resource allocation for NOMA, and MIMO-NOMA was discussed. Furthermore, we presented the key features and potential research challenges of NOMA.

References

[1] J. Andrews, S. Buzzi, W. Choi, et al., "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014. doi: 10.1109/JSAC.2014.2328098.
[2] G. Wunder, P. Jung, M. Kasparick, et al., "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 97–105, Feb. 2014. doi: 10.1109/MCOM.2014.6736749.
[3] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010. doi: 10.1109/TWC.2010.092810.091092.
[4] J. Zhu, D. W. K. Ng, N. Wang, R. Schober, and V. K. Bhargava, "Analysis and design of secure massive MIMO systems in the presence of hardware impairments," *arXiv preprint* arXiv:1602.08534, 2016.
[5] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, Jun. 2011. doi: 10.1109/MCOM.2011.5783993.
[6] T. Rappaport, S. Sun, R. Mayzus, et al., "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013. doi: 10.1109/ACCESS.2013.2260813.
[7] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, Apr. 2012. doi: 10.1109/JSAC.2012.120401.
[8] J. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, Mar. 2013. doi: 10.1109/MCOM.2013.6476878.
[9] D. Ramasamy, R. Ganti, and U. Madhow, "On the capacity of picocellular networks," in *IEEE International Symposium on Information Theory*, Istanbul, Turkey, Jul. 2013, pp. 241–245. doi: 10.1109/ISIT.2013.6620224.
[10] D. W. K. Ng, E. S. Lo, and R. Schober, "Robust beamforming for secure communication in systems with wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4599–4615, Aug. 2014. doi: 10.1109/TWC.2014.2314654.
[11] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3292–3304, Sept. 2012. doi: 10.1109/TWC.2012.072512.111850.
[12] 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, Dec. 2013. doi: 10.1109/TWC.2013.103113.130470.
[13] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, Sept. 2015. doi: 10.1109/MCOM.2015.7263349.
[14] J. M. Meredith, "Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13)," 3GPP Tech. Rep. TR 36.859, Dec. 2015.
[15] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008. doi: 10.1109/TSP.2007.909320.
[16] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *IEEE Vehicular Technology Conference*, Taiwan, China, May 2010, pp. 1–5. doi: 10.1109/VETECS.2010.5493941.
[17] R. Razavi, M. Al-Imari, M. A. Imran, R. Hoshyar, and D. Chen, "On receiver design for uplink low density signature OFDM (LDS-OFDM)," *IEEE Transactions on Communications*, vol. 60, no. 11, pp. 3499–3508, Nov. 2012. doi: 10.1109/TCOMM.2012.082812.110284.
[18] H. Nikopour and H. Baligh, "Sparse code multiple access," in *IEEE Personal, Indoor and Mobile Radio Communications Symposium*, London, United Kingdom, Sept. 2013, pp. 332–336. doi: 10.1109/PIMRC.2013.6666156.
[19] X. Dai, S. Chen, S. Sun, et al., "Successive interference cancelation amenable multiple access (SAMA) for future wireless communications," in *IEEE International Conference on Communication Systems*, Macau, China, Nov. 2014, pp. 222–226. doi: 10.1109/PIMRC.2013.6666156.
[20] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 76–88, Jan. 2016. doi: 10.1109/TSP.2015.2480042.
[21] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for MC-NOMA systems," *arXiv preprint*, arXiv:1603.08132, 2016.
[22] Q. Sun, S. Han, C.-L. I, and Z. Pan, "Energy efficiency optimization for fading MIMO non-orthogonal multiple access systems," in *IEEE International Conference on Communications*, London, United Kingdom, Jun. 2015, pp. 2668–2673. doi: 10.1109/ICC.2015.7248728.
[23] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015. doi: 10.1109/LSP.2015.2417119.
[24] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *IEEE Global Communications Conference*, Atlanta, USA, Dec. 2013, pp. 66–70. doi: 10.1109/GLOCOMW.2013.6824963.
[25] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *IEEE Personal, Indoor and Mobile Radio Communications Symposium*, London, United Kingdom, Sept. 2013, pp. 611–615. doi: 10.1109/GLOCOMW.2013.6824963.
[26] Z. Ding, Z. Yang, P. Fan, and H. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014. doi: 10.1109/LSP.2014.2343971.
[27] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "A new evaluation criterion for non-orthogonal multiple access in 5G software defined networks," *IEEE Access*, vol. 3, pp. 1633–1639, Sept. 2015. doi: 10.1109/ACCESS.2015.2480117.
[28] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cam-

bridge, United Kingdom: Cambridge University Press, 2005.
[29] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation," in *IEEE International Symposium on Wireless Communication Systems*, Paris, France, Aug. 2012, pp. 476–480. doi: 10.1109/ISWCS.2012.6328413.
[30] Y. Saito, Y. Kishiyama, A. Benjebbour, et al., "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *IEEE Vehicular Technology Conference*, Dresden, Germany, Jun. 2013, pp. 1– 5. doi: 10.1109/VTCSpring.2013.6692652.
[31] Z. Ding, P. Fan, and V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010– 6023, 2016. doi: 10.1109/TVT.2015.2480766.
[32] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 654–667, Feb. 2016. doi: 10.1109/TCOMM.2015.2511078.
[33] M. R. Hojeij, J. Farah, C. Nour, and C. Douillard, "Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access," in *IEEE Vehicular Technology Conference*, Glasgow, Scotland, May 2015, pp. 1–6. doi: 10.1109/VTCSpring.2015.7146056.
[34] M. R. Hojeij, J. Farah, C. Nour, and C. Douillard, "New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access," *Wireless Personal Communications*, vol. 87, no. 3, pp. 837– 867, May 2015. doi:10.1007/s11277-015-2629-2.
[35] B. Di, S. Bayat, L. Song, and Y. Li, "Radio resource allocation for downlink non-orthogonal multiple access (NOMA)networks using matching theory," in *IEEE Global Communications Conference*, San Diego, USA, Dec. 2015, pp. 1–6. doi: 10.1109/GLOCOM.2015.7417643.
[36] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems," in *IEEE Global Communications Conference*, San Diego, USA, Dec. 2015, pp. 1–6. doi: 10.1109/GLOCOM.2015.7417761.
[37] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 583–596, Feb. 2014. doi: 10.1109/TSP.2013.2293130.
[38] H. Kim, K. Kim, Y. Han, and S. Yun, "A proportional fair scheduling for multicarrier transmission systems," in *IEEE Vehicular Technology Conference*, Los Angeles, USA, Sept 2004, vol. 1, pp. 409– 413. doi: 10.1109/VETECF.2004.1400034.
[39] C. Wengerter, J. Ohlhorst, and A. G. E. von Elbwart, "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," in *IEEE Vehicular Technology Conference*, Stockholm, Sweden, May 2005, vol. 3, pp. 1903–1907. doi: 10.1109/VETECS.2005.1543653.
[40] F. Liu, P. Mahonen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *IEEE Personal, Indoor and Mobile Radio Communications Symposium*, Hong Kong, China, Aug. 2015, pp. 1127–1131. doi: 10.1109/PIMRC.2015.7343467.
[41] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink non-orthogonal multiple access with statistical channel state information," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4718–4731, Jul. 2016. doi: 10.1109/TWC.2016.2544922.
[42] Y. Lan, A. Benjebbouiu, X. Chen, A. Li, and H. Jiang, "Considerations on downlink non-orthogonal multiple access (NOMA) combined with closed-loop SU-MIMO," in *IEEE International Conference on Signal Processing and Communication System*, Gold Coast, Australia, Dec. 2014, pp. 1–5. doi: 10.1109/ICSPCS.2014.7021086.
[43] X. Chen, A. Benjebbour, Y. Lan, A. Li, and H. Jiang, "Impact of rank optimization on downlink non-orthogonal multiple access (NOMA) with SU-MIMO," in *IEEE International Conference on Communication Systems*, Macau, China, Nov. 2014, pp. 233–237. doi: 10.1109/ICCS.2014.7024800.
[44] B. Kim, S. Lim, H. Kim, et al., "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *IEEE Military Communications Conference*, San Diego, USA, Nov. 2013, pp. 1278–1283. doi: 10.1109/MILCOM.2013.218.
[45] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 791–800, Mar. 2015. doi: 10.1109/TCOMM.2015.2394393.
[46] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 405–408, Aug. 2015. doi: 10.1109/LWC.2015.2426709.
[47] Q. Sun, S. Han, Z. Xu, S. Wang, C.-L. I, and Z. Pan, "Sum rate optimization for MIMO non-orthogonal multiple access systems," in *IEEE Wireless Communications and Networking Conference*, New Orleans, USA, Mar. 2015, pp. 747–752. doi: 10.1109/WCNC.2015.7127563.
[48] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 537–552, Jan. 2016. doi: 10.1109/TWC.2015.2475746.
[49] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016. doi: 10.1109/TWC.2016.2542066.
[50] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 629–633, May 2016. doi: 10.1109/LSP.2016.2543025.
[51] T. Takeda and K. Higuchi, "Enhanced user fairness using non-orthogonal access with SIC in cellular uplink," in *IEEE Vehicular Technology Conference*, San Francisco, USA, Sept. 2011, pp. 1– 5. doi: 10.1109/VETECF.2011.6093272.
[52] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *IEEE International Symposium on Wireless Communications Systems*, Barcelona, Spain, Aug. 2014, pp. 781–785. doi: 10.1109/ISWCS.2014.6933459.
[53] S. Chen, K. Peng, and H. Jin, "A suboptimal scheme for uplink NOMA in 5G systems," in *IEEE International Wireless Communications and Mobile Computing Conference*, Dubrovnik, Croatia, Aug. 2015, pp. 1429–1434. doi: 10.1109/IWCMC.2015.7289292.
[54] Y. Chen, J. Schaepperle, and T. Wild, "Comparing IDMA and NOMA with superimposed pilots based channel estimation in uplink," in *IEEE Personal, Indoor and Mobile Radio Communications Symposium*, Aug. 2015, pp. 89–94. doi: 10.1109/PIMRC.2015.7343274.
[55] M. Al-Imari, P. Xiao, and M. A. Imran, "Receiver and resource allocation optimization for uplink NOMA in 5G wireless networks," in *IEEE International Symposium on Wireless Communication Systems*, Brussels, Belgium, Aug. 2015. doi: 10.1109/ISWCS.2015.7454317.
[56] B. Kim, W. Chung, S. Lim, et al., "Uplink NOMA with multi-antenna," in *IEEE Vehicular Technology Conference*, Glasgow, Scotland, May 2015, pp. 1–5. doi: 10.1109/VTCSpring.2015.7146149.
[57] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink non-orthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, Mar. 2016. doi: 10.1109/LCOMM.2016.2521374.
[58] H. Haci, H. Zhu, and J. Wang, "A novel interference cancellation technique for non-orthogonal multiple access (NOMA)," in *IEEE Global Communications Conference*, San Diego, USA, Dec. 2015, pp. 1– 6. doi: 10.1109/GLOCOM.2015.7417128.
[59] H. Haci, "Non-orthogonal multiple access (NOMA) with asynchronous interference cancellation," Ph.D. dissertation, Dept. Electron. Eng., Univ. of Kent, Kent, England, Mar. 2015.
[60] Z. Ding, M. Peng, and H. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462– 1465, Aug. 2015. doi: 10.1109/LCOMM.2015.2441064.
[61] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 416–419, Jun. 2016. doi: 10.1109/LWC.2016.2574709.
[62] Y. Liu, Z. Ding, M. Eikashlan, and H. Poor, "Cooperative non-orthogonal multiple access in 5G systems with SWIPT," in *European Signal Processing Conference*, Nice, France, Aug. 2015, pp. 1999– 2003. doi: 10.1109/EUSIPCO.2015.7362734.
[63] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Optimal design of non-orthogonal multiple access with wireless power transfer," *arXiv preprint* arXiv:1511.01291, 2015.
[64] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Non-orthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1– 1, Feb. 2016. doi: 10.1109/TVT.2016.2524694.
[65] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat, "Non-orthogonal multiple access for visible light communications," *IEEE Photonics Technology Letters*, vol. 28, no. 1, pp. 51– 54, Jan. 2016. doi: 10.1109/LPT.2015.2479600.
[66] Y. Zhang, H. M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Communications Letters*, vol. 20, no. 5, pp. 930–933, Mar. 2016. doi: 10.1109/LCOMM.2016.2539162.

Manuscript received: 2016-08-15

Biographies

WEI Zhiqiang (zhiqiang.wei@unsw.edu.au) received the BE degree from Northwestern Polytechnical University, China in 2012. He is currently pursuing the PhD degree in Wireless Communications Laboratory, University of New South Wales, Australia. His research interests include non-orthogonal multiple access and resource allocation.

YUAN Jinhong (j.yuan@unsw.edu.au) received the BE and PhD degrees in electronics engineering from Beijing Institute of Technology, China in 1991 and 1997, respectively. From 1997 to 1999, he was a reasearch fellow with the School of Electrical Engineering, University of Sydney, Australia. In 2000, he joined the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia, where he is currently a professor of telecommunications. He has authored two books, three book chapters, more than 200 papers in telecommunications journals and conference proceedings, and 40 industrial reports. His research interests include error control coding and information theory, communication theory, and wireless communications. He is a co-inventor of one patent on MIMO systems and two patents on low-density parity-check codes. He is currently serving as an associate editor for the *IEEE TCOM*. He served as the IEEE NSW Chair of Joint Communications/Signal Processions/Ocean Engineering Chapter from 2011 to 2014. He was the co-recipient of three best paper awards and one best poster award, including the Best Paper Award from the IEEE Wireless Communications and Networking Conference, Cancun, Mexico in 2011, and the Best Paper Award from the IEEE International Symposium on Wireless Communications Systems, Trondheim, Norway in 2007.

Derrick Wing Kwan Ng (w.k.ng@unsw.edu.au) received the bachelor degree with first class honors and the Master of Philosophy (M.Phil.) degree in electronic engineering from the Hong Kong University of Science and Technology (HKUST) in 2006 and 2008, respectively. He received his PhD degree from the University of British Columbia (UBC) in 2012. He was a senior postdoctoral fellow at the Institute for Digital Communications, University of Erlangen-Nuremberg, Germany. He is now working as a lecturer at the University of New South Wales, Australia. Dr. Ng has published more than 80 journal and conference papers and his publications have been cited over 2000 times in Google Scholar with an h-index of 20. Dr. Ng is currently an editor of *IEEE Communications Letters* and *IEEE Transactions on Green Communications and Networking*. He served as a Co-Chair for the Wireless Access Track of 2014 IEEE 80th Vehicular Technology Conference and 2016 IEEE GlobeCom Workshop on Wireless Energy Harvesting. He was also a co-organizer and guest editor of the special issue on Energy Harvesting Wireless Communications in EURASIP Journal on Wireless Communications and Networking in 2014.

Maged Elkashlan (maged.elkashlan@qmul.ac.uk) received the PhD degree in electrical engineering from the University of British Columbia, Canada in 2006. From 2007 to 2011, he was with the Wireless and Networking Technologies Laboratory, Commonwealth Scientific and Industrial Research Organization, Australia. During this time, he held an adjunct appointment with the University of Technology Sydney, Australia. In 2011, he joined the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. He currently holds visiting faculty appointments with the University of New South Wales, Australia, and the Beijing University of Posts and Telecommunications, China. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing for distributed data processing, heterogeneous networks, and massive MIMO. Dr. Elkashlan received the best paper award at the IEEE International Conference on Communications in 2014, the International Conference on Communications and Networking in China in 2014, and the IEEE Vehicular Technology Conference in 2013. He also received the Exemplary Reviewer Certificate of the IEEE CL in 2012. He serves as an editor of *IEEE TWC*, *IEEE TVT*, and *IEEE CL*. He also serves as a lead guest editor of the Special Issue on Green Media: The Future of Wireless Multimedia Networks of the *IEEE Wireless Communications Magazine* and the Special Issue on Millimeter Wave Communications for 5G of the *IEEE Communications Magazine*, and a guest editor of the Special Issue on Energy Harvesting Communications of the *IEEE Communications Magazine* and the Special Issue on Location Awareness for Radios and Networks of the *IEEE JSAC*.

DING Zhiguo (z.ding@lancaster.ac.uk) received his BEng from the Beijing University of Posts and Telecommunications, China in 2000, and the PhD degree from Imperial College London, U.K. in 2005. From Jul. 2005 to Aug. 2014, he was working in Queen's University Belfast, Imperial College and Newcastle University. Since Sept. 2014, he has been with Lancaster University as a Chair Professor in Signal Processing. From Sept. 2012 to Sept. 2017, he has also been an academic visitor in Princeton University. Dr Ding's research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He is serving as an editor for *IEEE TCOM*, *IEEE TVT*, *IEEE WCL*, and *IEEE CL*. He was the TPC Co-Chair for ICWMMN2015, and Symposium Chair for ICNC 2016 and WOCC 2015. He received the best paper award in ICWOC 2009 and WCSP 2015, IEEE CL Exemplary Reviewer 2012, and the EU Marie Curie Fellowship 2012-2014.

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 004
http://www.cnki.net/kcms/detail/34.1294.TN.20161018.1014.002.html, published online October 18, 2016

Unified Framework Towards Flexible Multiple Access Schemes for 5G

SUN Qi, WANG Sen, HAN Shuangfeng, and Chih-Lin I
(China Mobile Research Institute, Beijing 100032, China)

Abstract

Non-orthogonal multiple access (NOMA) schemes have achieved great attention recently and been considered as a crucial component for 5G wireless networks since they can efficiently enhance the spectrum efficiency, support massive connections and potentially reduce access latency via grant free access. In this paper, we introduce the candidate NOMA solutions in 5G networks, comparing the principles, key features, application scenarios, transmitters and receivers, etc. In addition, a unified framework of these multiple access schemes are proposed to improve resource utilization, reduce the cost and support the flexible adaptation of multiple access schemes. Further, flexible multiple access schemes in 5G systems are discussed. They can support diverse deployment scenarios and traffic requirements in 5G. Challenges and future research directions are also highlighted to shed some lights for the standardization in 5G.

Keywords

5G; non-orthogonal multiple access; unified framework; flexible multiple access

1 Introduction

Worldwide initiatives on the 5th generation (5G) wireless communication have been extensively carried out, starting with an investigation on user demands, scenarios, key performance indicators (KPIs) and enabling technologies. A global consensus is first forming that 5G network will be able to sustainably support 1000-fold mobile data traffic growth, improve energy efficiency (EE) and cost efficiency by over 100 times, provide fiber link access data rates and "zero" latency user experience, and be capable of connecting 100 billion devices and capable of delivering a consistent experience across a variety of scenarios including the cases of ultra-high traffic volume density, ultra-high connection density and ultra-high mobility [1]. Three typical usage scenarios of 5G are also identified: enhanced mobile broadband (eMBB), massive machine type communication (mMTC) and ultra-reliable low latency machine type communication (URLLC), targeting different 5G capabilities. Beyond that, the standardization organizations, e.g. 3GPP has started the new research on 5G, studying the new access technology to meet a broad range of use cases.

Multiple access schemes, the most fundamental aspect of the physical layer, to a large extent, are considered as the defining technical feature of each wireless communication generation and have continually evolved in each cellular generation from frequency division multiple access (FDMA), time division multiple access (TDMA) in 1G and 2G to code division multiple access (CDMA) in 3G and orthogonal frequency-division multiple access/single-carrier FDMA (OFDMA/SC-FDMA) for 4G. Facing the stringent demands of diverse scenarios in 5G, e.g., 1000x higher data rates, massive uplink connectivity and low access latency, the traditional pure orthogonal multiple access is not a good option. Some alternative non-orthogonal multiple access schemes have attracted considerable attention and been identified as a crucial technology component in 5G since they can serve multiple users in the same frequency and time resources via code domain multiplexing and/or power domain multiplexing to enhance system access performance. The non-orthogonal multiple access schemes are potentially able to support massive connections, improve spectrum efficiency and also reduce access latency via the grant free access. Currently, some potential alternative multiple access schemes are being actively studied in 3GPP for 5G, including superposition coding based non-orthogonal multiple access (SPC-NOMA) [2], multi user shared access (MUSA) [3], sparse code multiple access (SCMA) [4], pattern division multiple access (PDMA) [5], resource spread multiple access (RSMA) [6], non-orthogonal coded multiple access (NCMA) [7], and interleave-grid multiple access (IGMA) [8].

In this paper, the principles, advantages and application scenarios of different multiple access techniques are discussed

and compared. In addition, we introduce a unified framework that can merge a wide range of multiple access techniques, which helps to minimize the hardware functional module. Based on the unified framework, some initial work on flexible multiple access schemes is also introduced. Finally, the challenges and future directions are discussed.

2 Candidate Non-Orthogonal Multiple Access Solutions

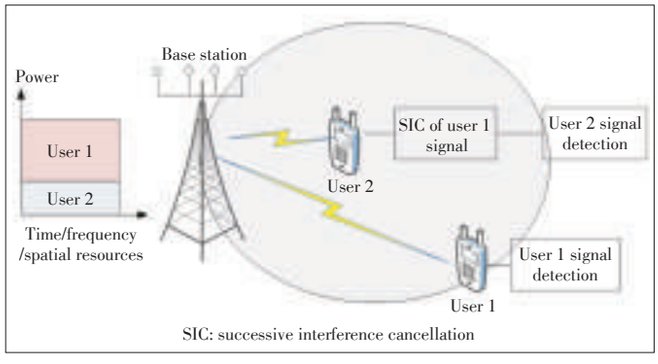
In this section, we introduce the typical candidate NOMA solutions for 5G, which can be basically divided into three categories, i.e., the power domain based, code domain based and interleaver based. Their principles and key features are discussed. At last, we provide their comparison in terms of application scenarios, system performance, receivers, etc.

2.1 Power Domain Based Solutions

2.1.1 SPC-NOMA

NOMA based on superposition coding utilizes power domain for user multiplexing and can be applied for both downlink and uplink. Established by network information theory, non-orthogonal access with successive interference cancellation (SIC)/dirty paper coding (DPC) can achieve the multiuser capacity region both in uplink and downlink. NOMA superposes multiple users in power-domain and exploits channel gain difference between the multiplexed users with the aid of advanced receiver, e.g. the SIC receiver, for user separation. **Fig. 1** shows signal transmission and receiving in downlink NOMA system with two users. Currently the NOMA technique is being discussed in the 3GPP under the study item of "study on downlink multiuser superposition transmission (MUST)" for release 13 [9]. For the study in 3GPP, the study scope of NOMA is very limited, e.g. only about downlink transmission, only for the intra-cell usage and only for data channels.

For 5G system, there are more application scenarios of NOMA technique, such as uplink and control channel, and more advanced NOMA techniques, such as combination with sophisticated multiple-input multiple-output (MIMO) techniques and inter-cell techniques. In [10]–[12], MIMO NOMA schemes have been studied. Network NOMA which considering the multi-cell scenarios are also studied from EE-SE co-design perspective in [13].



▲**Figure 1. Illustration of SPC -NOMA transmission.**

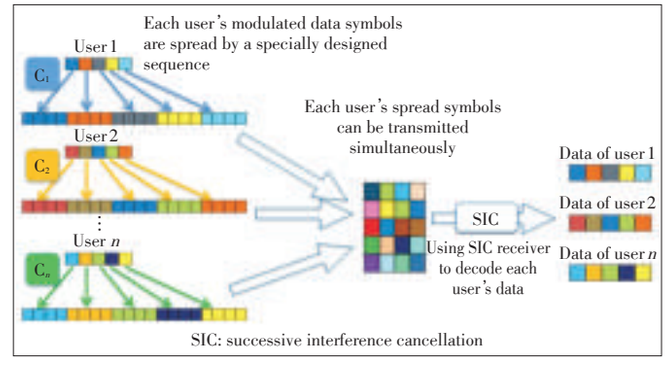
2.2 Code Domain Based Solutions

2.2.1 MUSA

MUSA is a non-orthogonal multiple access scheme operating in code domain. Conceptually, each user's modulated data symbols are spread firstly by a specially designed sequence which facilitates robust SIC implementation compared to the sequences employed by traditional direct-sequence CDMA (DS-CDMA). Then, each user's spread symbols are transmitted concurrently on the same radio resource by means of "Shared Access", which is essentially a superposition process. Finally, decoding of each user's data from superimposed signal can be performed at the base-station side using SIC technology.

The major processing blocks of MUSA transmitter and receiver are illustrated in **Fig. 2**. Symbols of each user are spread by a spreading sequence. Multiple spreading sequences constitute a pool from which each user can randomly pick one. Note that for the same user, different spreading sequences may also be used to different symbols. This may further improve the performance via interference averaging. Then, all spreading symbols are transmitted over the same time-frequency resources. The spreading sequences should have low cross-correlation and can be non-binary. At the receiver, codeword level SIC is used to separate data from different users. The complexity of codeword level SIC is less of an issue in the uplink as the receiver anyway needs to decode the data for all users. The only noticeable impact on the receiver implementation would be that the pipeline of processing may be changed in order to perform SIC operation.

MUSA relies on a special family of complex spread sequences that can enjoy relatively low cross-correlation even when they are very short, say, 8 or even 4. The real and imaginary parts of the complex spread sequence can be drawn from an M-ary real value set. For example, for a 3-value set {-1, 0, 1}, ev-



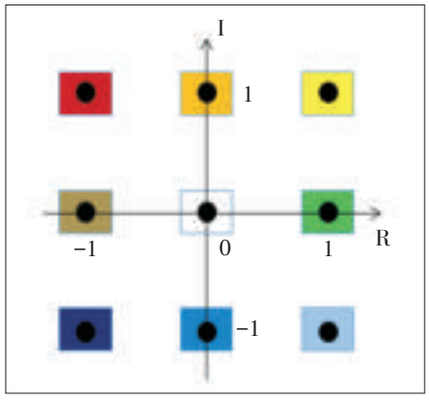
▲**Figure 2. An example of MUSA with four resources shared by multiple users [1].**

ery bit of the complex sequence is drawn from the constellation depicted in **Fig. 3** with equal probability.

It should be pointed out that the spread sequences used in MUSA are different from the spreading codes, in the sense that MUSA spreading does not have the low density property. Equipped with the well-optimized spreading sequence and state-of-the-art SIC technology, MUSA is capable of decoupling the multiuser mingled data even if those users are contending to access the system. Potentially a large number of devices are allowed to transmit data at their will, by randomly picking spread sequences, spread the data and send them. In other words, MUSA is suitable for the scenario where the uplink transmissions are not tightly scheduled, and the grants for transmission are not signaled per user basis, and with a high overloading. The relaxed UL synchronization requirement for MUSA allows simple derivation of UL time from a DL synchronization process, which can greatly cut down the battery consumption. Lastly, the code domain superposition nature of MUSA can turn the near-far problem into a near-far advantage. The disparity in the received signal to noise ratio (SNR) across the simultaneously transmitting users can be exploited in MUSA to facilitate SIC. Tight transmit power control is no longer needed, which can further lower the device cost and its power consumption.

2.2.2 SCMA

SCMA is a novel non-orthogonal multiple access scheme with sparse codebooks. The main idea of SCMA is to accommodate more users with limited resources and increase the total network throughput, without scarifying user experience, which can be overloaded to enable massive connectivity and support grant-free access. There are multiple layers in SCMA, which can be used for user multiplexing. Each layer has a predefined codebook, which consists of multiple codewords. The codewords are composed of multi-dimensional complex symbols, and the codewords in the same codebook have the same sparse pattern. For each layer, the coded bits are directly mapped to codewords, which are selected from layer-specific SCMA codebooks. The codewords of different layers are overlaid in code and power domains and carried over shared time-frequency resources. Typically, the layer multiplexing may become overloaded if the number of layers is more than the length of the codewords.



◀**Figure 3. The elements of the complex spreading sequence [3].**

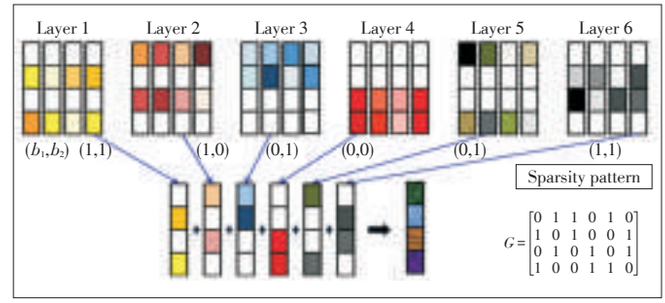
Fig. 4 shows an example of bits to codewords mapping in a SCMA system. The codebook design of SCMA has been studied in [14]; it has been shown that with multi-dimensional constellation, shaping and coding gain can be achieved. At the receiver, joint multiuser detection algorithms are needed. Due to the sparsity of the SCMA codeword structure, message passing algorithm (MPA) on factory graph with much lower complexity can be adopted to achieve a suboptimal performance. Some simplified algorithms are proposed [15]–[19] to further reduce the detection complexity.

Besides the codebook and the receiver design, some other challenging issues of SCMA, e.g., the energy efficiency optimization, uplink grant free access, downlink multiuser transmission, and the multi-cell transmission based on SCMA have also been studied. The energy efficiency performance and optimization of SCMA are investigated in [20] and [21]. In [22] and [23], uplink contention based grant-free access based on SCMA has been proposed for 5G radio access. [24] and [25] focus on the downlink multiuser SCMA (MU-SCMA) network. [24] theoretically derives the capacity for downlink Massive MIMO MU-SCMA systems. In [25], a weighted sum rate based user pairing and power sharing algorithm are introduced to the MU-SCMA network. It shows that SCMA can significantly increase the downlink spectral efficiency of 5G wireless cellular networks. Further, SCMA has also been introduced into multi-cell transmission. SCMA based uplink inter-cell interference cancellation technique and open loop joint coordinated multiple point transmission are studied in [26] and [27], respectively.

There are still many challenging issues for SCMA, which need to be solved in the future work. For example, the layer multiplexing in SCMA provides new degree-of-freedom for user scheduling. The algorithms for user grouping and power allocation need to be optimized. In addition, the combination of SCMA and MIMO can be further enhanced.

2.2.3 PDMA

PDMA introduces reasonable diversity between multiple users to promote the capacity, which can obtain higher multiuser



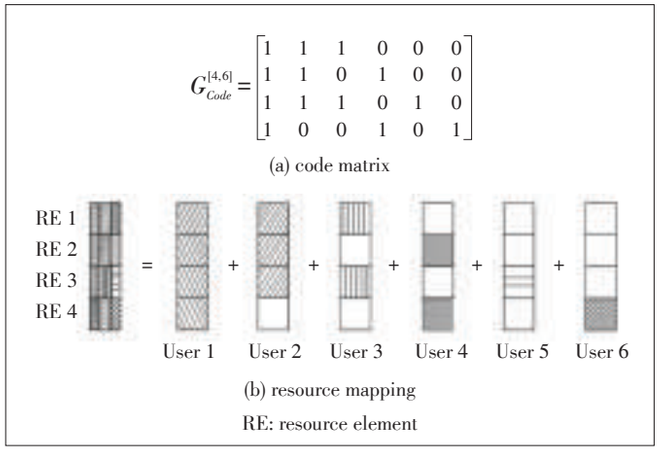
▲**Figure 4. Illustration of SCMA codebooks and the process of bit mapping [1].**

multiplexing and diversity gain. It considers the joint design of the transmitter and the receiver based on the optimization point of view for multiuser communication system. At the transmitter side, the non-orthogonal characteristic pattern is used to distinguish users based on the multiple signals domain (including time, frequency and the space domain). At the receiver side, sub-optimal multiuser detection by General SIC based on the features of the user pattern is utilized.

To alleviate the error propagation problem of the SIC receiver, the pattern used in PDMA is generally designed to ensure unequal transmission diversity for each user. In this way, the identical diversity order can be achieved after detection. Inspired by the idea of unequal transmission diversity and sparse coding, an example of pattern and the related resource mapping has been proposed (**Fig. 5**).

In the example, a code can also be seen as a pattern, which is used to define sparse mapping from data to a group of resources. The code could be represented by a binary vector. The dimension of the vector equals to the number of resources in a group. Each element in the vector corresponds to a resource in a resource group. A "1" means that data shall be mapped to the corresponding resource. Actually, the number of "1" in the code is defined as its transmission diversity order. A code matrix is constructed by all codes sharing on the same resource group. Assuming six users multiplexing on four resource elements (REs). The data for User 1 are mapped to all the four resources in the group, and the data for User 2 are mapped to the first three resources, etc. The order of transmission diversity of the six users is 4, 3, 2, 2, 1, and 1, which is obviously quite different from the SCMA scheme where all the users bear the same transmission diversity.

Generally, if N is the size of resource group (the row number of code matrix), there are 2^N-1 possible binary vectors for a code matrix. Assuming K is the column number determined based on overload factor, we can thus choose K patterns out from 2^N-1 candidates to construct code matrix. Selection of codes also gives impacts on performance and complexity.



▲**Figure 5. Users sharing on four resource elements [5].**

2.2.4 RSMA

RSMA combines the low rate channel code and the scrambling code (and optionally different interleavers) with good correlation properties to separate different transmitters. In RSMA system, all users use the same frequency and time resources to transmit messages to the base station, regardless of the number of concurrent users. In other words, each user's transmission power can be spread over all the available time and frequency resources.

RSMA can be coupled with various waveforms/modulation schemes depending on the design target. Generally, it includes the single carrier RSMA and the multi-carrier RSMA. The single carrier RSMA is optimized for battery power consumption and link budget extension by using single carrier waveforms. It allows grant-less transmission and potentially allows asynchronous access. The grant-less transmission using RSMA reduces the signaling overhead, while the single carrier waveform further reduces peak-to-average power ratio (PAPR) and achieves higher power amplifier efficiency. The pulse shaping block can further enhance the PAPR (e.g. potentially leading to constant envelope waveform), reducing out-of-band emission simultaneously. The multi-carrier RSMA is optimized for low latency access, where reducing access delay is the design priority. It is suitable for the scenario where a connected state device is already synchronized to the base station and not link budget limited (e.g., close to the base station). Such a device can use RSMA with OFDM-based multi-carrier waveform for grant-less transmission to reduce overall access delay.

2.2.5 NCMA

NCMA is a multiple access scheme based on the resource spreading by using non-orthogonal codewords, which is composed of the codewords obtained by Grassmannian line packing problem [28]. To minimize the MUI theoretically, the spreading codes are designed with the minimum correlation.

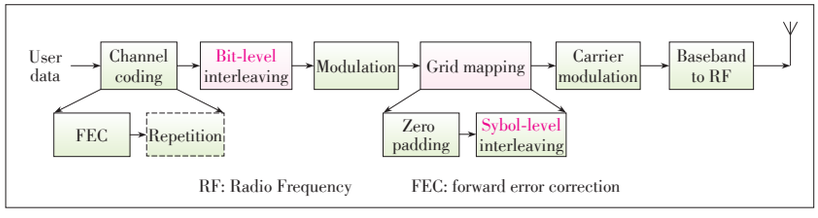
The non-orthogonal codebook is defined by $C=\left[c^1\cdots c^K\right]=\begin{bmatrix}c_1^1&\cdots&c_1^K\\ \vdots&\ddots&\vdots\\ c_N^1&\cdots&c_N^K\end{bmatrix}, C\in\mathbb{C}^{N\times K}$, where N is the spreading factor and K is the superposition factor. Then, the codebook design problem can be posed in terms of maximizing the minimum chordal distance between codeword pairs $\min_C\left(\max_{1\le k\le j\le K}\sqrt{1-\left|(c^k)^*.c^j\right|}\right)$ where $(c^k)^*$ is the conjugate codeword of c^k.

NCMA can provide the additional throughput or improved connectivity with a small loss of block error rate (BLER) in specific environments, by exploiting additional layers through the superposed symbol, while satisfying QoS constraints. Since the receiver of NCMA system is available for parallel interference

cancellation (PIC), the multiuser detection can be implemented with low complexity. In addition, the MUI level between codeword pairs is always similar due to the correlation characteristics mentioned above. Consequently, NCMA provides the potentials in terms of throughput or connectivity under special scenarios, e.g., huge connections with small packet in mMTC scenarios without changing the transmission block size, or for reducing the collision probability in contention based multiple access.

2.3 Interleaver Based Solution

IGMA is an interleaver-based MA scheme, The typical transmitter system structure using IGMA is shown in **Fig. 6**. Basically, the IGMA scheme distinguishes different users based on different bit-level interleavers, different grid mapping patterns and different combinations of bit-level interleaver and grid mapping pattern.



▲**Figure 6. The IGMA transmitter [8].**

Compared to the need of well-designed codewords or code sequences, the sufficient source of bit-level interleavers and/or grid mapping patterns are able to provide enough scalability for different connection densities, and also provide flexibility to achieve good balance between channel coding gain and benefit from sparse resource mapping. By proper selection, the low correlated bit-level interleavers is achieved. In the grid mapping process, sparse mapping based on zero padding and symbol-level interleaving is introduced, which provides another dimension for user multiplexing. Moreover, the density ρ of the grid mapping pattern is defined as the occupied RE numbers N_{used} dividing the total assigned RE numbers N_{all}, i.e. $\rho = N_{used}/N_{all}$. Different densities could be flexibly configured. It should be noted that the symbol sequence order is randomized after the grid mapping process due to symbol-level interleaving, which may further bring benefit in terms of combating frequency selective fading and inter-cell interference, compared to resource mapping using direct code-matrices/codebooks.

At the receiver side, the low complexity multiuser detector (MUD) and the elementary signal estimator (ESE) that takes advantage of the special property of interleaving can be utilized with a simple de-mapping operation on the top. It should be noted that lower density of the grid mapping pattern further reduces detection complexity of ESE for IGMA. In addition, MAP and MPA detectors are also applicable for IGMA, which can improve the detection performance a lot comparing to ESE at the cost of complexity. The complexity of MAP/MPA for IGMA probably can be alleviated when spare grid mapping is used, due to the similar property of LDS.

Fig. 7 shows an example of the grid mapping process of IGMA. The sparse symbol-to-RE mapping is performed based on an assigned grid mapping pattern. An exemplary operation can be mathematically formulated as a process by permutation matrix α_{GM}. According to the symbol-level interleave $\theta_{k,2}$ associated with the grid mapping pattern β_k with density $\rho_k (0<\rho_k \leq 1)$, the corresponding permutation matrix $\alpha_{GM} \in \mathbb{C}^{N\times L}$ can be obtained. Thus, the k_{th} user's symbol sequence s_k after zero padding and interleaving can be denoted by $s_k' = s_k \times \alpha_{GM} = [s_{k,1}', s_{k,2}', \cdots, s_{k,L}']$, where $L = N/\rho_k$ and ρ_k decides the number of zeros padded.

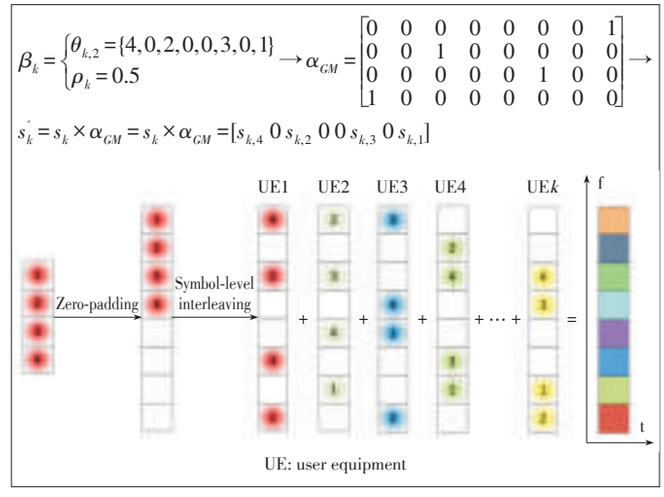
2.4 Summary of Multiple Access Techniques

The pros and cons of the multiple access techniques introduced above are summarized here in **Table 1.**

It's worth mentioning that some of these non-orthogonal schemes, such as SCMA MUSA and PDMA, can be implemented within a unified framework, and each of them corresponds to a different codebook mapping module. In this way, the air interface can handover between different multiple access schemes in a flexible way, and all the other modules can be reused. This helps to improve the resource utilization and reduce the cost. In the following section, we will provide a unified framework for the multiple access schemes.

3 Unified Framework of Multiple Access Schemes

Fig. 8 shows a unified framework of multiple access



▲**Figure 7. Example of the grid mapping process of IGMA when N = 4, ρ_k = 0.5 and L = 8 [8].**

▼Table 1. Summary of multiple access techniques

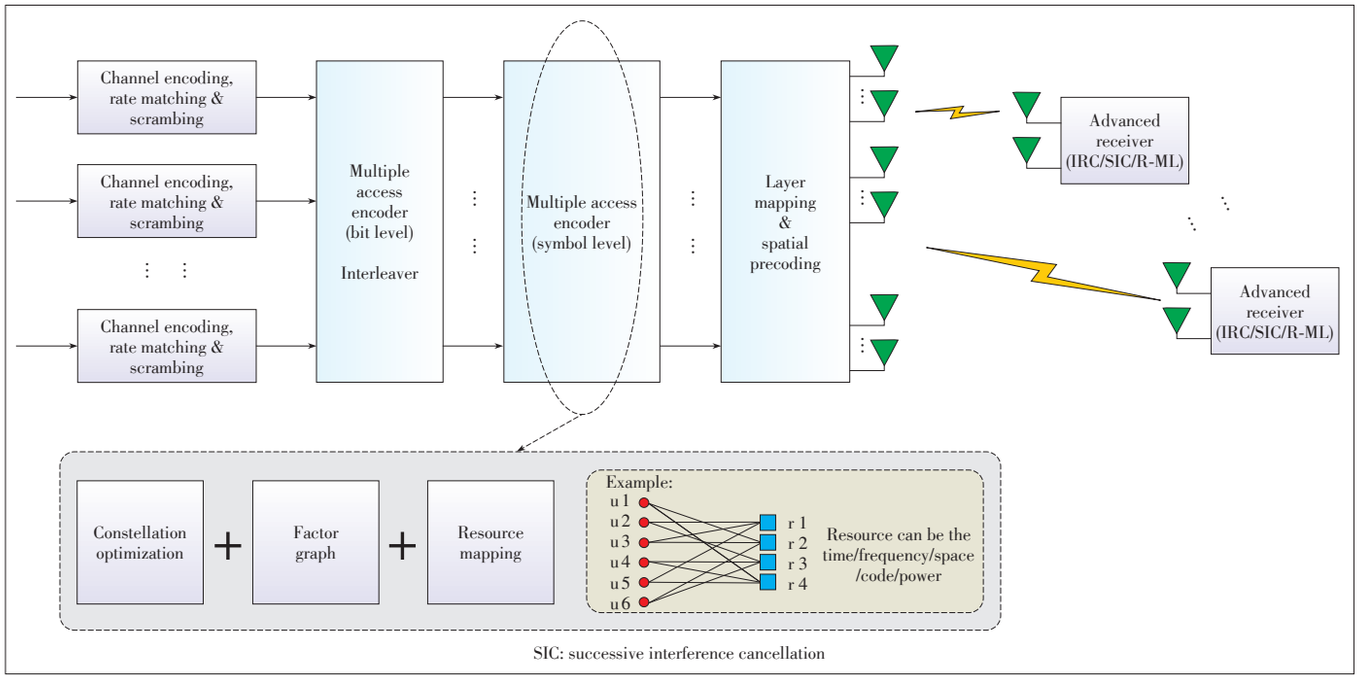
Category	Power domain based	Code domain based					Interleaver based
Scheme	SPC -NOMA	MUSA	SCMA	PDMA	RSMA	NCMA	IGMA
Scenario	DL: eMBB	UL: mMTC, URLLC DL: eMBB	UL: mMTC, URLLC DL: eMBB	UL: mMTC, URLLC DL: eMBB	UL: mMTC, URLLC	UL: eMBB, mMTC, URLLC	UL: eMBB, mMTC, URLLC
Multiplexing domain	Power	Code/Power	Code/Power	Code/Power/Spatial	Code/Power	code	Interleaver
Transmitter Overloading	Medium	High	High	High	High	High	High
Transmitter Spreading	No	Yes	Yes	Yes	Yes	Yes	Yes
Transmitter multi - dimension constellation	No	No	Yes	No	No	No	No
Receiver	SIC	SIC	MPA/SIC	SIC/MPA	SIC	PIC	MAP/MPA ESE MUD

DL: downlink
eMBB: enhanced mobile broadband
ESE: elementary signal estimator
IGMA: interleave-grid multiple access
mMTC: massive machine type communication
MPA: message passing algorithm
MUD: low complexity multiuser detector
MUSA: multi user shared access
NCMA: non-orthogonal coded multiple access
PDMA: pattern division multiple access
PIC: parallel interference cancellatio
RSMA: resource spread multiple access
SCMA: sparse code multiple access
SIC: successive interference cancellation
SPC-NOMA: superposition coding based non-orthogonal multiple access
UL: uplink
URLLC: ultra-reliable low latency machine type communication

schemes. The differences among these multiple access schemes lie in the different realization of interleaver, constellation optimization, factor graph and multiplexing domain. The detailed explanations are listed in **Table 2**.

4 Flexible Multiple Access in 5G

The above discussed advanced multiple access schemes as well as the traditional orthogonal multiple access scheme, e.g. OFDMA are all identified as potential candidates for 5G. There is no individual scheme can fulfill the requirements of all applications and scenarios in 5G system. A flexible adaptation of these multiple access schemes is needed to support the diverse deployment scenarios and traffic requirements. For example, in the case of massive connections, how to accommodate more users with limited resources has become a critical problem for next generation access network. With non-orthogonal multiple access schemes, e.g., SCMA, MUSA, PDMA and RSMA, the same resources are shared and reused by multiple users, thus the number of connections increases. To support the traffic with low latency requirement, non-orthogonal multiple access schemes help to realize grant-free multiple access, with which the latency is much lower, and the power consumption of the devices can be reduced. In other scenarios, such as downlink machine type traffic, the simple orthogonal multiple access schemes are better due to the device cost and implementation complexity. OFDMA can be utilized for the cell-center user with high data rate transmission applications.



▲Figure 8. Unified framework of multiple access schemes.

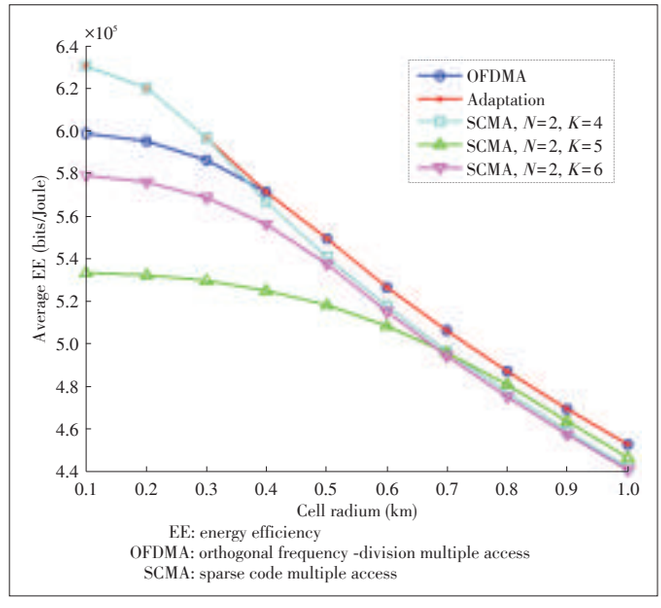
▼Table 2. Configuration methods of different multiple access schemes based on a unified framework

		Interleaver	Constellation mapping	Factor graph	Resource mapping (multiplexing domain)
OMA		Identity matrix	Gray-mapped legacy constellation	Identity matrix	Time/frequency/code/space
SPC -NOMA	MUST Cat 1 [9]	Identity matrix	non-Gray-mapped superposed constellation	Identity matrix	Power
	MUST Cat 2 [9]	Constraint permutation matrix	Gray-mapped superposed constellation	Identity matrix	Power/bit
	MUST Cat 3 [9]	Permutation matrix	Gray-mapped legacy constellation	Identity matrix	Bit
MUSA (uplink)		Identity matrix	Legacy constellation	Matrix composed of low cross-correlation and non-binary spreading sequence	Code
SCMA		Identity matrix	Joint optimization (multi -dimensional modulation + Sparse matrix[3])	Sparse matrix	Code/power
PDMA		Identity matrix	Legacy modulation	Sparse matrix with unequal diversity order	Code/power/space
RSMA		Optional	Legacy modulation	Matrix composed of scrambling code with good correlation properties	Code
NCMA		Identity matrix	legacy modulation	Matrix obtained by Grassmannian line packing problem	code
IGMA		Permutation matrix	legacy modulation	Sparse matrix	code

IGMA: interleave-grid multiple access
MUSA: multi user shared access
MUST: Downlink Multiuser Superposition Transmission
NCMA: non-orthogonal coded multiple access
OMA: orthogonal multiple acce
PDMA: pattern division multiple access
RSMA: resource spread multiple access
SCMA: sparse code multiple access
SPC-NOMA: superposition coding based non-orthogonal multiple access

Besides, the multiple access scheme should also be properly selected, taking the tradeoff of multiple conflicting objectives into account, e.g., complexity vs. performance, energy efficiency (EE) vs. spectral efficiency (SE) and coverage. In addition, because the channel conditions and service load may also dynamically vary, the multiple access schemes and their related parameters such as the number of codewords, length of codeword, spreading factor, max number of layers, need to be optimized based on the instant services and the link conditions. In the following, we provide two potential adaptive multiple access schemes in 5G.

In [21], the adaptive multiple access scheme is studied from EE-SE co-design perspective, taking the detection complexity into consideration. The SCMA and OFDMA schemes are taken as the candidate uplink multiple access schemes in the study. The problem is formulated to choose the optimal multiple access scheme and the related parameters simultaneously to maximize the EE under the total transmit power constraint, the quality of service (QoS) constraints and other specific requirements. The considered power consumption includes the transmit power consumption, the static circuit power consumption, and the SCMA decoding power consumption related which is proportional to the SCMA decoding complexity order $O(M^{d_f})$, where M is the constellation size, $d_f = \frac{N}{K}J$, K and N denotes the codeword length and non-zero entries in each codeword in SCMA, respectively, and $J = \binom{K}{N}$ is the maximum number of access users. **Fig. 9** shows the EE performance comparison of SCMA, OFDMA and the proposed link adaptation schemes with various cell radiuses. When the cell radius is small, the SCMA scheme has better EE performance; when the cell radius is large, the OFDMA scheme performs better than SCMA scheme. The reason is that the SCMA can access more users than OFDMA, and the increment of the number of access users per resource can improve the system EE when the cell radius is small since the user transmit power efficiency is large when the path loss is small. When the cell radius increases, the user transmit power efficiency decreases and the increment of the number of access users per resource will decrease the system EE. The adaptation scheme can obtain the overall good EE performance for all the cell radiuses (Fig. 9).

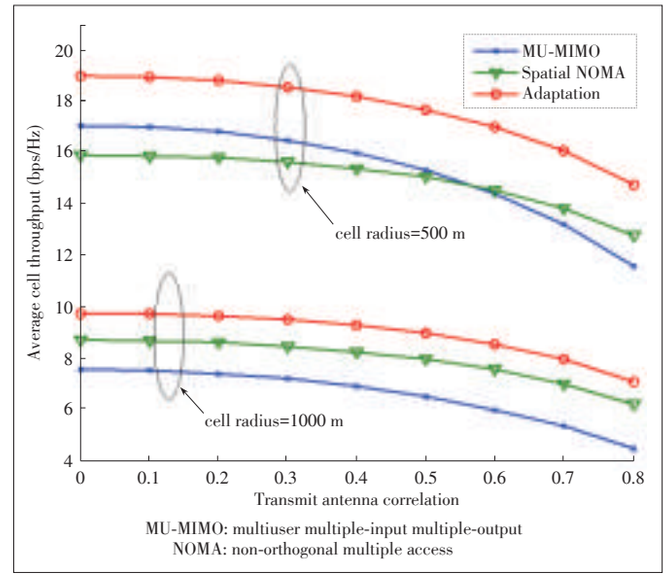


▲Figure 9. Average EE v.s. Cell Radiuses.

Another example of the adaptive multiple access is between the spatial NOMA (also known as MIMO NOMA) scheme and orthogonal the multiuser MIMO (MU - MIMO). **Fig. 10** shows

the concept of the orthogonal MU-MIMO and the spatial NOMA. In spatial NOMA, each two users can be served via one beam, and the interference between these two users will be large but can be cancelled via SIC decoding at the stronger user. This is unlike the orthogonal MU-MIMO precoding (e.g. zero-forcing MU-MIMO), in which there is no interference between users or it is usually small. Owing to this feature, when the channel has large correlations, spatial NOMA will have significantly higher throughput over orthogonal MU-MIMO since the MU-MIMO precoding needs to reduce the interference between users and will suffer large gain loss in that case. In addition, when the user channel gain difference is large, the spatial NOMA will also have better performance compared to orthogonal MU-MIMO due to the near-far effects. While in the high SNR regimes with low transmit correlation, the orthogonal MU-MIMO is preferred since it can approach the capacity bound of MIMO broadcast channel in high SNR regimes. Considering the time varying fading characteristics of MIMO channels and the random distribution feature of active users, in a multiuser scenario, the adaptation between orthogonal MU-MIMO and spatial NOMA is desired for both the cell average and cell edge throughput enhancement.

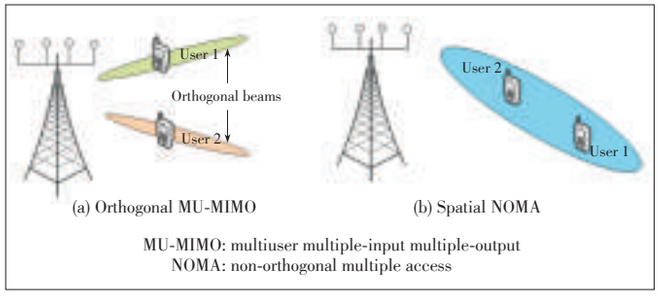
Fig. 11 shows the gain of the adaptation between orthogonal MU-MIMO based on zero-forcing scheme and spatial NOMA. When the transmit antenna correlation or the cell radius grows large, the spatial NOMA will have better performance than zero-forcing based MU-MIMO, and the adaptation gain will increase. The reason is that the higher transmit antenna correlation will lead to the higher probability that the user channels have high correlation, and the larger cell radium will lead to the higher probability that the user channels have large gain difference.



▲**Figure 11. Average cell throughput comparisons of various multiuser MIMO schemes (32 antenna base station at 6 GHz with four transceiver chains and four single antenna users)**

5 Conclusions

All the typical candidate NOMA solutions for 5G have different strength points and weakness points. None of them can surpass other schemes on all aspects. To fully exploit the advantages of these candidate technologies and traditional orthogonal multiple access solutions, a unified framework and a flexible multiple access schemes are required. Flexible switch among different NORMA schemes and the orthogonal multiple access technologies is expected to efficiently enhance the data rate and accommodate the necessary scalability for massive IoT connectivity and drastic reduction in access latency, and then to fully meet the diversified needs of 5G wireless communication systems. Some challenging problems need to be solved before NOMA schemes are put into use in 5G. In future, the impact of these candidate schemes on the existing systems, e.g. the grant free access procedure, reference signal, channel estimation and network assisted signaling, need to be carefully designed. What's more, the performance tradeoff of the code mapping manners in these schemes and their implementation complexity may need further evaluated. The adaptive mechanisms for part of these candidate schemes are also worth further study to meet the diversified requirements of 5G.



▲**Figure 10. Orthogonal MU-MIMO and Spatial NOMA.**

References

[1] FuTURE Mobile Communication Forum 5G SIG. (Nov. 2014). Rethink mobile communications for 2020+ [Online]. Available: http://www.future-forum.org/dl/141106/whitepaper.zip

[2] A. Benjebbovu, Y. Saito, Y. Kishiyama, et al., "Concept and practical considerations of non-orthogonal multiple access for future radio access," in *IEEE ISPACS 2013*, Naha, Japan, Nov. 2014, pp. 770–774. doi: 10.1109/ISPACS.2013.6704653.

[3] 3GPP, "Discussion on multiple access for new radio interface," TSG-RAN WG1 #84bis, Busan Korea, R1-162226, Apr. 11–15, 2016.

[4] H. Nikopour and H. Baligh, "Sparse code multiple access," in *IEEE PIMRC 2013*, London, England, Sept. 2013, pp. 332–336. doi: 10.1109/PIMRC.2013.6666156.

[5] 3GPP, "Candidate Solution for New Multiple Access", TSG -RAN WG1 #84bis, Busan, Korea, R1 -163383, Apr. 11–15, 2016.
[6] 3GPP, "Candidate NR multiple access schemes", TSG - RAN WG1 #84bis, Busan Korea, R1 -163510, Apr. 11–15, 2016.
[7] 3GPP, "Considerations on DL/UL multiple access for NR," TSG -RAN WG1 # 84bis, Busan Korea, R1 -162517, Apr. 11–15, 2016.
[8] 3GPP, "Considerations on DL/UL multiple access for NR," TSG -RAN WG1 # 84bis, Busan Korea, R1 -163992, Apr. 11–15, 2016.
[9] *Study on Downlink Multiuser Superposition Transmission (MUST) for LTE*, 3GPP TR36.859 v13.0.0, Jan. 2016.
[10] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non -orthogonal multiple access," *IEEE Transactions on Wireless Communications.*, to be published.
[11] Q. Sun, S. Han, C. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Communications Letters*, vol. 4, pp. 405–408, Aug 2015, doi: 10.1109/LWC.2015.2426709
[12] Q. Sun, Chih -Lin I, S. Han, et al., "Sum rate optimization for MIMO non -orthogonal multiple access systems," accepted by *IEEE WCNC*, 2015.
[13] S. Han, Chin -Lin I, Z. Xu, et al., "Energy efficiency and spectrum efficiency co -design: from NOMA to network NOMA," *IEEE MMTC E -Letter (Special Issue on 5G)*, vol. 9, no. 5, pp. 21–24, Sept. 2014.
[14] M. Taherzadeh, H. Nikopour, A. Bayesteh, et al., "SCMA codebook design," in *IEEE VTC Fall* 2014, Vancouver, Canada, Sept. 2013, pp. 1–5. doi: 10.1109/VTCFall.2014.6966170.
[15] K. Xiao, B. Xiao, S. Zhang, et al., "Simplified multiuser detection for SCMA with sum - product algorithm," in *IEEE WCSP* 2015, Nanjing, China, Oct. 2015, pp. 1–5. doi: 10.1109/WCSP.2015.7341328.
[16] Y. Wu, S. Zhang, Y. Chen, et al., "Iterative multiuser receiver in sparse code multiple access systems," in *Proc. IEEE ICC 2015*, London, England, Jun. 2015, pp. 2918–2923. doi: 10.1109/ICC.2015.7248770.
[17] A. Bayesteh, H. Nikopour, M. Taherzadeh, et al., "Low complexity techniques for SCMA detection," in *Proc. IEEE Globecom Workshops 2015*, San Diego, USA, Dec. 2015, pp. 1–6. doi: 10.1109/GLOCOMW.2015.7414184.
[18] H. Mu, Z. Ma, M. Alhaji, et al., "A fixed low complexity message pass algorithm detector for uplink SCMA system," *IEEE Wireless Communications Letters*, vol. 4. no. 6, pp. 585–588, 2015. doi: 10.1109/LWC.2015.2469668.
[19] Y. Du, B. Dong, Z. Chen, et al., "A fast convergence multiuser detection scheme for uplink SCMA systems," *IEEE Wireless Communications Letters*, to be published.
[20] S. Zhang, X. Xu, L. Lu, et al., "Sparse code multiple access: an energy efficient uplink approach for 5G wireless systems," in *Proc. IEEE Globecom 2014*, Austin, USA, Dec. 2014, pp. 4782–4787. doi: 10.1109/GLOCOM.2014.7037563.
[21] Q. Sun, Chih -Lin I, S. Han, et al., "Software defined air interface: a framework of 5G air interface," in *Proc. IEEE WCNC Workshop 2015*, New Orleans, USA, Mar. 2015, pp. 6–11. doi: 10.1109/WCNCW.2015.7122520.
[22] K. Au, L. Zhang, H. Nikopour, et al., "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE Globecom Workshop 2014*, Austin, USA, Dec. 2014, pp. 900–905. doi: 10.1109/GLOCOMW.2014.7063547.
[23] A. Bayesteh, E. Yi, H. Nikopour, et al., "Blind detection of SCMA for uplink Grant -Free Multiple Access," in *Proc. IEEE ISWCS 2014*, Barcelona, Spain, Aug. 2014, pp. 853–857. doi: 10.1109/ISWCS.2014.6933472.
[24] T. Liu, X. Li, and L. Qiu, "Capacity for downlink massive MIMO MU -SCM.A system," in *Proc. IEEE WCSP 2015*, Nanjing, China, Oct. 2015, pp. 1–5. doi: 10.1109/WCSP.2015.7341100.
[25] H. Nikopour, E. Yi, A. Bayesteh, et al., "SCMA for downlink multiple access of 5G wireless networks," in *Proc. IEEE Globalcom 2014*, Austin, USA, Dec. 2014, pp. 3940–3945. doi: 10.1109/GLOCOM.2014.7037423.
[26] U. Vilaipornsawai, H. Nikopour, and A. Bayesteh, "SCMA for open -loop joint transmission CoMP," in *Proc. IEEE VTC Fall 2015*, Boston, USA, Sept. 2015, pp. 1–5. doi: 10.1109/VTCFall.2015.7391126.
[27] Y. Li, X. Lei, P. Fan, et al., "An SCMA -based uplink inter -cell interference cancellation technique for 5G wireless systems," in *Proc. IEEE WCSP 2015*, Nanjing, China, Oct. 2015, pp. 1–5, doi: 10.1109/WCSP.2015.7341306
[28] A. Medra and T. N. Davidson, "Flexible codebook design for limited feedback systems via sequential smooth optimization on the grassmannian manifold," *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1305–1318, Mar. 2014. doi:10.1109/TSP.2014.2301137.

Manuscript received: 2016-06-30

Biographies

SUN Qi (sunqiyjy@chinamobile.com) received the BSE and PhD degree in information and communication engineering from Beijing University of Posts and Telecommunications, China in 2009 and 2014, respectively. After graduation, she joined the Green Communication Research Center of the China Mobile Research Institute. Her research focuses on 5G key technologies, including non -orthogonal multiple access, new waveforms, flexible duplex and UDN.

WANG Sen (wangsenyjy@chinamobile.com) received the PhD degree in information and communication engineering from Beijing University of Posts and Telecommunications, China in 2013. He joined the Green Communication Research Center of the China Mobile Research Institute after graduation. He is now working on the 5G key technologies and standardization. His research interests include massive MIMO, non -orthogonal multiple access, new waveforms and system level evaluation.

HAN Shuangfeng (hanshuangfeng@chinamobile.com) received his MS and PhD degrees in electrical engineering from Tsinghua University, China in 2002 and 2006 respectively. He joined Samsung Electronics as a senior engineer in 2006 working on MIMO, MultiBS MIMO etc. From 2012, he is a senior project manager in the Green Communication Research Center of the China Mobile Research Institute. His research interests are green 5G, massive MIMO, full duplex, NOMA and EE -SE codesign.

Chih-Lin I (icl@chinamobile.com) received her PhD degree in electrical engineering from Stanford University, USA. She has been working at multiple world -class companies and research institutes leading the R&D, including AT&T Bell Labs; Director of AT&T HQ, Director of ITRI Taiwan, and VPGD of ASTRI Hong Kong. She received the IEEE Trans. COM Stephen Rice Best Paper Award and is a winner of the CCCP National 1000 Talent Program. In 2011, she joined China Mobile as its Chief Scientist of wireless technologies, established the Green Communications Research Center, and launched the 5G Key Technologies R&D. She is spearheading major initiatives including 5G, C -RAN, high energy efficiency system architectures, technologies and devices; and green energy. She was an elected Board Member of IEEE ComSoc, Chair of the ComSoc Meetings and Conferences Board, and Founding Chair of the IEEE WCNC Steering Committee. She is currently an Executive Board Member of GreenTouch, a Network Operator Council Member of ETSI NFV, a Steering Board Member of WWRF, and a Scientific Advisory Board Member of Singapore NRF. Her current research interests center around "Green, Soft, and Open"

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 005
http://www.cnki.net/kcms/detail/34.1294.TN.20161011.1511.002.html, published online October 11, 2016

Multiple Access Rateless Network Coding for Machine-to-Machine Communications

JIAO Jian[1,2], Rana Abbas[2], LI Yonghui[2], and ZHANG Qinyu[1]
(1. Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China;
2. Center of Excellence in Telecommunications, University of Sydney, Sydney, NSW 2006, Australia)

Abstract

In this paper, we propose a novel multiple access rateless network coding scheme for machine-to-machine (M2M) communications. The presented scheme is capable of increasing transmission efficiency by reducing occupied time slots yet with high decoding success rates. Unlike existing state-of-the-art distributed rateless coding schemes, the proposed rateless network coding can dynamically recode by using simple yet effective XOR operations, which is suitable for M2M erasure networks. Simulation results and analysis demonstrate that the proposed scheme outperforms the existing distributed rateless network coding schemes in the scenario of M2M multicast network with heterogeneous erasure features.

Keywords

rateless network coding; multiple accesss; machine-to-machine communications (M2M)

1 Introduction

Machine-to-machine (M2M) communication system is expected to support a massive number of devices communicating with each other in a fully automated fashion with minimum or without human intervention [1]. Equipped with networked and real-time processing capabilities, these devices can implement a wide range of applications, such as intelligent transportation systems (ITS), healthcare monitoring, smart metering, energy management and smart grids.

M2M communication system is generally characterized by a massive number of machine-type communication (MTC) devices that have no/low mobility, low computational and storage capabilities, and low power budget [2]. Moreover, most MTC devices suffer from severe congestion and access delay in an M2M system with a large number of devices [3]–[5]. Therefore, the main motivation behind this paper is to propose a coding strategy that exploits the interference in the channel to increase data rates. Our work focuses on the cooperative joint network and coding strategy for MTC devices in multicast settings. These MTC devices disseminate messages to multiple receivers simultaneously with the help of relay nodes.

The rateless code, originally investigated in [6] for single source broadcasting in a single hop network, is deemed as a milestone for packet erasure codes. It can recover the original k information symbols from any $n = k + O(\sqrt{k}\ln^2(k/\theta))$ received coded symbols with the probability $1-\theta$ and the decoding cost of $O(k\ln(k/\theta))$ of operations, where θ is the allowable failure probability to recover the original message after n coded symbols have been received. In addition, the encoding and decoding process of the rateless code is complex, including logarithmic order for Luby Transform (LT) code and linear order for Raptor code. Furthermore, both LT and Raptor codes are able to provide practical capacity-achieving solutions, if their encoding degree distributions are sophisticated designed [7], [8].

The rateless code has been widely applied in cooperative communications [9]–[13]. In [9], the complexity, delay, and memory of different state-of-the-art rateless coding algorithms are analyzed for a multi-hop network. In [10], a superimposed on-the-fly recoding scheme is performed by each transport node in a multi-hop tree network, but it is difficult to implement due to the high decoding complexity. The first distributed LT (DLT) code is proposed in [11], and a new degree distribution, named deconvolved soliton distribution (DSD) is designed. However, all the source nodes and relays are assumed

This works was supported in part by Natural Scientific Research Innovation Foundation in Harbin Institute of Technology under Grant No. HIT. NSRIF 2017051, Shenzhen Basic Research Program under Grant Nos. JCYJ20150930150304185 and JCYJ2016 0328163327348, and National High Technology Research & Development Program of China under Grant No. 2014AA01A704.

to have exact information regarding the number of sources and encoding distributions to adapt relaying schemes. In [12], the authors utilized density evolution and linear programming frameworks to find an optimal combination at each relay node for any network architecture consisting of four sources. This manner can achieve the asymptotical error floor, but has intractable calculation complexity. A relaying protocol for Y-networks, namely Soliton-Like Rateless Coding (SLRC), is introduced in [13]. By enabling probabilistic forwarding and combining packets to reduce the overhead between relay and destination, the aggregate distribution at the destination can still maintain a near-ideal distribution, even if one source left the network. However, the lack of buffer utilization in SLRC relay limits the total encoding and decoding overheads.

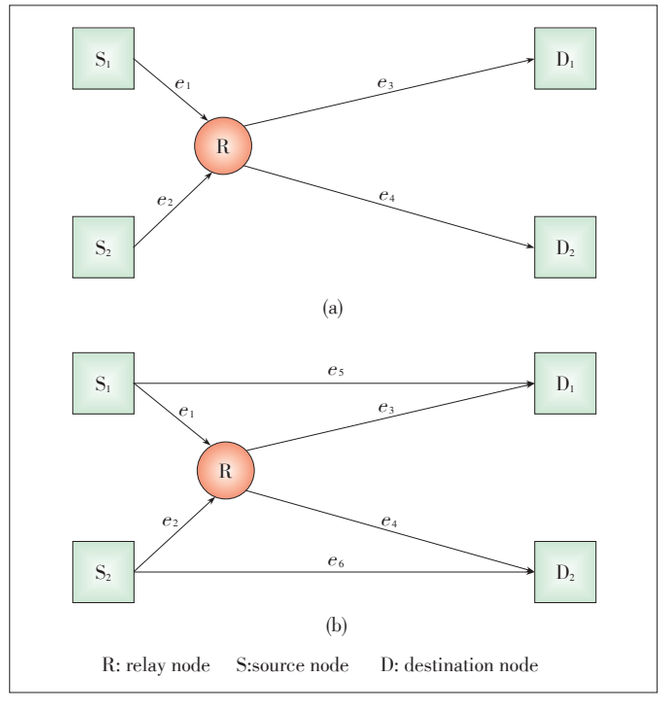
The destination cooperation in interference channels is another feature of M2M communication where one device can act as a relay for another device. A cooperative communication scheme for two mobile users is proposed in [4], which is potentially able to receive and decode each other's messages based on the signal-to-interference-plus-noise ratio (SINR). In [5], The received signal at the destination can be realized as a superposition of coded symbols sent from the relay, which is capacity approaching if an appropriate successive interference cancellation (SIC) is used for decoding [14].

In this paper, we propose an adaptive rateless network coding scheme in an M2M erasure network. First, an simple degree distribution is designed for rateless coding in all the source devices, and the collided devices transmit simultaneously. Then, an optimal relaying strategy is proposed to forward and combine the encoded packets with appropriate proportions, according to different erasure probabilities over the underlying edges. This is particularly suitable for M2M communications with strict power limitations, especially when the data size is small. By doing this, the total time slot of transmission is reduced obviously while the high decoding success rates are maintained. Moreover, we compared the current typical rateless coding relay schemes with our proposed scheme, with the aspects of the complexity and buffer memory. Simulation results show that the proposed rateless network coding scheme outperforms the existing distributed rateless coding schemes under various erasure probability scenarios.

2 Network Model and Rateless Code

2.1 Network Model

In [12] and [13], authors have introduced and optimized the applications of rateless coding in the Y-network model. We attempt to extend the Y-network model to a relay multicast model as shown in **Fig. 1a**. The relay R multicasts the data streams both to destination nodes D_1 and D_2 and guarantees the two source data to be recovered. Due to the special feature of rateless coding, the overheads at two destination nodes are apparently the same as the one in the Y-network and accordingly the DLT and SLRC algorithms are also appropriate for the network model in Fig. 1a.



▲**Figure 1. The proposed network models: (a) relay multicast model; (b) butterfly model.**

We also consider a butterfly network model (**Fig. 1b**), which has two sources, one relay and two destinations. Two direct edges are added to send two separate data streams form S_1 and S_2 respectively. The encoded packets should be processed (recoding) at R, and they can be converged at both D_1 and D_2 in the end. The model uses multicast from source to relay and directly from source to destination as well. This is its remarkable difference from the Y-network model. We define the edges in this model as e_1 to e_6. Each edge has an erasure probability ε_i, which is described as an independent-identical-distribution (i.i.d) Bernoulli variable. We assume that the packet size and the transmission rate of all the edges are equivalent (one time slot one packet). The rateless coding transmission scheme in this network model is described as follows.

1) Step 1: S_1 and S_2 generate the encoded packets with a rateless coding degree distribution [7];
2) Step 2: S_1 multicasts the encoded stream both to R and to D_1, and simultaneously, S_2 multicasts its stream to R and to D_2;
3) Step 3: R generates a new encoded stream by the relaying network coding (NC) scheme with the received encoded packets and multicasts them to D_1 and D_2 simultaneously;
4) Step 4: Once D_1 and D_2 receive enough encoded packets, they start to decode the encoded packets from the source and relay nodes to recover the two sources original packets;

5) Step 5: After successful decoding, D_1 (D_2) transmits a single acknowledgment (ACK) packet indicating the termination of the session.

2.2 Rateless Coding

Rateless coding is a modern forward error correction (FEC) technology. It protects from packet-loss, and can reduce the feedbacks for user acknowledgement due to rarely caring about the erasure probability of the channel. In a single-hop scenario, the original packets are defined as k data symbols, and the encoded packets required by the decoder are defined as n encoded symbols. Furthermore, the overhead is defined as $(n-k)/k$ and the number of encoded symbols sent by the sender is defined as N. Therefore, the expected code rate at the sender is conveyed as $R=\frac{k}{N}=\frac{(1-\varepsilon)k}{n}$, where ε is the erasure probability of the channel. On the other hand, the encoding and decoding complexity of rateless codes is very low, which are expressed as $O(\ln k)$ for LT codes and $O(k)$ for Raptor codes.

As an example of rateless coding, the LT code uses robust soliton distribution (RSD) to achieve the erasure channel capacity under the single hop network. The coding degree distribution is a key element for the successful recovery of data symbols. For a parameter δ and the length k RSD, $\mu(k)$ is defined as:

$$\mu(i)=\frac{\rho(i)+\tau(i)}{\beta},\ for\ i=1,\cdots,k, \tag{1}$$

where

$$\rho(i)=\begin{cases}1/i, & for\ i=1\\ 1/i(i-1), & for\ i=2,\cdots,k\end{cases}, \tag{2}$$

$$\tau(i)=\begin{cases}S/ik, & for\ i=1,\cdots,\lfloor k/S\rfloor-1\\ S\ln(S/\delta)/k, & for\ i=\lfloor k/S\rfloor\\ 0, & for\ i=\lfloor k/S\rfloor+1,\cdots,k\end{cases}, \tag{3}$$

$$\beta=\sum_{i=1}^{k}\mu(i)+\tau(i). \tag{4}$$

S is the average number of degree-one symbols, namely ripple size, which is defined by $S=c\ln(k/\delta)\sqrt{k}$, where $c>0$.

It is worth noting that the LT decoder performs the BP algorithm with the prior knowledge of the degree and associated neighbors. Given a block of encoded symbols, the decoder recursively decodes the data symbols from the bipartite graph connecting the information and encoded symbols. The BP algorithm starts from degree-one symbols, by removing their contributions from the graph in order to produce a smaller graph with another set of degree-one encoded symbols. Then, the new degree-one encoded symbols of this smaller graph are removed again, and iteratively the process continues to recover all data symbols, as described in [7] and [8].

3 Analysis of Relaying Schemes

As the rateless code is used in multi-source relay network, the erasure probability of different paths (multicast and unicast) may influence the relaying strategy and the corresponding performance with NC. Specifically, the relay R may receive no packet from S_1 or S_2 in a time slot due to packet loss in multi-source relay network. Hence, it is an interesting and significant topic to select proper rateless coding algorithms based on NC and relaying strategy for efficient transmissions on lossy network. In this section, we try to consider the conventional methods in Y-networks and butterfly networks, and compare their decoding performance at the destination nodes. Moreover, we propose a new optimized -NC scheme to trade off the decoding performance by selecting proper forwarding and combining probabilities.

3.1 Comparison of Typical Relaying Schemes

We assume that the number of original packets is k and the number of encoded packets generated is N at both the source nodes. In two fixed time slots, the destination nodes D_1/D_2 of butterfly network can receive one encoded packet from S_1/S_2 in the first slot, and then receive one from the relay R in the second slot. It is a limited condition that D_1 and D_2 only receive the maximum $2N$ packets when N encoded packets are sent from the source, if and only if all the edges are lossless. D_1 and D_2 use the BP decoding algorithm to decode the compilations of two encoded streams after $2N$ slots to recover $2k$ original packets, respectively. There are the following four typical relaying schemes in this butterfly network model:

1) Store-and-Forward (SF)

The relay R immediately forwards the packets to the next hop as soon as it receives packets. If two packets arrive simultaneously, R randomly forwards one of them and stores another into the buffer. If the relay R receives no packet, it waits for the next slot. Due to the uncertain storage of packets, this scheme may easily make congestion on R.

2) DLT

With S1 (S2) using DSD, R performs random decision protocol in [11] to combine and forward two received packets. Once the erasure event occurs at one of the edges between sources and relay, R directly forwards another received packet. If no packets arrive, the relay waits for the next slot. These waiting slots at relay lead to low efficiency due to a serious waste of sources. By using considerable low-degree encoded packets, this scheme could scarcely cover all the original packets of two sources, despite of its simple encoding complexity.

3) SLRC

With S_1 (S_2) using RSD, R uses the SLRC relaying scheme to operate the two encoded packets. It forwards most of the low degree packets (degree-one or degree-two) directly and com-

bines other high degree packets, in order to assure the aggregate distribution at destination nodes to be soliton-like. With the SLRC, the transmission time slot is not wasted if the packet on e_1 or e_2 is dropped, by R's choosing two packets from the buffers to combine and forward. Compared with the DLT scheme, this SLRC scheme obtains more gains.

4) eXclusive-OR (XOR)

This scheme is based on the simple eXclusive-OR (XOR) operation of classic NC in the butterfly model. The relay R combines the two received packets into one new packet. If only one packet arrives, the relay R sends the received packet directly, and if no packet arrives, R waits for the next time slot.

Since the relay carries the two received packets by forwarding and combining operations, the recoding complexity comes almost from XOR operation which depends mainly on the erasure probability of the edges. In addition, the relay requires two buffers to store the packets from different sources. Therefore, we compare the four schemes (**Table 1**). We can find that SF has the least recoding complexity, and the complexity of SLRC and DLT depends significantly on their forwarding probability. On the other hand, DLT and XOR need the smallest buffer size of only one packet for each source, while SLRC must store all the packets not forwarded for the next operation.

3.2 Analysis of Decoding Matrix

In this section, we make an intuitive analysis on recovery performance at the destination nodes by the BP decoding matrix, in which the BP decoding algorithm could be described as one unitization process, with the columns denoting the encoded symbols and the rows denoting the original data symbols.

We define A and B as the encoding matrix at S_1 and S_2 respectively, and denote the dimension of the matrix as the subscripts. The decoding matrices at D_1 and D_2 are the aggregation of the forwarding and combining sub-matrices A and B, with $2k$ rows due to the number of data symbols from two sources. Since the lost packets have been removed from the matrix, the number of columns reveals the received encoded symbols by destination nodes exactly.

For the SF scheme, the decoding matrices at D_1 and D_2 are defined as:

$$F^{D_1}_{2k\times[(1-\varepsilon_5)N+[(1-\varepsilon_1)+(1-\varepsilon_2)](1-\varepsilon_3)N/2]} = \begin{bmatrix} A^{S_1}_{k\times(1-\varepsilon_5)N} & A^{S_1}_{k\times(1-\varepsilon_1)(1-\varepsilon_3)N/2} & 0 \\ 0 & 0 & B^{S_2}_{k\times(1-\varepsilon_2)(1-\varepsilon_3)N/2} \end{bmatrix}, \quad (5)$$

$$F^{D_2}_{2k\times[(1-\varepsilon_6)N+[(1-\varepsilon_1)+(1-\varepsilon_2)](1-\varepsilon_4)N/2]} = \begin{bmatrix} 0 & A^{S_1}_{k\times(1-\varepsilon_1)(1-\varepsilon_4)N/2} & 0 \\ B^{S_2}_{k\times(1-\varepsilon_6)N} & 0 & B^{S_2}_{k\times(1-\varepsilon_2)(1-\varepsilon_4)N/2} \end{bmatrix}. \quad (6)$$

From the matrices F^{D_1} and F^{D_2} in (5) and (6), we know that the SF scheme is only appropriate for unicast like source to destination, since the dimensions of sub-matrices for A and B are extremely unequal. The lack of combination operations makes the destination nodes unable to recover the whole data symbols. Therefore, this scheme is inefficient for multicast in the butterfly network model.

For the SLRC scheme, as an optimized DLT, we only present its decoding matrices that are defined as:

$$H^{D_1}_{2k\times(1-\varepsilon_5+1-\varepsilon_3)N} = \begin{bmatrix} A^{S_1}_{k\times(1-\varepsilon_5)N} & A^{S_1}_{k\times\nu_1(1-\varepsilon_3)N} & 0 & A^{S_1}_{k\times(1-\sum_{i=1}^{2}\nu_i)(1-\varepsilon_3)N} \\ 0 & 0 & B^{S_2}_{k\times\nu_2(1-\varepsilon_3)N} & B^{S_2}_{k\times(1-\sum_{i=1}^{2}\nu_i)(1-\varepsilon_3)N} \end{bmatrix}, \quad (7)$$

$$H^{D_2}_{2k\times(1-\varepsilon_6+1-\varepsilon_4)N} = \begin{bmatrix} 0 & A^{S_1}_{k\times\nu_1(1-\varepsilon_4)N} & 0 & A^{S_1}_{k\times(1-\sum_{i=1}^{2}\nu_i)(1-\varepsilon_4)N} \\ B^{S_2}_{k\times(1-\varepsilon_6)N} & 0 & B^{S_2}_{k\times\nu_2(1-\varepsilon_4)N} & B^{S_2}_{k\times(1-\sum_{i=1}^{2}\nu_i)(1-\varepsilon_4)N} \end{bmatrix}, \quad (8)$$

where ν_i (i=1, 2) is the probability distribution of the relay forwarding packets from S_1 and S_2, while $\tilde{\nu}_i = 1-\nu_i$ is the distribution of the packets into the buffer.

For the XOR Scheme, the decoding matrices are defined as:

$$G^{D_1}_{2k\times[(1-\varepsilon_5)+Max\{(1-\varepsilon_1),(1-\varepsilon_2)\}(1-\varepsilon_3)]N} = \begin{bmatrix} A^{S_1}_{k\times(1-\varepsilon_5)N} & A^{S_1}_{k\times(1-\varepsilon_1)(1-\varepsilon_3)N} \\ 0 & B^{S_2}_{k\times(1-\varepsilon_2)(1-\varepsilon_3)N} \end{bmatrix}, \quad (9)$$

$$G^{D_2}_{2k\times[(1-\varepsilon_6)+Max\{(1-\varepsilon_1),(1-\varepsilon_2)\}(1-\varepsilon_4)]N} = \begin{bmatrix} 0 & A^{S_1}_{k\times(1-\varepsilon_1)(1-\varepsilon_4)N} \\ B^{S_2}_{k\times(1-\varepsilon_6)N} & B^{S_2}_{k\times(1-\varepsilon_2)(1-\varepsilon_4)N} \end{bmatrix}. \quad (10)$$

In (9) and (10), G can be segmented into four sub-matrices:

Table 1. The complexity and buffer for the four schemes

	Relay recoding complexity	Buffer size of relay
SF	0	$N/2$ packets for each source
DLT	$(1-\varepsilon_1)(1-\varepsilon_2)(1-\gamma_i)N$	One packet for each source
SLRC	$(1-\sum_{i=1}^{2}\nu_i)N$	$(1-\nu_i)N$ packets for each source
XOR	$(1-\varepsilon_1)(1-\varepsilon_2)N$	One packet for each source

* γ_i and ν_i are the distributions of the no-forwarding packets for DLT and SLRC

DLT: distributed Luby transform
SF: store-and-forward
SLRC: soliton-like rateless coding
XOR: eXclusive-OR

the two parts in the left are the forwarding matrix A or B, and the two right parts are the combined matrices. Since the combined symbols of this scheme may be lack of degree-one and degree-two packets, it needs enough columns of single sub-matrices A or B on the edge e_5 or e_6 in Fig. 1 to start the BP decoder. In the decoding process, only if all the data symbols of A have been recovered, B can begin to decode.

By comparing these decoding matrices, we can find that the forwarding matrix A or B has occupied such numerous columns in H, as $A^{S_1}_{k\times\nu_1(1-\varepsilon_3)N}$ or $B^{S_2}_{k\times\nu_2(1-\varepsilon_3)N}$. The combination part of encoded symbols as A and B $\begin{bmatrix} A^{S_1}_{k\times(1-\sum_{i=1}^{2}\nu_i)(1-\varepsilon_3)N} \\ B^{S_2}_{k\times(1-\sum_{i=1}^{2}\nu_i)(1-\varepsilon_3)N} \end{bmatrix}$ in the SLRC scheme has much less columns than that of $\begin{bmatrix} A^{S_1}_{k\times(1-\varepsilon_1)(1-\varepsilon_3)N} \\ B^{S_2}_{k\times(1-\varepsilon_2)(1-\varepsilon_3)N} \end{bmatrix}$ in the XOR scheme. SLRC can still recode new packets in the relay to transmit, even if no packets received due to enough large ε_1 and ε_2. However, it has not fully utilized the packets directly from D_1 and D_2 on e_5 and e_6. Note that, only N encoded packets at most could be sent by the sources and relay, which constrains the required time slots to be only $2N$ in the network model, given one packet at each time slot. It renders that the decoding matrix H has many single forwarding columns in A or B which obviously reduces the relevance of encoded symbols from S_1 and S_2. As a result, the large proportion of forwarding packets by the relay cannot give much help to improve the BP decoding performance, especially on the condition of the relatively small erasure probability ε_5 and ε_6.

On the other hand, when ε_5 and ε_6 become larger, the decoding performance is mostly decided by the proportion of forwarding the single packets and XOR combination of two sources' packets in the relay. In the XOR scheme, the number of recovery data symbols would decline very fast due to the lack of low degree encoded symbols for BP decoding. Besides, the XOR operations would be blocked and degraded since two separate packets from two sources could hardly arrive at the relay simultaneously in the large ε_5 and ε_6.

3.3 Proposed Optimized NC Scheme

On the basis of above analysis, we have found that the relaying schemes should forward the low-degree packets for starting BP decoder, by taking into consideration the erasure probabilities of the direct edges e_1 and e_2. On the other side, the relay also needs to remain the enough proportion of combinations of the packets in the buffers, in order to prevent from no received packets from S_1 and S_2 simultaneously. Accordingly, we propose a novel NC scheme with a self-adjusted forwarding probability associated with the variations of ε_5 and ε_6. The basic rule of the proposed scheme is as follows: if ε_1 and ε_2 increase, the relay immediately forwards more low-degree packets; if ε_5 and ε_6 decrease, the relay combines more packets.

The proposed algorithm, named Opt-NC scheme, has the comparable complexity and buffer requirements with the SLRC scheme. We denote the forwarding probabilities λ and θ for encoded packets from S_1 and S_2, which are predetermined to be equivalent to the erasure probabilities ε_5 and ε_6, respectively. Algorithm 1 shows the steps of Opt-NC algorithm.

Algorithm 1: Opt-NC Scheme (at one time slot)

p_1: received encoded packets from S_1;
p_2: received encoded packets from S_2;
d_1: degree of p_1;
d_2: degree of p_2;
λ: forwarding probability of the low degree packets from S_1;
θ: forwarding probability of the low degree packets from S_2;
a=rand();
b=rand();
if d_1=1∨2 and d_2=1∨2 and $a<\lambda$ and $b<\theta$
    forward p_1 or p_2 with equal probability;
    put another packet into the buffer of another source;
else if d_1=1∨2 and $a<\lambda$
    forward p_1 and put p_2 into the buffer of S_2;
else if d_2=1∨2 and $b<\theta$
    forward p_2 and put p_1 into the buffer of S_1;
else
    put the packets received into the buffers respectively;
    p_{b1}: random choose one packet in the buffer of S_1;
    p_{b2}: random choose one packet in the buffer of S_2;
    $p_{XOR}=p_{b1}$XOR p_{b2};
    forward p_{XOR};
end if

* ∨ means logical operator of OR

4 Simulation Results and Discussion

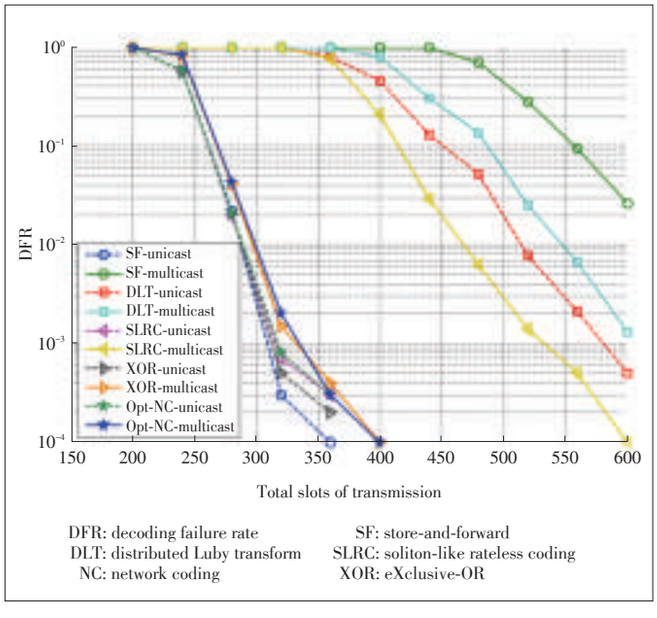
We analyze the performance of the above five algorithms in a butterfly network coding system as Fig. 1b. The encoding degree distribution is selected to be RSD with parameters δ=0.05, c=0.03. The number of data symbols k=100, and the number of encoded symbols from S_1 and S_2 is indicated to be the same as N. We emulate the encoding and decoding procedure using Monte Carlo experiments with 10,000 times. The ratio between the statistics of decoding failure times and total experiment times is defined as decoding failure rate (DFR). In this work, the lowest displayable DFR in our simulation is 10^{-4}. Given the time slots and erasure probabilities of edges, the lower DFR of relaying schemes means outstanding decoding performance. We give the unicast performance and multicast perfor-

mance respectively to discuss the influences between the single source and double sources.

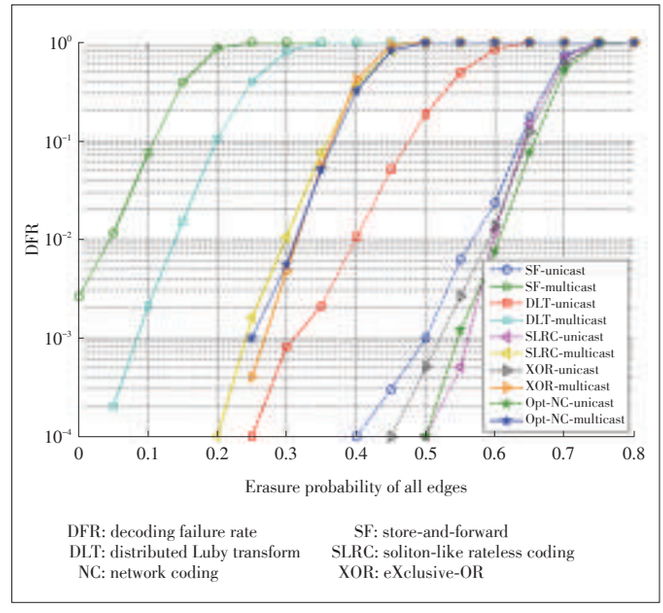
Fig. 2 gives the performance of five schemes in the erasure free network model. We can find that the Opt-NC scheme approaches the unicast and multicast curves of the XOR scheme, which obtain the lowest DFR of 10^{-4} at 400 time slots. The other schemes need at least 600 time slots to recover two sources' data, due to their inefficiency of buffer utilization at the relay. This simulation result proves that the Opt-NC scheme can apply as a typical NC scheme to get the remarkable multicast throughput gains in lossless network data transmission.

With vary erasure probabilities from 0 to 0.8 of all edges among ε_1 to ε_6, and the code length N=360, the decoding performance of five schemes is shown in **Fig. 3** . The unicast results of the schemes reveal the better performance than the multicast results, since the erasure probabilities of ε_1, ε_2, ε_3, ε_4 make the combination operations at the relay inappropriate. It is noted that the SLRC, XOR and Opt-NC schemes have similar multicast decoding performance, with the DFR lower than 10^{-4} and the erasure probability of 0.2. The simulation results indicate that our adaptive Opt-NC scheme integrates the advantages of SLRC and XOR, which also reveals outstanding decoding performance in lossy network.

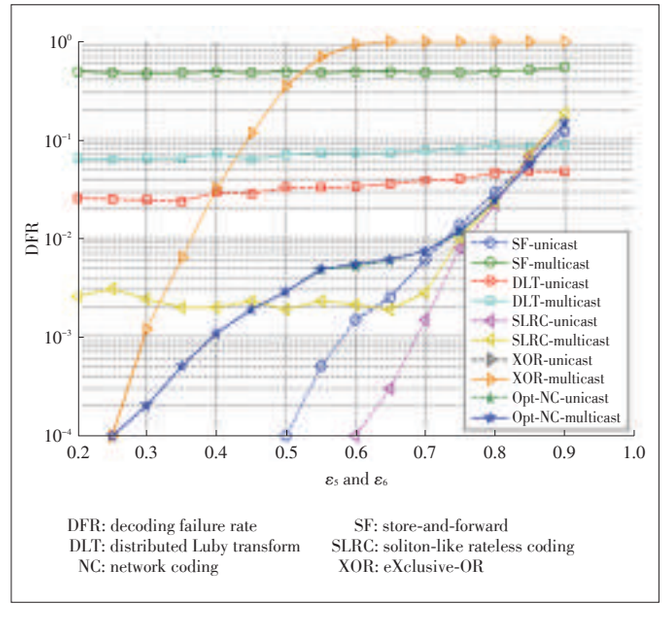
Fig. 4 shows the DFRs of five schemes with the encoded packets of N=250, ε_1 to ε_4 null, and ε_5 and ε_6 from 0.2 to 0.9. Since the multicasts of the SF and DLT are both restricted by the limited number of encoded packets from source, their decoding performance maintains at an inferior level in spite of ε_5 and ε_6 increasing. On the other side, the multicast DFR performance of XOR and that of Opt-NC are almost consistency with their unicasts. If the erasure probabilities of e_5 and e_6 are lower than 0.25, the Opt-NC scheme can get the similar DFR with



▲**Figure 2. All the edges are erasure free, k=100.**



▲**Figure 3. All the edges have the same erasure probability, k=100, N=360.**



▲**Figure 4. Edges e_5 and e_6 are lossy while other edges are lossless, k=100, N=250.**

that of the XOR scheme. If ε_5 and ε_6 increase from 0.25 to 0.9, the Opt-NC scheme outperforms the XOR scheme by its dynamic property. In addition, compared to the SLRC, the Opt-NC scheme also has a better performance as ε_5 and ε_6 are both lower than 0.45. However, the SLRC gives a lower decoding failure rate as the ε_5 and ε_6 are both in a range of 0.45 to 0.8. Once the erasure probability increases higher than 0.8 (the edges e_5 and e_6 are almost interrupted), the Opt-NC scheme approaches SLRC-multicast with a higher efficiency. In a word,

the proposed relaying scheme is an adaptive method to compromise the decoding performance of XOR and SLRC.

5 Conclusions

In this paper, we have studied rateless network coding applied in machine-to-machine communications for multiple access applications. A novel dynamic relaying scheme Opt-NC was proposed that exploits the forwarding and combining operations to obtain an enhanced decoding performance of the decoder at the destination nodes. The Opt-NC scheme has adaptive capability of responding to the vary erasure probability of direct edges. The simulation results show that the proposed relay scheme performs close to the optimal XOR scheme in lossless and lossy network, respectively. Furthermore, the Opt-NC scheme can be used in the physical layer by incorporating the XOR operation and superposition practical modulations.

References

[1] M. Shirvanimoghaddam, Y. Li, and M. Dohler, "Probabilistic rateless multiple access for machine-to-machine communication," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6815–6826, Dec. 2015.
[2] K. Zheng, S. Ou, J. Alonso-Zarate, et al., "Challenges of massive access in highly dense LTE-Advanced networks with machine-to-machine communications," *IEEE Wireless Communications Magazine*, vol. 21, no. 3, pp. 12–18, Jun. 2014.
[3] G. Durisi, T. Koch, and P. Popovski. (2015). "Towards massive, ultra-reliable, and low-latency wireless communication with short packets," *CoRR* [Online]. Available: https://arxiv.org/abs/1504. 06526
[4] B. W. Khoueiry and M. R. Soleymani, "A novel destination cooperation scheme in interference channels", in *Proc. IEEE 80th Vehicular Technology Conference (VTC 2014 - Fall)*, Vancouver, Canada, Sept., 2014. doi: 10.1109/VTCFall.2014.6965826.
[5] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson. (2016). "Massive multiple access based on superposition raptor codes for M2M communications," *CoRR* [Online]. Available: https://arxiv.org/abs/1602.05671
[6] J. Byers, M. Luby, M. Mitzenmacher, and A. Rege, "A digital fountain approach to reliable distribution of bulk data," in *Proc. ACM SIGCOMM'98*, Vancouver, Canada, Jan. 1998, pp. 56–67. doi:10.1145/285237.285258.
[7] M. Luby, "LT Codes," in *Proc. IEEE Symposium on the Foundations of Computer Science (STOC)*, Vancouver, Canada, 2002, pp. 271–280.
[8] A. Shokrollahi, "Raptor Codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006. doi: 10.1109/TIT.2006.874390.
[9] P. Pakzad, C. Fragouli, and A. Shokrollahi, "Coding schemes for line networks," in *Proc. ISIT*, Adelaide, Germany, 2005, pp. 1853–1857.
[10] R. Gummadi and R. S. Sreenivas, "Relaying a fountain code across multiple nodes," in *Proc. SIGCOMM'08*, Seattle, USA, 2008, pp. 483–484.
[11] S. Puducheri, J. Kliewer, and T. E. Fuja, "The design and performance of distributed LT codes," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3740–3754, Oct. 2007.
[12] D. Sejdinovic, R. Piechocki, and A. Doufexi, "AND-OR tree analysis of distributed LT codes," in *Proc. ITW*, Volos, Greece, 2009, pp. 261–265.
[13] A. Liau, S. Yousefi, and I. Kim, "Binary soliton-like rateless coding for the Y-network," *IEEE Transactions on Communications*, vol. 59, no. 12, pp. 3217–3222, Dec. 2011.
[14] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "On SINR-based random multiple access using codes on graph," in *IEEE Global Communications Conference (GLOBECOM)*, San Diego, USA, 2015. doi: 10.1109/GLOCOM.2015.7417013.

Manuscritp received: 2016-07-14

Biographies

JIAO Jian (jiaojian@hitsz.edu.cn) received his PhD degree in communication engineering from Harbin Institute of Technology (HIT) in 2011. He received his BS degree in electrical engineering from Harbin Engineering University in 2005, and his MASc degree in information and communication engineering from HIT Shenzhen Graduate School in 2007. He is an assistant research fellow in the Department of Electrical and Information Engineering of HIT Shenzhen Graduate School. His current interests include deep space communications, networking and channel coding.

Rana Abbas (rana.abbas@sydney.edu.au) is currently a PhD student at the Centre of Excellence in Telecommunications, School of Electrical And Information Engineering, The University Sydney, Australia, where she is a recipient of the Australian Postgraduate Awards scholarship and the Norman 1 Price scholarship. She received her bachelor's degree in electrical engineering from The University of Balamand, Lebanon in 2012 and her master's degree in electrical engineering from The University of Sydney, Australia in 2013. Her research interests include error control codes, machine-to-machine communications, random multiple access, and cooperative networks.

LI Yonghui (yonghui.li@sydney.edu.au) received his PhD degree in 2002 from Beijing University of Aeronautics and Astronautics. From 1999 to 2003 he was affiliated with Linkair Communication Inc., where he held the position of project manager with responsibility for the design of physical layer solutions for LAS-CDMA system. Since 2003 he has been with the Centre of Excellence in Telecommunications, the University of Sydney, Australia. He is now an associate professor at the School of Electrical and Information Engineering, University of Sydney. He is the recipient of the Australian Queen Elizabeth II Fellowship in 2008 and the Australian Future Fellowship in 2012. His current research interests are in the area of wireless communications, with a particular focus on MIMO, cooperative communications, coding techniques, and wireless sensor networks. He holds a number of patents granted and pending in these fields. He is an executive editor for *European Transactions on Telecommunications* (ETT). He received best paper awards at the IEEE International Conference on Communications (ICC) 2014 and the IEEE Wireless Days Conferences (WD) 2014.

ZHANG Qinyu (zqy@hit.edu.cn) received his bachelor's degree in communication engineering from Harbin Institute of Technology (HIT) in 1994, and PhD degree in biomedical and electrical engineering from the University of Tokushima, Japan, in 2003. From 1999 to 2003, he was an assistant professor with the University of Tokushima. From 2003 to 2005, he was an associate professor with the Shenzhen Graduate School, HIT, and was the founding director of the Communication Engineering Research Center with the School of Electronic and Information Engineering. Since 2005, he has been a full professor, and serves as the dean of the EIE School. He is on the Editorial Board of some academic journals, such as *The Journal on Communications*, *KSII Transactions on Internet and Information Systems*, and *Science China: Information Sciences*. He was the TPC Co-Chair of the IEEE/CIC ICCC'15, the Symposium Co-Chair of the IEEE VTC'16 Spring, an Associate Chair for Finance of ICMMT'12, and the Symposium Co-Chair of CHINACOM'11. He has been a TPC Member for INFOCOM, ICC, GLOBECOM, WCNC, and other flagship conferences in communications. He was the Founding Chair of the IEEE Communications Society Shenzhen Chapter. He has received the National Science Fund for Distinguished Young Scholars, the Young and Middle-Aged Leading Scientist of China, and the Chinese New Century Excellent Talents in University, and obtained three scientific and technological awards from governments. His research interests include aerospace communications and networks, wireless communications and networks, cognitive radios, signal processing, and biomedical engineering.

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 006
http://www.cnki.net/kcms/detail/34.1294.TN.20161024.1001.002.html, published online October 24, 2016

Multiple Access Technologies for Cellular M2M Communications

Mahyar Shirvanimoghaddam and Sarah J. Johnson
(School of Electrical Engineering and Computer Science, The University of Newcastle, NSW 2308, Australia)

Abstract

This paper reviews the multiple access techniques for machine-to-machine (M2M) communications in future wireless cellular networks. M2M communications aims at providing the communication infrastructure for the emerging Internet of Things (IoT), which will revolutionize the way we interact with our surrounding physical environment. We provide an overview of the multiple access strategies and explain their limitations when used for M2M communications. We show the throughput efficiency of different multiple access techniques when used in coordinated and uncoordinated scenarios. Non-orthogonal multiple access (NOMA) is also shown to support a larger number of devices compared to orthogonal multiple access techniques, especially in uncoordinated scenarios. We also detail the issues and challenges of different multiple access techniques to be used for M2M applications in cellular networks.

Keywords

Internet of Things (IoT); massive access; machine-to-machine (M2M) communications; multiple access

1 Introduction

Machine-to-machine (M2M) communications is expected to become an integral part of cellular networks in the near future. In M2M communications a large number of multi-role devices, such as sensors and actuators, wish to communicate with each other and with the underlying data transport infrastructure. To enable such massive communication in wireless networks, major shifts from current protocols and designs are necessary [1]. Current wireless networks that have been mainly designed and engineered for human-based applications, such as voice, video, and data, cannot be used for M2M communications due to the different nature of their traffic and service requirements [2]. These differences have posed many questions and challenges in the communication society, in both industry and research sectors.

M2M communications aims at providing the communication infrastructure for emerging Internet of Things (IoT) and involves the enabling of seamless information exchange between autonomous devices without any human intervention. M2M devices can be either stationary, such as smart meters, or mobile, such as fleet management devices, and they can connect to the network infrastructure using either wired or wireless links. Key challenges of massive M2M communications can be listed as follows [3]:

1) Device cost: For the mass deployment of M2M communications, low cost devices are necessary for most use cases.
2) Battery life: Most M2M devices are battery operated and replacing batteries is not practical for many applications.
3) Coverage: Deep indoor and regional connectivity is a requirement for many applications.
4) Scalability: Network capacity must be easily scaled to handle a large number of devices forecasted to arise in the near future.
5) Diversity: Cellular systems must be able to support diverse service requirements for different use cases, ranging from static sensor networks to tracking systems.

The wired solutions include cable, xDSL, and optical fiber, and can provide high reliability, high data rate, short delay, and high security. However, they are cost ineffective and do not support mobility and scalability; therefore, not appropriate for M2M applications [3]. On the other hand, Wireless capillary (i.e., short range) solutions, such as WLAN and ZigBee, can provide low cost infrastructure and scalability for most M2M applications, but they suffer from small coverage, low data rate, weak security, and severe interference. Wireless cellular, i.e., GSM, GPRS, 3G, LTE-A, WiMAX, etc., however offers excellent coverage, mobility and scalability support, and good security, and the fact that the infrastructure already exists

makes it a promising solution for M2M communications [3]. Therefore, our focus in this paper is on wireless cellular solutions for M2M communications.

The mobile industry is standardizing several low power technologies, such as extended coverage GSM (EC-GSM), LTE for machine-type communication (LTE-M), and narrow band IoT (NB-IoT). Since GSM is still the dominant mobile technology in many markets, it is expected to play a key role in the IoT due to its global coverage and cost advantages. EC-GSM enables coverage improvements of up to 20 dB with respect to GPRS on the 900 MHz band [4]. A combined capacity of up to 50,000 devices per cell on a single transceiver has been achieved by defining new control and data channels mapped over legacy GSM. LTE-M brings new power saving functionality suitable for serving a variety of IoT applications, which extend battery life to 10 years or more. NB-IoT is a self contained carrier that can be deployed with a system bandwidth of 200 kHz. These initiatives were undertaken in 3GPP Release 13 for M2M specific applications [3].

Despite all these efforts, further improvements are required in the way that devices communicate with the base station to support a large number of devices and not jeopardizing the human-based communication quality. The multiple access (MA) techniques have been identified as a key area where improvements for M2M communications are needed. The fact that the radio access strategy in LTE is still based on random access mechanisms turns it into a potential bottleneck for the performance of cellular networks when the number of M2M devices grows [5]. Moreover, radio resources are orthogonally allocated to the users/devices in the current LTE standards, which is not effective for M2M communications when the number of devices goes very large, due to the limited number of radio resources [6].

In this paper, we consider several multiple access technologies and show their performance in coordinated and uncoordinated scenarios. Overall, coordinated strategies outperform uncoordinated ones as in coordinated strategies the base station can optimally allocate the radio resources between the devices and support a larger number of devices. We also show that the non-orthogonal multiple access (NOMA) scheme achieves the highest throughput in both coordinated and uncoordinated strategies, whereas frequency division multiple access (FDMA) has comparable performance in coordinated scenarios. This suggests that FDMA can be effectively used in coordinated scenarios to achieve maximum throughput (this has been considered by 3GPP for M2M communications in the NB-IoT solution), while in uncoordinated scenarios, NOMA strategies must be considered to effectively support a large number of devices and use the available radio resources in an efficient manner.

The remainder of the paper is organized as follows. Section 2 represents the unique characteristics of M2M communications and its challenges in cellular networks. In Section 3, we provide an overview on different multiple access technologies. Coordinated and uncoordinated MA techniques are represented in Section 4 and 5, respectively, where we characterize their maximum achievable throughput. Practical issues for implementing MA techniques for M2M communications are presented in Section 6. Finally, Section 7 concludes the paper.

2 M2M Communications: Characteristics and Challenges

Until recently, cellular systems have been designed and engineered for human based applications, such as voice, video, and data, with a higher demand on downlink. M2M communications however has different traffic characteristics that include small and infrequent data generated from a very large number of devices, which imposes a higher traffic volume on the uplink. In addition, M2M applications have very diverse service requirements. For instance, in alarm signal applications, a small-size message must be delivered to the base station (BS) within 10 ms, while in other applications, such as smart metering, the delay of up to several hours or even a day is tolerable [7].

Due to limited radio resources and the large number of devices involved in M2M communications, wireless networks should minimize the time wasted due to collisions or exchanging control messages. The throughput must be large enough to support a large number of devices. Control overhead must be minimized as the payload data in many M2M applications is of small size and the control overhead of conventional approaches in current cellular systems results in an inefficient M2M communications [8]. In fact, if the control overhead of a protocol is large, the effective throughput is degraded even though the physical data rate may not be affected. It is also required that the effective throughput remain high irrespective of the traffic level [9].

Scalability is another challenge in M2M communications as it is expected that a large number of devices arise in M2M scenarios. These devices have dynamic behaviour, i.e., entering and leaving the network frequently; thus the network must easily tolerate the changes in the node density with little control information exchange. Energy efficiency is also one of the most important challenges in M2M communications, as devices in many M2M applications are battery operated and long life times are expected for these devices [10]. More specifically, the energy spent on radio access and data transmission in M2M communications must be minimized to improve the energy efficiency in a large scale. For instance, in high load scenarios, exchanging control information may consume more than 50% of the total energy in IEEE 802.11 MAC protocol, which shows its ineffectiveness in dense M2M applications [9].

In many M2M applications, the network latency is a critical factor that determines the effectiveness of the service. For instance, in intelligent transportation systems and healthcare monitoring, it is highly important to make the communication

reliable and fast. Channel access delay then needs to be minimized to reduce the overall latency in M2M communications. Moreover, in cellular systems, human-to-human (H2H) devices coexist with M2M devices, and the communication protocol must be designed in such a way to not jeopardize the quality of human-based communications. Resource management and allocation are challenging tasks in M2M communications which coexist with H2H applications, as H2H applications have completely different service requirements [11].

These unique characteristics of M2M communications introduce a number of networking challenges in cellular networks. The fundamental issue arises from the fact that most M2M applications involve a huge number of devices. The question is then how the available radio resources have to be shared among devices such that their service requirements are simultaneously met.

3 Overview of Multiple Access Techniques for M2M Communications

Multiple access techniques can be divided into two broad categories, depending on how the radio resources are allocated to the devices. These include 1) uncoordinated, where the devices transmit data using slotted random access and there is no need to establish dedicated resources, and 2) coordinated, where devices transmit on separate resources pre-allocated by the base station. In coordinated MA, the base station knows a priori the set of devices that have data to transmit. The BS can also acquire channel state information (CSI) of these devices based on which it allocates resources to optimize system throughput. CSI to the devices can be obtained by each device sending an upload pilot signal.

Multiple access techniques can be also divided into orthogonal and non-orthogonal approaches. In orthogonal MA (OMA), radio resources are orthogonally divided between devices, where the signals from different devices are not overlapped with each other. Instances of OMA (**Fig. 1**) are time division multiple access (TDMA), frequency division multiple access (FDMA), orthogonal frequency division multiple access (OFDMA), and single carrier FDMA (SCFDMA). First and second generation cellular systems are mainly developed using OMA approaches, which avoid intra-cell interference and simplify air interface design. However, OMA approaches have no ability to combat the inter-cell interference; therefore careful cell planning and interference management techniques are required to solve the interference problem [12].

Non-orthogonal MA (NOMA) techniques have been adopted in second and third generation cellular systems. NOMA allows overlapping among the signals from different devices by exploiting power domain, code domain, and interleaver pattern. Code division multiple access (CDMA) is the well-known example of NOMA which has been adopted in second and third generation cellular systems. CDMA is robust against inter-cell interference, but suffers from intra-cell interference [12]. CDMA is also not suitable for data services which require high single-user rates. Rather than CDMA which exploits code domain, NOMA in current study in general exploits power domain. NOMA is also shown to provide better performance than OMA [12]. In NOMA, signals from multiple users are superimposed in the power-domain and successive interference cancellation (SIC) is used at the BS to decode the messages. It is also shown that NOMA can achieve the multiuser capacity region both in the uplink and downlink [12].

In this paper, we compare NOMA and OMA strategies in both coordinated and uncoordinated scenarios, and show that NOMA can provide the system with higher capacity to support M2M devices, especially in the uncoordinated scenario. This is achieved by exploiting the power domain, rather than frequency-domain or time-domain as in FDMA and TDMA, respectively.

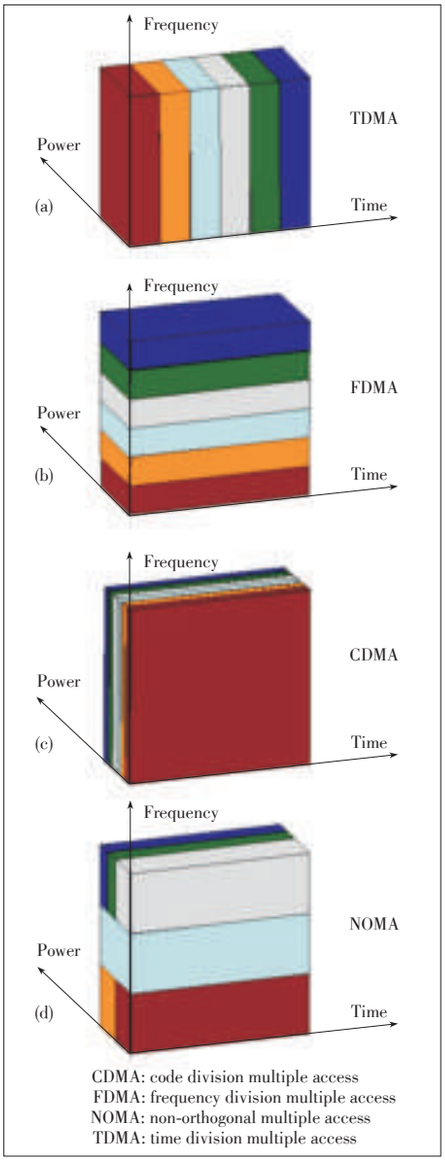


Figure 1. Different multiple access schemes.
CDMA: code division multiple access
FDMA: frequency division multiple access
NOMA: non-orthogonal multiple access
TDMA: time division multiple access

For the analysis in this paper, we consider a single cell centered by base station and devices uniformly distributed around it in a circular region with radius R. The uplink load seen by the base station is modeled by a Poisson point process with mean λ arrivals per second. We further assume a time slotted system with a slot duration of τ_s. We perform our analysis on a typical radio resource with slot duration τ_s and bandwidth W. Each device packet is assumed to have a payload of L bits.

The channel from a device located at distance r from the base station is modelled by $g=(r/R)^{-\gamma}$, where γ denotes the path loss exponent and we ignore shadowing and small scale fading [13]. The received signal-to-noise ratio (SNR) for a device transmitting with power P_t over bandwidth W is then given by [14]:

$$\mu_r = \frac{P_t}{P_{\max}} \frac{W}{W_t} \mu g , \tag{1}$$

where $P_{\max}$ is the maximum transmit power and μ is the reference SNR, defined as the average received SNR from a device transmitting at maximum power $P_{\max}$ over bandwidth W located at the cell edge. Without loss of generality, we assume ordered channel gain $g_1 \geqslant g_2 \geqslant \cdots \geqslant g_K$, where K is the number of devices.

4 Coordinated Multiple Access Strategies

In this section, we consider the coordinated multiple access strategies, i.e., TDMA, FDMA, and NOMA, and compare their throughput efficiency. In this section, we assume that the BS has perfect CSI to all the devices.

4.1 Optimal Throughput FDMA Strategy

In FDMA, the spectrum is partitioned between the devices and each device will transmit in a portion of the spectrum. Fig. 1b shows the FDMA strategy, where the whole spectrum has been divided between 6 devices, and each device will use its allocated bandwidth for the data transmission.

Using Shannon's capacity formula, the minimum bandwidth required for the transmission of L bits by the ith device over time τ_s is given by the solution of the following equation [13]:

$$\frac{L}{\tau_s W_{\min_i}} = \log_2\left(1+\mu\frac{W}{W_{\min_i}}g_i\right). \tag{2}$$

The maximum load that can be supported in a resource block of duration τ_s and bandwidth W is given by:

$$K_{\max} = \max\left\{K:\sum_{i=1}^{K} W_{\min_i} \leqslant W\right\}. \tag{3}$$

4.2 Optimal Throughput TDMA Strategy

In TDMA, the whole spectrum is used by each device in separate time instances. Fig. 1a shows the TDMA scheme, where the same time duration is allocated for 6 devices, and each device will only transmit in its allocated time slot using the whole spectrum. TDMA is an interesting MA strategy due to its simplicity, but it is not efficient for M2M applications with a large number of devices. Moreover, with increasing the number of devices, each device's transmission will be delayed which is not appropriate for delay-sensitive M2M applications.

Assuming a capacity approaching code and using Shannon's capacity equation, the time required for a device located at distance r from the base station to deliver its packet to the destination is given by [13]:

$$\tau \geqslant \frac{L}{W\log_2(1+\mu_r)}, \tag{4}$$

and the minimum time required to deliver the message is obtained when the device is transmitting with full power $P_{\max}$:

$$\tau_{\min_i} = \frac{L}{W\log_2(1+\mu g_i)}. \tag{5}$$

Similar to FDMA, the maximum number of devices which can be supported in a resource block of duration τ_s and bandwidth W can then be found as follows:

$$K_{\max} = \max\left\{K:\sum_{i=1}^{K}\tau_{\min_i} \leqslant \tau_s\right\}. \tag{6}$$

4.3 Optimal Throughput NOMA Strategy

Unlike TDMA and FDMA, devices in the NOMA strategies are assumed to transmit in the same resource block and their transmissions interfere with each other. We assume that the BS perform successive interference cancellation (SIC), where it starts the decoding with the device with the largest channel gain and treats the signals from other devices as additive noise. After decoding the first device, its signal will be removed from the received signal and the BS continues the decoding for the second device and treats the remainder as additive noise. This process is continued until all the devices are successfully decoded. Under this decoding strategy, the Shannon Capacity formula for the ith device is given by:

$$L = W\tau_s \log_2\left(1+\frac{P_i\mu g_i}{1+\sum_{j=i+1}^{K} P_i\mu g_j}\right), \tag{7}$$

and the required transmit power can be calculated as follows:

$$P_i\mu g_i = \left(2^{\frac{L}{W\tau_s}}-1\right)\left(1+\sum_{j=i+1}^{K} P_i\mu g_j\right). \tag{8}$$

By substituting, $i=K$, we have:

$$P_K = \frac{2^{\frac{L}{W\tau_s}}-1}{\mu g_K}, \tag{9}$$

and by going backwards and finding the transmit power for the ith device, we have:

$$P_i = \frac{2^{\frac{(K-i)L}{W\tau_s}}\left(2^{\frac{L}{W\tau_s}}-1\right)}{\mu g_i}. \quad (10)$$

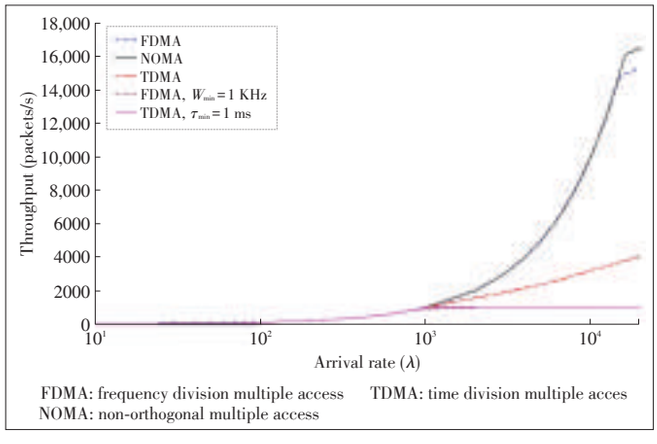
The maximum load that the BS can support in a resource block of bandwidth W and duration τ_s can be found as follows:

$$K_{\max} = \max\left\{K : P_i \leqslant P_{\max} \text{ for } i = 1,2,\cdots,K\right\}. \quad (11)$$

4.4 Comparison Between Coordinated MA Techniques

Fig. 2 shows the maximum throughput versus arrival rates for different coordinated MA techniques. NOMA can achieve very high throughput when the arrival rate is very large. FDMA performs very close to the NOMA strategy and can support all the active devices for the arrival rates up to 14,000 packets per second. The advantage of NOMA comes from the fact that the devices can use the whole spectrum thus achieving a higher throughput compared to FDMA. Only a fraction of the spectrum is used by each device in FDMA. Also, TDMA cannot support many devices, which shows that it is not an effective MA strategy for M2M communications.

It is clear that the time slot duration τ_i and subchannel bandwidth W_i cannot be arbitrarily small in TDMA and FDMA, respectively. As can be seen in Fig. 2, if we put some constraints on the minimum time slot duration or subchannel bandwidth, the number of devices which can be supported by FDMA and TDMA would be limited. For example, if the minimum time slot duration for TDMA is set to be 1 ms, the maximum number of devices which can be supported in a time slot of duration 1 s is 1000. Similarly, if the minimum subchannel bandwidth in FDMA is set to be 1 kHz, the maximum number of devices which can be supported by the BS will be 1000. This shows that in practical systems where the minimum subchannel bandwidth and time slot duration cannot be very small, the maximum throughput of TDMA and FDMA will be limited. In such cases, NOMA can bring more benefits to the system as it can support a larger number of devices without dividing the radio resource into subchannels or time slots.



▲**Figure 2. Average throughput versus arrival rates for different coordinated MA techniques. Total available bandwidth is $W = 1$ MHz, time slot duration is $\tau_s = 1$ sec, and the packet length is $L = 1000$ bits.**

5 Uncoordinated Multiple Access Strategies

In this section, we assume that the base station does not have CSI to the devices, which is particularly the case for M2M communications with a large number of devices, where it is almost impractical for the base station to estimate the channel to every device with random activities. The only information we assume is available at the BS is the traffic load which can be obtained using different load estimation algorithms.

5.1 Uncoordinated FDMA

In this scheme, we assume that the base station chooses a selection probability p_c and broadcasts this information to the devices. Each device which has data to transmit only switches on its transmitter with probability p_c. We refer to these devices as active devices. Let N_c denote the number of active device. We further assume that the BS uniformly divides the spectrum into N_w subchannels, and each device randomly chooses a subchannel for its transmission. We also assume that each device only transmits on a selected subchannel if the maximum transmit power required to deliver its message to the BS is less than $P_{\max}$, assuming no collision on the selected subchannel. More specifically, the ith device is transmitting in a subchannel if the following condition holds:

$$\left(2^{\frac{LN_w}{W\tau_s}}-1\right) \leqslant N_w \mu g_i. \quad (12)$$

Therefore, the probability that a device is transmitting can be calculated as follows:

$$P\left(\left(2^{\frac{LN_w}{W\tau_s}}-1\right) \leqslant N_w \mu g_i\right) = \left(\frac{N_w \mu}{2^{\frac{LN_w}{W\tau_s}}-1}\right)^{\frac{2}{\gamma}}, \quad (13)$$

which is due to the fact that the devices are uniformly distributed in the cell and the probability that a device is located at distance r is given by $2r/R^2$. The average number of active devices which can deliver their messages, considering no collision, can be found as follows:

$$N_p = N_c\left(\frac{N_w \mu}{2^{\frac{LN_w}{W\tau_s}}-1}\right)^{\frac{2}{\gamma}}. \quad (14)$$

As the devices randomly choose a sub-channel for their

transmission, more than one device can select the same subchannel, which leads to collision. The base station cannot decode any of the devices that are simultaneously transmitting on that particular subchannel. The probability of collision can be calculated as follows [14]:

$$P_c = 1 - \left(1 - \frac{1}{N_w}\right)^{N_p - 1}. \quad (15)$$

The average number of devices which can successfully deliver their messages to the BS is given by $N_P(1-P_c)$. We assume that the BS finds the optimal values for p_c and N_w such that the number of devices which can be supported by the BS is maximized.

5.2 Uncoordinated TDMA

Similar to FDMA, we assume that the BS assigns an access probability p_c to the devices. Let N_c denote the number of active device. We further assume that the BS uniformly divides the time into N_t time slots, and each device randomly chooses a time slot for its transmission. We also assume that the each device only transmits in a selected time slot if the maximum transmit power required to deliver its message to the BS is less than $P_{\max}$, assuming no collision on the selected time slot. More specifically, the ith device is transmitting in a time slot, if the following condition holds:

$$\left(2^{\frac{LN_t}{W\tau_s}} - 1\right) \leqslant \mu g_i. \quad (16)$$

Therefore, the probability that a device is transmitting can be calculated as follows:

$$p\left(\left(2^{\frac{LN_t}{W\tau_s}} - 1\right) \leqslant \mu g_i\right) = \left(\frac{\mu}{2^{\frac{LN_t}{W\tau_s}} - 1}\right)^{\frac{2}{\gamma}}, \quad (17)$$

which is due to the fact that the devices are uniformly distributed in the cell and the probability that a device is located at distance r is given by $2r/R^2$. The average number of active devices which can deliver their messages, considering no collision, can be found as follows:

$$Np = Nc\left(\frac{\mu}{2^{\frac{LN_t}{W\tau_s}} - 1}\right)^{\frac{2}{\gamma}}. \quad (18)$$

The average number of devices which can successfully deliver their messages to the BS is given by $N_P(1-P_c)$, where P_c is given by (15) by replacing N_w with N_t. We assume that the BS finds the optimal values for P_c and N_t such that the number of devices which can be supported by the BS is maximized.

5.3 Uncoordinated NOMA

We consider that each device performs power control such that the received SNR at the BS for each device is γ_0. A device will only transmit if and only if the transmit power required to achieve the SNR γ_0 at the base station is less than $P_{\max}$. Let N_p denote the number of devices which can transmit, i.e., their required transmit power is less than $P_{\max}$. The achievable rate for the devices considering the successive interference cancellation at the BS can be calculated as follows:

$$R_{\min} = \log_2\left(1 + \frac{\gamma_0}{1 + (N_p - 1)\gamma_0}\right). \quad (19)$$

A message of length L can be delivered by N_p devices if $W\tau_s R_{\min} \geqslant L$. Using (19), the required SNR γ_0 for successfully delivering a message of length L at the BS is derived as follows:

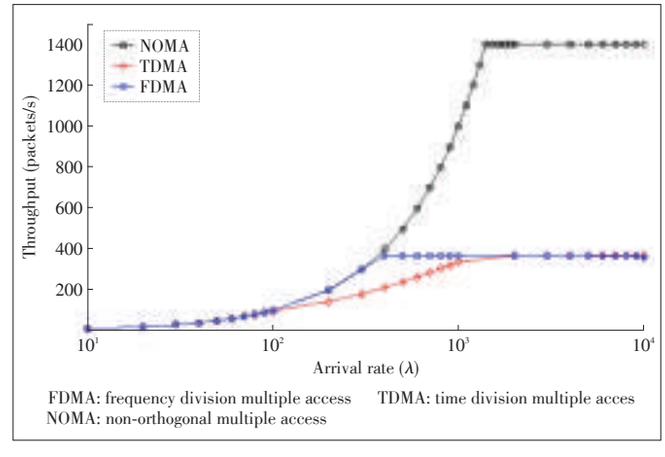
$$\gamma_0 = \frac{1}{\frac{1}{2^{\frac{L}{W\tau_s}} - 1} - N_p}. \quad (20)$$

Accordingly, the number of devices which can be supported at the BS is upper bounded as follows:

$$N_p \leqslant \frac{1}{2^{\frac{L}{W\tau_s}} - 1}. \quad (21)$$

5.4 Comparison Between Uncoordinated MA Techniques

Fig. 3 shows the maximum number of devices which can be supported by the base station versus different arrival rates for uncoordinated MA strategies. The minimum time slot duration



▲Figure 3. Average throughput versus arrival rates for different uncoordinated MA techniques. Total available bandwidth is $W = 1$ MHz, time slot duration is $\tau_s = 1$ s, and the packet length is $L = 1000$ bits. The minimum time slot duration for TDMA is considered to be 1 ms and the minimum subchannel bandwidth in FDMA is considered to be 1 kHz.

for TDMA is considered to be 1 ms, which corresponds to $N_t = 1000$, and the minimum subchannel bandwidth in FDMA is considered to be 1 kHz, which corresponds to $N_w = 1000$. As shown in this figure, NOMA can support much larger number of devices compared to the FDMA and TDMA strategies. This is due to the high collision probability in uncoordinated FDMA and TDMA in high arrival rates, while in NOMA a large number of devices can simultaneously transmit in the same resource block by exploiting the power domain. This shows the advantage of NOMA in uncoordinated scenarios. Therefore, NOMA can be an excellent choice for M2M applications with a large number of devices and random traffic. Moreover, FDMA outperforms TDMA in moderate loads but they perform similarly in low and high arrival rates.

It is important to note that in NOMA the constraints on minimum time slot duration or subchannel bandwidth do not affect the throughput efficiency. This is due to the fact that in NOMA all the devices are transmitting in the whole bandwidth in all slot duration. One could consider some limitations in the minimum power difference between the devices, which mostly depends on the hardware capability to distinguish different power levels which is out of scope of this paper.

6 Practical Considerations of Massive NOMA for M2M Communications and Future Directions

NOMA can bring many benefits to cellular systems which include, but are not limited to, the following. NOMA can effectively use the spectrum and provide higher throughput by exploiting power domain and non-orthogonal multiplexing. It also provides robust performance gain in high mobility scenarios. NOMA is also compatible with OFDMA and can be applied on top of OFDMA for downlink and SC-FDMA for uplink. It can be also combined with multi-antenna techniques to improve the system performance. Using NOMA, multiple users can simultaneously transmit in the same subband without being identified by the destination a priori. The devices can attach their terminal identities to their messages and the base station can identify the devices after decoding their messages. The RA procedure can be eliminated and therefore the access delay and signaling overhead will be significantly reduced [12].

Although NOMA can improve spectrum efficiency and system capacity, there are many practical challenges for this technology to be potentially used in real wireless systems for M2M communications. Here, we outline the main practical consideration of massive NOMA for M2M communications.

First, in uncoordinated strategies the base station needs to estimate the arrival rate to effectively detect the devices. In uncoordinated FDMA, the BS needs to know the number of devices to find the optimal access probability and the number of subbands. In NOMA, the problem is much more complicated as the BS runs the SIC and needs to know the number of devices with different power levels. For simplicity, one could consider that the devices perform power control such that only one power level is received at the BS, but this may have some implications on the actual performance of the system as the overall system data rate will be dominated by the device with the lowest SINR; and thus will not effectively use the available spectrum. However, even with this simplification and suboptimal power allocations, NOMA outperforms FDMA in uncoordinated scenarios and can support a large number of devices under high loads.

Second, channel estimation at the devices is necessary in uncoordinated strategies employing NOMA techniques. This is due to the fact that the devices are not identified by the BS beforehand and they are simultaneously transmitting at the same resource block. To enable the BS to detect the devices and decode their messages, the devices need to perform channel estimation and adjust their power so the BS only deals with some known power levels rather than unknown channel gains. On the other hand, to effectively perform SIC, the multipath effect must be carefully taken into consideration as multipath will spread the signal over time, which decreases the effective signal to noise ratio for each device, and makes the BS unable to perform SIC. One can consider several techniques, such as time reversion [15], to eliminate the multipath effect by treating the channel between each device and the BS as the natural match filter. This has been shown an effective way to combat multipath effect for several fixed location M2M applications [16].

Third, NOMA requires synchronization among the devices at the symbol level. This is very challenging as providing time synchronicity between a large number of devices distributed in a large environment is tedious. However, the devices in many M2M applications are deployed in fixed locations, so each device can determine its propagation delay using different distance estimation strategies or using control information periodically sent by the BS.

Fourth, as the number of devices transmitting in each resource block in uncoordinated NOMA is random, the physical data rate cannot be determined beforehand. One could consider a very low rate code at each device, but it might be inefficient when used in low-to-moderate loads. An effective strategy is then to use rateless codes to automatically adapt to the traffic condition. Authors in [17] have proposed to use analog fountain codes to enable massive multiple access for M2M communications and achieve very high throughput even in high loads. Moreover, as shown in [18], binary rateless codes can be effectively used to enable NOMA for M2M communications. These coding strategies were mainly proposed to maximize the throughput in M2M communications and for delay sensitive applications with very short messages, more advanced coding techniques should be combined with rateless ideas to enable low latency massive multiple access in M2M communications.

Last but not least, NOMA is still in its early stage of its development and more research work must be done to clearly identify its effectiveness in real scenarios. From an information theoretic point of view, it achieves the capacity region of the multiple access channel and thus is optimal in terms of throughput. But in real M2M applications when NOMA is jointly considered with medium access control layer in real world scenarios, it might not be as efficient as OMA techniques, which have been considered as effective multiple access techniques for a long time and several issues and challenges have been solved over the years.

7 Conclusions

In this paper, we provided an overview of multiple access techniques for emerging machine-to-machine communications in cellular systems. The unique challenges of M2M communications were represented, where we identified scalability, energy efficiency, and reliability, as the most important features for every potential multiple access technology which is considered for M2M communications. We provided a simple study on the throughput efficiency of multiple access techniques in both coordinated and uncoordinated scenarios. NOMA was shown to provide the highest throughput in both coordinated and uncoordinated scenarios, whereas FDMA has shown comparable performance with NOMA in coordinated scenarios. NOMA is shown to be scalable in uncoordinated scenarios and can support a large number of devices. It can be also combined with different access management schemes to control the load over the base station. We also provided some of the practical issues in NOMA which needed to be considered for the use of NOMA strategies for M2M communications in future cellular systems.

References

[1] H. Shariatmadari, R. Ratasuk, S. Iraji, et al., "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commununications Magazine*, vol. 53, no. 9, pp. 10–17, Sept. 2015. doi: 10.1109/MCOM.2015.7263367.

[2] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2M: From mobile to embedded internet," *IEEE Commununications Magazine*, vol. 49, no. 4, pp. 36–43, Apr. 2011. doi: 10.1109/ MCOM.2011.5741144.

[3] Ericsson. (Jan. 2016). Cellular networks for massive IoT. Tech. Rep. Uen 284 23-3278 [Online]. Available: https://www.ericsson.com/res/docs/whitepapers/wp iot.pdf

[4] *Service Requirements for Machine-Type Communications (MTC); Stage 1*, 3GPP TS 22.368 V.13.0.0, Jun. 2014.

[5] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? a survey of alternatives," *IEEE Commununications Surveys & Tutorials*, vol. 16, no. 1, pp. 4–16, 2014. doi: 10.1109/SURV.2013.111313.00244.

[6] G. Naddafzadeh-Shirazi, L. Lampe, G. Vos, and S. Bennett, "Coverage enhancement techniques for machine-to-machine communications over LTE," *IEEE Communications Magazine*, vol. 53, no. 7, pp. 192–200, Jul. 2015. doi: 10.1109/MCOM.2015.7158285.

[7] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Jounral*, vol. 1, no. 1, pp. 22–32, Feb. 2014. doi: 10.1109/JIOT.2014.2306328.

[8] D. Wiriaatmadja and K. W. Choi, "Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 33–46, Jan. 2015. doi: 10.1109/TWC.2014.2328491.

[9] A. Rajandekar and B. Sikdar, "A survey of MAC layer issues and protocols for machine-to-machine communications," *IEEE Internet of Things Journal*, vol. 2, no. 2, pp. 175–186, Apr. 2015. doi: 10.1109/JIOT.2015.2394438.

[10] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 86–93, Jun. 2013. doi: 10.1109/MCOM.2013.6525600.

[11] K. Zheng, S. Ou, J. Alonso-Zarate, et al., "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Communications Magazine*, vol. 21, no. 3, pp. 12–18, Jun. 2014. doi: 10.1109/MWC.2014.6845044.

[12] A. Li, Y. Lan, X. Chen, and H. Jiang, "Non-orthogonal multiple access (NOMA) for future downlink radio access of 5G," *China Communications*, vol. 12, Supplement, pp. 28–37, Dec. 2015.

[13] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela, "Power-efficient system design for cellular-based machine-to-machine communications," *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5740–5753, Nov. 2013. doi: 10.1109/TWC.2013.100713.130025.

[14] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela, "Fundamentals of throughput maximization with random arrivals for M2M communications," *IEEE Transactions on Communications*, vol. 62, no. 11, pp. 4094–4109, Nov. 2014. doi: 10.1109/TCOMM.2014.2359222.

[15] B. Wang, Y. Wu, F. Han, Y. H. Yang, and K. J. R. Liu, "Green wireless communications: A time-reversal paradigm," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1698–1710, Sept. 2011. doi: 10.1109/JSAC.2011.110918.

[16] Y. Chen, F. Han, Y. H. Yang, et al., "Time-reversal wireless paradigm for green internet of things: An overview," *IEEE Internet of Things Jounral*, vol. 1, no. 1, pp. 81–98, Feb. 2014. doi: 10.1109/JIOT.2014.2308838.

[17] M. Shirvanimoghaddam, Y. Li, M. Dohler, B. Vucetic, and S. Feng, "Probabilistic rateless multiple access for machine-to-machine communication," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6815–6826, Dec. 2015. doi: 10.1109/TWC.2015.2460254.

[18] M. Shirvanimoghaddam, S. J. Johnson, and M. Dohler. (2016). An efficient massive access strategy based on superposition Raptor codes for M2M communications. *CoRR* [Online]. Available: http://arxiv.org/pdf/1602.05671v1.pdf

Manuscript received: 2016-06-30

Biographies

Mahyar Shirvanimoghaddam (fmahyar.shirvanimoghaddam@newcastle.edu.au) received the BSc degree with 1st Class Honours from University of Tehran, Iran, in September 2008, the MSc degree with 1st Class Honours from Sharif University of Technology, Iran, in October 2010, and the PhD degree from The University of Sydney, Australia, in January 2015, all in electrical engineering. He then held a research assistant position at the Centre of Excellence in Telecommunications, School of Electrical and Information Engineering, The University of Sydney, before coming to the University of Newcastle, Australia, where he is now a postdoctoral research associate at the School of Electrical Engineering and Computer Science. His general research interests include channel coding techniques, cooperative communications, compressed sensing, machine-to-machine communications, and wireless sensor networks.

Sarah Johnson (sarah.johnsong@newcastle.edu.au) received the BE (Hons) degree in electrical engineering in 2000, and PhD in 2004, both from the University of Newcastle, Australia. She then held a postdoctoral position with the Wireless Signal Processing Program, National ICT Australia before returning to the University of Newcastle where she is now an Australian Research Council Future Fellow. Her research interests are in the fields of error correction coding and network information theory. She is the author of a book on iterative error correction published by Cambridge University Press.

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 007
http://www.cnki.net/kcms/detail/34.1294.TN.20160928.1048.002.html, published online September 28, 2016

Software Defined Optical Networks and Its Innovation Environment

LI Yajie[1], ZHAO Yongli[1], ZHANG Jie[1], WANG Dajiang[2], and WANG Jiayu[2]
(1. Beijing University of Posts and Telecommunications, Beijing 100876, China)
2. ZTE Corporation, Shenzhen 518057, China)

Abstract

Software defined optical networks (SDONs) integrate software defined technology with optical communication networks and represent the promising development trend of future optical networks. The key technologies for SDONs include software-defined optical transmission, switching, and networking. The main features include control and transport separation, hardware universalization, protocol standardization, controllable optical network, and flexible optical network applications. This paper introduces software defined optical networks and its innovation environment, in terms of network architecture, protocol extension solution, experiment platform and typical applications. Batch testing has been conducted to evaluate the performance of this SDON testbed. The results show that the SDON testbed has good scalability in different sizes. Meanwhile, we notice that controller output bandwidth has great influence on lightpath setup delay.

Keywords

optical networks; software defined networking; innovation environment

1 Introduction

With the emerging of new network services, the interaction of all kinds of information grows day by day. It is an eternal theme for optical networks to satisfy the transmission demands for high speed, wide broadband, large capacity, and long-distance transmission. The changes of servics properties brings a new challenge: intelligence of optical networks. For example, high burst services require optical networks to have dynamic adaptability; large-scale networking requires optical networks to be scalable; and variable bandwidth provisioning requires optical networks to be flexible. To realize the intelligent optical network, the industry has carried out a long-term research and exploration. So far, the intelligent optical network has gone through three important stages of development.

1) Automatic Switching Optical Networks (ASON)

An ASON is divided into three planes: the transmission plane, control plane and management plane. With the control plane based on Generalized Multi-Protocol Label Switching (GMPLS) protocol, ASON adopts distributed signaling and routing to solve the connection control problem and satisfy the function demands of automatic switching [1]–[4]. However, ASON has obvious limitations in many aspects, including large-scale connection control, complex path calculation, network openness, devices interworking, and cost reduction. Besides, the GMPLS standard is very complex, which greatly affects the application and promotion of ASON.

2) Path Computation Element (PCE) Architecture for ASON

In order to better adapt to the characteristics of multi-layer multi-domain large-scale optical networks, the Internet Engineering Task Force (IETF) separates the path calculation function from the control layer and develops an independent unit, i.e., PCE [5]–[8]. In order to satisfy the function demands of large-scale multi-layer/domain, PCE adopts the distributed signaling and centralized routing to solve the problem of path selection and calculation for inter-layer and inter-domain path. However, with unitary function of path calculation, PCE needs to cooperate with other technology in applications.

3) Software Defined Optical Networks (SDONs)

SDONs can offer a unified schedule and control for various kinds of optical layer resources according to the requirements of users or operators. With programmable software and dynamic customization, the SDON solves the problem of function extension and therefore realizes rapid response to requests, efficient utilization of resources and flexible service provisioning. The SDON well supports service processing, control strategy and programmable transmission device, which achieves programmable tuning of optical network elements [9]. Therefore, the SDON is more suitable for multi-layer/domain and multi-constraint optical networks, and it can effectively improve operational efficiency and reduce cost.

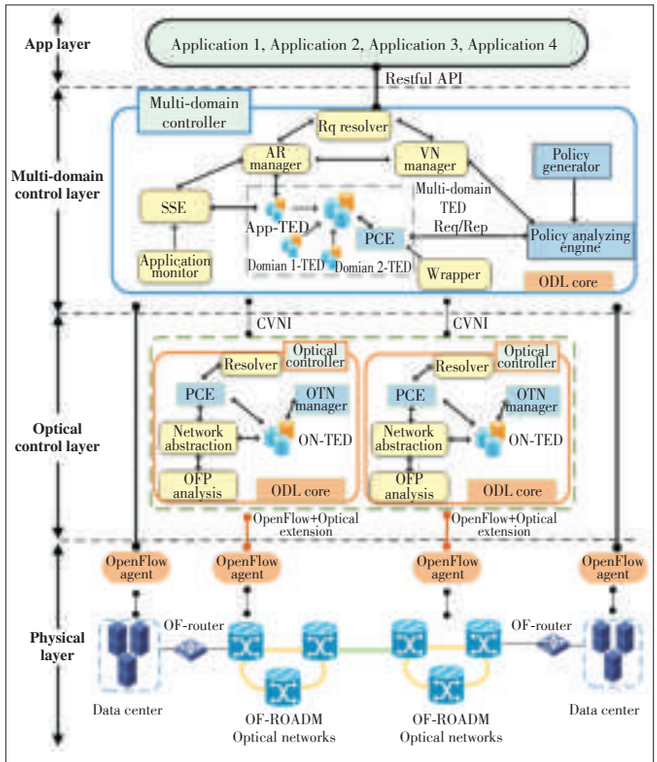
The article introduces the SDON innovation environment from the perspectives of architecture, protocol extension, experimental platform and typical applications. Section 2 describes the hierarchical control architecture and the process of cross-

This work was supported by ZTE Industry-Academia-Research Cooperation Funds under Grant No. Surrey-Ref-9953.

domain connection provisioning in detail. Section 3 depicts the workflow of connection provisioning in multi-domain optical networks. Section 4 shows the necessary extension work of OpenFlow 1.3 protocol for optical networks. The experimental environment and typical applications of SDON are respectively discussed in section 5 and section 6. Section 7 is performance evaluation of the SDON testbed and the last section summarizes the paper.

2 Hierarchical Architecture for SDON

As shown in **Fig. 1**, considering the cross-layer distribution of multi-domain optical network resources, this paper proposes an OpenFlow enabled hierarchical control architecture in order to solve the problem of programmable control in optical network. With the advantage of software-defined networking, the architecture uniformly abstracts optical transmission network resources and content resources, and provides them with the multi-domain controller through the northbound interface. In this way, the uniform control of cross-layer resources is realized. The hierarchical architecture consists of three layers: the physical layer, control layer and application layer.



▲**Figure 1. Hierarchical network architecture.**

1) The physical layer mainly includes data-center and inter-data-center optical transmission networks. OpenFlow-enabled IP Routers (OF-Router) and optical transmission equipment with OpenFlow agents (OF-ROADM) are deployed in the network.

2) The application layer mainly includes various applications such as dynamic migration of virtual machines, virtual network provisioning, and spectrum defragmentation. It is connected with the control layer through the Restful API interface. All the service requests are triggered from this layer.

3) The control layer is mainly composed of optical controllers and multi-domain controller.

In the optical controller, the protocol analysis module analyzes the underlying optical transmission equipment via the OpenFlow protocol extended for optical transmission devices. It collects the status of OF-ROADM in the optical network and abstracts the network topology information. Then the abstracted topology information is sent to the network abstraction module and stored in the optical database (ON-TED). The OTN manager manage the optical transmission equipment, such as lightpath setup and deletion, and resources allocation. The optical network controller packages the network status and topology information via the protocol encapsulation module and makes a notification to the multi-domain controller.

The multi-domain controller integrates the network information collected by the optical controller through the southbound Control Virtual Network Interface (CVNI) interface and monitors the network status. With the northbound Restful-API, it parses the application requests sent by the application layer. It consists of nine function modules and one resource integration module. Resource integration is completed by the heterogeneous network database (Het-TED). The application database (App-TED) and ON-TED in network are set into the same database, with the purpose of supplying the network resources information for the corresponding module in the network. The nine function modules are respectively described as follows.

1) Application monitor: It monitors the computing resources in the network and reports the information of computing resources to service selection engine.

2) Service selection engine (SSE): According to the status of application resources and network resources requests, it selects the most appropriate application resources to meet the virtual requests.

3) Application resource manager: It manages the application resources in the network, and keeps real-time synchronous update with the resources information in the application resources database (App-TED).

4) Request resolver: It parses the requests sent by the application layer and forward them to the corresponding module for processing.

5) Virtual network manager: It manages the virtual network requests sent by the application layer according to the status of application resources and network resources, and selects the

most appropriate network links to meet the virtual network requests with the corresponding strategy.

6) Policy generator: According to the network requests from the application layer, it generates the corresponding strategy information of resource provisioning for the heterogeneous network controller.

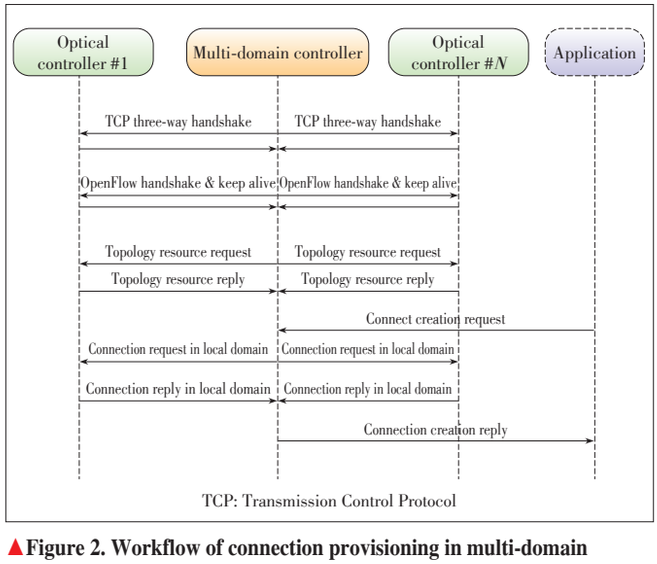
7) Policy analyzing engine: It parses the strategy generated by the strategy generator module, and sends it to the corresponding module for network resource allocation.

8) PCE: As a core component of the multi-domain controller, PCE is used in response to the request of path computation. The calculation is based on the input path information, strategy information and request information, etc. It may return two kinds of calculation results, the appropriate path information computed by multi-domain or the failure information.

9) Wrapper: It packages the resource allocation information with the CVNI protocol and sends it to the optical controller via the CVNI interface.

3 Workflow of Connection Provisioning

Fig. 2 shows the workflow of providing an end-to-end connection in the multi-domain optical networks. After a Transmission Control Protocol (TCP) connection is established, the multi-domain controller completes the handshake with the optical controller by using OpenFlow messages, then periodically sends packages to keep the connection alive. The multi-domain controller requests the abstract topology as well as the detailed port information. After receiving a request for connection setup the optical controller completes the path computation and resource allocation in the local domain via the domain-specified protocol. Once the process is finished, the optical controller sends a "success" reply to the multi-domain controller. When the multi-domain controller collects all the "success" messages from optical domains, a "success" notification will be sent to the application layer. At this point, a connection or lightpath is considered to be established successfully.



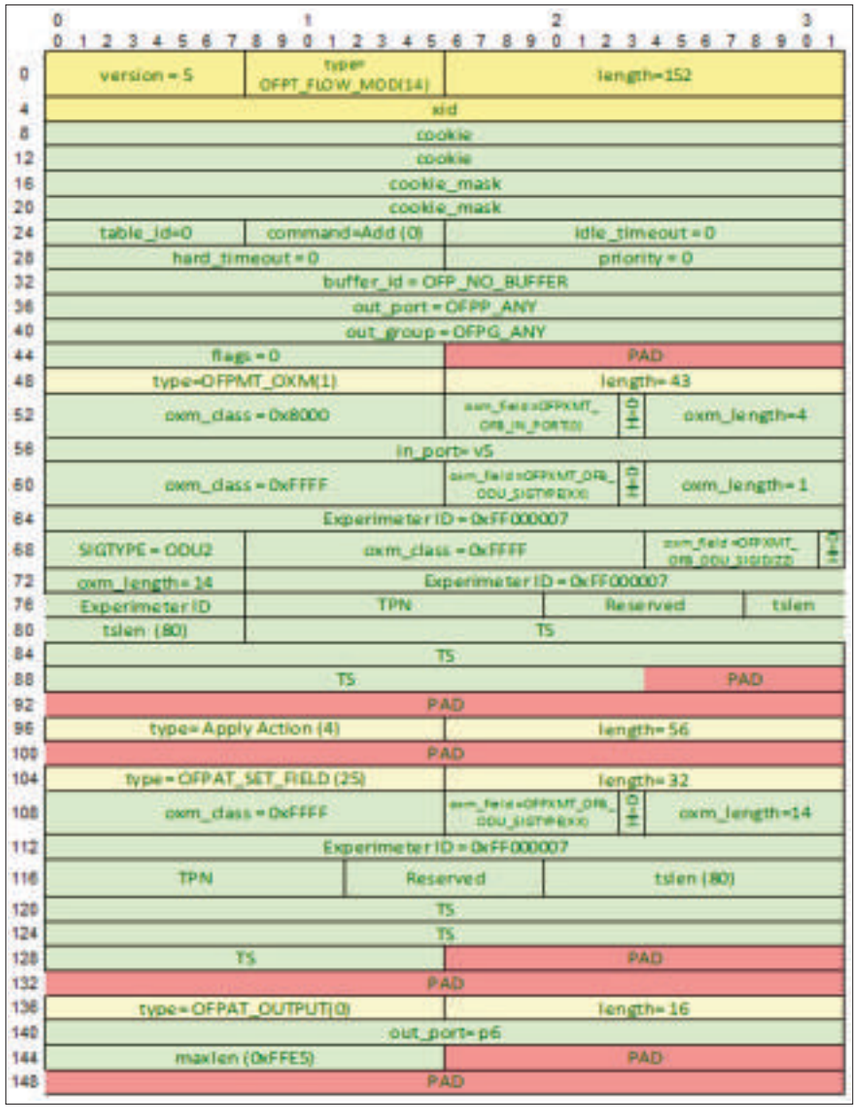
▲**Figure 2. Workflow of connection provisioning in multi-domain optical networks.**

4 Protocol Extensions for SDON

Based on OpenFlow 1.3 protocol, CVNI is an interface protocol between the multi-domain controller and optical layer controller. Several OpenFlow messages in CVNI have been extended to satisfy the requirements of optical networks. The multi-domain controller sends a GET_CONFIG_REQUEST message to the optical controller to get the location of network nodes and the optical controller replies a GET_CONFIG_REPLY message. The MULTIPART_REQUET messages is used by the multi-domain controller to obtain topology resources including ports and links information. The MULTIPART_REPLY message carries topology information from the optical controller to the multi-domain controller. The multi-domain controller employs FLOW_MOD messages to complete connection setup and deletion. The match field and action field in an extended FLOW_MOD message respectively represent the input optical port and output optical port. Note that the multi-domain controller sends a BARRIER_REQUEST message to the single-domain controller in order to verify whether the optical cross connection is deployed successfully. The single-domain controller then sends a BARRIER_REPLY message to notify the multi-domain controller that the connection is created or deleted successfully. Due to space limitation, only FLOW_MOD message extension is illustrated in **Fig. 3**.

5 Experimental Platform for SDON

As shown in **Fig. 4**, an all-optical network innovation (AONI) experimental platform for SDON is distributed in three geography locations connected by optical fiber links. Two of them are located in Room 342 and Room 423 in the Science&Research building of Beijing University of Posts and Telecommunications (BUPT), and the third location is at 21Vianet Company in Jiuxianqiao, Beijing. Two data centers are respectively deployed in Room 342 and 21Vianet Company, and Room 423 serves as the access network for users, which composes a typical network environment with the application of data center. The AONI platform supports three typical network scenarios, i.e., the inter-data center network, user access to data center network and intra-data center network. The AONI platform focuses on how to embody the advantages of optical switching network in these three scenarios. The platform supports both optical burst switching and optical circuit switching, and supports both flexible grid high-speed optical transmission and fixed grid transmission. Therefore, the AONI platform not only provides efficient transmission and switching in the future but also remains compatible with traditional networks. The high capacity optical burst switching (OBS) is mainly used for



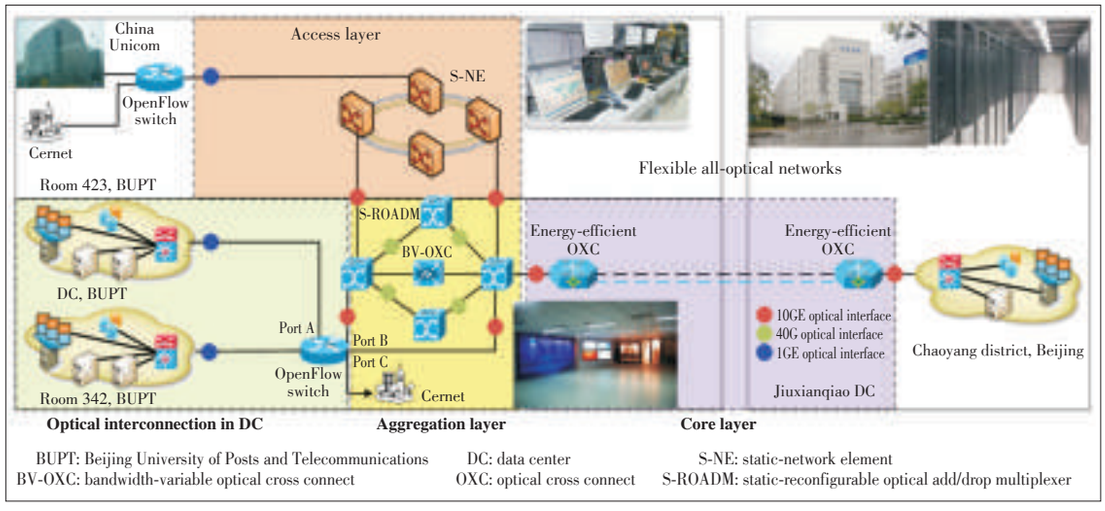
▲Figure 3. FLOW_MOD message extension of CVNI protocol.

the adaption to the high burstiness characteristic of intra-data center services. The flexible grid high-speed optical transmission andoptical circuit switching are mainly applied to the inter-data center to realize large-grained variable bandwidth switching. The all-optical access and convergence layer uses fixed grid transmission and switching to achieve flexible access of broadband services. Thus the architecture of AONI includes intra-data center all-optical interconnection, all-optical access layer, all-optical convergence layer and all-optical core layer. Such a platform can highly simulate real scenarios of all-optical switching wide area network (WAN) in the future.

6 Typical Applications of SDON

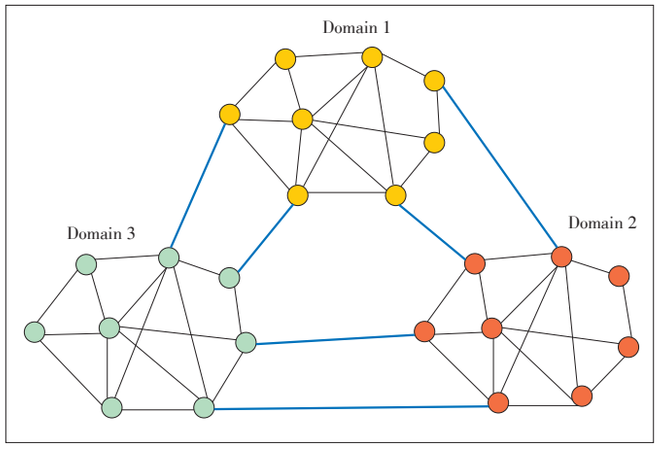
The SDON is a promising solution to high intelligence of next generation optical network and has broad application prospects. The typical applications include bandwidth on demand (BoD) provisioning, virtual machine (VM) online migration, spectrum defragmentation, and virtual optical networks (VON) provisioning. The homepage of AONI applications is shown in **Fig. 5**.

BoD applications and VM migration are implemented based on the physical topology shown in Fig. 4. For lack of flexible-grid optical devices, a multi-domain logical topology (**Fig. 6**) is designed for VON provisioning and spectrum defragmentation. Both the physical topology and the logical topology are under control of the SDN controller. Each domain in-

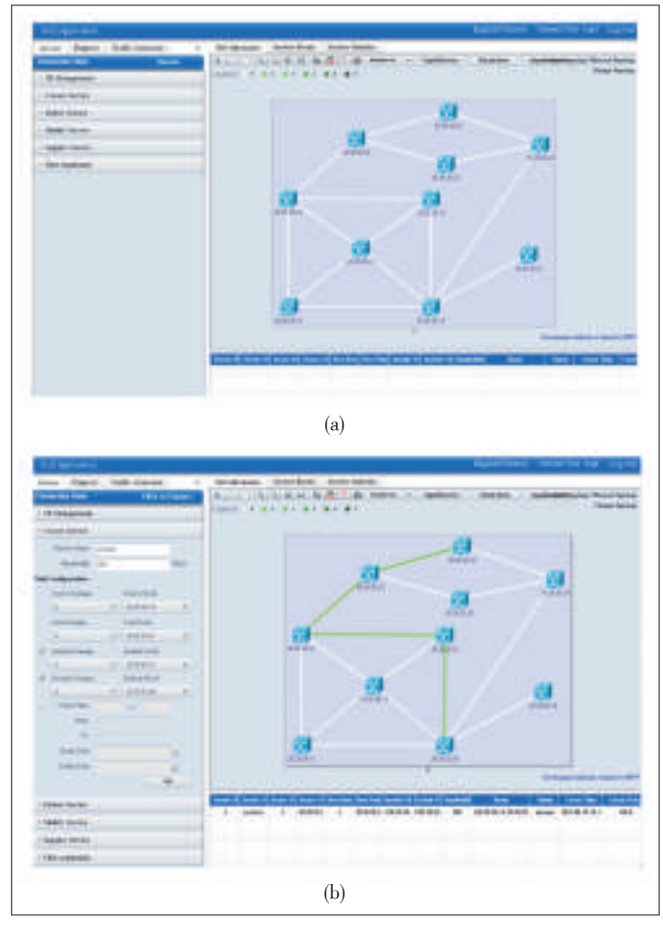


◀Figure 4. AONI: all optical network innovation environment.

▲Figure 5. Homepage of AONI applications.



▲Figure 6. Multi-domain logical topology.



▲Figure 7. Web view of BoD application: (a) before connection setup; (b) after connection setup.

cludes eight standalone OpenFlow-Agents (OF-AGs). Running on high-performance Linux servers, each OF-AG is programmed based on Open-vSwitch.

6.1 BoD Applications

BoD applications help users have a global understanding of underlying optical networks and accomplish a series of operations in optical networks, including connection setup, connection deletion, connection query, connection modification and so on. A lightpath connection is built and on-demand bandwidth is allocated according to users requirements. Besides the instant operation, users are able to make an appointment to carry out above operations by setting starting and ending time. **Fig. 7** shows a connection named "service 1" is created from node 20.20.20.14 to node 20.20.20.21 with required bandwidth. The detailed information about this connection is listed in the lower part of Fig. 7b, including routing, current status, creation delay and so on.

6.2 VM Online Migration

VM migration plays an important role in data backup and load balance of data centers. A VM migration application enables online migration of virtual machines among different data centers. With transmission advantages of optical networks, it just takes a short time to complete the migration process. In addition, the online migration pattern has no impact on users' access to resources in the migrating virtual machine. In **Fig. 8**, a VM, 863VM, is migrated from server 10.108.50.40 to server 10.108.51.124 and the migration path is 20.20.20.14 - 20.20.20.15 - 20.20.20.12. Meanwhile, users can query resource utilization information of the selected servers, such as CPU and memory status.

6.3 VON application

Optical network virtualization technologies support the dynamic provisioning of VONs in the same network infrastructure and achieve high-efficiency utilization of network resources. Because of its centralized control manner, software-defined networking (SDN) is regarded as a promising technology for realizing VON provisioning. In the AONI testbed, network operators can provide virtual optical networks for different customers. The topology of VON can either be pre-configured by oper-

(a) (b)

▲Figure 8. Web view of VM migration: (a) before the migration; (b) after the migration.

(a) (b)

▲Figure 9. Web view of VON provisioning: (a) before the process; (b) after the process.

ators or be customized by users. In **Fig. 9**, a triangle VON topology is successfully mapped to multi - domain networks. Meanwhile, 1 + 1 protection is available for services deployed in the VON. The green path in Fig. 9b stands for the working route of the service while the purple path represents the protection path.

6.4 Spectrum Defragmentation

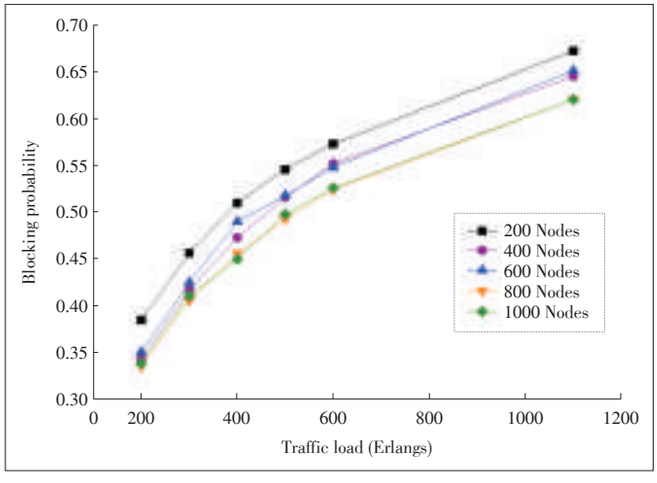
The frequent setup and release of lightpaths in a dynamic network scenario will fragment the optical spectrum into non-aligned, isolated and small - sized spectrum segments. Spectrum fragments result in low spectrum utilization and high blocking probability since these fragments could hardly be occupied for new incoming requests. With the application of spectrum defragmentation, users can have a good knowledge of spectrum utilization and trigger spectrum defragmentation if necessary to optimize spectrum resources. In **Fig. 10**, there are 50 connections or lightpaths deployed in the multi-domain network shown in Fig. 6. It is obvious that the spectrum utilization is effectively improved with the implementation of spectrum defragmentation

7 Performance Evaluation of SDON Testbed

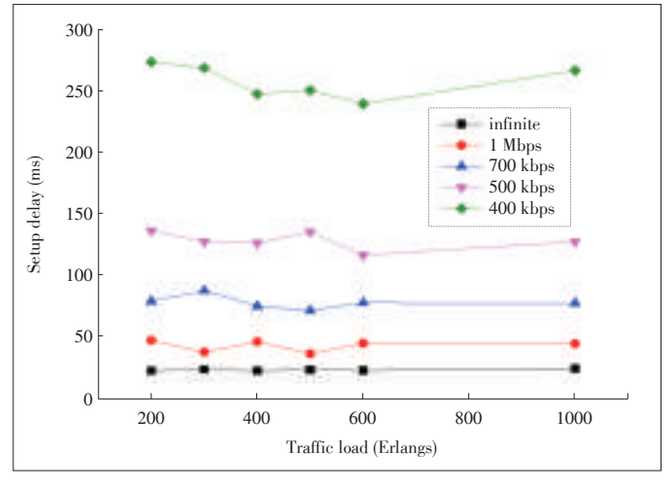
Batch testing has been conducted to evaluate the performance of this SDON testbed. Ten thousands lightpath requests are generated following Poisson distribution, and their source-destination pairs per execution are randomly chosen. The holding time of lightpath requests follows exponential distribution.

To verify the scalability of the SDON testbed, we compare the blocking probabilities of different network sizes. As shown in **Fig. 11**, the number of network nodes ranges from 200 to

▲Figure 10. Web view of spectrum defragmentation: (a) before defragmentation; (b) after DFefragmentation.



▲Figure 11. Blocking probabilities of different network sizes.



▲Figure 12. Lightpath setup delay of different output bandwidth.

1000. For each network size, the blocking probability increases with traffic load. With the same traffic load, 1000-nodes network has the lowest blocking probability since it has the highest network capacity.

In addition, the relationship between the controller output bandwidth and the lightpath setup delay is studied. The controller output bandwidth can be adjusted by the VMware Ethernet bandwidth modulator. As shown in **Fig. 12**, the output bandwidth of controller is set to five different values, including 400 kbps, 500 kbps, 700 kbps, and 1 Mbps. The average delay of lightpath setup is calculated for each case. We can see that the output bandwidth of controller has great influence on lightpath setup delay. With the growth of output bandwidth, the average setup delay decreases significantly from 300 ms to 50 ms.

8 Conclusions

With the advantage of programmable network elements, the SDON realizes service customization, adaptive modulation format, flexible bandwidth allocation and dynamic provisioning of virtual network resources with centralized control manner. This paper introduces SDON and its innovation environment—AONI in terms of network architecture, protocol extension solution, experiment platform, typical applications and performance evaluation. The SDON represents the development direction of optical networks and has broad application prospects in the future.

References

[1] *Requirements for Generalized Multi-Protocol Label Switching (GMPLS) Routing for the Automatically Switched Optical Network (ASON)*, IETF RFC4258, Nov. 2005.

[2] *Requirements for Generalized MPLS (GMPLS) Signaling Usage and Extenstions for Automatically Switched Optical Network (ASON)*, IETF RFC4139, Jul. 2005.
[3] *Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)*, IETF RFC5212, Jul. 2008.
[4] Y. Ji, D. Ren, H. Li, X. Liu, and Z. Wang, "Analysis and experimentation of key technologies in service-oriented optical internet," *Science China Information Sciences*, vol. 54 no. 2, pp. 215–226, Feb. 2011. doi: 10.1007/s11432-010-4168-5.
[5] *A Path Computation Element (PCE)-Based Architecture*, IETF RFC4655, Aug. 2006.
[6] *Path Computation Element (PCE) Communication Protocol Generic Requirements*, IETF RFC4657, Sept. 2006.
[7] *Requirements for Path Computation Element (PCE) Discovery*, IETF RFC4674, Oct. 2006.
[8] *Path Computation Element Communication Protocol (PCECP) Specific Requirements for Inter-Area MPLS and GMPLS Traffic Engineering*, IETF RFC4927, Jun. 2007.
[9] J. Zhang, H. Yang, Y. Zhao, et al., "Experimental demonstration of elastic optical networks based on enhanced software defined networking (eSDN) for data center application," *Optics Express*, vol. 21, no. 22, pp. 26990–27002, Nov. 2013. doi:10.1364/OE.21.026990.

Manuscript received: 2016-03-31

Biographies

LI Yajie (yajieli@bupt.edu.cn) is a PhD candidate in State Key Laboratory of Information Photonics and Optical Communication, Beijing University of Posts and Telecommunications (BUPT), China. His research interest is software defined optical networks.

ZHAO Yongli (yonglizhao@bupt.edu.cn) received his PhD degree from BUPT. He is an associate professor in State Key Laboratory of Information Photonics and Optical Communication, BUPT. His research interest is optical transport networks.

ZHANG Jie (lgr24@bupt.edu.cn) received his PhD degree from BUPT. He is a professor in State Key Laboratory of Information Photonics and Optical Communication, BUPT. His research interest is optical transport networks.

WANG Dajiang (wang.dajiang@zte.com.cn) works in wireline product operation of BN product team, ZTE Corporation. His research interest is optical transport networks.

WANG Jiayu (wang.jiayu1@zte.com.cn) received his master degree from BUPT. He is a SDON R&D representative from BN product team, ZTE Corporation. His research interest is optical transport networks.

New Members of *ZTE Communications* Editorial Board

Dr. CHEN Yan is a professor in the Department of Electrical Engineering and Computer Science at Northwestern University, USA. He is also an adjunct professor in the College of Computer Science at Zhejiang University, China. He got his PhD in Computer Science at University of California at Berkeley, USA in 2003. His research interests include network security, measurement and diagnosis for large scale networks and distributed systems. He won the Department of Energy (DoE) Early CAREER award in 2005, the Department of Defense (DoD) Young Investigator Award in 2007, and the Best Paper nomination in ACM SIGCOMM 2010. Based on the Google Scholar, his papers have been cited for over 9000 times and his h-index is 40.

Dr. SONG Wenzhan is the Georgia Power Mickey A. Brown Professor of Engineering in the University of Georgia, USA. Dr. Song is a distinguished scientist and educator on cyber-physical systems informatics and security in energy, environment and health applications, where decentralized sensing, computing, communication and security play a critical role and need a transformative study. He has an outstanding record of leading large multidisciplinary research projects on those issues with multi-million grant support from NSF, NASA, USGS, and industry, and his research was featured in *MIT Technology Review*, *Network World*, *Scientific America*, *New Scientist*, *National Geographic*, etc. Dr. Song is a recipient of NSF CAREER Award (2010), Outstanding Research Contribution Award (2012) at GSU, Chancellor Research Excellence Award (2010) at WSU. He was also a recipient of 2004 National Outstanding Oversea Student Scholarship by China (only 40 in USA) during PhD study. Dr. Song also has a outstanding publication record and serves many premium conferences and journals as editor, chair or TPC member. He is also an inaugural member of OpenFog consortium involving industry and academic leaders.

Research Paper

DOI: 10.3969/j. issn. 1673-5188. 2016. 04. 008
http://www.cnki.net/kcms/detail/34.1294.TN.20160802.1732.004.html, published online August 2, 2016

Depth Enhancement Methods for Centralized Texture-Depth Packing Formats

YANG Jar-Ferr, WANG Hung-Ming, and LIAO Wei-Chen

(Department of Electrical Engineering, Institute of Computer and Communication Engineering, National Cheng Kung University, 1 University Road, Taiwan 701, China)

Abstract

To deliver three-dimension (3D) videos through the current two-dimension (2D) broadcasting systems, the frame-compatible packing formats properly including one texture frame and one depth map in various down-sampling ratios have been proposed to achieve the simplest and most effective solution. To enhance the compatible centralized texture-depth packing (CTDP) formats, in this paper, we further introduce two depth enhancement algorithms to further improve the quality of CTDP formats for delivering 3D video services. To compensate the loss of color YCbCr 444 to 420 conversion of colored-depth, two efficient depth reconstruction processes based on texture and depth consistency are proposed. Experimental results show that the proposed enhanced CTDP depacking process outperforms the 2DDP format and the original CTDP depacking procedure in synthesizing virtual views. With the help of the proposed efficient depth reconstruction processes, more correct reconstructed depth maps and better synthesized quality can be achieved. Before the available 3D broadcasting systems, which adopt truly depth and texture dependent coding procedure, we believe that the proposed CTDP formats with depth enhancement could help to deliver 3D videos in the current 2D broadcasting systems simply and efficiently.

Keywords

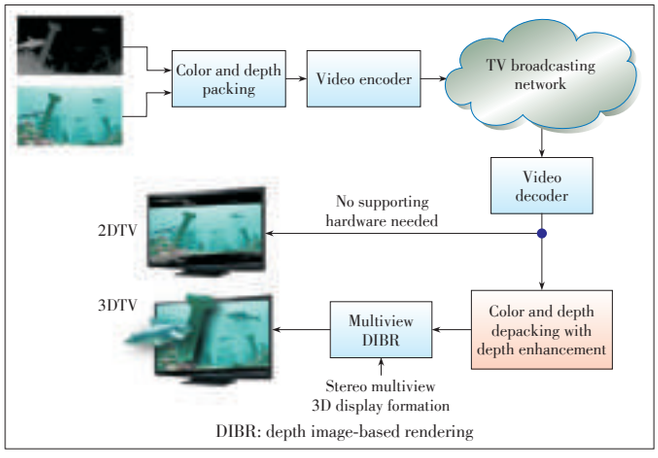
3D videos; frame-compatible; 2D-plus-depth; CTDP

1 Introduction

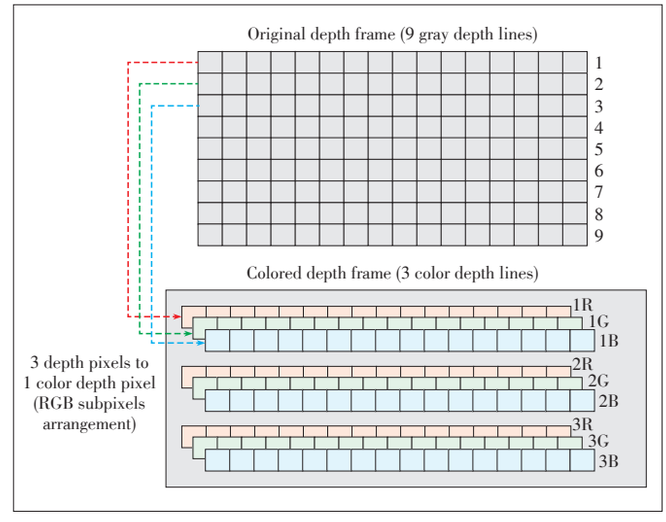
Over past decades, more and more three-dimensional (3D) videos have been produced in the formats of stereo or multiple views with their corresponding depth maps. People desire to have more truthful and exciting experience through the true 3D visualizations. In order to fit the traditional two-dimensional (2D) television (TV) programs, we need to modify the 3D videos to accommodate the certain constraints. Frame-packing is one of possible solutions to introduce 3D services in the current cable and terrestrial 2D TV systems. There are several well-known formats for packing the stereo views into 2D frame such as side-by-side (SbS), top-and-bottom (TaB), and checkerboard frame-compatible formats [1]–[4]. However, there exist two major problems, which slow down the development of the 3D TV services, in the existing frame-packing methods. The frame-compatible packing 3D videos of the stereo views mean that two texture images are gathered in one frame, which may make serious annoying effects on traditional 2D displays. Besides, stereo packing formats cannot support multi-view naked-eye 3D displays unless the stereo videos are further processed by real-time stereo matching methods [5], [6] and depth image-based rendering (DIBR) algorithms [7], [8]. To support multiview 3D displays, the 2D-plus-depth packing (2DDP) frame-compatible format, which arranges the texture in the left and the depth in the right, is suggested [9]. Once the color texture and depth arranged in the SbS fashion, the 2DDP format will bring even worse annoying visualization in 2D displays than the stereo packing formats. Recently, MPEG JCT-3V team proposed the latest coding standard for 3D video with depth [9]. However, it still needs some time to be deployed in current digital video broadcasting systems, which are with 2D and 3D capabilities.

To deal with the above problems, a novel frame—compatible centralized texture-depth packing (CTDP) formats for delivering 3D video services is proposed [10]. With AVS2 and HEVC video coders, the proposed CTDP formats [10] show better objective and subjective visual quality in 2D and 3D displays than the 2DDP format. In the CTDP format, the sub-pixel is utilized to store the depth information, while the texture information is arranged in the center of the frame to raise the 2D-compatible visual quality. However, the rearrangement will degrade the quality of the reconstructed depth map, especially when the video format with YCbCr space is 420 format with 4 Y components, one Cb component and one Cr component for each 4 color pixels. To further increase the visual quality, an efficient depth reconstruction process is also proposed in this paper. The frame structure of the CTDP method in cooperation with the current broadcasting system is shown in **Fig. 1**. Without any extra hardware, the 2D TV displays can also exhibit an acceptable 2D visual quality. For glasses or naked-eye 3D displays, we only need a simple CTDP depacking circuit followed by DIBR kernel to synthesize stereo or multiple views if the view-related sub-pixel formation of a naked-eye 3D display is given.

The rest of the paper is organized as follows. The CTDP formats are overviewed in Section 2. The proposed depth reconstruction process is described in Section 3. Experimental results to demonstrate the effectiveness of the proposed system are shown in Section 4. Finally, we conclude this paper in Sec-



▲Figure 1. The broadcasting architecture by using the proposed enhanced CTDP format.



▲Figure 2. Rearrangement of the depth frame into RGB colored depth frame in vertical direction.

tion 5.

2 Centralized Texture-Depth Packing Formats

To achieve system compatibility, the basic concept of the CTDP method [10] is similar to frame compatible concept to pack texture and depth information together while keeping the same resolution as 2D videos. To solve the 2D visualization issue, we can arrange the texture in the center and the depth in two sides of the packed frame.

2.1 Colored-Depth Frame

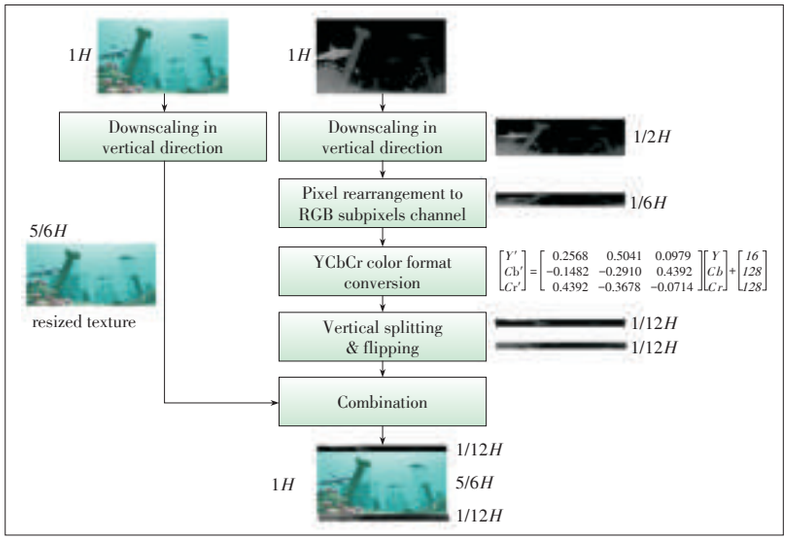
The depth frame is only a gray image with Y components. To pack the depth frame, the colored-depth frame is suggested to represent it [10]. Thus, the colored-depth frame can be treated as the normal color texture frame, which can be directly encoded by any 2D video encoders with three times efficiency. As shown in **Fig. 2**, three depth horizontal lines are treated as horizontal R, G, and B subpixel lines in the RGB colored-depth frame. Since the nearby depth values are very close, the RGB colored-depth frame will exhibit nearly gray visual sensation. After color subpixels packing in the vertical direction, the vertical resolution of RGB colored-depth frame becomes one third of the original resolution. In Fig. 2, for example, the nine depth lines have been packed into three RGB colored- depth lines. For the most video coders, the coding and decoding processes are conducted in YCbCr color space. Therefore, we apply the RGB to YCbCr color space conversion as

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.2568 & 0.5041 & 0.0979 \\ -0.1482 & -0.2910 & 0.4392 \\ 0.4392 & -0.3678 & -0.0714 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

to transfer it to the YCbCr colored-depth frame [11]. It is noted that the sub-pixels in RGB space are with full resolution of (4, 4, 4). If the YCbCr space is with (4, 4, 4) format, the color space transformation will not change the depth results with about +/− 0.5 error due to the round-off errors in color space conversions. However, for the most video coders, the sub-pixels in YCbCr space could be in (4, 2, 0) or (4, 2, 2) format, where Cb and Cr components will be further downsampled. Even without coding errors, the YCbCr colored-depth frame might have slightly translation errors.

2.2 Centralized Texture-Depth Packing

Without loss of generality for frame-compatible packing, we assume that the vertical CTDP packing formats are desired. Then, we need to reduce the vertical resolutions of texture and depth separately such that the total packed resolution will remind the same, where the original horizontal resolution is H. If the reduction factors for texture and depth resolutions are a and b, we should choose reduction factors to satisfy $\alpha+(1/3)\beta=1$ to achieve the frame compatible requirement [10]. For example, the reduction factors (a = 3/4, b = 3/4) , (a = 5/6, b = 1/2), (a = 7/8, b = 3/8), (a = 11/12, b = 1/4), and (a = 15/16, b = 3/16) will satisfy the above frame compatible requirement. **Fig. 3** shows the flowchart of the computation of generating the texure-5/6 CTDP format. First, we downscale the vertical resolution of texture and depth frames into five-sixths and one-second of the original resolution, respectively. By using the colored-depth concept, the resized depth frame with 1/2H can be further represented into RGB subpixels as suggested in Section 2.1 to reduce the vertical size to 1/6H. Then, we can split the depth frame evenly into two separated parts with the size of 1/12H. To make better coding efficiency and better 2D visualization, these two split colored-depth frames should be flipped vertically. The flipped depth frames will have better alignments to the texture frame and better visualization for 2D displays with visual shadow sensation. Finally, we obtain the texture-5/6 CTDP frame by combining the first



▲Figure 3. The computation of the proposed frame compatible texture-5/6 CTDP format.

flipped depth part (1/12*H*), the resized texture frame (5/6*H*), and the other flipped depth part (1/12*H*) from top to bottom sequentially.

The ratio of downscaling can also be changed to generate the other CTDP formats [12]–[15]. For example, the reduction ratio of the texture frame could be 7/8 or 15/16. For texture-7/8 and texture-15/16 reduction ratios, the vertical resolutions of depth frames will be respectively downscaled to 3/8 and 3/16 to satisfy (2). Except the resizing factor, the packing procedures for texture-7/8 and texture-15/16 are similar to that of texture-5/6. If we want to attain horizontal CTDP formats, all the resizing of texture and depth frame, the color-packed depth frame, slipping, and flipping procedures should be performed in the horizontal direction. The packed frame can be obtained by combining the first flipped depth part, the resized texture frame, and the other flipped depth part from left to right sequentially. The outlooks of the original texture, depth, and the CTDP frames with different ratios and different orientations are shown in **Fig. 4**. It is noted that in the proposed CTDP format, the width/height of the flipped depth part will be always in the horizontal/vertical CTDP format, which helps avoid the compression artifact in texture and depth boundary. Please refer to [13] for more details of the arrangement.

2.3 Depacking CTDP Formats

With respect to the packing procedure in Fig. 3, the flow diagram for depacking the texture-5/6 CTDP format is shown in **Fig. 5**. Once we receive the CTDP format, we should first split the packed frame into three parts: the top flipped depth part, the central texture, and the bottom flipped depth part. For two flipped depth parts, we perform another vertical flipping and combined them into the whole texture-packed depth frame. The YCrCb colored-depth frame might need to upsample Cr and Cb components back to (4, 4, 4) format first. Then, we can convert it to (4, 4, 4) RGB colored-depth frame by

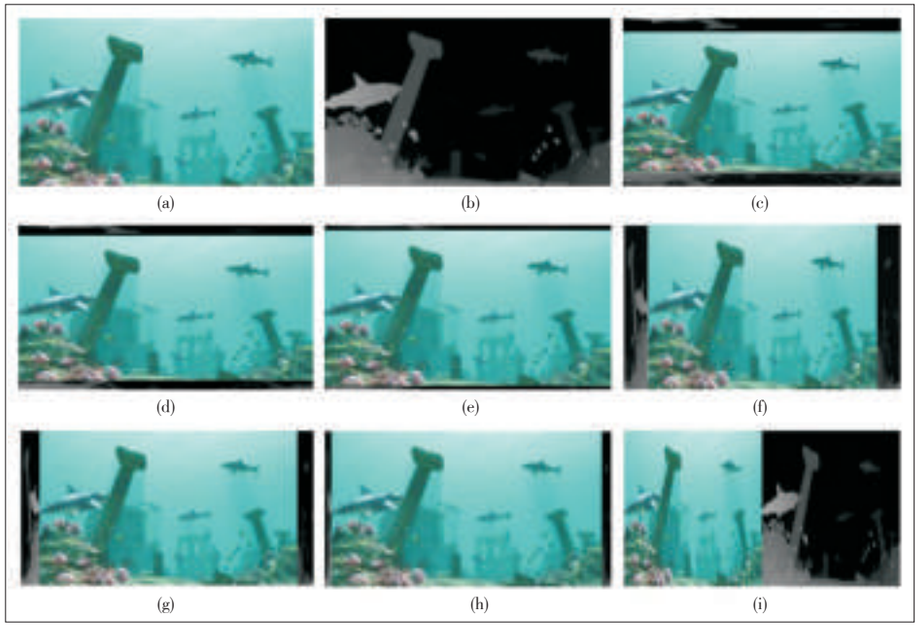
$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1.1644 & -0.0001 & 1.5960 \\ 1.1644 & -0.3917 & -0.8130 \\ 1.1644 & 2.0173 & -0.0001 \end{bmatrix} \left(\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} - \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \right). \quad (2)$$

After the color space conversion, The RGB colored-depth frame (1/6*H*) can be finally recovered to the resized depth frame (1/2*H*).

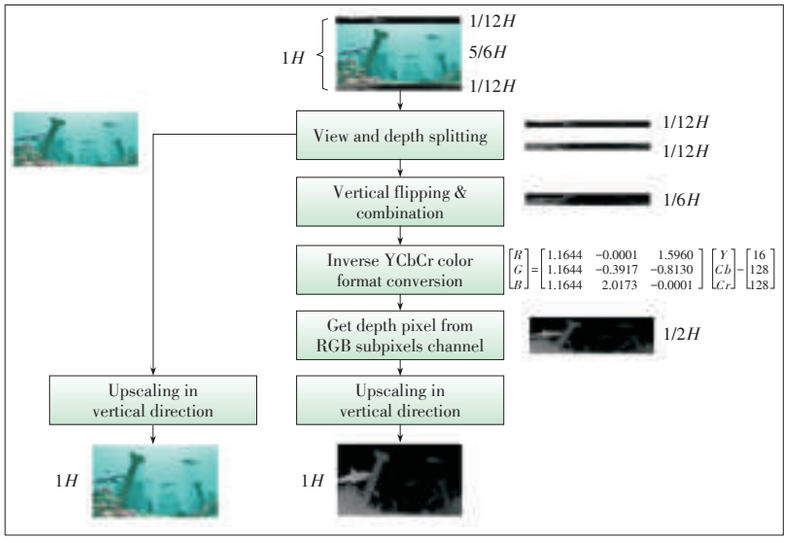
After 6/5 upscaling texture and 2/1 depth frames in the vertical direction, we finally depack the original texture and depth frames. Of course, a possible DIBR method should be used to generate all the necessary views. As for the other texture reduction ratios such as 7/8 and 15/16, all the procedures will be the same except the resizing factors of depth will be 3/8 and 3/16, respectively.

3 Depth Enhancement Algorithms

From the previous section, it is known that when the YCbCr space is (4, 2, 0) or (4, 2, 2) format, the YCbCr colored-depth frame will induce translation errors along the depth edges. To



▲Figure 4. Schematics of original (a) texture, (b) depth; (c) vertical texture-5/6 CTDP; (d) vertical texture-7/8 CTDP; (e) vertical texture-15/16 CTDP; (f) horizontal texture-5/6; (g) CTDP, horizontal texture-7/8 CTDP; (h) horizontal texture-15/16 CTDP; and (i) 2DDP frame compatible formats.



▲Figure 5. The computation of the proposed texture-5/6 CTDP depacking procedure.

further reduce the depth edge errors, in this paper, we propose two efficient depth enhancement processes. The enhancement processes can be incorporated with the original depacking process as shown in **Fig. 6.** The enhancement processes include YCbCr calibration, texture-similarity-based depth up-sampling and pattern-based down-sampling. Details of the enhancement algorithms are addressed in the following subsections.

3.1 YCbCr Calibration

When the YCbCr color space is (4, 4, 4), the color space transformation between RGB color space and YCbCr color space will only contain round-off errors in color space conversions. However, for the most video coders, the subpixels in the YCbCr color space might be (4, 2, 0) or (4, 2, 2) formats, where Cb and Cr components will be further down-sampled in order to save the bandwidth in broadcasting systems. At the depacking side, we need to calibrate the translation errors between YCbCr (4, 4, 4) and YCbCr (4, 2, 0) and (4, 2, 2). For simplicity, we will illustrate our proposed system in YCbCr (4, 2, 0), however, the similar manner can still be applied for YCbCr (4, 2, 2).

Before we start to calibrate the YCbCr data, we first define some anchor pixels, which are shown in **Fig. 7**. The anchor pixels denote the pixels which have the correct Cb and Cr subpixel values.

The diagram of missing components in YCbCr (4, 2, 0) for all surrounding pixels is shown in **Fig. 8**. Each color means a set which Cb and Cr subpixel components are down-sampled. The black area means the missing Cb and Cr subpixels and they can be given by:

$$Cb_{\text{cal}}(a,b) = \arg_{Cb_c} \min |Y_C - Y(a,b)|, \tag{3}$$

and

$$Cr_{\text{cal}}(a,b) = \arg_{Cr_c} \min |Y_C - Y(a,b)|, \tag{4}$$

where Y_C is a vector of the neighbor anchor pixels of the pixels $Y(a, b)$.

3.2 Texture-Similarity-Based Depth Up-Sampling

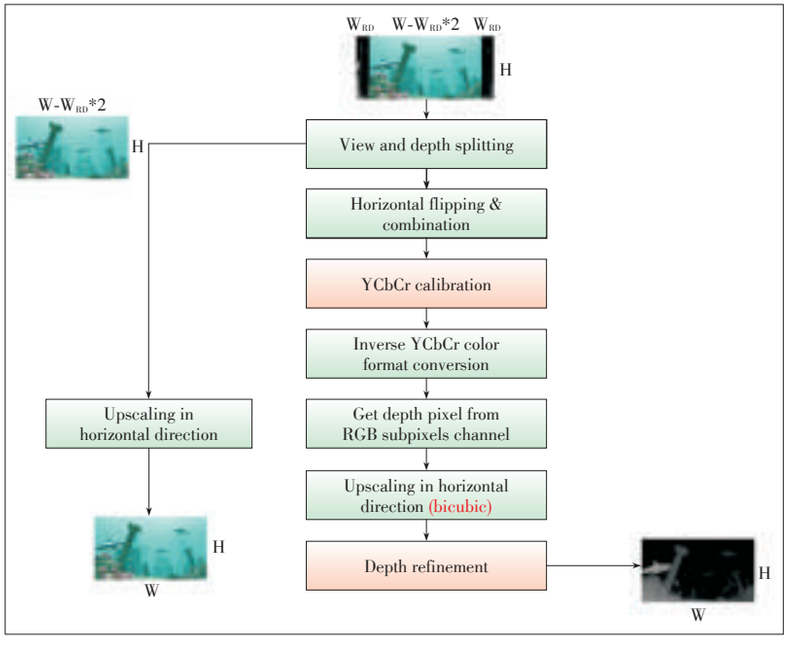
In order to preserve the continuity of the edge, the directional vectors are utilized to calculate the edge direction in the low-resolution (LR) depth and the corresponding high resolution (HR) texture image. The directional vectors of LR depth image and HR texture image can be formed as:

$$\overrightarrow{Vd_L} = \sum_{\Omega} \exp\left(-\frac{D_E(x_L,y_L) - D_\Omega}{\sigma_V}\right) \times \vec{u}_\Omega, \tag{5}$$

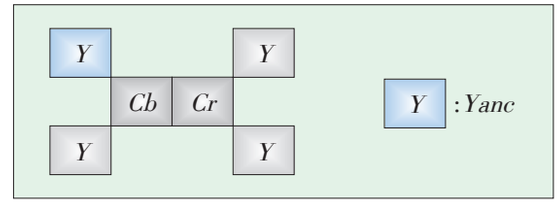
and

$$\overrightarrow{Vc} = \sum_{\Omega} \exp\left(-\frac{Y(x,y) - Y_\Omega}{\sigma_V}\right) \times \vec{u}_\Omega, \tag{6}$$

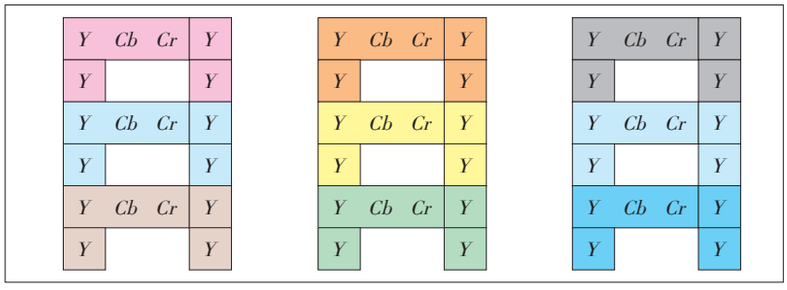
where $\overrightarrow{Vd_L}$ and $\overrightarrow{Vc}$ denote the directional vectors of the pixels in LR depth image and HR texture image, respectively, σ_V represents the standard deviation



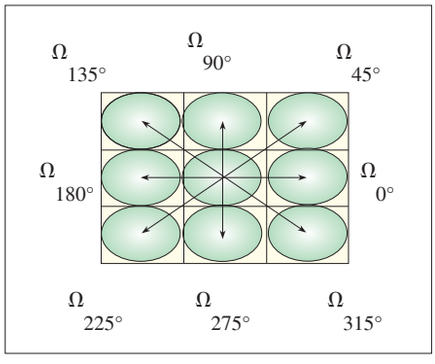
▲Figure 6. Flowchart of the depth-enhanced CTDP de-packing system.



▲Figure 7. Anchor pixels in YCbCr (4, 2, 0).



▲Figure 8. Missing Cb and Cr subpixels in YCbCr (4, 2, 0).



◀Figure 9. The diagram of the neighbor pixels, Ω.

of the directional vector function, Ω denotes the 8 neighbor pixels of the target pixel (**Fig. 9**), D_E represents the combined depth, which is obtained from previous step, Y is the brightness of the texture image, and $\vec{u}_\Omega$ is the unit vector corresponding to the neighbor pixels Ω in 8 directions.

Before up-sampling the depth image, the directional vectors are first transformed from Cartesian coordinate system to Spherical coordinate system. The transform function is given by:

$$r = \sqrt{x_\partial^{\,2} + y_\partial^{\,2}}, \tag{7}$$

and

$$\theta = \arctan\left(\frac{y_\partial}{x_\partial}\right), \tag{8}$$

where x_∂ and y_∂ denote the coordinate of reconstructed depth at high resolution. For example, at vertical texture-11/12 CTDP, $x_\partial = 4x$ and $y_\partial = y$. However, the resolution of directional vectors in depth image is smaller than the resolution of directional vectors in texture image. The bilinear interpolation [16] is utilized to scale up the depth directional vector to the resolution of the texture image. After that, The interpolated depth image is formed as:

$$D_{up} = \begin{cases} \frac{1}{T_{up}} \sum D_E(q) \times \psi(D_E(p) - D_E(q)), & if \|Vd(\theta) - Vc(\theta)\| < \frac{\pi}{8} \; or \; Vc(r) < 1 \\ hole, & else \end{cases}, \tag{9}$$

where T_{up} denotes the normalized factor, p is the target pixel which needs to be scale up, q is the neighbor pixels of the target pixel, and $Vd(\theta)$ is the value of θ in the scaled $Vd_L(\theta)$. ψ denotes the Gaussian weight function and can be given as:

$$\psi(n) = \exp\left(-\frac{n^2}{\sigma_\psi}\right). \tag{10}$$

The basic concept of the depth interpolation is to compare the directional vectors of the depth image and the texture image. The weighted summation of the LR depth is utilized to interpolate the HR depth if the directional vectors of the depth image and the texture image are similar. Otherwise, the pixels in HR depth are regarded as holes, which are filled in the step of hole-filling. The function of hole-filling is given as:

$$D_{hole-filling}(x,y) = \begin{cases} \arg_{D_w} \xi(\min(\Delta Pc(\theta))), & if\,(x,y) \ni holes \\ D_{up}(x,y), & else \end{cases}, \tag{11}$$

where $\Delta Pc(\theta)$ denotes the difference of the degree between Pc_θ and 8 neighbor pixels. ξ represents the selection function of the hole-filling and it can be formed as:

$$\xi(m) = \begin{cases} Y(m), & if \|Y - Y(m)\| < TH_Y \\ \xi(m) + 1, & else \end{cases}, \tag{12}$$

where Y denotes the brightness of the target pixel, $Y(m)$ denotes the brightness of the neighbor pixels in m direction, TH_Y is the threshold to control the selection range, and $\xi(m) + 1$ represents the next pixel in m direction.

3.3 Pattern-Based Down-Sampling

In order to contain texture image and depth image in one single frame, both depth image and texture image need to be down-sampled. For the depth image, the bilinear and bi-cubic convolution methods are utilized to down-sample the depth image. However, the weighted summation strategy in bilinear and bi-cubic convolution leads to the blur of the down-sampled data. Hence, we propose two sampling patterns to down-sample the depth image without fusing the data. There are the direct line pattern and slant line pattern.

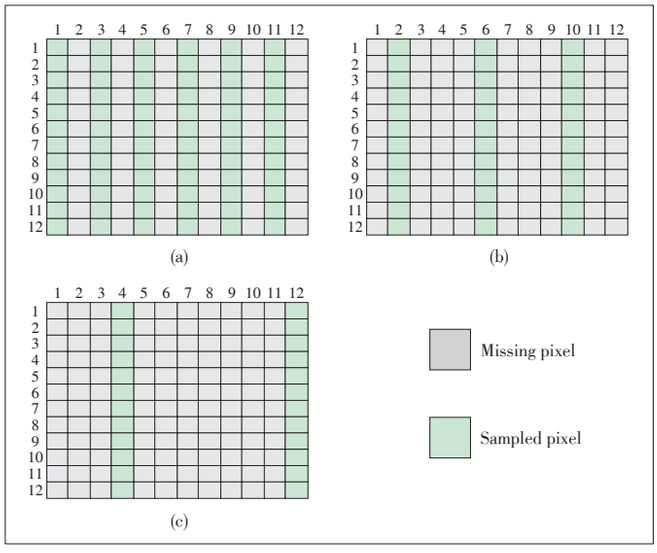
1) Direct line pattern

The sampling strategy of direct line pattern is to grab pixels in the straight line direction. According to the characteristic of the CTDP format, the reduction of the resolution is only in either horizontal or vertical direction. The function of direct line pattern is given as:

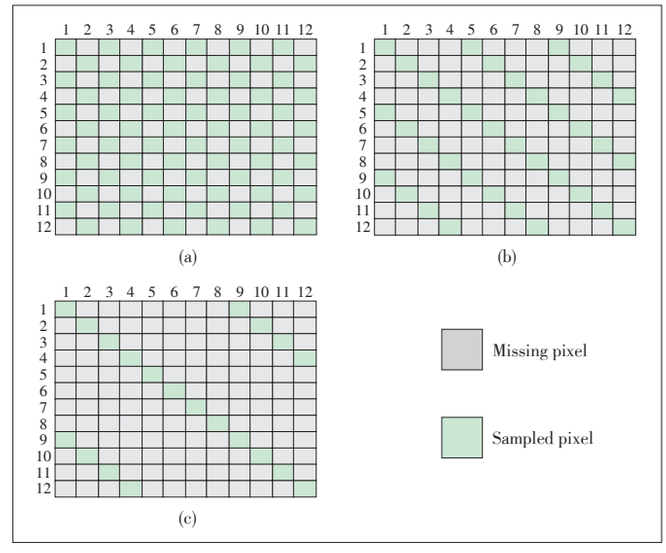
$$D_{down}(x,y) = D_{origin}(\partial_{hor} \times x - [\partial_{hor}/2], \partial_{ver} \times y - [\partial_{ver}/2]), \tag{13}$$

where ∂_{hor} and ∂_{ver} are the factors of down-sampling ratios in horizontal direction and vertical direction, respectively. For CTDP format usage, either ∂_{hor} or ∂_{ver} is equal to 1, while the other one denotes the down-sampling ratio in packing procedure. $[x]$ is the floor function, which means the largest integer not greater than x. The direct line pattern in horizontal direction with 2, 4, 8 down-sampling ratio is shown in **Fig. 10**.

2) Slant line pattern



▲Figure 10. The direct line pattern in horizontal direction of (a) down-sampling factor 2; (b) down-sampling factor 4; and (c) down-sampling factor 8.



▲Figure 11. The slant line pattern of (a) down-sampling factor 2; (b) down-sampling factor 4; (c) down-sampling factor 8.

The sampling strategy of slant line is to grab pixels in 45 degree direction. The function of direct line pattern is given as:

$$D_{down}(x,y)=D_{origin}(\partial_{hor}\times x-(\partial_{hor}-y),y)\,, \tag{14}$$

or

$$D_{down}(x,y)=D_{origin}(x,\partial_{ver}\times y-(\partial_{ver}-x))\,. \tag{15}$$

Equ. (14) is utilized to down-sample the depth image in horizontal direction, while the down-sampling of the vertical direction follows (15). The slant line sampling pattern is suitable for down-sampling the depth image both in vertical and horizontal direction, which is shown in **Fig. 11** with 2, 4, 8 down-sampling ratios.

With the down-sampling by the direct line pattern, the up-sampling function in de-packing procedure needs to be modified as:

$$D_{up}=\begin{cases}D_E(p), & if\ p\in sampled\ data\\ \frac{1}{T}\sum D_E(q)\times\psi(D_E(p)-D_E(q)), & else\ if\ \|Vd(\theta)-Vc(\theta)\|<\frac{\pi}{8}\ or\ Vc(r)<1.\\ hole, & else\end{cases} \tag{16}$$

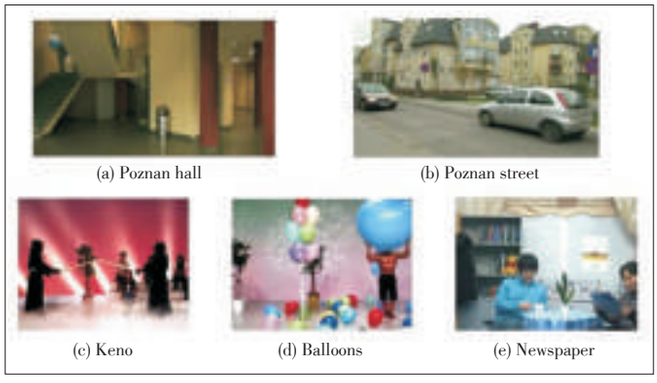
Because of the pattern-based sampling strategy, the pixels of the up-sampled depth are directly copied from the LR depth if there are located at position of the direct line pattern.

4 Experimental Results

4.1 Performance Evaluation of CTDP Format with Respect to 2DDP Format

In order to verify the coding performances of the proposed CTDP formats with respect to the 2DDP format, we conducted a set of experiments to evaluate performances of packing methods in cooperation with a specific video coder (AVS2) in terms of the peak signal-to-noise ratio (PSNR), bitrate qualities of the depacked texture and depacked depth frames, and their synthesized virtual views. In the experimental simulations, we use five MPEG 3D video sequences, which are Poznan Hall, Poznan Street, Kendo, Balloons, and Newspaper sequences as shown in **Figs. 12a–12e**, respectively.

The AVS2 coding conditions are followed by the instruction suggested by the AVS workgroup while the QPs are set to 27, 32, 38, and 45 for Intra frames [17]. Under All Intra (ai), Low Delay P (ldp), Random Access (ra) test conditions, **Tables 1** and **2** show the average BDPSNR and BDBR [18] performance for different kinds of CTDP formats with respect to the 2DDP format achieved by AVS2. For calculating the PSNR of the 2DDP format, we first separate the texture and depth frames from the 2DDP frame and upsample them to the original image size $W\times H$. By using the recovered texture and depth frames from 2DDP frame and the original uncompressed texture and depth frames, the PSNR can therefore be calculated. Similarly, the PSNR of CTDP format is calculated by using the texture and depth frames recovered from CTDP frame and the original uncompressed texture and depth frames. From Tables 1 and 2, we can see that the proposed texture-5/6, 7/8, and 15/16 CTDP formats have much better PSNR and bitrate saving in texture when comparing with the 2DDP format, which means our CTDP format can achieve better visual quality in 2D displays when only texture frames are viewed. In addition, the depth quality for CTDP formats will become worse while the resizing factors getting bigger. Besides the comparisons of original texture and depth achieved by different packing formats, we also compare the quality of synthesized virtual view with respect to the 2DDP format. It is noted that the reference synthesized vir-



▲Figure 12. Five texture and depth frame compatible packing formats.

▼Table 1. Averaged BDPSNR performances

BDPSNR						
	Recovered texture performance			Synthesized virtual view performance		
	5/6	7/8	15/16	5/6	7/8	15/16
ai	2.2693	2.2587	2.34168	0.6150	0.3664	-0.7732
ra	2.5868	2.71322	2.86354	0.7371	0.7489	-0.2754
ldp	2.2079	2.42228	2.54206	0.3920	0.6005	-0.3928
avg	2.3546	2.464733	2.582427	0.5813	0.5720	-0.4805

▼Table 2. Averaged BDBR performances

BDBR						
	Recovered texture performance			Synthesized virtual view performance		
	5/6	7/8	15/16	5/6	7/8	15/16
ai	-55.3282	-46.5149	-47.9069	-22.9490	-7.92186	56.0625
ra	-61.0706	-59.5852	-61.6926	-24.6686	-24.6019	30.0742
ldp	-48.7147	-54.3501	-56.2303	-9.9111	-19.6368	36.5666
avg	-55.0378	-53.4834	-55.2766	-19.1762	-17.3869	40.9011

tual view for calculating the PSNR is also obtained by the original uncompressed texture and depth frames. The DIBR setting for virtual view synthesis is shown in **Table 3**. As to the quality of the synthesized virtual view, the texture-5/6 and 7/8 CTDP formats after the DIBR process show better BDPSNR and BDBR performances than 2DDP format. It is noted that all synthesized views do not perform any depth enhancement and depth preprocessing, and the hole filling used in the DIBR process is the simple background extension.

▼Table 3. DIBR settings for virtual view synthesis

Sequence	Resolution	Frames	Coded view	Synthesized view
Poznan hall	1920*1088	200	6	5
Poznan street	1920*1088	250	4	3
Kendo	1024*768	300	3	4
Balloons	1024*768	300	3	5
Newspaper	1024*768	300	4	6

In summary, the texture qualities BDPSNR and BDBR in Tables 2 and 3 can be treated as the objective quality indices in 2D displays, while the virtual view qualities can be the objective quality indices in 3D displays. The results show that the proposed texture-5/6 and 7/8 CTDP format will be the better choices for the broadcasters. The texture-3/4 CTDP format has better 3D performance while texture-7/8 CTDP format achieves better 2D performance.

4.2 Performance Evaluation of Depth Enhancement for CTDP Format

To verify the proposed depth enhancement mechanism, we first show the reconstructed depth from original and depth-enhanced CTDP formats. The RD curves for different ratios of CTDP formats are shown in **Fig. 13**. It can be seen that the proposed refined CTDP format can always achieve better performance. The gains between the depth-enhanced CTDP and the original CTDP formats are increased while the ratio of texture is increased.

For the subjective evaluation, the partial portions of the reconstructed depth for Shark sequence are shown in **Fig. 14**. It can be seen that the depth can be reconstructed well especially for the edge region by using the depth enhancements.

In the following, we will compare the synthesis results. The partial portions of the generated views are shown in **Fig. 15**. From the results, the proposed CTDP format can successfully preserve the edges well of the synthesis views without the jaggy noise.

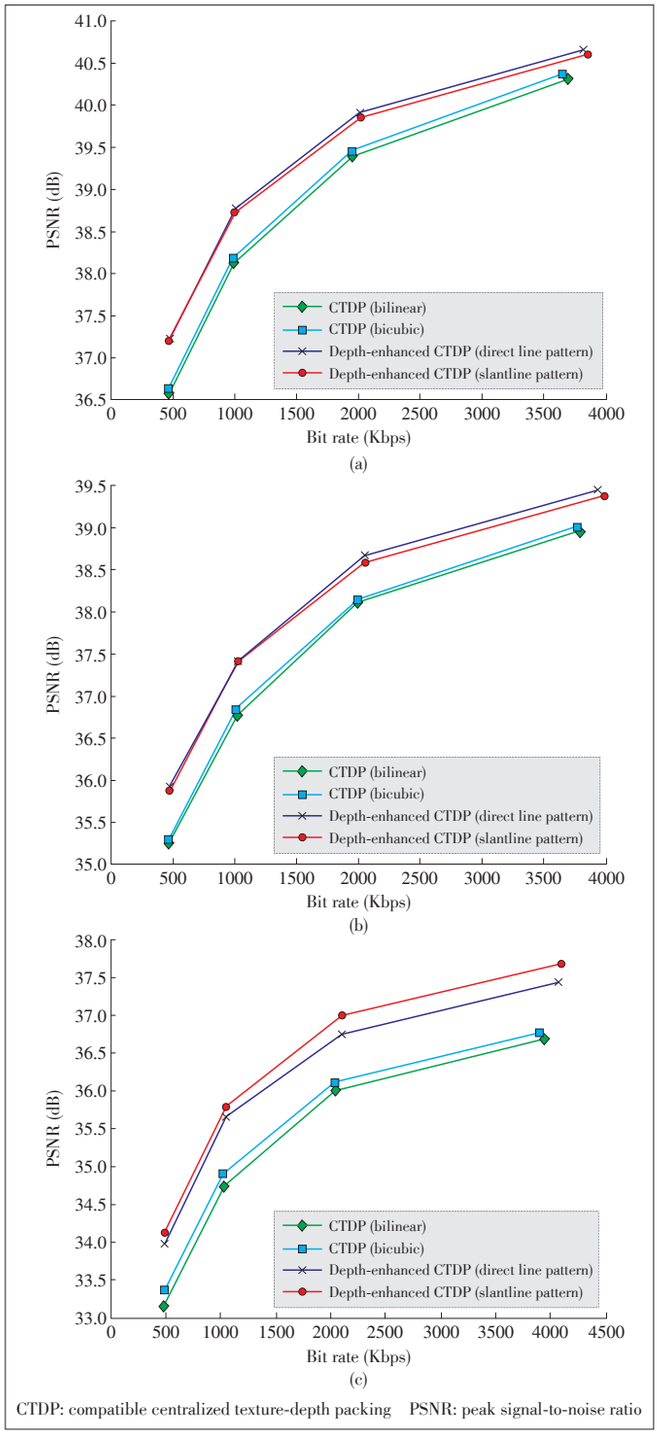
4.3 Comparison with Different Depth Interpolation Methods

The comparison results of different depth interpolation methods are shown in **Table 4** for Shark sequence at all-intra (ai) coding condition with QP=32. The symbols of Bi and BC denote the bilinear and bi-cubic convolution interpolation methods, respectively. The methods of JBU [19] and FEU [20] are the texture-similarity based depth interpolation methods. The proposed depth up-sampling method has better PSNR and SSIM results for reconstructed depth images in vertical-11/12 CTDP and vertical-23/24 CTDP formats. For the vertical-5/6 CTDP format, the proposed depth up-sampling method can also provide better reconstructed depth images.

The comparison results of partial reconstructed depth with different depth interpolation methods are shown in **Fig. 16**. The reconstructed depth images of bilinear and bi-cubic convolution interpolation methods have serious jaggy noise among the edges. It can be seen that the proposed depth up-sampling method can outperform other methods with better edges.

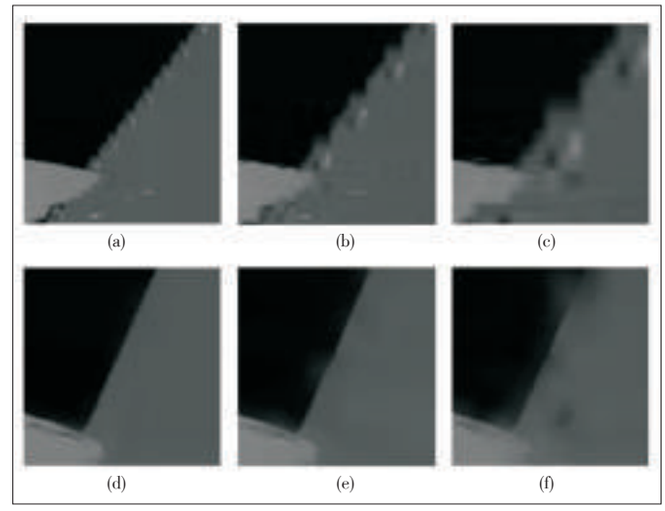
5 Conclusions

In this paper, we proposed depth enhancement processes for

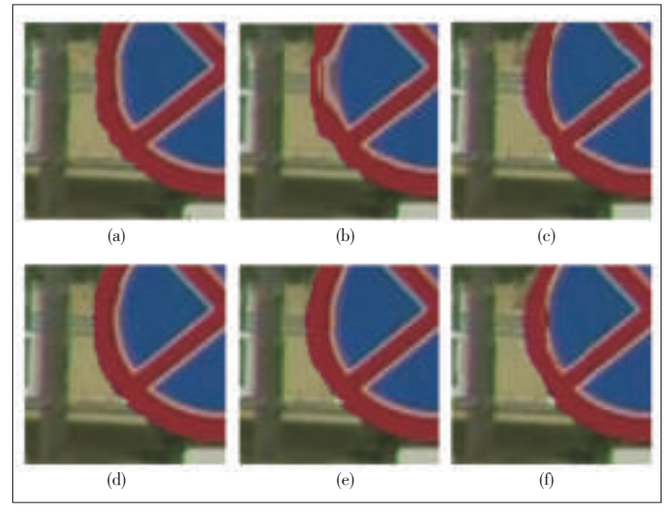


▲Figure 13. RD curves of reconstructed depth from the original CTDP depacking process and the proposed depth-enhanced CTDP depacking process for (a) texture-5/6; (b) texture-11/12; (c) texture-23/24 formats.

CTDP formats [10]. The CTDP formats can be comfortably and directly viewed in 2DTV displays without the need of any extra computation. However, the CTDP formats slightly suffer from the depth discontinuities for high texture ratios. Comparing to



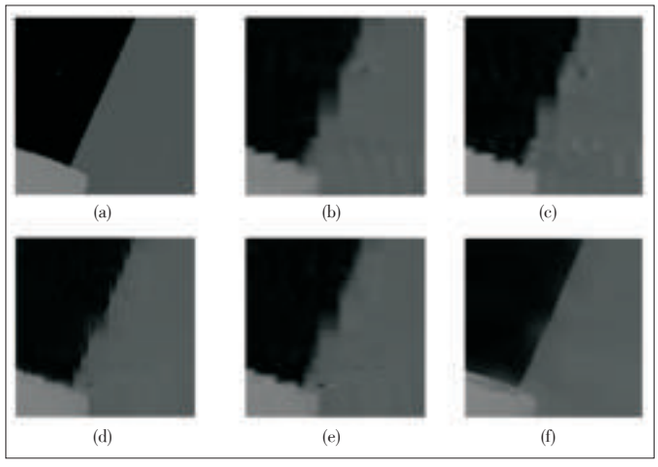
▲Figure 14. Partial portions of reconstructed depth in S10 Shark with the original CTDP depacked: (a) texture-5/6; (b) texture-11/12; (c) texture-23/24 formats and the proposed depth enhanced CTDP depacked; (d) texture-5/6; (e) texture-11/12; (f) texture-23/24 formats.



▲Figure 15. Partial synthesis views of S02 Poznan Street with original CTDP depacking: (a) texture-5/6; (b) texture-11/12; (c) texture-23/24 and the proposed depth-enhanced CTDP depacking; (d) texture-5/6; (e) texture-11/12; (f) texture-23/24.

▼Table 4. The PSNR and SSIM comparison of different depth interpolation methods in S10 Shark at All Intra (ai) QP=32

PSNR (dB)	Bi	BC	JBU	FEU	Proposed
Vertical-5/6 CTDP	33.7164	33.5138	33.6252	33.7135	33.5239
Vertical-11/12 CTDP	31.7651	31.6581	32.0857	31.7850	32.4422
Vertical-23/24 CTDP	29.8525	29.7836	30.3490	29.8953	30.5411
SSIM	Bi	BC	JBU	FEU	Proposed
Vertical-5/6 CTDP	0.9361	0.9334	0.9361	0.9368	0.9397
Vertical-11/12 CTDP	0.9147	0.9122	0.9192	0.9158	0.9270
Vertical-23/24 CTDP	0.8914	0.8879	0.8959	0.8925	0.9068



▲**Figure 16. The comparison of partial reconstructed depth with different depth interpolation methods for vertical-11/12 CTDP format: (a) ground truth; (b) bilinear; (c) bi-cubic convolution; (d) JBU [19]; (e) FEU [20]; (f) proposed.**

the 2DDP format, the CTDP formats with the same video coding systems, such as AVS2 (RD 6.0) and HEVC [10], show better coding performances in texture and depth frames and synthesized virtual views. To further increase the visual quality, in this paper, the depth enhancement methods, including YCbCr calibration and texture-similarity-based depth up-sampling, are proposed. Experimental results reveal that the proposed depth enhancement can efficiently help to increase the depacking performances of the CTDP formats to achieve better reconstructed depth images and better synthesis views as well. With the aforementioned simulation results, we believe that the proposed depth enhanced CTDP depacking methods will be a greatly-advanced system for current 2D video coding systems, which can provide 3D video services effectively and simply.

References

[1] J.-F. Yang, H.-M. Wang, K.-I. Liao, L. Yu, and J.-R. Ohm, "Centralized texture-depthpacking formats for effective 3D video transmission over current video broadcasting systems," *IEEE Transactions on Circuits and Systems for Video Technology*, submitted for publication.
[2] Dolby Laboratories, Inc. (2015). *Dolby Open Specification for Frame-Compatible 3D Systems* [Online]. Available: http://www.dolby.com
[3] ITU. (2015). *Advanced Video Coding for Generic Audio—Visual Services* [Online]. Available: http://www.itu.int
[4] G. Sullivan, T. Wiegand, D. Marpe, and A. Luthra, "Text of ISO/IEC 14496-10 advanced video coding (third edition)," ISO/IEC JTC 1/SC 29/WG11, Redmond, USA, Doc. N6540, Jul. 2004.
[5] G. J. Sullivan, A. M. Tourapis, T. Yamakage, and C. S. Lim, "ISO/IEC 14496-10: 200X/FPDAM 1," ISO/IEC JTC 1/SC 29/WG11, Apr. 2009.
[6] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," *IEEE Transactions on Pattern Analysis and Matching Intelligence*, vol. 16, no. 9, pp.920–932, Sept. 1994. doi:10.1109/34.310690.
[7] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 7, pp.1073– 1079, Jul. 2009. doi: 10.1109/TCSVT.2009.2020478.
[8] S.-C. Chan, H.-Y. Shum, and K.-T. Ng, "Image-based rendering and synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 22–33, Nov. 2007. doi: 10.1109/MSP.2007.905702.
[9] T.-C. Yang, P.-C. Kuo, B.-D. Liu, and J.-F. Yang, "Depth image-based rendering with edge-oriented hole filling for multiview synthesis," in *Proc. International Conference on Communications, Circuits and Systems*, Chengdu, China, Nov. 2013, vol. 1, pp. 50–53. doi: 10.1109/ICCCAS.2013.6765184.
[10] Philips 3D Solutions, "3D interface specifications, white paper," Eindhoven, The Netherlands, Dec. 2006.
[11] *Studio Encoding Parameters of Digital Television for Standard 4:3 and Wide-Screen 16:9 Aspect Ratios*, ITU-R BT.601-5, 1995.
[12] J.-F. Yang, K.-Y. Liao, H.-M. Wang, and Y.-H. Hu, "Centralized texture-depth packing (CTDP) SEI message syntax," Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Strasbourg, France, Doc. no. JCT3V-J0108, Oct. 2014.
[13] J.-F. Yang, K.-Y. Liao, H.-M. Wang, and C.-Y. Chen, "Centralized texture-depth packing (CTDP) SEI message," Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Geneva, Switzerland, Doc. no. JCT3V-K0027, Feb. 2015.
[14] J.-F. Yang, H.-M Wang, Y.-A. Chiang, and K. Y. Liao, "2D frame compatible centralized color depthpacking format (translated from Chinese)," AVS 47th Meeting, Beijing, China, AVS-M3225, Dec. 2013.
[15] J.-F. Yang, H.-M. Wang, K.-Y. Liao, and Y.-A. Chiang, "AVS2 syntax message for 2D frame compatible centralized color depth packing formats (translated from Chinese)," AVS 50th Meeting, Nanjing, China, AVS-M3472, Oct. 2014.
[16] H. C. Andrews and C. L. Patterson, "Digital interpolation of discrete images," *IEEE Transaction on Computers*, vol. 25, no. 2, 1976.
[17] X.-Z. Zheng, "AVS2-P2 common test conditions (translated from Chinese)," AVS 46th Meeting, Shenyang, China, AVS-N2001, Sep. 2013.
[18] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," Austin, USA, Doc. VCEG-M33 ITU-T Q6/16, Apr. 2001.
[19] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transaction on Graphics*, vol. 26, no. 3, Article 96, Jul. 2007. doi:10.1145/1275808.1276497.
[20] S.-Y. Kim and Y.-S. Ho, "Fast edge-preserving depth image upsampler," *Journal of Consumer Electronics*, vol. 58, no. 3, pp. 971–977, Aug. 2012. doi: 10.1109/TCE.2012.6311344.

Manuscript received: 2015-11-12

Biographies

YANG Jar-Ferr (jefyang@mail.ncku.edu.tw) received his PhD degree from the University of Minnesota, USA in 1988. He joined the National Cheng Kung University (NCKU) started from an associate professor in 1988 and became a full professor and distinguished professor in 1995 and 2007. He was the chairperson of Graduate Institute of Computer and Communication Engineering during 2004–2008 and the director of the Electrical and Information Technology Center 2006–2008 in NCKU. He was the associate vice president for Research and Development of the NCKU. Currently, he is a distinguished professor and the director of Technologies of Ubiquitous Computing and Humanity (TOUCH) Center supported by National Science Council (NSC), Taiwan, China. Furthermore, he is the director of Tomorrow Ubiquitous Cloud and Hypermedia (TOUCH) Service Center. During 2004–2005, he was selected as a speaker in the Distinguished Lecturer Program by the IEEE Circuits and Systems Society. He was the secretary, and the chair of IEEE Multimedia Systems and Applications Technical Committee and an associate editor of *IEEE Transaction on Circuits and Systems for Video Technology*. In 2008, he received the NSC Excellent Research Award. In 2010, he received the Outstanding Electrical Engineering Professor Award of the Chinese Institute of Electrical Engineering, Taiwan, China. He was the chairman of IEEE Tainan Section during 2009–2011. Currently, he is an associate editor of *EURASIP Journal of Advances in Signal Processing* and an editorial board member of *IET Signal Processing*. He has published 104 journal and 167 conference papers. He is a fellow of IEEE.

WANG Hung-Ming (ming@video5.ee.ncku.edu.tw) received the BS and PhD degrees in electrical engineering from National Cheng Kung University (NCKU), Taiwan, China in 2003 and 2009, respectively. He is currently a senior engineer of Novatek Microelectronics Corp., Taiwan, China. His major research interests include 2D/3D image processing, video coding and multimedia communication.

LIAO Wei-Chen (a800812momo@gmail.com) received the BS and MS degrees in electrical engineering from National Cheng Kung University (NCKU), Taiwan, China in 2013 and 2015, respectively. His major research interests include image processing, video coding and multimedia communication.

ZTE Communications Guidelines for Authors

• Remit of Journal

ZTE Communications publishes original theoretical papers, research findings, and surveys on a broad range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world.

• Manuscript Preparation

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 3000 to 8000, and no more than 8 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

• Abstract and Keywords

Each manuscript must include an abstract of approximately 150 words written as a single paragraph. The abstract should not include mathematics or references and should not be repeated verbatim in the introduction. The abstract should be a self-contained overview of the aims, methods, experimental results, and significance of research outlined in the paper. Five carefully chosen keywords must be provided with the abstract.

• References

Manuscripts must be referenced at a level that conforms to international academic standards. All references must be numbered sequentially in-text and listed in corresponding order at the end of the paper. References that are not cited in-text should not be included in the reference list. References must be complete and formatted according to ZTE Communications Editorial Style. A minimum of 10 references should be provided. Footnotes should be avoided or kept to a minimum.

• Copyright and Declaration

Authors are responsible for obtaining permission to reproduce any material for which they do not hold copyright. Permission to reproduce any part of this publication for commercial use must be obtained in advance from the editorial office of *ZTE Communications*. Authors agree that a) the manuscript is a product of research conducted by themselves and the stated co-authors, b) the manuscript has not been published elsewhere in its submitted form, c) the manuscript is not currently being considered for publication elsewhere. If the paper is an adaptation of a speech or presentation, acknowledgement of this is required within the paper. The number of co-authors should not exceed five.

• Content and Structure

ZTE Communications seeks to publish original content that may build on existing literature in any field of communications. Authors should not dedicate a disproportionate amount of a paper to fundamental background, historical overviews, or chronologies that may be sufficiently dealt with by references. Authors are also requested to avoid the overuse of bullet points when structuring papers. The conclusion should include a commentary on the significance/future implications of the research as well as an overview of the material presented.

• Peer Review and Editing

All manuscripts will be subject to a two-stage anonymous peer review as well as copyediting, and formatting. Authors may be asked to revise parts of a manuscript prior to publication.

• Biographical Information

All authors are requested to provide a brief biography (approx. 100 words) that includes email address, educational background, career experience, research interests, awards, and publications.

• Acknowledgements and Funding

A manuscript based on funded research must clearly state the program name, funding body, and grant number. Individuals who contributed to the manuscript should be acknowledged in a brief statement.

• Address for Submission

magazine@zte.com.cn
12F Kaixuan Building, 329 Jinzhai Rd, Hefei 230061, P. R. China

ZTE COMMUNICATIONS

ZTE Communications has been indexed in the following databases:

- Cambridge Scientific Abstracts (CSA)
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Inspec
- Ulrich's Periodicals Directory
- Wanfang Data—Digital Periodicals